Measuring Diffusion over a Large Network

Xiaoqi He and Kyungchul Song

Central University of Finance and Economics and University of British Columbia

ABSTRACT. This paper introduces a measure of the diffusion of binary outcomes over a large, sparse network, when the diffusion is observed in two time periods. The measure captures the aggregated spillover effect of the state-switches in the initial period on their neighbors' outcomes in the second period. This paper introduces a causal network that captures the causal connections among the cross-sectional units over the two periods. It shows that when the researcher's observed network contains the causal network as a subgraph, the measure of diffusion is identified as a simple, spatio-temporal dependence measure of observed outcomes. When the observed network does not satisfy this condition, but the spillover effect is nonnegative, the spatio-temporal dependence measure serves as a lower bound for diffusion. Using this, a lower confidence bound for diffusion is proposed and its asymptotic validity is established. The Monte Carlo simulation studies demonstrate the finite sample stability of the inference across a range of network configurations. The paper applies the method to data on Indian villages to measure the diffusion of microfinancing decisions over households' social networks.

KEY WORDS. Diffusion; Causal Inference; Network Interference; Latent Networks; Treatment Effects; Cross-Sectional Dependence; Common Shocks; Dependency Graphs

JEL CLASSIFICATION: C10, C13, C31, C33

1. Introduction

The phenomenon of diffusion over a network, where a person's change of state leads to another change in her neighborhood, has been of interest in epidemiology, economics, sociology, and marketing research.[1] Examples include diffusion of disease, technology, innovation, or product recommendations over various social, industrial, or community networks. The main purpose of this paper is to propose a measure of the diffusion of state changes over a large network as a causal parameter and to develop statistical inference on the measure.

There are challenges in developing such an empirical measure of diffusion. First, the diffusion of actions tends to occur constantly in real time, whereas the researcher makes multiple snapshot observations during the diffusion process. Second, diffusion can begin locally around a fraction of units and spread over the entire cross-section of units over time. This propagation can be fast over a short period, even if the contact network is sparse, and induces extensive cross-sectional dependence of outcomes, which can hamper asymptotic inference. Third, the researcher rarely observes the true underlying contact network accurately, but often observes only its proxy. In such a situation, the underlying contact network, not the proxy, determines the cross-sectional dependence structure of the observed outcomes. Without knowing this structure, developing statistical inference on diffusion is a nontrivial task.

This paper considers a generalized diffusion model over a large latent contact network, where the researcher observes the states of people twice, first in the initial period and then in the second period. The researcher also observes a network over them, where the observed network does not necessarily coincide with or even "approximate" the contact network. In this setting, we introduce an aggregate measure of diffusion and provide asymptotic inference on the measure. Here we explain several key features of our proposal.

First, we adopt the Neyman-Rubin potential outcome approach in the program evaluations literature (Imbens and Rubin (2015)) and measure diffusion as the expected increase in the number of people until the second observation when one additional randomly chosen person switches the state in the initial period. We call this measure the *Average Diffusion at the Margin (ADM)* in the paper. The potential outcome approach in the paper regards the initial switches in the first period as the "treatment", and the states observed in the second period as the "outcomes". Our approach requires that we can accurately estimate the probability of the initial switch. Thus, the paper's framework fits the situation where many initial switches are triggered, for example, by targeted advertisements based on their observed characteristics

---

[1]To name but a few examples, see, e.g., Coleman, Katz, and Menzel (1957), Conley and Udry (2001) and Conley and Udry (2010) for diffusion of technology and innovation; Banerjee, Chandrasekhar, Duflo, and Jackson (2013)) for diffusion of microfinancing decisions over villages; Leskovec, Adamic, and Huberman (2007) for diffusion of product recommendations; and de Matos, Ferreira, and Krackhardt (2014) for peer influence on iPhone purchases among students.

and networks. It does not apply to a situation where diffusion begins with a single switch or a few "seeds".

The potential outcome approach in this paper has both weaknesses and strengths. The weakness is that its focus is rather narrow, which is on the measurement of the average causal effect of the initial state-switches on neighbors' outcomes by a later time. The approach does not aim to recover from data any aspects of the causal mechanism between the initial state-switches and subsequent outcomes. Learning such aspects from data can be useful in generating predictions in a counterfactual scenario about the diffusion process. The strength of the potential outcome approach is that, as we show later, we can adopt a generalized diffusion model that encompasses a wide range of diffusion models used in the literature as special cases.

Second, we explicitly consider that the researcher's observation of diffusion can be asynchronous with the actual process of diffusion in real time. More specifically, we assume that the researcher observes the states in the initial period and in a later period, without observing the actual process of diffusion between these periods. Although our setting appears similar to the two-period network interference models studied in the literature (e.g., Aronow and Samii (2017)), it has distinctive features. First, we distinguish between the contact network and the network that governs the causal relations between the observed states. For the latter network, we introduce the notion of the *causal graph* which assigns an edge between two people if and only if we can trace one back to the other along a walk in the contact network. Second, we allow the unobserved heterogeneity in the potential outcomes to be cross-sectionally dependent with a dependence structure shaped by the contact network.

Third, we allow the observed network to be different from the contact network. It is well known that observing a large contact network accurately is difficult in practice (see, e.g., Breza, Chandrasekhar, and Tahbaz-Salehi (2018) and Banerjee, Chandrasekhar, Duflo, and Jackson (2019) for references and discussion.) We often observe only a proxy network measured through various survey questions. In dealing with the discrepancy between the contact network and the observed network, our development is divided into two cases. First, we consider a situation where the observed network contains the causal graph as a subgraph - which we call the *subgraph hypothesis* in this paper. In this case, while the observed network does not capture the details of the causal relations among the cross-sectional units, it encodes conditional independence among the observed outcomes.[2] We show that this is sufficient to identify the ADM as a simple spatio-temporal dependence measure between the actions in

---

[2]It suffices for the observed graph to capture the dependence structure, not the details on the causal relations, because the researcher rarely obtains accurate information on the direction of influence between two people from survey data in practice. For example, when there is a recorded edge from one farmer to another whenever the former borrows money from the latter, it is not obvious which direction between the two will be the correct direction of influence in studying the diffusion of agricultural technology. In such a case, our framework allows

the observed graph. Second, we consider the case where the subgraph hypothesis fails, but the spillover effect of each person on her neighbors is nonnegative. In this case, we show that the spatio-temporal dependence measure still serves as a lower bound for the ADM. Later in the paper, we provide a method for directional testing of the subgraph hypothesis.

Fourth, the researcher using our framework can be entirely agnostic about the cross-sectional dependence structure of the covariates or the relationship between the covariates and the contact network. Our framework allows a wide class of network formation models for the contact network. This flexibility is due to our defining the ADM as a quantity that is conditional on the contact network. When the contact network is formed homophilously, that is, people with similar characteristics are more likely to become neighbors with each other, the covariates exhibit cross-sectional dependence conditional on the contact network. However, the researcher rarely knows the cross-sectional dependence structure of the covariates because they do not observe the contact network. The unknown cross-sectional dependence structure poses a challenge in developing statistical inference on the ADM. To address this challenge, we follow the approach of Kuersteiner and Prucha (2013) in linear panel models. Thus, we treat the covariates, contact network, and observed graph as part of *common shocks* so that the asymptotic validity of the inference is not affected by the details of the way the contact network or the observed network is formed or by the way the covariates are cross-sectionally dependent.

The literature has shown that similarity in the observed actions among the cross-sectional units may stem from the similarity in their attributes, and this similarity can produce what seems like diffusion, even when there is no diffusion of outcomes over the network. We call this *spurious diffusion*.[3] The severity of spurious diffusion depends on the cross-sectional dependence structure of omitted attributes. To control for this effect of similarity in attributes, it is important to measure diffusion after controlling for the covariates. The ADM does precisely this. Our identification result shows that the ADM is identified as a spatial-temporal measure among the residuals after *controlling for the covariates*.

To illustrate the usefulness of our measure, we apply it to estimate the diffusion of microfinancing decisions over various measures of social networks using data on Indian villages used by Banerjee, Chandrasekhar, Duflo, and Jackson (2013). We consider two definitions of the initial state-switch. First, we define a person's initial state-switch to be an event that the

---

the researcher simply to take the undirected version of the observed graph (by turning all the directed edges to undirected ones) and use it as the observed graph.

[3]Spurious diffusion arises especially when the network is formed with homophily, so that correlated actions among neighbors primarily come from shared characteristics. An early recognition of this was made by Manski (1993) in a linear-in-means model when he distinguished between correlated effects and endogenous effects. For more recent literature recognizing this issue, see Aral, Muchnik, and Sundararajan (2009) and Shalizi and Thomas (2011).

person is a "leader", which is determined based on the expected connectedness in the social network, such as teachers, shopkeepers, and saving group leaders. According to the experiment design of Banerjee, Chandrasekhar, Duflo, and Jackson (2013), the leaders are those who first learned about the microfinancing program through private meetings with credit officers. Hence the initial state-switch by a person is equivalent to the person's having access to the information on the micro-financing decisions in the initial period. The second definition of the initial state-switch is an event in which the person was a leader and participated in the program. We call those people "leader-adopters".

Our first study focuses on the causal effect of "leader-adopters" on other households' participation in micro-financing along the network. For the observed networks, our directional tests find no evidence against the subgraph hypothesis. We find that the estimated ADM is positive with statistical significance at 5%. We also estimate the ADM by redefining the initial triggers to be the "leaders" instead. In this case, the estimated ADM turns out to be statistically insignificant even at 10%. Our results demonstrate the importance of being precise about the initial triggers in the study of diffusion.

Related Literature

Diffusion of disease, information, and technology, has received a great deal of attention in the literature. The traditional literature on the diffusion of disease uses a parsimonious model in which every individual is assumed to meet every other person randomly, and the dynamics of the aggregate measure of diffusion are characterized by a few parameters such as the infection rate and meeting rate. (See Newman (2010), chapter 17, for a review of the models and literature.) This approach extends to various approaches of mean-field approximation. (See Jackson and Rogers (2007), Jackson and Yariv (2007) and Young (2009), among others.)

In the literature of statistics, computer science, and economics, a great deal of attention has been paid to the diffusion of information or actions, and to its interaction with network structures (see DeMarzo, Vayanos, and Zwiebel (2003), Leskovec, Adamic, and Huberman (2007), Golub and Jackson (2010), Campbell (2013), and Banerjee, Chandrasekhar, Duflo, and Jackson (2013), to name but a few). A focus from the policy perspective is often the problem of optimal seeding, that is, the problem of finding a group of people who can be seeded with information to diffuse it to the whole population most effectively. For example, motivated by marketing research, Kempe, Kleinberg, and Tardos (2003) provide approximation guarantees for the problem of finding the influence maximizing subset of nodes in threshold and cascade diffusion models. Recent contributions in the economics literature include Banerjee, Chandrasekhar, Duflo, and Jackson (2019), Akbarpour, Malladi, and Saberi (2020), and Beaman, BenYishay, Magruder, and Mobarak (2021). Furthermore, a stream of literature on economic

theory studies strategic interactions in the diffusion process. For example, see Morris (2000), Jackson and Yariv (2007) and Sadler (2020) and references therein.

The potential outcome approach we adopt in this paper is closely related to the literature on network interference in program evaluations. Since treatment always precedes the observed outcomes, a program evaluation set-up can be viewed as a two-period panel environment, where the first period decisions are the treatments and the second period decisions are the outcomes. Recent research in this area includes van der Laan (2014), Aronow and Samii (2017), Athey, Eckles, and Imbens (2018), and Leung (2020). Using a dynamic structural model, van der Laan (2014) proposes a method of causal inference with network interference. In his setting, the cross-sectional dependence of potential outcomes arises solely from overlapping treatment exposures. Aronow and Samii (2017) distinguish treatment assignment (controlled by the random experiment design) and treatment exposure (not controlled by the design). They develop asymptotically valid inference for average treatment effects for network interference, under the assumption that the exposure map is known yet flexible. Athey, Eckles, and Imbens (2018) consider various hypothesis testing problems on treatment effects with network interference. They propose various tests that are exact in finite samples, assuming randomized experiments on the network effects. Leung (2020) studies identification of causal parameters under the assumption that the potential outcomes do not depend on the labels of the nodes in the network. He develops asymptotic inference under weak assumptions that account for the randomness of the underlying graphs. Leung (2022) proposes asymptotic inference on treatment spillover effects which permits a treated unit to have an impact on all other units, with the impact becoming weaker as the other units are farther away from the treated unit in a network.

The literature on social interactions and networks has recognized that recovering the underlying network accurately is difficult in practice (see Banerjee, Chandrasekhar, Duflo, and Jackson (2019) for a discussion and references.) Choi (2017) proposes methods of inference for the average treatment effect under network interference, assuming that the individual treatment effects are monotone, but without requiring any information on the network. de Paula, Rasul, and Souza (2020) propose identification and estimation of the network structure in a linear social interactions model with panel data. Lewbel, Qu, and Tang (2021) consider a linear social interactions model with many disjoint groups of people, where each group is connected under a social network that is not observed. They provide identification analysis and develop estimation methods. Zhang (2020) studies identification and estimation of spillover effects with mismeasured networks, employing the measurement error approach of Hu (2008). Li, Sussman, and Kolacyzyk (2021) analyze the bias and variance of causal effect estimators with mismeasured networks in the framework of Aronow and Samii (2017) and propose a method of moments estimator that reduces the bias.

This paper is organized as follows. Section 2 introduces a generalized diffusion model and the ADM as a causal parameter from the potential outcome perspective. The section provides identification results, and discusses spurious diffusion. Section 3 presents an inference method under the subgraph hypothesis and establishes its asymptotic validity. The section also presents a directional test of the subgraph hypothesis. Section 4 develops inference for the case where the subgraph hypothesis fails. Section 5 discusses the Monte Carlo simulation results. Section 6 presents an empirical application that uses the microfinancing data from Banerjee, Chandrasekhar, Duflo, and Jackson (2013). Section 7 concludes. The Supplemental Note provides the mathematical proofs of the results established in the paper.

2. Measuring Diffusion over a Network

2.1. A Generalized Diffusion Model

Suppose that a set N of people are connected along a simple directed graph called a **contact network** $G_{\text{ctt}} = (N, E_{\text{ctt}})$ (where subscript "ctt" is a mnemonic for "contact"). Edge set E_{ctt} consists of edges ij, where edge ij represents person i being exposed to the influence of person j. Each person $i \in N$ at each time t has two possible states, the default state 0 that everyone is endowed with in the beginning, and the state 1 to which a person may switch. When person i has not yet switched her state, she is exposed to the influence of the previous state-switches of people in her neighborhood (denoted by $N_{\text{ctt}}(i)$) in the contact network.

We assume that once the state is switched, it is *irreversible*. Hence, each person is allowed to switch her state at most once during the observation periods. For example, the binary state-switch may represent an investment in a technology which cannot be reversed during the observed periods. Another example is the decision to abandon a technology permanently. The irreversibility of the state-switch is satisfied trivially by definition when the focus of interest is the diffusion of a *first-time* event, such as the first-time purchase of a product of a new brand.

More formally, initial state-switches, $A_{j,0} \in \{0, 1\}$, are realized for each person j at time $t = 0$, and then the diffusion of state-switches arises subsequently along discrete time $t \in \{1, 2, ...\}$. The state-switch of person i at time t, denoted by $A_{i,t} \in \{0, 1\}$, is a function of $(A_{j,t-1})_{j \in N_{\text{ctt}}(i)}$ and her own state vector $U_{i,t} \in \mathbf{R}^d$: for $t = 1, 2, 3, ...$,

$$
(2.1) \qquad A_{i,t} = \begin{cases} \rho_{i,t}\left((A_{j,t-1})_{j \in N_{\text{ctt}}(i)}, U_{i,t}; G_{\text{ctt}}\right), & \text{if } A_{i,s} = 0, \text{ for all } s = 0, 1, ..., t-1 \\ 0, & \text{otherwise,} \end{cases}
$$

for some map $\rho_{i,t}$, which may depend on the contact network G_{ctt}. If $A_{i,s} = 1$ for some $s = 0, 1, ..., t-1$, we automatically have $A_{i,t} = 0$, by the irreversibility of the state-switch, that is, person i cannot switch her state, once she has switched it previously. For example, in the

context of technology adoption in agriculture, when a farmer i adopted a new technology at time $t-1$ and kept using the technology at time t, we record the state-switches as $A_{i,t-1} = 1$ and $A_{i,t} = 0$. The indicator $A_{i,t-1} = 1$ does *not* represent the state of using the technology, but the *switch* of the state. Hence what triggers the state-switches in our model is neighbors' previous *state-switches*, not their states. Our diffusion model excludes the possibility of state-switches depending on the persistence of a certain state of a neighbor. At each time t, we call each node i a **switcher** at t if $A_{i,t} = 1$, and a **stayer** at t if $A_{i,t} = 0$.

In the literature, diffusion models that accommodate network information can be divided into two categories: threshold diffusion models and cascade diffusion models (Leskovec, Adamic, and Huberman (2007)). In a threshold diffusion model, an individual adopts an action if there are enough people in the person's neighborhood who adopt the same action. This adoption behavior can be due to pressure for conformity. A special case of such a model is a linear threshold model of diffusion (Granovetter (1978))[4]

$$(2.2) \qquad \rho_{i,t}\left((A_{j,t-1})_{j \in N_{\mathrm{ctt}}(i)}, U_{i,t}; G_{\mathrm{ctt}}\right) = 1\left\{ \sum_{j \in N_{\mathrm{ctt}}(i)} p_{ij} A_{j,t-1} \geq U_{i,t} \right\},$$

where p_{ij} is a nonnegative weight that person i assigns to person j. In the linear threshold model (2.2) with $p_{ij} = 1/|N_{\mathrm{ctt}}(i)|$, person i switches the state if she has not yet done so and a large enough fraction of her contact neighbors have switched their states previously.

A cascade model assumes that for each person, her neighbors take turns influencing her action. The model specifies the probability that a neighbor successfully influences a person. Its special case is an independent cascade model, where the probability of a person's successful influence does not depend on the prior history of successes or failures by her neighbors (Goldenberg, Libai, and Muller (2001)). However, as noted by Kempe, Kleinberg, and Tardos (2003), the threshold model and the cascade model, once generalized, can be shown to be equivalent. Our diffusion model in (2.1) encompasses this generalized version.

The generality of our diffusion model allows the view that the map $\rho_{i,t}$ in (2.1) arises from an equilibrium in a strategic game. Consider the diffusion game introduced by Sadler (2020). In this game, at each period t, a player is said to *get exposed*, if one of her neighbors in the contact network switched the state in the previous period. At the beginning of the diffusion process, each player i learns about her private value U_i and the number of her neighbors. Then, she decides whether to be a *potential adopter* or not, by maximizing her expected payoff. When the player becomes a potential adopter, she switches her state once she gets exposed, and this state-switch becomes irreversible. Player i's strategy is represented by map σ_i, which assigns 0 or 1 to (u_i, d_i), so that $\sigma(u_i, d_i) = 1$ indicates that player i with private

---

[4]See Acemoglu, Ozdaglar, and Yildiz (2011) and Banerjee, Chandrasekhar, Duflo, and Jackson (2013) for more recent applications of this model.

value $U_i = u_i$ and $|N_{ctt}(i)| = d_i$ decides to be a potential adopter. At the end of the diffusion process, each player receives a payoff that depends on the private value, the number of her contact neighbors, and the total number of switchers in her contact neighborhood.

Once strategy σ_i is assumed to arise as a perfect Bayesian equilibrium as in Sadler (2020), each player i's actual state-switch is determined as in (2.1) with

$$\rho_{i,t}\left((A_{j,t-1})_{j \in N_{ctt}(i)}, U_i; G_{ctt}\right) = \sigma_i\left(U_i, |N_{ctt}(i)|\right) 1\left\{\sum_{j \in N_{ctt}(i)} A_{j,t-1} > 0\right\}.$$

When player i is a potential adopter (i.e., $\sigma_i(U_i, |N_{ctt}(i)|) = 1$), she switches her state only when she has never switched the state before and gets exposed, that is, $\sum_{j \in N_{ctt}(i)} A_{j,t-1} > 0$.[5] In our context, we assume that we observe diffusion over a single large network. This means that if there are multiple equilibria in this diffusion game, only one equilibrium is involved in generating the data. Our causal inference then focuses on this equilibrium.

Despite its generality, our diffusion model excludes some models of diffusion that are used in the literature. For example, it excludes the model considered in Jackson and Yariv (2007). They use mean-field approximation, where each person's decision to switch the state depends on the previous actions of her contact neighbors only through the probability of her neighbors' actions in the previous period. In such a model, decisions about state-switches are realized through random meetings, rather than through exposure to neighbors' influence on a given network.

2.2. The Researcher's Observation

Throughout the paper, we assume that the researcher observes the state of each person at two points in time: at time $t = 0$ and at time $t = t_1$. We denote person i's state at time 0 as $Y_{i,0}$ and her state at time $t = t_1$ as $Y_{i,1}$. We can relate the observed states to the state-switches as follows:

(2.3) $$Y_{i,0} = A_{i,0}, \text{ and } Y_{i,1} = \sum_{t=0}^{t_1} A_{i,t}.$$

Since each person can switch the state at most once, $Y_{i,1}$ takes a value of either 1 or 0, depending on whether person i has switched the state at some time $t = 0, 1, ..., t_1$ or not. In other words, $Y_{i,1} = 1$ if and only if person i has switched the state by time t_1.

The causal relations between $Y_0 = (Y_{j,0})_{j=1}^n$ and $Y_1 = (Y_{i,1})_{i=1}^n$ are determined by the model (2.1) of diffusion. To express these causal relations, we define a graph $G_{cau} = (N, E_{cau})$, where $ij \in E_{cau}$ if and only if i and j are on a walk from j to i of length less than or equal to t_1 in

---

[5]Since our generalized diffusion model allows the unobserved heterogeneity $U_{i,t}$ to be time-varying, its time-invariance in this game is accommodated as a special case.

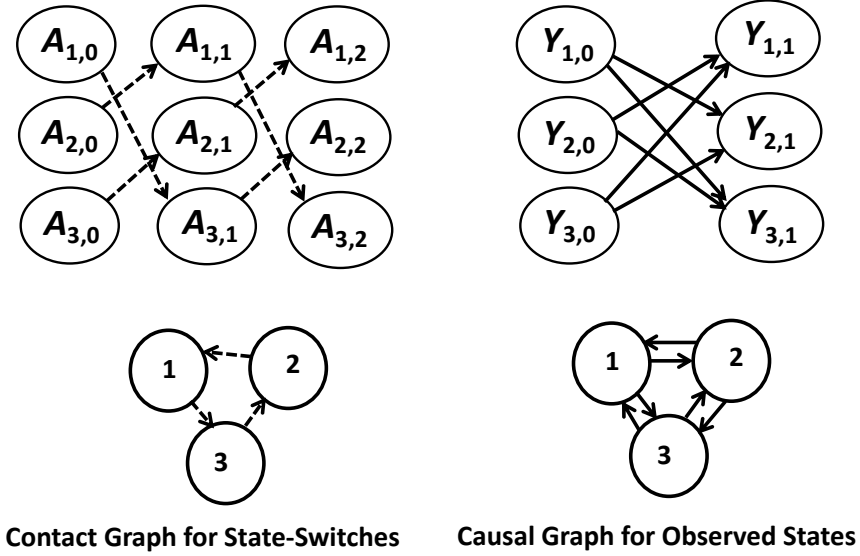**Contact Graph for State-Switches**   **Causal Graph for Observed States**

FIGURE 1. Illustration of a Contact Network and Its Causal Graph: The figure illustrates a contact network with three nodes, $N = \{1, 2, 3\}$, its causal graph at the second observation time being $t_1 = 2$. The edge set of the contact network is given by $\{12, 23, 31\}$. For example, the influence on $A_{1,2}$ is traced back to $A_{3,0}$ and that of $A_{1,1}$ is traced back to $A_{2,0}$. Hence the causal in-neighborhood of node 1 is $\{2, 3\}$. Reasoning similarly with nodes 2 and 3, we obtain the causal graph between (Y_0, Y_1) as one with the edge set $\{12, 21, 23, 32, 31, 13\}$.

the contact network G_{ctt}.[6] We call the graph G_{cau} the ***causal graph*** of (Y_0, Y_1), because, when $ij \notin E_{\text{cau}}$, the state $Y_{i,1}$ does not depend on $Y_{j,0}$ and hence there is no causal effect of $Y_{j,0}$ on $Y_{i,1}$. (See Figure 1 for an illustration of a contact network and its causal graph when $t_1 = 2$.)

In practice, the researcher rarely observes the causal graph G_{cau} perfectly. Throughout the paper, we assume that the researcher does not observe the causal graph G_{cau}, but there is an ***observed graph*** $G_{\text{obs}} = (N, E_{\text{obs}})$ which is a simple, directed or undirected graph on N. In this paper, we require that the causal graph and the observed graph are sparse, as stated in Assumption 3.4 later. On the relationship between the two graphs, we consider two scenarios: one in which the causal graph is a subgraph of the observed graph, and the other in which a monotone-treatment-effect type condition is satisfied. Section 2.4 provides more details on this.

To summarize, the researcher observes $(Y_1, Y_0, X, G_{\text{obs}})$, where $X = (X_i)_{i \in N}$ is a profile of covariate vectors. We regard graphs G_{ctt} and G_{obs} as stochastic. Since we allow the covariates to be arbitrarily dependent across people, each covariate vector X_i may include group specific characteristics, such as the population size of person i's village, or it may include the size of person i's neighborhood or the average of some covariates over person i's neighbors in G_{obs}.

---

[6]Person i at time t_1 can be traced back to person j at time 0 through the contact network, that is, there exist $i_1, i_2, ..., i_{t_1-1}, i_{t_1} = i$ such that $j \in N_{\text{ctt}}(i_1)$ and $i_{s-1} \in N_{\text{ctt}}(i_s)$, for $s = 2, ..., t_1$. The same person can appear repeatedly in the sequence.

Let $\mathcal{C}$ be a vector of common shocks which represents part of the environment shared by all the agents and yet unobserved by the researcher. We take $\mathcal{F}$ to be the σ-field generated by $(G_{\text{ctt}}, G_{\text{obs}}, X, \mathcal{C})$. The σ-field $\mathcal{F}$ represents the environment in which the diffusion of state-switches among people begins.

2.3. Average Diffusion at the Margin

We adopt a potential outcome approach to introduce a measure of diffusion as a causal parameter. Our measure is inspired by the weighted average treatment effects in the program evaluation literature (Hirano, Imbens, and Ridder (2003)). In the standard setting with no spillover effects, the treatment of a person affects the person only. However, in our case with spillover effects, the "treatment" by a person, which corresponds to her initial state-switch from 0 to 1, potentially influences multiple people's outcomes. Thus, we first define the individual influence of each person's initial state-switch as the expected increase in the number of switchers due to the person's initial switch, and then take the measure of diffusion to be a weighted average of the individual influences.

We begin by defining a potential outcome for each person i, when the initial state-switch of a person j in her neighborhood in G_{cau} is counterfactually *fixed* to be $d \in \{0, 1\}$. To define the potential outcome, we first relate $Y_{i,1}$ to her neighbors' initial state-switches by iterating (2.1) recursively up to t_1 so that for some map $\tilde{\rho}_i$, we can write

$$(2.4) \qquad Y_{i,1} = \tilde{\rho}_i\left((Y_{k,0})_{k \in \overline{N}_{\text{cau}}(i)}, V_i; G_{\text{ctt}} \right),$$

where V_i denotes the vector of $U_{k,s}$'s with $k \in \overline{N}_{\text{cau}}(i)$ and $s \leq t_1$, and

$$\overline{N}_{\text{cau}}(i) = N_{\text{cau}}(i) \cup \{i\},$$

with $N_{\text{cau}}(i) = \{j \in N : ij \in E_{\text{cau}}\}$, that is, the in-neighborhood of i in G_{cau}. For $k, j \in N$ and $d \in \{0, 1\}$, we define

$$Y_{k,0}^{[d,j]} = \begin{cases} d, & \text{if } k = j \\ Y_{k,0}, & \text{if } k \neq j, \end{cases}$$

so that $Y_{k,0}^{[d,j]}$ is the same as $Y_{k,0}$ except that it is fixed to be d when $k = j$. The potential outcome for person i with person j's initial state-switch fixed at the "treatment status" d is defined to be

$$Y_{ij}^*(d) = \begin{cases} \tilde{\rho}_i\left((Y_{k,0}^{[d,j]})_{k \in \overline{N}_{\text{cau}}(i)}, V_i; G_{\text{ctt}} \right), & \text{if } j \in \overline{N}_{\text{cau}}(i) \\ Y_{i,1}, & \text{if } j \notin \overline{N}_{\text{cau}}(i). \end{cases}$$

Using the potential outcomes $Y_{ij}^*(d)$, we introduce a measure of diffusion as follows. First, for $j = 1, ..., n$, we define

$$\Delta_j = \sum_{i \in N} \mathbf{E}[Y_{ij}^*(1) - Y_{ij}^*(0) \mid \mathcal{F}],$$

where we recall that $\mathcal{F}$ is the σ-field of $(G_{\text{ctt}}, G_{\text{obs}}, X, \mathcal{C})$, which reflects the environment at the start of diffusion. Then Δ_j represents the expected increase in the number of switchers by time t_1 just due to one person j's being a switcher at $t = 0$, when other people, say, k, choose $Y_{k,0}$, at time $t = 0$.[7] Hence Δ_j measures the individual influence of the initial state-switch by person j on the outcomes of other people by t_1.

Our measure of diffusion is taken to be a weighted average of the individual influences Δ_j over $j \in N$. More specifically, we define

$$\text{ADM} = \sum_{j \in N} w_j \Delta_j,$$

where $w_j \geq 0$ are weights such that $\sum_{j \in N} w_j = 1$. We call this quantity the ***Average Diffusion at the Margin (ADM)***. The ADM measures the expected increase in the number of switchers *just due to* one additional person's being an initial switcher when the person is randomly selected with probability w_j, while the rest of the people, say, k, switch their states according to $Y_{k,0}$ in the initial period.

We recommend using equal weights ($w_j = 1/n$) in most cases. However, it may be of interest to consider non-equal weights in some cases, especially to compare the spillover effect between two demographic groups. For example, the setup could be that for each $j = 1, ..., n$,

$$w_{j,\text{M}} = \frac{1}{n_\text{M}} 1\{j \text{ is male }\} \text{ and}$$

$$w_{j,\text{F}} = \frac{1}{n_\text{F}} 1\{j \text{ is female }\},$$

where n_M and n_F denote the number of males and females, respectively. We define the weight vectors $w_\text{M} = (w_{1,\text{M}}, ..., w_{n,\text{M}})$ and $w_\text{F} = (w_{1,\text{F}}, ..., w_{n,\text{F}})$. Then, by comparing the two ADM's with weights w_M and w_F, we can see which gender group causes more spillover effects in terms of the ADM.

The ADM tends to decrease in the second observation period t_1 if t_1 is large enough. Since there are other initial switchers (other than the additional initial switcher j) and state-switches are irreversible, most people who are susceptible to a switch may have already switched their states after a long while, leaving only a few people susceptible to a switch, regardless of whether the additional person j initially switched her state or not. Therefore,

---

[7] The conditional expectation given $\mathcal{F}$ involves the joint distribution of other people's initial state-switches $Y_{k,0}$, because $Y_{ij}^*(d)$ depends on $Y_{k,0}$'s, $k \in N_{\text{cau}}(i) \setminus \{j\}$.

when t_1 is large, the ADM tends to be close to zero. This means that the ADM with large t_1 may not be an effective measure of the impact of people's initial switch of the state on others.[8]

Unlike a typical two-period network interference setting (e.g., Rosenbaum (2007), Aronow and Samii (2017), van der Laan (2014), and Leung (2020)), we explicitly consider that there may have been progress in diffusion between the two observation periods. In this case, we cannot assume that the unobserved heterogeneity V_i in (2.4) is cross-sectionally independent conditional on $\mathcal{F}$. To see this, suppose that $t_1 = 2$, and

$$(2.5) \qquad i_1' \in N_{\mathrm{ctt}}(i_1), i_2' \in N_{\mathrm{ctt}}(i_2), \text{ and } i'' \in N_{\mathrm{ctt}}(i_1') \cap N_{\mathrm{ctt}}(i_2'),$$

for some $i_1', i_2', i'' \in N$. From model (2.1), we have

$$A_{i,1} = \rho_{i,1}\left((A_{j,0})_{j \in N_{\mathrm{ctt}}(i)}, U_{i,1}; G_{\mathrm{ctt}}\right), \text{ for } t = 1, \text{ and}$$

$$A_{i,2} = \rho_{i,2}\left((A_{j,1})_{j \in N_{\mathrm{ctt}}(i)}, U_{i,2}; G_{\mathrm{ctt}}\right), \text{ for } t = 2.$$

By composing maps $\rho_{i_1,2}$ and $\rho_{j,1}$, $j \in N_{\mathrm{ctt}}(i_1)$, we find that

$$Y_{i_1,1} = \tilde{\rho}_{i_1}\left((Y_{j,0})_{j \in \overline{N}_{\mathrm{cau}}(i_1)}, V_{i_1}; G_{\mathrm{ctt}}\right),$$

for some map $\tilde{\rho}_{i_1}$, where V_{i_1} is a vector that contains $U_{i'',0}$ by (2.5). Arguing similarly for i_2, we obtain $Y_{i_2,1} = \tilde{\rho}_{i_2}\left((Y_{j,0})_{j \in \overline{N}_{\mathrm{cau}}(i_2)}, V_{i_2}; G_{\mathrm{ctt}}\right)$, for some map $\tilde{\rho}_{i_2}$, where again, V_{i_2} contains $U_{i'',0}$. Since both V_{i_1} and V_{i_2} contain $U_{i'',0}$, they cannot be assumed to be independent.

2.4. Identification of the ADM

For the initial state-switches and unobserved heterogeneity in (2.1), we introduce the following assumption.

Assumption 2.1. $(U_j, Y_{j,0})$'s are conditionally independent across j's given $\mathcal{F}$, where $U_j = (U_{j,1}, ..., U_{j,t_1})$.

Assumption 2.1 is satisfied if each individual's decision to switch the state depends only on $\mathcal{F}$ and idiosyncratic shocks that are cross-sectionally independent conditional on $\mathcal{F}$. For the V_i's in (2.4), the assumption allows V_{i_1} and V_{i_2} to be cross-sectionally dependent given $\mathcal{F}$ as long as the in-neighborhoods of i_1 and i_2 in causal graph G_{cau} overlap.

As a causal parameter, the ADM is generally not identified if the initial actions $Y_{j,0}$ are allowed to be correlated with potential outcomes conditional on observables. We introduce

---

[8]This is analogous to a situation where a biologist tries to measure the impact of a chemical on the proliferation rate of bacteria by comparing two petri dishes of bacteria, one treated with the chemical and the other is the control. When measuring the impact of the chemical, the timing of measurement matters; the impact cannot be measured well when the measurement is made after the bacteria have fully multiplied in both dishes. (We thank Jungsun Ghil for helping us with this analogy.)

an analogue of the "unconfoundedness condition" used in the program evaluation literature. (See Imbens and Wooldridge (2009) for an explanation of this condition.)

Assumption 2.2. (i) For all $i, j \in N$ such that $i \neq j$, $(Y_{ij}^*(1), Y_{ij}^*(0), G_{\text{ctt}}, \mathcal{C})$ is conditionally independent of $Y_{j,0}$ given G_{obs}, X.

(ii) For some $c \in (0, 1/2)$, $c \leq \mathbf{E}[Y_{j,0} \mid G_{\text{obs}}, X] \leq 1 - c$, for all $j \in N$ and $n \geq 1$.

Assumption 2.2(i) is satisfied if each person's action $Y_{j,0}$ at time $t = 0$ arises only as a function of (G_{obs}, X) and some random events that are independent of all other components. Assumption 2.2(ii) is a version of the overlap condition, where the "propensity score" $\mathbf{E}[Y_{j,0} \mid G_{\text{obs}}, X]$ is required to be bounded away from zero uniformly over $j \in N$ and over n. This overlap condition is not plausible, for example, if nearly all the agents are initial switchers, or there are only a small number of initial switchers.

For example, suppose that people's initial decisions are triggered by targeted advertisements to which they are exposed, and these advertisements are made based only on their observed characteristics, their network positions in the contact network, and some random, independent shocks. Then, the correlation among their initial decisions is fully explained by the covariates and network positions. In this case, the conditional independence conditions in Assumptions 2.1 and 2.2(i) are plausible. However, these conditions are not satisfied, for example, if a person's initial action $Y_{j,0}$ is based on her observation of other people's characteristics that are not observed by the researcher.

We allow for unobserved common shock $\mathcal{C}$ yet in a limited way. While we permit the common shock $\mathcal{C}$ to be arbitrarily correlated with the potential outcomes, covariates, and both the contact and observed graphs, we require that it is conditionally independent of initial state-switches $Y_{j,0}$ given G_{obs} and X. In other words, if the common shock affects the initial state-switches, it does so only through affecting G_{obs} and X. When this is not the case, Assumption 2.2 can fail. For example, suppose that $Y_{j,0}$'s are generated as follows:

$$Y_{j,0} = 1\{X_j'\gamma_0 + V_0 \geq \varepsilon_j\},$$

where V_0 is an unobserved random variable that is part of the common shock $\mathcal{C}$, and the ε_j's are unobserved random variables that are independent of X and V_0. Furthermore, we assume that V_0 is not a function of X or G_{obs}. Thus, there is an unobserved common shock that influences people's initial state-switches and is not captured by the observed covariates or the observed graph. In this case, the common shock $\mathcal{C}$ fails to be conditionally independent of $Y_{j,0}$ given G_{obs} and X, violating Assumption 2.2.

To describe the condition for the observed graph, it is convenient to introduce the following notion.

Definition 2.1. We say that the observed graph G_{obs} satisfies the ***subgraph hypothesis***, if it contains the causal graph G_{cau} as a subgraph.

If the observed graph satisfies the subgraph hypothesis, it captures all the potential dependencies among observed state-switches that are relevant in measuring the diffusion.[9] The subgraph hypothesis does not require that the observed graph should "approximate" the causal graph G_{cau} in any sense. The hypothesis is substantially weaker than the assumption for the networks used in the linear-in-means models in the literature (e.g., Manski (1993) and Bramoullé, Djebbari, and Fortin (2009)). In these models, the observed network that is used to define the endogenous peer effects should be equal to, or at least "approximate" the true network; it is not enough to assume that the observed network contains the true network as a subgraph. (See de Paula, Rasul, and Souza (2020) and Lewbel, Qu, and Tang (2021) for approaches that do not require information on the networks in the data in the context of linear interactions models.) In Section 3.5, we develop a directional test of the subgraph hypothesis.

We define the following:

$$(2.6) \qquad \mathsf{C} = \sum_{j \in N} \sum_{i \in N : ij \in E_{\text{obs}}} \frac{\text{Cov}(Y_{j,0}, Y_{i,1} \mid \mathcal{F}) w_j}{\text{Var}(Y_{j,0} \mid \mathcal{F})},$$

where E_{obs} denotes the edge set of G_{obs}. The quantity C captures the spatio-temporal dependence of states between the two observation periods. It can be viewed as a measure of the empirical relevance of an observed network G_{obs} in explaining the *residual* cross-sectional covariance pattern between Y_1 and Y_0 after "controlling for covariates and network characteristics". We present our identification result as follows.

Lemma 2.1. *Suppose that Assumptions 2.1 and 2.2 hold.*
(i) If the observed graph G_{obs} satisfies the subgraph hypothesis, then

$$\text{ADM} = \mathsf{C}.$$

(ii) If for each $ij \in E_{\text{cau}}$,

$$(2.7) \qquad \mathbf{E}\left[Y_{ij}^*(1) - Y_{ij}^*(0) \mid \mathcal{F} \right] \geq 0,$$

---

[9]In the Supplemental Note, we introduce weaker conditions based on what we call the *Dependency Causal Graph (DCG)* which captures this notion more precisely. By simply taking G_{obs} to be a complete graph, that is, a graph with an edge for every pair $i, j \in N$, $i \neq j$, we can fulfill the subgraph hypothesis automatically. However, such an option is excluded by a sparsity condition in Assumption 3.4. We exclude such a graph primarily because many observed edges may induce extensive dependence among observed outcomes in a way that invalidates the asymptotic inference we develop later.

then

$$\text{ADM} \geq \mathsf{C}.$$

Lemma 2.1 shows that when the observed graph satisfies the subgraph hypothesis, the ADM is identified as a weighted sum of covariances under Assumption 2.2. When Assumption 2.2 is in doubt, the inference result on C is interpreted in non-causal terms, only as a measure of covariation between the outcomes in one period with their neighboring outcomes in the next period. The quantity C is related to the population version of the well-known inverse probability weighted estimator in the program evaluation literature:

$$(2.8) \qquad \mathsf{C} = \sum_{j \in N} \sum_{i \in N : ij \in E_{\text{obs}}} \mathbf{E}\left[\frac{Y_{j,0} Y_{i,1}}{\mu_{j,0}} - \frac{(1 - Y_{j,0}) Y_{i,1}}{1 - \mu_{j,0}} \mid \mathcal{F} \right] w_j,$$

where $\mu_{j,0} = \mathbf{E}[Y_{j,0} \mid \mathcal{F}]$. Here we can see that $Y_{j,0}$ plays the role of the treatment indicator and $\mu_{j,0}$ that of the propensity score in a program evaluation setting.

When the observed graph does not satisfy the subgraph hypothesis, but condition (2.7) is satisfied, C is a lower bound for the ADM. Condition (2.7) is a variant of the monotone treatment effect assumption used by Manski (1997) and Choi (2017). The condition requires that the spillover effect be nonnegative in the sense that when one of person i's in-neighbors in the causal graph becomes a switcher at time $t = 0$, this should not make it less likely that person i becomes a switcher by time t_1.

The intuition behind the result of Lemma 2.1(ii) is as follows. By Assumptions 2.1 and 2.2, the conditional covariance between $Y_{j,0}$ and $Y_{i,1}$ given $\mathcal{F}$ is non-zero only if $ij \in E_{\text{cau}} \cap E_{\text{obs}}$. (For all $ij \in E_{\text{obs}} \setminus E_{\text{cau}}$, $Y_{j,0}$ and $Y_{i,1}$ have zero conditional covariance.) Furthermore, the terms in the sum in C do not include those j such that $ij \in E_{\text{cau}} \setminus E_{\text{obs}}$, which influence i non-negatively by (2.7) and contribute to the ADM. Since C does not include these nonnegative effects, it is a lower bound for the ADM.

2.5. Spurious Diffusion

In the literature on diffusion or network interference, it has been noted that when the networks are formed homophilously, that is, people with similar characteristics becoming neighbors with each other, the observed correlation of actions between people who are adjacent in the network may not represent the spillover effect.[10] For example, suppose that $Y_{i,1}$ represents the indicator of the i-th student purchasing a new smartphone, and G_{obs} represents

---

[10]For example, Aral, Muchnik, and Sundararajan (2009) develop methods to distinguish between influence-based contagion and homophily-based diffusion. They find that the peer influence is generally overestimated when the homophily effect is ignored. Shalizi and Thomas (2011) point out challenges arising from the confounding of social contagion, homophily, and the influence of individual traits.

the friendship network of the students. Suppose further that the friendship network exhibits homophily along the income of the student's parents, and that a student from a high-income household is more likely to purchase a new smartphone than a student from a low-income household. In this case, even if there is no diffusion of new smartphone purchases along a friendship network, the estimated diffusion without controlling for heterogeneity in income may be significantly different from zero. When the measured diffusion is not zero only because of failings to condition on certain covariates, we refer to this phenomenon as **spurious diffusion**.

The primary reason for spurious diffusion under omitted covariates is that the unconfoundedness condition in Assumption 2.2 may fail when we omit covariates. However, the way the consequence of omitting covariates arises in the context of diffusion over a network is different from the standard setting of causal inference. In the standard setting of causal inference with no spillover effects, omitting covariates that affect both the potential outcomes and the treatment variable can cause asymptotic bias in the estimated causal effect. In the context of diffusion over a network, however, the treatment variable ($Y_{j,0}$) and the outcome variable ($Y_{i,1}$) do not belong to the same cross-sectional unit, that is, $i \neq j$. Hence, as we show below, spurious diffusion due to omitted variables arises only when the covariates exhibit cross-sectional dependence conditional on the contact and observed graphs.

To formalize this observation, let $\mathbb{S} = \{1, ..., p\}$, where p denotes the dimension of X_i. For some $S \subset \mathbb{S}$, we let the vector X_i be decomposed into two subvectors $X_{i,S}$ and $X_{i,-S}$, where $X_{i,S}$ denotes the subvector of X_i having its entry indices restricted to S, and $X_{i,-S}$ the subvector of X_i whose entries are not in $X_{i,S}$. Suppose further that the researcher observes only $X_{-S} = (X_{i,-S})_{i \in N}$. Let C_{-S} be the same as C, except that X_{-S} replaces X, that is, the covariates in X_S are omitted. Without observing X_S, we can recover C_{-S}, but not C, from data. The following proposition shows that when there is no cross-sectional dependence among the $(U_i, Y_{i,0}, X_i)$'s given $(G_{\mathrm{obs}}, G_{\mathrm{ctt}}, \mathcal{C})$, there is no spurious diffusion.

Proposition 2.1. *Suppose that there is no diffusion in the sense that state-switches, $A_{i,t}$, are generated as follows: for $t = 1, 2, 3, ...,$*

$$(2.9) \qquad A_{i,t} = \begin{cases} \rho_{i,t}\left(U_{i,t}; G_{\mathrm{ctt}}\right), & \text{if } A_{i,s} = 0, \text{ for all } s = 0, 1, ..., t-1 \\ 0, & \text{otherwise.} \end{cases}$$

Suppose further that $(U_i, Y_{i,0}, X_i)$'s are conditionally independent across i's given $(G_{\mathrm{obs}}, G_{\mathrm{ctt}}, \mathcal{C})$. Then there is no spurious diffusion from omitting X_S, that is,

$$\mathsf{C}_{-S} = 0.$$

The requirement that the $(U_i, Y_{i,0}, X_i)$'s are conditionally independent given $(G_{\mathrm{obs}}, G_{\mathrm{ctt}}, \mathcal{C})$ is stronger than Assumption 2.1, as it includes the cross-sectional conditional independence

of the X_i's given $(G_{\text{obs}}, G_{\text{ctt}}, \mathcal{C})$. The proposition shows the role of cross-sectional dependence of covariates in creating spurious diffusion.[11]

Cross-sectional dependence of the covariates is not merely a theoretical concern. If the contact network is formed homophilously, that is, with edges formed between people with similar observed characteristics, then, conditional on the two people being neighbors, their observed characteristics are correlated. Hence, when the contact network is formed homophilously, it creates cross-sectional dependence conditional on the network, and omitting part of the covariates can create spurious diffusion.

3. Inference on Diffusion under the Subgraph Hypothesis

3.1. Estimation of the ADM

We saw that when G_{obs} satisfies the subgraph hypothesis, the ADM is identified as C which was defined in (2.6). Here we consider the estimation of C. First, recall the definition: $\mu_{j,0} = \mathbf{E}[Y_{j,0} \mid \mathcal{F}]$. By Assumption 2.2(i), we have

$$\mu_{j,0} = \mathbf{E}[Y_{j,0} \mid G_{\text{obs}}, X].$$

We let $\sigma_{j,0}^2 = \mu_{j,0}(1 - \mu_{j,0})$ and rewrite

$$\mathsf{C} = \sum_{j \in N} \frac{w_j}{\sigma_{j,0}^2} \sum_{i \in N: ij \in E_{\text{obs}}} \mathbf{E}\big[(Y_{i,1} - c_i)(Y_{j,0} - \mu_{j,0}) \mid \mathcal{F}\big],$$

where c_i is any random variable that is $\mathcal{F}$-measurable, taking values in $[0,1]$. Later, we will provide specifications of $\mu_{j,0}$ and c_i which we can estimate consistently.[12] Once we have consistent estimators $\hat{c}_i$ and $\hat{\mu}_{j,0}$ of c_i and $\mu_{j,0}$, we construct a sample analogue estimator of C as follows:

$$\hat{\mathsf{C}} = \sum_{j \in N} \frac{w_j}{\hat{\sigma}_{j,0}^2} \sum_{i \in N: ij \in E_{\text{obs}}} (Y_{i,1} - \hat{c}_i)(Y_{j,0} - \hat{\mu}_{j,0}),$$

where $\hat{\sigma}_{j,0}^2 = \hat{\mu}_{j,0}(1 - \hat{\mu}_{j,0})$. In the next subsection, we develop inference on C using this estimator.

---

[11]See Song (2022) for a decomposition method that can be used in empirically investigating the role of covariates in creating spurious diffusion.

[12]It would be ideal if we could find a specification of $\mathbf{E}[Y_{i,1} \mid \mathcal{F}]$ that is consistently estimable and use it in place of c_i. However, it is difficult to find such a specification that is compatible with the generalized diffusion model in (2.1) because $Y_{i,1}$ arises through the diffusion process over a complex, unobserved contact network.

Alternatively, we could define $\hat{C}$ simply by setting $\hat{c}_i = 0$, and prove that this alternative estimator is consistent for C. Then this becomes the sample analogue of (2.8):

$$\sum_{j \in N} \sum_{i \in N : ij \in E_{\text{obs}}} \left(\frac{Y_{j,0} Y_{i,1}}{\hat{\mu}_{j,0}} - \frac{(1 - Y_{j,0}) Y_{i,1}}{1 - \hat{\mu}_{j,0}} \right) w_j.$$

However, our unreported simulation studies show that the choice of c_i and $\hat{c}_i$ that we propose later (see (3.9) and (3.10) below) leads to substantially shorter confidence intervals than this alternative estimator. Furthermore, the use of c_i and $\hat{c}_i$ that we propose does not complicate the asymptotic inference procedure much because the estimation error in $\hat{c}_i$ does not influence the asymptotic variance of estimator $\hat{C}$ after an appropriate scale-location normalization.

3.2. Confidence Intervals for C

The ADM, as identified as C under Lemma 2.1(i), measures the diffusion among a finite set of people, N, and naturally depends on the sample size n. To construct a confidence interval for C, we employ asymptotic approximation as the size of the contact network n grows to infinity. In other words, we aim to construct a confidence interval whose coverage probability is bounded from below by a given level $1 - \alpha$ up to an error that goes to zero as $n \to \infty$.

We first provide an overview of the main idea of constructing confidence intervals. Details on the estimators used here will follow in Section 3.3, where we present a concrete example of the conditional mean specification for $Y_{j,0}$ and a choice of c_i. Define

$$(3.1) \qquad a_i = \sum_{j \in N : ij \in E_{\text{obs}}} \frac{n w_j (Y_{j,0} - \mu_{j,0})}{\sigma_{j,0}^2} \text{ and } \hat{a}_i = \sum_{j \in N : ij \in E_{\text{obs}}} \frac{n w_j (Y_{j,0} - \hat{\mu}_{j,0})}{\hat{\sigma}_{j,0}^2},$$

where estimators $\hat{\mu}_{j,0}$ and $\hat{\sigma}_{j,0}^2$ are those used in the construction of $\hat{C}$. Then, under the conditions stated below, we can show the asymptotic linear representation:

$$(3.2) \qquad \sqrt{n}(\hat{C} - C) = \frac{1}{\sqrt{n}} \sum_{i \in N} (q_i - \mathbf{E}[q_i \mid \mathcal{F}]) + o_P(1),$$

as $n \to \infty$, where

$$q_i = (Y_{i,1} - c_i) a_i + \psi_i,$$

and ψ_i is a random variable that appears in the following representation:

$$(3.3) \qquad \frac{1}{\sqrt{n}} \sum_{i \in N} (Y_{i,1} - c_i)(\hat{a}_i - a_i) = \frac{1}{\sqrt{n}} \sum_{i \in N} (\psi_i - \mathbf{E}[\psi_i \mid \mathcal{F}]) + o_P(1).$$

Representation (3.2) suggests the construction of confidence intervals as follows.

Using an estimator $\hat{\psi}_i$ of ψ_i, we define[13]

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i_1, i_2 \in N : \overline{N}_{\text{obs}}(i_1) \cap \overline{N}_{\text{obs}}(i_2) \neq \varnothing} (\hat{q}_{i_1} - \hat{h}_{i_1})(\hat{q}_{i_2} - \hat{h}_{i_2}),$$

where $\overline{N}_{\text{obs}}(i) = N_{\text{obs}}(i) \cup \{i\}$ with $N_{\text{obs}}(i) = \{j \in N : ij \in E_{\text{obs}}\}$,

$$\hat{q}_i = (Y_{i,1} - \hat{c}_i)\hat{a}_i + \hat{\psi}_i, \text{ and } \hat{h}_i = \hat{\lambda}' X_i,$$

and

$$\hat{\lambda} = \left(\frac{1}{n} \sum_{i \in N} X_i X_i' \right)^{-1} \frac{1}{n} \sum_{i \in N} X_i \hat{q}_i.$$

Then, the $(1 - \alpha)$-level confidence interval for C is given by

$$\mathbb{C}_{1-\alpha} = \left[\hat{C} - \frac{z_{1-\alpha/2} \hat{\sigma}}{\sqrt{n}}, \ \hat{C} + \frac{z_{1-\alpha/2} \hat{\sigma}}{\sqrt{n}} \right],$$

where $z_{1-\alpha/2}$ is the $1 - (\alpha/2)$ percentile of $N(0, 1)$.

To see the motivation behind this method of constructing confidence intervals, note first that we would ideally like to estimate the following quantity consistently:

$$(3.4) \qquad \sigma^2 = \text{Var}\left(\frac{1}{\sqrt{n}} \sum_{i \in N} q_i \mid \mathcal{F} \right).$$

However, since G_{ctt} is not observed, we cannot estimate $\mathbf{E}[q_i \mid \mathcal{F}]$ consistently without making further assumptions. Such assumptions will have implications for the underlying process of diffusion over the contact network. Instead, we consider the following linear projection

$$(3.5) \qquad h_i = \lambda' X_i,$$

where X_i includes an intercept term, and

$$\lambda = \left(\frac{1}{n} \sum_{i \in N} X_i X_i' \right)^{-1} \frac{1}{n} \sum_{i \in N} X_i \mathbf{E}[q_i \mid \mathcal{F}].$$

While this use of linear projection makes the inference conservative, it makes the procedure relatively simple and seems to work well in our simulation study.[14]

---

[13]As noted by Kojevnikov (2021), $\hat{\sigma}^2$ is always positive semidefinite.

[14]Instead of using a linear projection on X_i, one may also use $\Phi(X_i)$, where $\Phi : \mathbf{R}^{d_X} \to \mathbf{R}^K$ is a vector of nonlinear functions, such as basis functions in the series estimation. As we are not estimating a nonparametric function, the number K of basis functions does not need to increase with the sample size n.

3.3. Specification of the Conditional Mean of $Y_{j,0}$ and Choice of c_i

To construct a consistent estimator of $\mu_{j,0}$, we make the following assumption: for all $j \in N$,

$$\text{(3.6)} \qquad \mathbf{E}[Y_{j,0} \mid G_{\text{obs}}, X] = F_0(X_j' \gamma_0),$$

where F_0 is a known distribution function. The covariate X_j can include network characteristics of G_{obs}, such as the average degree of agent j or that of her neighbors. A part of the covariate vector X_j can also include the average of certain characteristics of person j's neighbors in G_{obs}. Then we obtain the estimator of $\mu_{j,0}$ as follows:

$$\text{(3.7)} \qquad \hat{\mu}_{j,0} = F_0(X_j' \hat{\gamma}),$$

where $\hat{\gamma}$ is estimated using maximum likelihood estimation (MLE), that is,

$$\text{(3.8)} \qquad \hat{\gamma} = \arg\max_{\gamma} \sum_{j \in N} \left(Y_{j,0} \log F_0(X_j' \gamma) + (1 - Y_{j,0}) \log(1 - F_0(X_j' \gamma)) \right).$$

Under the usual conditions, we can show that $\hat{\gamma}$ is consistent and asymptotically normal. (See Section C in the Supplemental Note for details.)

For c_i, we consider the following choice:

$$\text{(3.9)} \qquad c_i = F_1(X_i' \beta^*),$$

where F_1 is again a known distribution function (such as the distribution function of $N(0,1)$), and[15]

$$\beta^* = \arg\max_{\beta} \sum_{i \in N} \mathbf{E}\left[Y_{i,1} \log F_1(X_i' \beta) + (1 - Y_{i,1}) \log(1 - F_1(X_i' \beta)) \mid \mathcal{F} \right].$$

Then we take the following as an estimator of c_i:

$$\text{(3.10)} \qquad \hat{c}_i = F_1(X_i' \hat{\beta}),$$

where $\hat{\beta}$ is a quasi-MLE defined as

$$\text{(3.11)} \qquad \hat{\beta} = \arg\max_{\beta} \sum_{i \in N} \left(Y_{i,1} \log F_1(X_i' \beta) + (1 - Y_{i,1}) \log(1 - F_1(X_i' \beta)) \right).$$

In this case, the quantity ψ_i in (3.3) takes the following form:

$$\text{(3.12)} \qquad \psi_i = \Gamma_n \times \pi_{i,0} f_0(X_i' \gamma_0) X_i,$$

---

[15]It important to note that this choice of c_i is *not* based on any parametric specification of $\mathbf{E}[Y_{i,1} \mid \mathcal{F}]$. As mentioned before, any choice of c_i that is $\mathcal{F}$-measurable can be used as long as we can use its estimator to construct a consistent and asymptotically normal estimator of C. Our choice in (3.9) is motivated by its simplicity and good finite sample behavior (e.g., as compared to the choice of $c_i = 0$).

where, with f_0 denoting the density of F_0, $\pi_{i,0} = (Y_{i,0} - \mu_{i,0})/\sigma_{i,0}^2$,

$$(3.13) \qquad \Gamma_n = \left(\frac{1}{n} \sum_{i \in N} \mathbf{E}\left[(Y_{i,1} - c_i)\kappa_i \mid \mathcal{F} \right] \right) \times \left(\frac{1}{n} \sum_{i \in N} \frac{f_0^2(X_i'\gamma_0)}{\sigma_{i,0}^2} X_i X_i' \right)^{-1},$$

and

$$\kappa_i = \sum_{j \in N_{\text{obs}}(i)} \frac{n w_j f_0(X_j'\gamma_0)X_j'}{\sigma_{j,0}^2} \left(\pi_{j,0}(2\mu_{j,0} - 1) - 1 \right).$$

We can obtain the estimator $\hat{\psi}_i$ of ψ_i by taking

$$(3.14) \qquad \hat{\psi}_i = \hat{\Gamma}_n \times \hat{\pi}_{i,0} f_0(X_i'\hat{\gamma}) X_i,$$

where, with $\hat{\sigma}_{j,0}^2 = \hat{\mu}_{j,0}(1 - \hat{\mu}_{j,0})$, $\hat{\pi}_{i,0} = (Y_{i,0} - \hat{\mu}_{i,0})/\hat{\sigma}_{i,0}^2$,

$$\hat{\Gamma}_n = \left(\frac{1}{n} \sum_{i \in N} (Y_{i,1} - \hat{c}_i)\hat{\kappa}_i \right) \times \left(\frac{1}{n} \sum_{i \in N} \frac{f_0^2(X_i'\hat{\gamma})}{\hat{\sigma}_{i,0}^2} X_i X_i' \right)^{-1},$$

and

$$\hat{\kappa}_i = \sum_{j \in N_{\text{obs}}(i)} \frac{n w_j f_0(X_j'\hat{\gamma})X_j'}{\hat{\sigma}_{j,0}^2} \left(\hat{\pi}_{j,0}(2\hat{\mu}_{j,0} - 1) - 1 \right).$$

3.4. Asymptotic Validity

Here we present the result that shows that the confidence intervals $\mathbb{C}_{1-\alpha}$ are asymptotically valid and introduce conditions for the estimators.

Assumption 3.1.

$$\frac{1}{\sqrt{n}} \sum_{i \in N} (Y_{i,1} - c_i)(\hat{a}_i - a_i) = \frac{1}{\sqrt{n}} \sum_{i \in N} (\psi_i - \mathbf{E}[\psi_i \mid \mathcal{F}]) + o_P(1), \text{ and}$$

$$\frac{1}{\sqrt{n}} \sum_{i \in N} (\hat{c}_i - c_i)a_i = o_P(1),$$

where ψ_i is a random variable such that for some $C > 0$ that does not depend on n,

$$(3.15) \qquad \max_{i \in N} \mathbf{E}\left[\psi_i^4 \mid \mathcal{F} \right] \le C \max_{i \in N} (|N_{\text{ctt}}(i)| + |N_{\text{obs}}(i)|)^4,$$

and ψ_i is measurable with respect to $\sigma\left(Y_{j,0} : j \in \overline{N}_{\text{cau}}(i) \cup \overline{N}_{\text{obs}}(i) \right) \vee \mathcal{F}$.[16]

Assumption 3.2. (i) There exists $C > 0$ such that $|\hat{\mu}_{j,0}| \le C$ for all $j \in N$ and all $n \ge 1$.

---

[16]Given two σ-fields, $\mathcal{F}_1$ and $\mathcal{F}_2$, we define $\mathcal{F}_1 \vee \mathcal{F}_2$ to be the smallest σ-field that contains both $\mathcal{F}_1$ and $\mathcal{F}_2$.

(ii) There exists a constant $\kappa > 0$ such that as $n \to \infty$,

$$\max_{j \in N} |\hat{\mu}_{j,0} - \mu_{j,0}| + \max_{i \in N} |\hat{c}_i - c_i| + \max_{i \in N} |\hat{\psi}_i - \psi_i| = O_P(n^{-\kappa}).$$

(iii) There exists a constant $\overline{w} > 0$ such that $w_j \le \overline{w}/n$ for all $j \in N$ and all $n \ge 1$.

Assumption 3.1 is a set of high level conditions for $\hat{\mu}_{j,0}$ and $\hat{c}_i$. These conditions are satisfied by the specifications of $\mu_{j,0}$ and c_i which admit $\sqrt{n}$-consistent and asymptotically normal estimators. Asymptotic normality is obtained when the contact network is not too dense. (For the precise condition of the denseness of the contact network and the observed graph, see Assumption 3.4, and the discussion that follows.) Assumption 3.2(i) requires that $\hat{\mu}_{j,0}$ is bounded uniformly over j and $n \ge 1$. This is a mild condition that is mostly satisfied, when $\mu_{j,0}$ is parametrically specified as explained previously. For example, the choice $\hat{\mu}_{j,0}$ as in (3.7) immediately satisfies this condition. Assumption 3.2(ii) requires that the estimation errors of $\hat{\mu}_{j,0}$, $\hat{c}_i$ and $\hat{\psi}_i$ vanish in probability at a certain polynomial rate in n uniformly over $i \in N$. We provide lower level conditions for Assumptions 3.1 and 3.2 in the Supplemental Note, using the setting in Section 3.3.

Here we introduce an assumption that ensures nondegeneracy of the asymptotic distribution. Recall the definition σ^2 in (3.4).

Assumption 3.3. (i) There exists small $c' > 0$ such that for all $n \ge 1$,

$$\sigma^2 > c' \text{ and } \lambda_{\min}\left(\frac{1}{n}\sum_{i \in N} X_i X_i'\right) > c',$$

where $\lambda_{\min}(A)$ for a symmetric matrix A denotes the smallest eigenvalue of A.

(ii) There exists $C > 0$ such that for all $n \ge 1$ and $i \in N$, $\|X_i\|^2 \le C$.

Assumption 3.3(i) ensures the nondegeneracy of the distribution of the leading term in the asymptotic linear representation in (3.2). This condition is violated, if the randomness of q_i (conditional on $\mathcal{F}$) disappears as $n \to \infty$. Since the conditional distribution of q_i given $\mathcal{F}$ is highly unlikely to be degenerate in finite samples, it seems reasonable to rely on its asymptotic approximation using Assumption 3.3(i). The bounded support condition in Assumption 3.3(ii) has been used in the literature (see, for example, Hirano, Imbens, and Ridder (2003).) This condition can be relaxed to conditions involving moment conditions for the covariates, yet without adding insights.

The following assumption requires the contact network and the observed graph to be sparse.

Assumption 3.4. (i) For some $C > 0$,

$$\max_{i \in N} \left(|N_{\text{ctt}}(i)| + |N_{\text{obs}}(i)| \right) = O_P \left((\log n)^C \right).$$

(ii) There exists $\tilde{c} > 0$ such that

$$\frac{1}{n} \sum_{i \in N} |N_{\text{ctt}}(i)| \geq \tilde{c},$$

for all $n \geq 1$.

Assumption 3.4(i) requires that the maximum degrees of the contact network and the observed graph are allowed to grow at most at a polynomial rate of the logarithm of the network size. While it is possible to relax this condition to a more complex, weaker condition, we nonetheless require this type of constraint on the denseness of the networks to build asymptotic inference for the ADM. Assumption 3.4(ii) excludes the case where the contact network becomes dominated by isolated nodes asymptotically. This assumption simplifies the notation and is innocuous in our setting because we can simply redefine the contact network to consist of non-isolated nodes.

We study this assumption in terms of the primitives in a general network formation model. For simplicity, we focus on the contact network only. (The remarks below apply to the observed graph similarly.) Suppose that for each $t = 1, 2, ..., t_1$, the contact neighborhoods $N_{\text{ctt}}(i)$, $i \in N$, are formed by the following rule: for $i, j \in N$,

(3.16) $$j \in N_{\text{ctt}}(i) \text{ if and only if } \varphi_{n,ij}(X, \eta) \geq v_{ij},$$

where $\eta = (\eta_i)_{i \in N}$ is a profile of individual-specific shocks, $\varphi_{n,ij}$ are nonstochastic maps, and v_{ij} are edge-specific shocks that are conditionally independent across $i, j \in N$ given (X, η). Suppose that for some $k > 0$,

(3.17) $$\max_{i \in N} \sum_{j \in N \setminus \{i\}} P \left\{ \varphi_{n,ij}(X, \eta) \geq v_{ij} \mid X, \eta \right\} = O_P \left((\log n)^k \right).$$

This rate condition requires that the maximum expected number of neighbors in the contact network should not grow too fast, as $n \to \infty$. Then, we can use a result from Kojevnikov, Marmer, and Song (2021) and show that Assumption 3.4 is satisfied for the contact network G_{ctt}. (See Lemma B.5 in the Supplemental Note.) The network formation model (3.16) is generic, encompassing various network formation models considered in the literature. One example is the following specification:

$$\varphi_{n,ij}(X, \eta) = u_n(X_i, X_j) + \eta_i + \eta_j,$$

where $u_n(X_i, X_j)$ is a term that involves the homophily component in the payoff function of the agents. (See Section 2.3.2 in Kojevnikov, Marmer, and Song (2021) for further discussion and references on this model.)[17]

The following theorem establishes that the confidence interval $\mathbb{C}_{1-\alpha}$ is asymptotically valid.

Theorem 3.1. *Suppose that Assumptions 2.1-3.4 hold, and G_{obs} satisfies the subgraph hypothesis. Then,*

$$\liminf_{n \to \infty} P\left\{\text{ADM} \in \mathbb{C}_{1-\alpha}\right\} \geq 1 - \alpha.$$

3.5. A Directional Test of the Subgraph Hypothesis

3.5.1. **Directional Test.** A crucial condition for the observed graph G_{obs} that delivers the point-identification of the ADM as C is that the observed graph G_{obs} satisfies the subgraph hypothesis. In this subsection, we develop a test of the following hypothesis:

H_0: The observed graph G_{obs} satisfies the subgraph hypothesis.
H_1: The observed graph G_{obs} does not satisfy the subgraph hypothesis.

The main challenge for testing this hypothesis is that there are many ways in which the null hypothesis can be violated. This means that we may not be able to develop a test that ensures nontrivial power against all the ways in which the null hypothesis is violated.[18] Hence, we develop a simple directional test of H_0, which is designed to have nontrivial power against only certain violations of the subgraph hypothesis that lead to ADM $\neq$ C.

The main idea is as follows. First, under Assumptions 2.1 and 2.2, we can write

(3.18)
$$\text{ADM} = \text{C} + \Delta\left(E_{\text{cau}} \setminus E_{\text{obs}}\right),$$

where, for a set E of edges, we define

$$\Delta(E) = \sum_{i,j \in N : ij \in E} \frac{\text{Cov}(Y_{j,0}, Y_{i,1} \mid \mathcal{F}) w_j}{\text{Var}(Y_{j,0} \mid \mathcal{F})}.$$

(Recall that E_{cau} denotes the set of edges in the causal graph G_{cau} and E_{obs} the set of edges in the observed graph G_{obs}.) The decomposition in (3.18) shows that any discrepancy between

---

[17]This network formation model does not accommodate the model in Ridder and Sheng (2020), where the payoff specification captures preference externalities from indirect friends, and the payoff parameters do not depend on the sample size.

[18]It is well known in the literature of nonparametric tests that any omnibus test has nontrivial power against all but a finite number of directions of violations (see Janssen (2000)). Given such a result, we are not confident that it is possible to develop a test that shows a nontrivial power against all possible violations of the null hypothesis that G_{obs} satisfies the subgraph hypothesis.

ADM and C arises due to the presence of unobserved edges ij in G_{cau}. Under the subgraph hypothesis, we have $E_{\text{cau}} \setminus E_{\text{obs}} = \varnothing$ so that

$$\Delta\left(E_{\text{cau}} \setminus E_{\text{obs}}\right) = 0.$$

Therefore, if $\Delta\left(E_{\text{cau}} \setminus E_{\text{obs}}\right) \neq 0$ (that is, ADM $\neq$ C), this indicates a violation of the subgraph hypothesis due to the presence of edges in $E_{\text{cau}} \setminus E_{\text{obs}}$. An ideal directional test would be one that tests whether $\Delta(E_{\text{cau}} \setminus E_{\text{obs}}) = 0$ or not. However, this test is infeasible because we do not observe the causal graph G_{cau}. Hence, we propose choosing a surrogate graph $G' = (N, E')$ that is disjoint with the observed graph $G_{\text{obs}} = (N, E_{\text{obs}})$ (that is, $E' \cap E_{\text{obs}} = \varnothing$). (We will discuss how we construct such a graph G' later.)

Let C$'$ be the same as C in (2.6) except that E_{obs} is replaced by E'. Then, we can write

$$\text{C}' = \Delta\left(E_{\text{cau}} \cap E'\right).$$

Since $E_{\text{cau}} \cap E'$ is contained in $E_{\text{cau}} \setminus E_{\text{obs}}$, under the null hypothesis, we have $E_{\text{cau}} \cap E' = \varnothing$ and, hence, C$' = 0$. If C$' \neq 0$, it implies a violation of the subgraph hypothesis. Thus we propose testing the subgraph hypothesis by testing whether C$' = 0$ or not. More specifically, our directional test proceeds as follows:

$$\text{Reject } H_0, \text{ if and only if } \left|T(G')\right| > z_{1-\alpha/2},$$

where

$$T(G') = \frac{\sqrt{n}\hat{\text{C}}'}{\hat{\sigma}'},$$

and $\hat{\text{C}}'$ and $\hat{\sigma}'$ are the same as $\hat{\text{C}}$ and $\hat{\sigma}$ except that G' replaces G_{obs}.

This test is a directional test intended to detect only such an alternative hypothesis that G' contains edges in the causal graph G_{cau} that are not recorded in the observed graph G_{obs}. Roughly speaking, the directional test detects a violation of the subgraph hypothesis by checking whether there are missing causal relationships between $Y_{j,0}$ and $Y_{i,1}$ with $ij \in E'$, which lead to nonzero covariances $\text{Cov}(Y_{j,0}, Y_{i,1} \mid \mathcal{F})$.[19] From here on, we call the alternative graph G' the **direction** of the test.

3.5.2. **Choice of a Direction.** Here, we consider the method of choosing G' in practice. Intuitively, it is desirable to choose a direction G' such that if the subgraph hypothesis fails, it is primarily due to the edges in G'. Even when the subgraph hypothesis fails, the observed graph may still contain information on the causal graph. Hence, it is likely that even if an

---

[19]See Stute (1997) and Escanciano and Song (2010) for the approach of directional tests in nonparametric and semiparametric testing. Being a directional test, our test is not designed to detect all kinds of violations of the null hypothesis. As a consequence, the main merit of this test is that when one rejects the null hypothesis, one knows that it is because of the edges in G' that belong to G_{cau} but are not captured by G_{obs}.
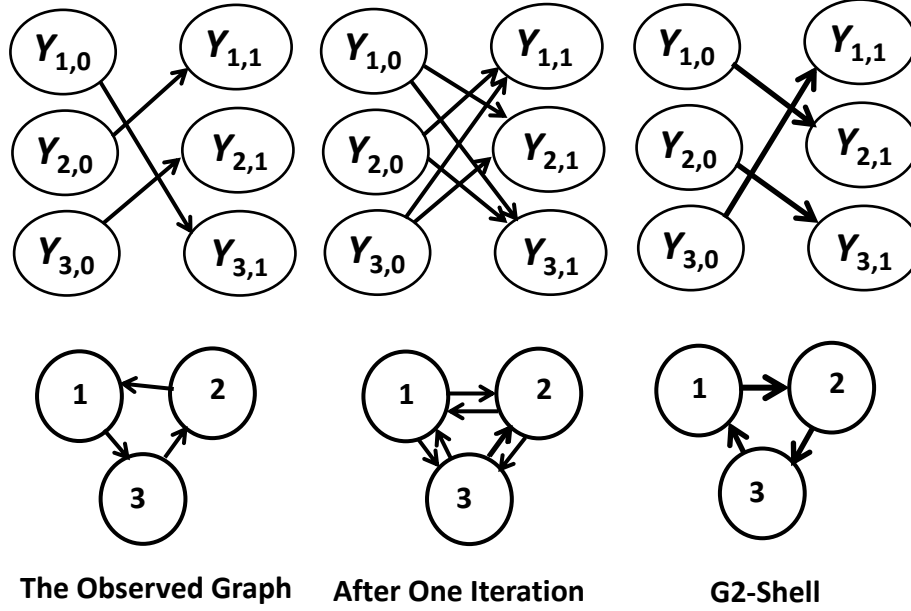
FIGURE 2. Constructing the G2-Shell from the Observed Graph: The figure illustrates the construction of the G2-shell from the observed graph, G_{obs}, with three nodes, $N = \{1, 2, 3\}$ and the edge set $E_{\text{obs}} = \{12, 23, 31\}$. First, we construct $G_{\text{obs},2}$ by adding an edge between two nodes that are two-edges away from each other in G_{obs}. For example, we add edge 21 (that is, directed edge from 1 to 2) because, after one iteration of G_{obs}, vertex 2 is linked back to vertex 3 which is linked back to vertex 1. Then the G2-shell of the observed graph G_{obs} is taken to be the graph $G_{\text{obs},2}$ after removing all the edges in the original observed graph G_{obs}.

edge ij is not in the observed graph, the nodes i and j might still be causally related if they are indirectly connected in the observed graph. For example, suppose that there are two people i and j such that there exists a third person k, where j influences k and k influences i. Suppose further that $ik, kj \in E_{\text{obs}}$, but $ij \notin E_{\text{obs}}$. In this case, by the transitivity of causality, j influences i indirectly, and hence $ij \in E_{\text{cau}}$, despite that $ij \notin E_{\text{obs}}$. We propose to construct direction G' by collecting such edges ij. More specifically, we take the direction of the test to be the graph $G' = (N, E')$, where $ij \in E'$ if and only if $ij \notin E_{\text{obs}}$ and for some $k \in N$, $ik, kj \in E_{\text{obs}}$. Thus, the edges ij in E' are such that i and j are connected through two edges but not adjacent in G_{obs}. We call the graph G' the **G2-shell** of the observed graph G_{obs}. See Figure 2 for an illustration of a G2-shell.

In practice, the edge set of the G2-shell of an observed graph can be large and it can weaken the power of the test through a large variance estimator $\hat{\sigma}^2$. We propose using subgraphs, $G_1, ..., G_R$, that are randomly drawn from the G2-shell as follows. First, we set the cutoff degree $\bar{d}_n$ and let $G_{\text{2shell}} = (N, E_{\text{2shell}})$ denote the G2 shell constructed from G_{obs} as previously. Then, we implement the following algorithm.

Subgraph Generation Algorithm

    for $r = 1 : R$,

        Initialize graph $G_r = (N, E_r)$ to be an empty graph.

        for $i = 1 : n$,

            Find the in-degree $d_n(i)$ of i in the G2-shell, that is, $d_n(i) = |\{j \in N : ij \in E_{2\text{shell}}\}|$.

            If $d_n(i) \leq \overline{d}_n$, place all its edges, $ij \in E_{2\text{shell}}$, into E_r.

            If $d_n(i) > \overline{d}_n$, randomly select $\overline{d}_n$ number of edges from its edges, $ij \in E_{2\text{shell}}$, and

                place them in E_r.

        end for

        Store G_r

    end for

In practice, we suggest setting the cutoff degree $\overline{d}_n = 5$ or choosing the cutoff degree $\overline{d}_n$ to be the smallest integer not greater than the average in-degree of G_{obs}. For the number of subgraphs, R, we suggest choosing $R = 100$. We find that these choices work well in our simulations.

Once we have generated subgraphs, say, $G_1, ..., G_R$, using the subgraph generation algorithm, we proceed to perform a multiple testing procedure as follows. For each subgraph G_r, we construct the following hypothesis testing problem:

$$H_{0,r} : \mathsf{C}_r = 0, \text{ vs. } H_{1,r} : \mathsf{C}_r \neq 0,$$

where C_r is the same as C except that G_r replaces G_{obs}. For each $r = 1, ..., R$, we construct the individual test statistic $|T(G_r)|$. From the absolute value of the standard normal distribution, we find individual p-values and perform a multiple testing procedure that controls the familywise error rate (FWER) at the level of the directional test. For example, the procedure proposed by Holm (1979) suggests that we start with the smallest p-value and continue to reject the hypotheses H_{0,m_r}, $r = 1, ..., R$, as long as[20]

(3.19)
$$\hat{p}_{(r)} < \frac{\alpha}{R - r + 1},$$

---

[20]See also Section 9.1 in Lehmann and Romano (2005). It is known that the Holm's step down procedure tends to be conservative because it does not take into account the dependence structure among the individual test statistics. Romano and Wolf (2005) and Romano and Shaikh (2010) propose an improved step-down procedure that considers the dependence structure. To apply their ideas, we should be able to derive the asymptotic distribution of the statistics of the form $\sqrt{n} \max_{m \in M'} \hat{\mathsf{C}}_m$ for all subsets $M' \subset M$. In our case, the index m corresponds to each randomly selected subgraph of the G2-shell of the large observed graph. To the best of our knowledge, it is not obvious how one can derive such an asymptotic distribution without making further assumptions on the formation of the observed graph. We leave this question to future research.

where $\hat{p}_{(1)} < \hat{p}_{(2)} < ... < \hat{p}_{(R)}$ are the p-values ordered from the smallest to the largest. Then we reject H_0 if any of the individual hypotheses are rejected. We investigate the finite sample performance of our directional testing procedure in the next section.

4. Inference on Diffusion When the Subgraph Hypothesis Fails

When the subgraph hypothesis, with the nonnegative spillover condition (2.7) holds, the spatio-temporal measure C serves as a lower bound for the ADM as shown in Lemma 2.1. In this section, we consider this setting and propose a lower confidence bound for the ADM. We assume that G_{obs} does not satisfy the subgraph hypothesis, but the nonnegative spillover condition (2.7) holds. Then, as we saw before, C serves as a lower bound for the ADM. The lower confidence bound at level $1 - \alpha$ for the ADM that we propose is

$$\hat{L}_{1-\alpha} = \hat{C} - \frac{z_{1-\alpha}\hat{\sigma}}{\sqrt{n}},$$

where $\hat{C}$ and $\hat{\sigma}$ are as previously defined.

Here, we consider the conditions under which this lower confidence bound is asymptotically valid. The main challenge for inference when G_{obs} does not satisfy the subgraph hypothesis is that the cross-sectional dependence structure (shaped by the causal graph G_{cau}) is unknown and the estimated long run variance $\hat{\sigma}^2$ is not consistent for the true long run variance based on causal graph G_{cau}. We address this challenge by introducing a mild, technical condition on causal graph G_{cau} as follows.

Assumption 4.1. There exist $\nu \in (0, 1/2)$ and $c_1 > 0$ such that for each $ij \in E_{\text{ctt}} \setminus E_{\text{obs}}$,

$$\text{(4.1)} \qquad \qquad \text{Cov}\left(Y_{j,0}, Y_{i,1} \mid \mathcal{F}\right) \geq c_1 n^{-1/2+\nu},$$

for all $n \geq 1$.

Assumption 4.1 is satisfied if for all $ij \in E_{\text{ctt}}$, $\text{Cov}(Y_{i,1}, Y_{j,0} \mid \mathcal{F})$ is bounded away from zero by a sequence that decreases to zero not too fast. We can weaken this condition so that condition (4.1) is satisfied by a large fraction of ij's in $E_{\text{ctt}} \setminus E_{\text{obs}}$. (See the Supplemental Note for details.) This condition is violated when there are many pairs ij in $E_{\text{ctt}} \setminus E_{\text{obs}}$ such that $\text{Cov}(Y_{i,1}, Y_{j,0} \mid \mathcal{F})$ is positive and yet very close to zero. If the contact network is an empty graph (as in the case of no diffusion), or G_{obs} contains the contact network as a subgraph, which is satisfied under the subgraph hypothesis, Assumption 4.1 is trivially satisfied.

To see the plausibility of Assumption 4.1, consider the example where each state-switch represents an irreversible technology adoption by a farmer in a contact network, and the

adoption decision is made based on linear threshold diffusion of the following form:

$$\rho_{i,t}\left((A_{j,t-1})_{j\in N_{\text{ctt}}(i)}, U_{i,t}; G_{\text{ctt}}\right) = 1\left\{\frac{\beta}{|N_{\text{ctt}}(i)|}\sum_{j\in N_{\text{ctt}}(i)} A_{j,t-1} \geq U_{i,t}\right\},$$

where the $U_{i,t}$'s are independent and identically distributed (i.i.d.) across $t \geq 1$ and the i's, follow the uniform distribution on $[0,1]$, and are conditionally independent of the $A_{i,0}$'s given $\mathcal{F}$. The parameter β captures the influence from the previous actions by the farmers in the neighborhood in the contact network. Assume that β does not depend on n. In this linear threshold diffusion model, Assumption 4.1 is satisfied, if $\beta > 0$. (This is shown in Section D in the Supplemental Note.)[21] We are ready to present a theorem that establishes the asymptotic validity of the lower confidence bound for the ADM.

Theorem 4.1. *Suppose that the condition of Lemma 2.1(ii) and Assumptions 2.1-3.4 and 4.1 hold. Then*

$$\liminf_{n\to\infty} P\left\{\text{ADM} \geq \hat{L}_{1-\alpha}\right\} \geq 1-\alpha.$$

As we mentioned before, the main challenge for inference due to G_{obs} failing the subgraph hypothesis is that the cross-sectional dependence structure of the observations is not precisely known. Here, we provide the intuition for how we deal with this challenge to obtain Theorem 4.1. First, with σ^2 defined in (3.4), we have

(4.2)
$$\frac{\sqrt{n}(\hat{C} - C)}{\sigma} \to_d N(0,1).$$

Since we do not observe G_{cau}, consistent estimation of σ^2 is not feasible. However, we can show that $\hat{\sigma}^2$ is consistent for $\tilde{\sigma}^2_{\text{obs}}$, where

$$\tilde{\sigma}^2_{\text{obs}} = \frac{1}{n}\sum_{i_1 i_2 : \overline{N}_{\text{obs}}(i_1)\cap\overline{N}_{\text{obs}}(i_2)\neq\varnothing} \mathbf{E}[(q_{i_1} - h_{i_1})(q_{i_2} - h_{i_2}) \mid \mathcal{F}].$$

To resolve the gap between σ^2 and $\tilde{\sigma}^2_{\text{obs}}$, we define $\overline{N}(i) = N(i) \cup \{i\}$ with $N(i) = N_{\text{obs}}(i) \cup N_{\text{cau}}(i)$, and

$$\tilde{\sigma}^2 = \frac{1}{n}\sum_{i_1 i_2 : \overline{N}(i_1)\cap\overline{N}(i_2)\neq\varnothing} \mathbf{E}[(q_{i_1} - h_{i_1})(q_{i_2} - h_{i_2}) \mid \mathcal{F}].$$

Then, it follows that

(4.3)
$$\sigma^2 \leq \tilde{\sigma}^2.$$

---

[21]If $\beta = 0$, this model is the same as one that has an empty contact network. Once we replace the contact network by an empty network, Assumption 4.1 is satisfied trivially.

Hence using $\tilde{\sigma}$ in place of σ will make the limiting distribution of the statistic in (4.2) less dispersed than $N(0, 1)$, making the inference conservative yet asymptotically valid. On the other hand, $\hat{\sigma}^2$, which we use for the lower confidence bound is consistent for $\tilde{\sigma}^2_{\text{obs}}$. Thus, it remains to deal with the gap between $\tilde{\sigma}^2$ and $\tilde{\sigma}^2_{\text{obs}}$. This is where Assumption 4.1 plays a role. To see this, first, we write

$$(4.4) \qquad P\left\{\text{ADM} \geq \hat{L}_{1-\alpha}\right\} = P\left\{\frac{\sqrt{n}(\text{C} - \hat{\text{C}})}{\tilde{\sigma}} \geq -z_{1-\alpha} + R_{1n} + R_{2n}\right\},$$

where

$$R_{1n} = \frac{z_{1-\alpha}(\tilde{\sigma} - \tilde{\sigma}_{\text{obs}}) - \sqrt{n}(\text{ADM} - \text{C})}{\tilde{\sigma}}, \text{ and}$$

$$R_{2n} = -\frac{z_{1-\alpha}(\hat{\sigma} - \tilde{\sigma}_{\text{obs}})}{\tilde{\sigma}}.$$

We can show that $R_{2n} = o_P(1)$. For R_{1n}, we can show that under Assumption 4.1,

$$P\left\{z_{1-\alpha}(\tilde{\sigma} - \tilde{\sigma}_{\text{obs}}) \leq \sqrt{n}(\text{ADM} - \text{C})\right\} \to 1,$$

as $n \to \infty$. Thus we have $P\{R_{1n} \leq 0\} \to 1$, as $n \to \infty$. This yields that the probability on the right-hand side of (4.4) is bounded from below by

$$P\left\{\frac{\sqrt{n}(\text{C} - \hat{\text{C}})}{\tilde{\sigma}} \geq -z_{1-\alpha}\right\} + o(1).$$

The leading probability is asymptotically bounded from below by $1 - \alpha$ by (4.2) and (4.3). Thus, we obtain the asymptotic validity of the lower confidence bound $\hat{L}_{1-\alpha}$.

5. Monte Carlo Simulations

5.1. Simulation Design

In this section, we present the results from our Monte Carlo study on the finite sample behavior of our inference procedures. For the study, we first generate the contact network G_{ctt}. Here, we mimic the network that we used in the empirical section. The adjacency matrix of the contact network is chosen as a block diagonal matrix and each block matrix is generated first by a Barabási-Albert model. Then we randomly rewire one link for each isolated node. We treat each block matrix as a village. We experiment on different numbers of villages to balance the sparsity and clustering of the network.

After generating the contact network, we generate binary actions $A_t = (A_{i,t})_{i \in N}$ that percolate through the contact network over time as follows. For each $j \in N$, we specify

$$A_{j,0} = 1 \left\{ F_0(X_j' \gamma_0) \geq U_{j,0} \right\},$$

where the $U_{j,0}$'s follow the uniform distribution on $[0,1]$ and are independent of the other random variables, and F_0 is the distribution function of $N(0,1)$. The covariates in the vector X_j are drawn i.i.d. from the uniform distribution on $[0,1]$. We set $p = 5$ and $\gamma_0 = [-1.5, 0.6, -0.5, -0.4, 0.7, 0.5]'$, and the mean of $Y_{j,0}$ is around 0.16.

For each $i \in N$ and $t = 1, ..., t_1$, we specify

$$A_{i,t} = \begin{cases} 1 \left\{ \delta_0 \overline{A}_{i,t-1} + X_i' \beta_0 - U_{i,t} > 0 \right\}, & \text{if } A_{i,t-1} = A_{i,t-2} = ... = 0, \\ 0, & \text{otherwise,} \end{cases}$$

where the $U_{i,t}$'s are i.i.d. following the uniform distribution on $[0,1]$, and

$$\overline{A}_{i,t-1} = \frac{1}{|N_{\text{ctt}}(i)|} \sum_{j \in N_{\text{ctt}}(i)} A_{j,t-1},$$

and $N_{\text{ctt}}(i)$ is the in-neighborhood of i in the contact network G_{ctt}. In the simulation, we consider the choice of observation period $t_1 \in \{1, 2, 3\}$. We take the covariates to be time-invariant. We set $\beta_0 = [-1, 0.3, -0.4, -0.1, 0.4, 0.8]'$. The parameter δ_0, when it takes a nonzero value, generates diffusion in our model over a given graph. We explore different values of $\delta_0 \in \{0, 1.0, 1.5\}$ which represent varied spillover effects.

To generate the observed graph G_{obs}, we endow each pair (i, j) of people an edge in the observed graph G_{obs} if and only if there exists a walk from j to i that has length less than or equal to t_1. The observed graph generated this way is the same as the causal graph G_{cau} and hence satisfies the subgraph hypothesis.

For the Monte Carlo simulations, we first draw the contact network, the observed graph, and the covariates, and store their values outside the Monte Carlo loop. In this simulation design, we do not allow for the unobserved common shock $\mathcal{C}$ in the definition of $\mathcal{F}$.

A summary of the network characteristics is given in Table 1. The observed graphs are denser than the contact networks when $t_1 = 2, 3$. The observed graph becomes denser as t_1 increases. The true value of the ADM is reported in Table 2. The table shows that the value of the true ADM decreases over time. This is expected because one additional switcher in the initial period makes less difference at a later stage of diffusion.

TABLE 1. The Characteristics of the Networks

		G_{ctt}		G_{obs} for $t_1 = 2$		G_{obs} for $t_1 = 3$	
		$n_V = 20$	$n_V = 30$	$n_V = 20$	$n_V = 30$	$n_V = 20$	$n_V = 30$
max. deg.	$n_H = 50$	11	12	27	30	37	39
	$n_H = 200$	29	28	90	84	145	134
ave. deg.	$n_H = 50$	1.703	1.727	4.807	4.971	9.682	9.798
	$n_H = 200$	1.923	1.923	6.744	6.583	17.42	17.06
cluster	$n_H = 50$	0.043	0.048	0.283	0.268	0.404	0.368
	$n_H = 200$	0.024	0.017	0.140	0.146	0.240	0.269

Notes: The table presents the network characteristics of the contact network G_{ctt} and the observed graph G_{obs}. For $t_1 = 1$, the observed graph G_{obs} is set to be the same as the contact network G_{ctt}. We considered two different groups of villages, $n_V = 20$ and $n_V = 30$, and two different numbers of households in each village, $n_H = 50$ and $n_H = 200$.

TABLE 2. The True Value of the ADM over Time

			$\delta_0 = 1.0$			$\delta_0 = 1.5$		
			$t_1 = 1$	$t_1 = 2$	$t_1 = 3$	$t_1 = 1$	$t_1 = 2$	$t_1 = 3$
True ADM	$n_V = 20$	$n_H = 50$	0.3109	0.1923	0.0840	0.4443	0.3060	0.1158
		$n_H = 200$	0.3088	0.1915	0.0736	0.4426	0.3007	0.1115
True ADM	$n_V = 30$	$n_H = 50$	0.3087	0.1912	0.0824	0.4441	0.3063	0.1228
		$n_H = 200$	0.3081	0.1896	0.0743	0.4419	0.3012	0.1208

Notes: This table shows the true values of the average diffusion at the margin (ADM) we used for the Monte Carlo study. The ADM decreases over time for each given network. We consider three periods with the number of villages $n_V = 20$ and $n_V = 30$ and cross-sectional dependency parameters $\delta_0 = 1.0$ and $\delta_0 = 1.5$. The true ADMs were computed from 100,000 simulations.

5.2. Results

The finite sample performance of asymptotic inference on the ADM is shown in Table 3. The table reports the empirical coverage probability and mean length of confidence intervals based on the standard normal distribution.

First, the finite sample coverage probability of the asymptotic inference appears overall stable over a wide range of networks. For example, when $n_V = 30$ and $n_H = 200$, the maximum degree of the contact network is 29 and that of the observed graph is 145. In this case, the empirical coverage probability at the nominal level of 95% is around $0.932 \sim 0.974$, depending on δ_0 and t_1, whereas the length of the confidence interval is around $0.104 \sim 0.273$.

Second, as the sample size increases, the length of the confidence interval decreases. This result holds across the range of graphs considered in the study. This reflects the law of large numbers, which arises due to the weak cross-sectional dependence among the outcomes.

TABLE 3. The Empirical Coverage Probability and the Mean Length of Confidence Intervals from the Normal Distribution at 95% Nominal Level

The Empirical Coverage Probabilities

δ_0		$n_V = 20$			$n_V = 30$		
		$t_1 = 1$	$t_1 = 2$	$t_1 = 3$	$t_1 = 1$	$t_1 = 2$	$t_1 = 3$
0	$n_H = 50$	0.9532	0.9333	0.9243	0.9562	0.9394	0.9310
	$n_H = 200$	0.9523	0.9351	0.9316	0.9527	0.9423	0.9359
1.0	$n_H = 50$	0.9581	0.9365	0.9271	0.9649	0.9390	0.9298
	$n_H = 200$	0.9590	0.9397	0.9309	0.9656	0.9477	0.9349
1.5	$n_H = 50$	0.9738	0.9382	0.9239	0.9711	0.9432	0.9297
	$n_H = 200$	0.9688	0.9404	0.9280	0.9744	0.9467	0.9322

The Mean-Length of Confidence Intervals

δ_0		$n_V = 20$			$n_V = 30$		
		$t_1 = 1$	$t_1 = 2$	$t_1 = 3$	$t_1 = 1$	$t_1 = 2$	$t_1 = 3$
0	$n_H = 50$	0.2438	0.3867	0.4963	0.2017	0.3254	0.4115
	$n_H = 200$	0.1276	0.2295	0.3303	0.1035	0.1858	0.2732
1.0	$n_H = 50$	0.2900	0.3996	0.4798	0.2409	0.3334	0.3979
	$n_H = 200$	0.1594	0.2447	0.3276	0.1288	0.1958	0.2670
1.5	$n_H = 50$	0.3262	0.3987	0.4514	0.2684	0.3334	0.3744
	$n_H = 200$	0.1841	0.2543	0.3137	0.1482	0.2008	0.2558

Notes: The simulation study was based on a single generation of the random graphs. We performed $1,000$ simulations and compared the 20 villages case with the 30 villages case. The 30 villages case shows better coverage probabilities and smaller mean lengths because these two networks share similar network characteristics and only differ in the number of total nodes.

However, as the graph becomes denser, the confidence interval becomes longer. This is expected because, as the graph becomes denser, the cross-sectional dependence becomes extensive, increasing the uncertainty of the estimators.

We also perform a Monte Carlo study of the directional test for the subgraph hypothesis developed in Section 3.5. For this, we first generate a contact network G_{ctt} as before, using a block-diagonal adjacency matrix based on the Barabási-Albert random graph. Let $G_{c,m}$ be a graph constructed by adding an edge between any nodes that are within the geodesic distance m from each other in contact network G_{ctt}. For the DGP for outcomes, we use $G_{\text{ctt}}^0 = G_{\text{ctt}}$ as the contact network under the null hypothesis H_0 and $G_{\text{ctt}}^1 = G_{c,2}$ as the contact network under the alternative hypothesis. Under both the null hypothesis and the alternative hypothesis, we set the observed graph as follows: $G_{\text{obs}} = G_{c,1}$ for $t_1 = 1$ and $G_{\text{obs}} = G_{c,2}$ for $t_1 = 2$. Thus, by the design of the contact networks, G_{obs} satisfies the subgraph hypothesis under the null hypothesis but not under the alternative hypothesis. The graph generation for the directional tests is summarized in Table 4.

TABLE 4. Generation of Graphs for Directional Tests

	$t_1 = 1$			$t_1 = 2$		
H_0	$G_{\text{ctt}} = G_{c,1}$	$G_{\text{cau}} = G_{c,1}$	$G_{\text{obs}} = G_{c,1}$	$G_{\text{ctt}} = G_{c,1}$	$G_{\text{cau}} = G_{c,2}$	$G_{\text{obs}} = G_{c,2}$
H_1	$G_{\text{ctt}} = G_{c,2}$	$G_{\text{cau}} = G_{c,2}$	$G_{\text{obs}} = G_{c,1}$	$G_{\text{ctt}} = G_{c,2}$	$G_{\text{cau}} = G_{c,4}$	$G_{\text{obs}} = G_{c,2}$

Notes: This table reports the contact network and observed graph specification from $G_{c,m}$.

For the direction of the test, we set $n_V = 30$ villages, each village with either $n_H = 50$ or $n_H = 200$ households. We apply Holm's multiple testing procedure by randomly drawing 100 subgraphs from the $G2$-shell of G_{obs} using the subgraph generation algorithm introduced earlier. We consider two ways of choosing the cutoff degree $\overline{d}_n$: one way is to set $\overline{d}_n$ to be 5 and the other way is to set $\overline{d}_n$ to be the smallest integer not greater than the average degree of G_{obs}.

We compute the FWER using 1,000 Monte Carlo simulations. For brevity, we report only the results with the choice of $\overline{d}_n = 5$. The results are similar for the case with $\overline{d}_n$ equal to the smallest integer not greater than the average degree of G_{obs}. The finite sample FWER's are shown in Table 5. The FWERs are around the nominal level α under both the null hypothesis and the alternative hypothesis. The power of the test is reduced in the case of $t_1 = 2$. This is due to the fact that the observed graph is denser, resulting in a dense $G2$-shell from which we draw 100 subgraphs. When the $G2$-shell is dense, the 100 subgraphs may contain many edges that do not effectively capture the dependencies between nodes under the alternative hypothesis. This explains the weaker power of the directional test when the observed graph is denser.

6. Empirical Application: Diffusion of Microfinancing Decisions over Social Networks

6.1. Data

We apply our procedure to estimate the ADM of microfinancing decisions in 43 rural villages in southern India. The data set used here originated from the data used by Banerjee, Chandrasekhar, Duflo, and Jackson (2013). In 2006, social network data were collected for 75 rural villages in Karnataka, a state in southern India. The network was measured along the twelve dimensions, such as: visiting each other; kinship; borrowing or lending money, rice, or kerosene; giving or exchanging advice; or going to the place of prayer together. The data contain information on microfinancing decisions observed over the trimesters after the

TABLE 5. The Familywise Error Rate

δ_0			H_0			H_1		
			$\alpha = 1\%$	$\alpha = 5\%$	$\alpha = 10\%$	$\alpha = 1\%$	$\alpha = 5\%$	$\alpha = 10\%$
1.0	$n_H = 50$	$t_1 = 1$	0.006	0.039	0.080	1.000	1.000	1.000
		$t_1 = 2$	0.021	0.072	0.126	0.030	0.106	0.202
1.5	$n_H = 50$	$t_1 = 1$	0.011	0.059	0.092	1.000	1.000	1.000
		$t_1 = 2$	0.010	0.059	0.107	0.115	0.320	0.473
1.0	$n_H = 200$	$t_1 = 1$	0.009	0.033	0.076	1.000	1.000	1.000
		$t_1 = 2$	0.017	0.067	0.116	0.045	0.135	0.239
1.5	$n_H = 200$	$t_1 = 1$	0.005	0.038	0.076	1.000	1.000	1.000
		$t_1 = 2$	0.013	0.045	0.095	0.172	0.491	0.703

Notes: This table reports the results of the directional tests, with the cutoff degree $\overline{d}_n = 5$. Each sample consists of 30 villages with 50 or 200 households in each village. The FWERs were computed according to Holm's multiple testing procedure over 100 hypotheses, using $1,000$ simulations.

program began operations as well as other individual characteristics such as caste, village leader indicator, savings indicator, and education. In this empirical study, the cross-sectional units are households.

We distinguish between two definitions of initial state-switchers. First, we define the **leaders** as those who first learned about the micro-financing program. In 2007, a microfinancing institution, Bharatha Swamukti Samsthe (BSS), began operations in 43 of the villages, first by inviting a pre-determined set of people (such as teachers, shopkeepers, and saving group leaders) to a private meeting with credit officers, and asking the leaders to spread the information on the microfinancing program. The leaders were pre-determined based on their expected well-connectedness within the village. Hence, they fit well with the definition of the leaders in this data. Second, we define the **leader-adopters** as those who are leaders and participated in the micro-financing program in the first trimester.

First, we focus on the ADM based on the initial switchers as households, one of whose members is a leader-adopter, that is, we define $Y_{j,0} = 1$ if and only if household j has a leader-adopter. Then, we take $Y_{i,1} = 1$ if and only if household i adopted microfinancing in the first trimester or later. This ADM measures the diffusion of microfinancing decisions triggered by the village leaders participating in the microfinancing program. Second, we define the initial switchers to be those households that have a leader, that is, we define $Y_{j,0} = 1$ if and only if the household j has a village leader, and take $Y_{i,1}$ as before. The ADM in this case measures the diffusion triggered by the village leaders' obtaining the information on the microfinancing program.

TABLE 6. Social Network Characteristics from Survey Data on Indian Villages

Networks	Size	Maximum Degree	Average Degree	Median Degree
G_{ee}	4413	42	3.7088	3
G_{sc}	4413	76	5.6646	5
G_{all}	4413	78	6.1854	5

Notes: The network G_{ee} is defined based on material exchanges between households (such as borrowing rice, kerosene or money). The network G_{sc} is defined based on social activities (such as seeking advice or going to the same temple or church). The network G_{all} is the union of the two networks G_{ee} and G_{sc}.

For the study, we construct directed networks at the household level. Similar to Leung (2015), we obtain directed links based on the individual survey questions that collected social network data along the twelve dimensions.[22] We define graph G_{ee} to be the social network where two households are linked if and only if material exchanges (borrowing/lending rice, kerosene or money) occurred, and graph G_{sc} to be the social network, where two households are linked if and only if some social activities (such as seeking advice, or going to the same temple or church) occurred. We also consider graph G_{all}, which is the union of G_{ee} and G_{sc} so that two households are linked if and only if any of the 12 dimensions in the social network data (as mentioned before) occurred between the households. The summary statistics for the three networks are listed in Table 6.

We include similar covariates from the survey data included in Banerjee, Chandrasekhar, Duflo, and Jackson (2013). The demographic controls in Banerjee, Chandrasekhar, Duflo, and Jackson (2013) include the number of households, self-help group participation in the village, savings participation in the village, caste composition, and village households that are designated as leaders. Except for the number of the households which is the same across households in the same village, all the other controls are dummy variables. We construct the set of covariates from the household averages. The household average covariates are constructed by averaging the individual covariates from the same household. For example, if any member of a household participates in a self-help group(SHG)/has savings/is a leader in the village, then the dummy for SHG/savings/leaders for the household average equals 1. The caste composition dummy is an indicator of minority group, which includes the scheduled caste and scheduled tribes, the relatively disadvantaged groups. In the following experiment, we use the leaders dummy to construct the initial outcome and therefore exclude it from the covariates.

The data appears to fit our framework well. The leader status was pre-determined primarily based on individuals' positions in the social networks. They were the first people who were

---

[22]The networks are constructed as follows. For each graph, the household-level adjacency matrix is constructed from the individual data, where a node is a household. One household is linked to another if any of its members is linked through the relationships indicated by the social network.
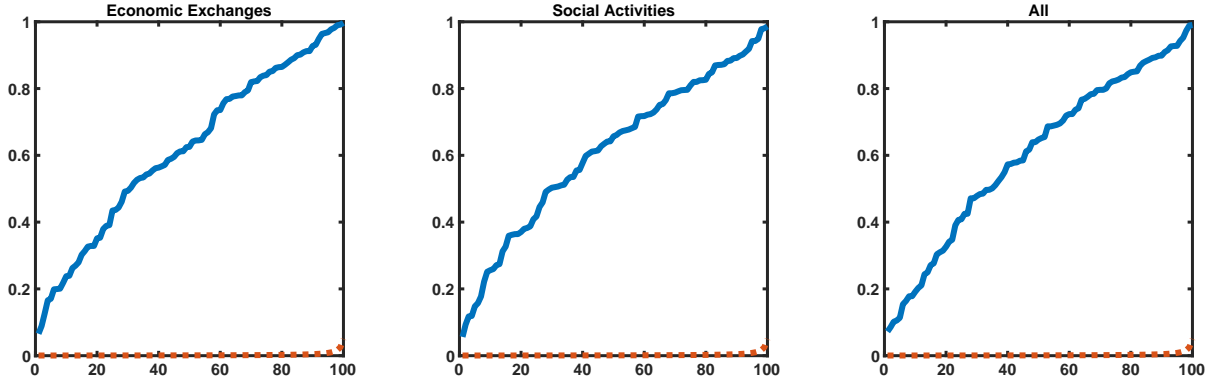
FIGURE 3. Directional Tests of the Subgraph Hypothesis Using the ADM Based on Leader-Adopters and Multiple Subgraphs of the G2-Shell with the Cutoff Degree Equal to 5: The three panels show the directional testing results for three types of networks, based on economic exchanges, social activities, and the union of all relations, using the original graph in the data. The tests are based on the ADM with leader-adopters as initial state-switchers. The cutoff degree is set to be 5 for all the three types of networks. The solid line is the sorted p-values and the dotted line is the Holm thresholds at FWER control at 5%. We reject the null hypothesis of the subgraph hypothesis, if the minimum value of the dotted line lies above the blue line at some point. The null hypothesis is not rejected for all three networks with the original graph. Thus, the results show no evidence against the subgraph hypothesis.

exposed to the information on the microfinancing program, and learned about the program through private meetings. Therefore, it is plausible that their leadership status and initial adoptions are conditionally independent across different households, and independent from the potential outcomes of their neighbors, given the networks and observed characteristics. Furthermore, the researcher does not observe the actual process of diffusion. The administrative data provided by BSS contained information on microfinancing decisions over several trimesters. It is plausible that within each trimester, the diffusion progressed without the researcher observing it. Our framework accommodates this feature of misalignment between the actual timing of the diffusion and the researcher's observation.

6.2. Results

6.2.1. **Directional Tests of the Subgraph Hypothesis.** Our proposal offers two routes to the researcher to make inference on the ADM. First, she can construct the confidence interval for the ADM under the subgraph hypothesis for the observed graph used. Second, she can replace the subgraph hypothesis by the assumption of nonnegative spillover effect in (2.7), and construct the lower confidence bound for the ADM.

TABLE 7. Directional Tests of the Subgraph Hypothesis Using the G2-Shell

	ADM	CI		
		99%	95%	90%
G_{ee}	0.113	[-0.400,0.625]	[-0.277,0.503]	[-0.215,0.440]
G_{sc}	-0.170	[-1.087,0.747]	[-0.868,0.527]	[-0.756,0.415]
G_{all}	-0.088	[-0.926,0.750]	[-0.726,0.550]	[-0.623,0.447]

Notes: The subgraph hypothesis implies that the ADM measured with the G2-shell of the observed graph is zero. The confidence intervals for the ADM contain zero at 99%, 95% and 90%. This suggests no evidence against the subgraph hypothesis in the directional test using the ADM measured with the G2-shell.

In this empirical application, we first investigate whether there is any evidence against the subgraph hypothesis for the three observed networks, G_{ee}, G_{sc}, and G_{all}, using the directional tests we introduced earlier. We performed this test in two different ways: first, using the multiple testing approach (the Holm procedure) based on 100 random subgraphs from the G2-shell of the observed graphs, and second, using a single test based on the G2-shell of the observed graph.

The results from the multiple testing approach are shown in Figure 3. Here, the tests are based on the ADM with leader-adopters as initial state-switchers. The cutoff degree is set as $\overline{d}_n = 5$ for all the three kinds of networks. The solid lines indicate the sorted p-values and the dotted lines represent the Holm's thresholds (the term on the right-hand side of (3.19)). The p-values for the different networks are all substantially and uniformly greater than the Holm's thresholds at 5% of FWER. Hence we do not reject any of the individual hypotheses at 5% of FWER. This result shows no evidence against the subgraph hypothesis for the observed graphs, G_{ee}, G_{sc}, and G_{all}.[23]

We also perform a single test of the subgraph hypothesis by checking if the ADM is significantly different from zero when the G2-shell is used as the observed graph. Under the subgraph hypothesis, the ADM is zero. Hence, if there are links not captured by the observed graph but captured by the G2-shell, the ADM constructed using the G2-shell may be different from zero. The results of this directional test are reported in Table 7. The estimated ADM using the G2-shell is not significantly different from zero at 99%, 95%, and 90% across the different definitions of social networks. This result suggests no evidence against the subgraph hypothesis.

In summary we investigated whether the subgraph hypothesis is plausible or not in our data, using both multiple-testing and single-testing approaches. We have not found any evidence against the subgraph hypothesis for the observed graphs.

---

[23]We also performed multiple testing, setting the cut-off degree to be the ceiling of the average degree of the observed graph, and with leaders as initial state-switchers. The results are similar, showing no evidence against the subgraph hypothesis at the 5% significance level. These results are omitted for brevity.

6.2.2. **Estimated ADM.** We measure the ADM in two different ways. First, we measure the ADM, taking the initial state-switchers as households having a leader-adopter. Second, we measure the ADM, taking the initial state-switchers as households having a leader. The results are shown in Table 8.

The estimated ADMs with leader-adopters as initial switchers are significantly different from zero at 95% except for G_{ee} when we use the subgraph hypothesis. When we drop the subgraph hypothesis but employ the assumption of nonnegative spillover effects, the estimated ADM is significantly different from zero at 95% for all three graphs. This shows the role of social networks in channeling the diffusion of microfinancing decisions that were triggered by the leader's microfinancing decisions. The expected increase in the number of switchers by the end of the observation period due to one additional leader-adopter is estimated to be around 0.22-0.43, depending on the definition of the graphs.

When we considered the ADM using the leaders as initial switchers, the estimated ADMs were insignificant across all the graphs, in terms of the lower confidence bound or the confidence intervals. This shows that even if there was a diffusion of information on the microfinancing programs, their actual impacts may be significantly different across the leaders, depending on whether the leader actually participated in the program or not. For example, it is conceivable that the influence of the leaders who participated in the microfinancing program was stronger than the influence of the leaders who learned about the program but did not participate in it. Our results demonstrate that the measured diffusion can depend crucially on how we define the initial triggers.

7. Conclusion

In this paper, we proposed a measure of diffusion of binary outcomes over a large network using the potential outcome approach in causal inference. We introduced a spatio-temporal dependence measure of outcomes and showed that it is identified as the measure of diffusion under a certain condition for the observed network. We showed that when the condition fails but the spillover effect is nonnegative, the dependence measure is a lower bound for the measure of diffusion. Then, we developed an inference method and established its asymptotic validity, despite the fact that the observations exhibit heterogeneity and a complex form of cross-sectional dependence. Our empirical application demonstrated the usefulness of our framework.

There are different directions for extensions that appear potentially promising. First, we may consider a setting where the unconfoundedness condition fails, because the initial switches are observed a while after the diffusion already progressed. It would be interesting to see if we can develop an instrumental variable approach in this situation. Second, it would be

TABLE 8. The Estimated ADM of Microfinancing over Social Networks

The Case with the Initial Switchers Being Leader-Adopters.

	ADM	LB			CI		
		99%	95%	90%	99%	95%	90%
G_{ee}	0.215	-0.075	0.010	0.055	[-0.106,0.536]	[-0.029,0.460]	[0.010,0.420]
G_{sc}	0.434	0.035	0.152	0.214	[-0.008,0.875]	[0.098,0.770]	[0.152,0.716]
G_{all}	0.435	0.020	0.141	0.206	[-0.025,0.895]	[0.085,0.785]	[0.141,0.728]

The Case with the Initial Switchers Being Leaders.

	ADM	LB			CI		
		99%	95%	90%	99%	95%	90%
G_{ee}	-0.052	-0.137	-0.112	-0.099	[-0.146,0.041]	[-0.123,0.019]	[-0.112,0.007]
G_{sc}	-0.049	-0.165	-0.131	-0.113	[-0.178,0.080]	[-0.146,0.049]	[-0.131,0.033]
G_{all}	-0.051	-0.167	-0.133	-0.115	[-0.180,0.079]	[-0.149,0.048]	[-0.133, 0.032]

Notes: The estimated ADM of microfinancing decisions depends on how we define initial switchers. When we take the initial switchers to be leader-adopters, the results are significantly positive overall. When we take them to be just leaders, the significance disappears. The household covariates include four basic controls: number of households in the village, dummies for self-help group participation, savings participation, and caste composition.

useful to explore methods that can be used to assess the role of the covariates in the causal interpretation of the spatio-temporal dependence measure. While Song (2022) provides a method of decomposition to capture the spurious diffusion, the method assumes the subgraph hypothesis. It would be interesting to characterize the behavior of the spatio-temporal dependence measure as a lower bound when the subgraph hypothesis fails.

References

ACEMOGLU, D., A. OZDAGLAR, AND E. YILDIZ (2011): "Diffusion of Innovations in Social Networks," *IEEE Conference on Decision and Control and European Control Conference, 12/2011*, pp. 2329–2334.

AKBARPOUR, M., S. MALLADI, AND A. SABERI (2020): "Just a Few Seeds More: Value of Network Information for Diffusion," *Working Paper*.

ARAL, S., L. MUCHNIK, AND A. SUNDARARAJAN (2009): "Distinguishing Influence-Based Contagion from Homophily-Driven Diffusion in Dynamic Networks," *Proceedings of the National Academy of Sciences of the United States of America*, 106(51), 21544–21549.

ARONOW, P., AND C. SAMII (2017): "Estimating Average Causal Effects under General Interference, with Application to a Social Network Experiment," *Annals of Applied Statistics*, 11, 1912 – 1947.

ATHEY, S., D. ECKLES, AND G. W. IMBENS (2018): "Exact P-values for Network Interference," *Journal of the American Statistical Association*, 113, 230–240.

BANERJEE, A., A. G. CHANDRASEKHAR, E. DUFLO, AND M. O. JACKSON (2013): "The Diffusion of Microfinance," *Science*, 341(6144), 1236498.

——— (2019): "Using Gossips to Spread Information: Theory and Evidence from Two Randomized Controlled Trials," *Review of Economic Studies*, 86, 2453–2490.

BEAMAN, L., A. BENYISHAY, J. MAGRUDER, AND A. M. MOBARAK (2021): "Can Network Theory-based Targeting Increase Technology Adoption?," *American Economic Review*, 111(6), 1918–43.

BRAMOULLÉ, Y., H. DJEBBARI, AND B. FORTIN (2009): "Identification of Peer Effects through Social Networks," *Journal of Econometrics*, 150, 41–55.

BREZA, E., A. G. CHANDRASEKHAR, AND A. TAHBAZ-SALEHI (2018): "Seeing the Forest for the Trees? An Investigation of Network Knowledge," .

CAMPBELL, A. (2013): "Word-of-Mouth Communication and Percolation in Social Networks," *American Economic Review*, 103, 2466–2498.

CHOI, D. (2017): "Estimation of Monotone Treatment Effects in Network Experiments," *Journal of the American Statistical Association*, 112, 1147–1155.

COLEMAN, J., E. KATZ, AND H. MENZEL (1957): "The Diffusion of an Innovation Among Physicians," *Sociometry*, 20, 253–273.

CONLEY, T., AND C. UDRY (2001): "Social learning through networks: The adoption of new agricultural technologies in Ghana," *American Journal of Agricultural Economics*, 83(3), 668–673.

CONLEY, T. G., AND C. R. UDRY (2010): "Learning about a New Technology: Pineapple in Ghana," *American Economic Review*, pp. 35–69.

DE MATOS, M. G., P. FERREIRA, AND D. KRACKHARDT (2014): "Peer Influence in the Diffusion of the iPhone 3G over a Large Social Network," *Management Information System Quarterly*, 38, 1103–1133.

DE PAULA, A., I. RASUL, AND P. C. SOUZA (2020): "Identifying Network Ties from Panel Data: Theory and an Application to Tax Competition," .

DEMARZO, P. M., D. VAYANOS, AND J. ZWIEBEL (2003): "Persuasion Bias, Social Influence, and Unidimensional Opinions," *Quartely Journal of Economics*, 118, 909–968.

ESCANCIANO, J. C., AND K. SONG (2010): "Testing Single-Index Restrictions with a Focus on Average Derivatives," *Journal of Econometrics*, 156, 377–391.

GOLDENBERG, A., B. LIBAI, AND E. MULLER (2001): "Talk of the Network: A Complex Systems Look at the Underlying Process of Word-of-Mouth," *Marketing Letters*, 12(3), 211–223.

GOLUB, B., AND M. O. JACKSON (2010): "Naïve Learning in Social Networks and the Wisdom of Crowds," *American Economic Journal*, 2(1), 112–149.

GRANOVETTER, M. (1978): "Threshold Models of Collective Behavior," *American Journal of Sociology*, 83, 1420–1443.

HIRANO, K., G. IMBENS, AND G. RIDDER (2003): "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score," *Econometrica*, 71, 1161 – 1189.

HOLM, S. (1979): "A Simple Sequentially Rejective Multiple Test Procedure," *Scandinavian Journal of Statistics*, 6, 65–70.

HU, Y. (2008): "Identification and Estimation of Nonlinear Models with Misclassification Error Using Instrumental Variables: A General Solution," *Journal of Econometrics*, 144, 27–61.

IMBENS, G. W., AND D. B. RUBIN (2015): *Causal Inference: for Statistics, Social, and Biomedical Sciences. An Introduction*. Cambridge University Press, New York.

IMBENS, G. W., AND J. M. WOOLDRIDGE (2009): "Recent Developments in the Econometrics of Program Evaluation," *Journal of Economic Literature*, 47, 5–86.

JACKSON, M. O., AND B. W. ROGERS (2007): "Meeting Strangers and Friends of Friends: How Random Are Social Networks?," *American Economic Review*, 97, 890–915.

JACKSON, M. O., AND L. YARIV (2007): "Diffusion of Behavior and Equilibrium Properties in Network Games," *American Economic Review*, 97, 92–98.

JANSSEN, A. (2000): "Global Power Function of Goodness of Fit Tests," *Annals of Statistics*, 28, 239–253.

KEMPE, D., J. KLEINBERG, AND E. TARDOS (2003): "Maximizing the Spread of Influence through a Social Network," *KDD '03: Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 137–146.

KOJEVNIKOV, D. (2021): "The Bootstrap for Network Dependent Processes," *arXiv preprint arXiv:2101.12312*.

KOJEVNIKOV, D., V. MARMER, AND K. SONG (2021): "Limit Theorems for Network Dependent Random Variables," *Journal of Econometrics*, 222(2), 882–908.

KUERSTEINER, G. M., AND I. R. PRUCHA (2013): "Limit Theory for Panel Data Models with Cross Sectional Dependence and Sequential Exogeneity," *Journal of Econometrics*, 174(2), 107–126.

LEHMANN, E. L., AND J. P. ROMANO (2005): *Testing Statistical Hypotheses*. Springer, New York.

LESKOVEC, J., L. A. ADAMIC, AND B. A. HUBERMAN (2007): "The Dynamics of Viral Marketing," *ACM Transactions on the Web (TWEB)*, 1(1), 5–es.

LEUNG, M. P. (2015): "Two-Step Estimation of Network-Formation Models with Incomplete Information," *Journal of Econometrics*, 188, 182–195.

———— (2020): "Treatment and Spillover Effects Under Network Interference," *Review of Economics and Statistics*, 102, 368–380.

LEUNG, M. P. (2022): "Causal Inference Under Approximate Neighborhood Interference," *Econometrica*, 90(1), 267–293.

LEWBEL, A., X. QU, AND X. TANG (2021): "Social Networks with Unobserved Links," *Working Paper*.

LI, W., D. SUSSMAN, AND E. KOLACYZYK (2021): "Causal Inference under Network Interference with Noise," *arXiv:2105.04518v1 [stat.ME]*.

MANSKI, C. F. (1993): "Identification of Endogenous Social effects: The Reflection Problem," *Review of Economic Studies*, 60(3), 531–542.

———— (1997): "Monotone Treatment Response," *Econometrica*, 65(6), 1311–1334.

MORRIS, S. (2000): "Contagion," *Review of Economic Studies*, 67, 57–78.

NEWMAN, M. E. J. (2010): *Networks: An Introduction*. Oxford University Press, New York, USA.

RIDDER, G., AND S. SHENG (2020): "Estimation of Large Network Formation Games," *arXiv preprint arXiv:2001.03838*.

ROMANO, J. P., AND A. M. SHAIKH (2010): "Inference for the Identified Set in Partially Identified Econometric Models," *Econometrica*, 78, 169–211.

ROMANO, J. P., AND M. WOLF (2005): "Stepwise Multiple Testing as Formalized Data Snooping," *Econometrica*, 73(4), 1237–1282.

ROSENBAUM, P. (2007): "Interference Between Units in Randomized Experiments," *Journal of the American Statistical Association*, 102(477), 191–200.

SADLER, E. (2020): "Diffusion Games," *American Economic Review*, 110, 225–270.

SHALIZI, C. R., AND A. C. THOMAS (2011): "Homophily and Contagion Are Generically Confounded in Observational Social Newtwork Studies," *Sociological Methods and Research*, 40, 211–239–535.

SONG, K. (2022): "A Decomposition Analysis of Diffusion Over a Large Network," *Econometric Theory*, 38, 1221–1251.

STUTE, W. (1997): "Nonparametric Model Checks for Regression," *Annals of Statistics*, 25, 613–641.

VAN DER LAAN, M. (2014): "Causal Inference for a Population of Causally Connected Units," *Journal of Causal Inference*, 2, 13–74.

YOUNG, P. H. (2009): "Innovation Diffusion in Heterogeneous Populations: Contagion, Social Influence, and Social Learning," *American Economic Review*, 99, 1899–1924.

ZHANG, L. (2020): "Spillovers of Program Benefits with Mismeasured Networks," *Working Paper*.

SUPPLEMENTAL NOTE TO "MEASURING DIFFUSION OVER A LARGE NETWORK"

Xiaoqi He and Kyungchul Song

Central University of Finance and Economics and University of British Columbia

This note is the Supplemental Note to He and Song (2022). It is divided into four parts. Part A explains the general notion of dependency causal graph which captures the stochastic dependence between variables that arises due to their causal connections. This notion is used to produce a condition that is weaker than the subgraph hypothesis in the paper. We prove all the results in the paper under this general condition. Part B presents the proofs of the results in the paper along with auxiliary results. Part C provides low level conditions for Assumptions 3.1 - 3.2. Part D provides some further details on Assumption 4.1 using a linear threshold diffusion model.

A. Dependency Causal Graphs

The main proposal of He and Song (2022) for measuring diffusion is built on several crucial assumptions. In particular, we focus on two assumptions used in the paper. The first assumption is that the initial switches and the unobserved heterogeneities are cross-sectionally independent conditional on the contact network, the observed graph, the entire profile of covariates, and the unobserved common shock. Second, for the point-identification of the ADM, the paper uses the subgraph hypothesis. In this section, we show how these two conditions can be relaxed into the requirement that both the causal graph and the observed graph be what we call Dependency Causal Graphs.

For any set $A \subset N$, we write $Y_{A,1} = (Y_{i,1})_{i \in A}$, and $Y_{A,0} = (Y_{j,0})_{j \in A}$, and for any graph $G = (N, E_G)$, let

$$N_G(A) = \bigcup_{i \in A} N_G(i) \text{ and } \overline{N}_G(A) = N_G(A) \cup A,$$

where $N_G(i) = \{j \in N : ij \in E_G\}$. The set $\overline{N}_G(A)$ represents the set of people in A and their in-neighbors in G. Recall our notation $Y_1 = (Y_{i,1})_{i \in N}$ and $Y_0 = (Y_{j,0})_{j \in N}$.

Definition A.1. Given a graph G that is $\mathcal{F}$-measurable, we say that G is a **Dependency Causal Graph (DCG) for** (Y_1, Y_0) **given** $\mathcal{F}$, if for any subsets $A, B, A', B' \subset N$ such that

$$(\overline{N}_G(A) \cup B) \cap (\overline{N}_G(A') \cup B') = \varnothing,$$

we have

$$(Y_{A,1}, Y_{\overline{N}_G(A) \cup B, 0}) \perp\!\!\!\perp (Y_{A',1}, Y_{\overline{N}_G(A') \cup B', 0}) \mid \mathcal{F},$$

where $Z_1 \perp\!\!\!\perp Z_2 \mid \mathcal{F}$ denotes the conditionally independence of Z_1 and Z_2 given $\mathcal{F}$.

When (Y_1, Y_0) has graph G as a DCG given $\mathcal{F}$, this implies that $Y_{j,0}$'s are conditionally independent given $\mathcal{F}$, that is, any source of cross-sectional dependence among the initial period outcomes is already included in $\mathcal{F}$. Furthermore, conditional on the *same* $\mathcal{F}$, any state-switches by next observation time t_1, $Y_{i_1,1}$ and $Y_{i_2,1}$, are allowed to be correlated only when they share certain people in their in-neighborhoods in G. The definition of DCG also involves sets B and B', so that $Y_{j,0}$ is conditionally independent of $Y_{i,1}$ given $\mathcal{F}$, whenever j is not in the in-neighborhood of i, that is, the initial state-switch of person j does not influence that of person i by period t_1, and $Y_{j_1,0}$ and $Y_{j_2,0}$ are conditionally independent given $\mathcal{F}$ whenever $j_1 \neq j_2$.

We introduce a condition that is weaker than Assumption 2.1.

Assumption A.1. The causal graph G_{cau} is a DCG for (Y_1, Y_0) given $\mathcal{F}$.

This assumption means that if there is any dependence between $Y_{j,0}$ and $Y_{i,1}$, this must be because person j is in the in-neighborhood of i in the causal graph G_{cau}. Recall the notation $U_i = (U_{i,1}, \ldots, U_{i,t_1})$, a vector of the unobserved attributes for person i which affects her actions from time $t = 1$ until time $t = t_1$. We obtain the following result.

Lemma A.1. *Suppose that Assumption 2.1 holds, namely that $(U_i, Y_{i,0})$'s are conditionally independent across i's given $\mathcal{F}$. Then, G_{cau} is a DCG for (Y_1, Y_0) given $\mathcal{F}$.*

Proof: Note that $Y_{i,1}$ in (2.4) is determined by $(U_j, Y_{j,0})_{j \in \overline{N}_{\mathrm{cau}}(i)}$ and G_{ctt}, where $\overline{N}_{\mathrm{cau}}(i) = N_{\mathrm{cau}}(i) \cup \{i\}$ and $N_{\mathrm{cau}}(i)$ is the in-neighborhood of i in G_{cau}. Take subsets A, B, A', B' of N such that $\overline{N}_{\mathrm{cau}}(A) \cup B$ and $\overline{N}_{\mathrm{cau}}(A') \cup B'$ are disjoint. We observe from (2.4) that $(Y_{A,1}, Y_{\overline{N}_{\mathrm{cau}}(A) \cup B, 0})$ is a function of $((U_j, Y_{j,0})_{j \in \overline{N}_{\mathrm{cau}}(A) \cup B}, G_{\mathrm{ctt}})$, and similarly, $(Y_{A',1}, Y_{\overline{N}_{\mathrm{cau}}(A') \cup B', 0})$ is a function of $((U_j, Y_{j,0})_{j \in \overline{N}_{\mathrm{cau}}(A') \cup B'}, G_{\mathrm{ctt}})$. Since $\overline{N}_{\mathrm{cau}}(A) \cup B$ and $\overline{N}_{\mathrm{cau}}(A') \cup B'$ are disjoint, $(Y_{A,1}, Y_{\overline{N}_{\mathrm{cau}}(A) \cup B, 0})$ and $(Y_{A',1}, Y_{\overline{N}_{\mathrm{cau}}(A') \cup B', 0})$ are conditionally independent given $\mathcal{F}$, by the assumption of the lemma. Thus, G_{cau} is a DCG for (Y_1, Y_0) given $\mathcal{F}$. ∎

The lemma says that the causal graph G_{cau} is a DCG, if the unobserved heterogeneities and initial state-switches are cross-sectionally conditionally independent given $\mathcal{F}$. It is also worth

noting that a DCG is not uniquely determined by a conditional joint distribution of (Y_1, Y_0) given $\mathcal{F}$. If G_{cau} is a DCG, then any graph, say, G that contains G_{cau} as a subgraph is also a DCG. This is because any two sets A and B with non-overlapping neighborhoods in G are also non-overlapping in G_{cau} when G_{cau} is a subgraph of G. This observation yields the following lemma saying that the subgraph hypothesis is stronger than the hypothesis that G_{obs} is a DCG for (Y_1, Y_0) given $\mathcal{F}$.

Lemma A.2. *Suppose that G_{cau} is a DCG for (Y_1, Y_0) given $\mathcal{F}$, and that G_{obs} satisfies the subgraph hypothesis. Then, G_{obs} is a DCG for (Y_1, Y_0) given $\mathcal{F}$.*

The subgraph hypothesis is simpler but stronger than the DCG assumption. For example, some people in the neighborhood in the contact network G_{ctt} may stop exerting causal influence after a certain period (such as in the case where the function $\rho_{i,t}$ in (2.1) no longer depends on such people after some time). In this case, the observed graph G_{obs} can fail the subgraph hypothesis and yet still be a DCG. This is because the presence of an edge between two nodes in the causal graph G_{cau} *does not necessarily mean that there is causal influence from one to the other*. It only indicates that such causal influence is allowed in the model.

The literature of probabilistic graphical models associates a causal graph with conditional independence restrictions. (See Koller and Friedman (2009) and references therein.) The conditional independence restrictions in this paper are different from those used in this literature. Suppose that the intersection of the in-neighborhoods of vertices i_1 and i_2 is equal to $\{j\}$. Then a probabilistic graphical model assumes that $Y_{i_1,1}$ and $Y_{i_2,1}$ are conditionally independent given $Y_{j,0}$. While such conditional independence can be accommodated in our framework, it is not used in the definition of a DCG here.

B. Mathematical Proofs

For the mathematical proofs of all the results in the paper, we use Assumption A.1 in place of Assumption 2.1, and the assumption that G_{obs} is a DCG for (Y_1, Y_0) given $\mathcal{F}$ in place of the subgraph hypothesis for G_{obs}.

B.1. Proofs of Lemma 2.1 and Proposition 2.1

Lemma B.1. *Suppose that $G_1 = (N, E_1)$ and $G_2 = (N, E_2)$ are DCGs for (Y_1, Y_0) given $\mathcal{F}$. Then, for each $j \in N$,*

$$\sum_{i \in N : ij \in E_1} Cov(Y_{i,1}, Y_{j,0} \mid \mathcal{F}) = \sum_{i \in N : ij \in E_2} Cov(Y_{i,1}, Y_{j,0} \mid \mathcal{F}) = \sum_{i \in N : ij \in E_1 \cap E_2} Cov(Y_{i,1}, Y_{j,0} \mid \mathcal{F}).$$

Proof: The result follows because whenever $ij \notin E_1$ or $ij \notin E_2$, $\text{Cov}(Y_{i,1}, Y_{j,0} \mid \mathcal{F}) = 0$, by the condition that G_1 and G_2 are DCGs. ∎

Lemma B.2. *Suppose that Assumption 2.2 holds. Then, for each pair $i, j \in N$,*

$$\mathbf{E}[Y_{ij}^*(1) - Y_{ij}^*(0) \mid \mathcal{F}] = \frac{\text{Cov}(Y_{i,1}, Y_{j,0} \mid \mathcal{F})}{\mu_{j,0}(1 - \mu_{j,0})},$$

where we recall the definition: $\mu_{j,0} = \mathbf{E}[Y_{j,0} \mid \mathcal{F}]$.

Proof: Since $Y_{j,0} \in \{0, 1\}$, we write

$$Y_{i,1} Y_{j,0} = Y_{ij}^*(1) Y_{j,0} \quad \text{and} \quad Y_{i,1}(1 - Y_{j,0}) = Y_{ij}^*(0)(1 - Y_{j,0}).$$

Hence, taking conditional expectations given $\mathcal{F}$, and, using Assumption 2.2,

$$\frac{\mathbf{E}[Y_{i,1} Y_{j,0} \mid \mathcal{F}]}{\mu_{j,0}} - \frac{\mathbf{E}[Y_{i,1}(1 - Y_{j,0}) \mid \mathcal{F}]}{1 - \mu_{j,0}} = \mathbf{E}[Y_{ij}^*(1) - Y_{ij}^*(0) \mid \mathcal{F}].$$

We write the difference on the left-hand side as

$$\frac{(1 - \mu_{j,0}) \mathbf{E}[Y_{i,1} Y_{j,0} \mid \mathcal{F}] - \mu_{j,0} \mathbf{E}[Y_{i,1}(1 - Y_{j,0}) \mid \mathcal{F}]}{\mu_{j,0}(1 - \mu_{j,0})} = \frac{\text{Cov}(Y_{i,1}, Y_{j,0} \mid \mathcal{F})}{\mu_{j,0}(1 - \mu_{j,0})}.$$

The equality follows because $\mu_{j,0} = \mathbf{E}[Y_{j,0} \mid \mathcal{F}]$. ∎

Proof of Lemma 2.1: (i) By Lemma B.2, we obtain that

$$\text{ADM} = \sum_{j \in N} \frac{w_j}{\sigma_{j,0}^2} \sum_{i \in N: ij \in E_{\text{cau}}} \text{Cov}(Y_{i,1}, Y_{j,0} \mid \mathcal{F})$$

$$= \sum_{j \in N} \frac{w_j}{\sigma_{j,0}^2} \sum_{i \in N: ij \in E_{\text{obs}}} \text{Cov}(Y_{i,1}, Y_{j,0} \mid \mathcal{F}) = \mathsf{C}.$$

The second equality follows from Lemma B.1 because both G_{cau} and G_{obs} are DCGs.

(ii) Using Lemma B.2, we write

$$\text{ADM} = \sum_{j \in N} \frac{w_j}{\sigma_{j,0}^2} \sum_{i \in N: ij \in E_{\text{cau}} \setminus E_{\text{obs}}} \text{Cov}(Y_{i,1}, Y_{j,0} \mid \mathcal{F}) + \sum_{j \in N} \frac{w_j}{\sigma_{j,0}^2} \sum_{i \in N: ij \in E_{\text{cau}} \cap E_{\text{obs}}} \text{Cov}(Y_{i,1}, Y_{j,0} \mid \mathcal{F})$$

$$= \sum_{j \in N} \frac{w_j}{\sigma_{j,0}^2} \sum_{i \in N: ij \in E_{\text{cau}} \setminus E_{\text{obs}}} \text{Cov}(Y_{i,1}, Y_{j,0} \mid \mathcal{F}) + \mathsf{C},$$

because for each $ij \in E_{\text{obs}} \setminus E_{\text{cau}}$, $\text{Cov}(Y_{i,1}, Y_{j,0} \mid \mathcal{F}) = 0$ by the assumption that G_{cau} is a DCG. The desired result follows by (2.7) and Lemma B.2. ∎

We let $\varepsilon_{i,1} = Y_{i,1} - \mathbf{E}[Y_{i,1} \mid \mathcal{F}]$ and $\varepsilon_{j,0} = Y_{j,0} - \mathbf{E}[Y_{j,0} \mid \mathcal{F}]$.

Proof of Proposition 2.1: Let $\mathcal{F}_{-S} = \sigma(G_{\text{ctt}}, G_{\text{obs}}, X_{-S}, \mathcal{C})$. Define

$$\varepsilon_{j,0,-S} = Y_{j,0} - \mathbf{E}[Y_{j,0} \mid \mathcal{F}_{-S}] \text{ and } \varepsilon_{i,1,-S} = Y_{i,1} - \mathbf{E}[Y_{i,1} \mid \mathcal{F}_{-S}].$$

Let us write

$$\varepsilon_{j,0,-S} = \varepsilon_{j,0} + \mu_{j,0,-S}^{\Delta} \text{ and } \varepsilon_{i,1,-S} = \varepsilon_{i,1} + \mu_{i,1,-S}^{\Delta},$$

where

$$\mu_{j,0,-S}^{\Delta} = \mathbf{E}[Y_{j,0} \mid \mathcal{F}] - \mathbf{E}[Y_{j,0} \mid \mathcal{F}_{-S}] \text{ and } \mu_{i,1,-S}^{\Delta} = \mathbf{E}[Y_{i,1} \mid \mathcal{F}] - \mathbf{E}[Y_{i,1} \mid \mathcal{F}_{-S}].$$

Then we can write C_{-S} as

(B.1) $$\sum_{j \in N} \sum_{i \in N : ij \in E_{\text{cau}}} \frac{\text{Cov}(Y_{i,1}, Y_{j,0} \mid \mathcal{F}_{-S}) w_j}{\text{Var}(Y_{j,0} \mid \mathcal{F}_{-S})}$$

$$= \sum_{j \in N} \sum_{i \in N : ij \in E_{\text{cau}}} \frac{\text{Cov}(\varepsilon_{i,1}, \varepsilon_{j,0} \mid \mathcal{F}_{-S}) w_j}{\text{Var}(Y_{j,0} \mid \mathcal{F}_{-S})} + \sum_{j \in N} \sum_{i \in N : ij \in E_{\text{cau}}} \frac{\text{Cov}\left(\mu_{i,1,-S}^{\Delta}, \mu_{j,0,-S}^{\Delta} \mid \mathcal{F}_{-S}\right) w_j}{\text{Var}(Y_{j,0} \mid \mathcal{F}_{-S})}.$$

The equality above follows because $\mathcal{F}_{-S} \subset \mathcal{F}$, $\mathbf{E}[\varepsilon_{j,0} \mid \mathcal{F}] = 0$, and $\mathbf{E}[\varepsilon_{i,1} \mid \mathcal{F}] = 0$. The leading term on the right-hand side is zero because whenever $i \neq j$, $Y_{i,1}$ and $Y_{j,0}$ are conditionally independent given $\mathcal{F}$ due to the no-diffusion assumption. (Note that the conditional independence assumption in this proposition implies Assumption 2.1.) For the last covariance, since $Y_{j,0}$ is conditionally independent of X_{-j} given $(G_{\text{obs}}, G_{\text{ctt}}, \mathcal{C})$, we find that $\mathbf{E}[Y_{j,0} \mid \mathcal{F}]$ is a function of $(X_j, G_{\text{obs}}, G_{\text{ctt}}, \mathcal{C})$ only, and so is $\mu_{j,0,-S}^{\Delta}$.

Similarly, since $(U_i, Y_{i,0})$ is conditionally independent of X_{-i} given $(G_{\text{obs}}, G_{\text{ctt}}, \mathcal{C})$, using (2.3) and (2.9), we find that $\mathbf{E}[Y_{i,1} \mid \mathcal{F}]$ is also a function of $(X_i, G_{\text{obs}}, G_{\text{ctt}}, \mathcal{C})$ only, and so is $\mu_{i,1,-S}^{\Delta}$. Since X_i's are conditionally independent given $(G_{\text{obs}}, G_{\text{ctt}}, \mathcal{C})$, the conditional covariance in the last term in (B.1) is zero. ■

B.2. Preliminary Results on Moments

For the rest of the proofs, notation C represents a positive constant that does not depend on n. Constant C may take different values in different places. For each $i \in N$, let

$$N(i) = \{j \in N : ij \in E_{\text{cau}} \text{ or } ij \in E_{\text{obs}}\}.$$

Hence $N(i)$ is the union of the in-neighborhoods of i in the causal graph G_{cau} and the observed graph G_{obs}. Let $G = (N, E)$ be the graph defined by the neighborhoods $N(i)$, that is, $ij \in E$ if and only if $j \in N(i)$. Define

(B.2) $$d_{mx} = \max_{i \in N} |N(i)| \text{ and } d_{av} = \frac{1}{n} \sum_{i \in N} |N(i)|.$$

Hence d_{mx} and d_{av} are the maximum and average degrees of the union of G_{cau} and G_{obs}. Let $\overline{N}(i) = N(i) \cup \{i\}$. From here on, we assume that Assumption A.1 (in place of Assumption 2.1) and Assumptions 2.2-3.4 hold.

For each $i \in N$, define

$$e_{i,1} = Y_{i,1} - c_i.$$

Recall that we have chosen c_i to be an $\mathcal{F}$-measurable random variable.

Lemma B.3. *(i)*

$$Var\left(\frac{1}{\sqrt{n}} \sum_{i \in N} a_i \mid \mathcal{F} \right) = O_P\left(d_{av} d_{mx} \right).$$

(ii)

$$Var\left(\frac{1}{\sqrt{n}} \sum_{i \in N} e_{i,1} a_i \mid \mathcal{F} \right) = O_P\left(d_{av} d_{mx}^3 \right).$$

Proof: (i) We write

(B.3) $$Var\left(\frac{1}{\sqrt{n}} \sum_{i \in N} a_i \mid \mathcal{F} \right) = \frac{1}{n} \sum_{i_1, i_2 \in N} \sum_{j \in N_{\text{obs}}(i_1) \cap N_{\text{obs}}(i_2)} \frac{n^2 w_j^2 \mathbf{E}\left[\varepsilon_{j,0}^2 \mid \mathcal{F} \right]}{\sigma_{j,0}^4}.$$

The last equality follows because $\mathbf{E}[Y_{j,0} \mid \mathcal{F}] = \mathbf{E}[Y_{j,0} \mid G_{\text{obs}}, X]$, and $Y_{j,0}$'s are conditionally independent across j's given $\mathcal{F}$ by Assumption 2.1. By Assumptions 2.2(ii) and 3.2(iii), we have

$$\max_{j \in N} \frac{n^2 w_j^2 \mathbf{E}\left[\varepsilon_{j,0}^2 \mid \mathcal{F} \right]}{\sigma_{j,0}^4} \leq \max_{j \in N} \frac{n^2 w_j^2}{\sigma_{j,0}^4} \leq \frac{\overline{w}^2}{c^4}.$$

Hence the last term in (B.3) is bounded by

$$\frac{\overline{w}^2}{c^4} \times \frac{1}{n} \sum_{i_1, i_2 \in N} |N_{\text{obs}}(i_1) \cap N_{\text{obs}}(i_2)|.$$

The last sum over $i_1, i_2 \in N$ is bounded by the number of the pairs (i_1, i_2) such that $N_{\text{obs}}(i_1) \cap N_{\text{obs}}(i_2) \neq \emptyset$. This number is bounded by

(B.4) $$\sum_{i_1 \in N} \sum_{k \in N_{\text{obs}}(i_1)} |N_{\text{obs}}(k)| \leq n d_{av} d_{mx}.$$

This gives the desired rate.

(ii) Whenever i_1, i_2 are such that $\overline{N}_{\text{cau}}(i_1) \cap \overline{N}_{\text{cau}}(i_2) = \varnothing$, for any $j_1 \in N_{\text{obs}}(i_1)$ and $j_2 \in N_{\text{obs}}(i_2)$, we have $\text{Cov}\left(\varepsilon_{j_1,0} e_{i_1,1}, \varepsilon_{j_2,0} e_{i_2,1} \mid \mathcal{F}\right) = 0$, because G_{cau} is a DCG. We write

$$\text{Var}\left(\frac{1}{\sqrt{n}} \sum_{i \in N} e_{i,1} a_i \mid \mathcal{F}\right)$$

$$= \frac{1}{n} \sum_{i_1,i_2 \in N : \overline{N}_{\text{cau}}(i_1) \cap \overline{N}_{\text{cau}}(i_2) \neq \varnothing} \sum_{j_1 \in N_{\text{obs}}(i_1), j_2 \in N_{\text{obs}}(i_2)} \frac{n^2 w_{j_1} w_{j_2} \text{Cov}\left(\varepsilon_{j_1,0} e_{i_1,1}, \varepsilon_{j_2,0} e_{i_2,1} \mid \mathcal{F}\right)}{\sigma_{j_1,0}^2 \sigma_{j_2,0}^2}$$

$$\leq \frac{\overline{w}^2}{c^4} \frac{1}{n} \sum_{i_1,i_2 \in N : \overline{N}_{\text{cau}}(i_1) \cap \overline{N}_{\text{cau}}(i_2) \neq \varnothing} |N_{\text{obs}}(i_1)||N_{\text{obs}}(i_2)| \leq \frac{\overline{w}^2 d_{av} d_{mx}^3}{c^4}.$$

The last inequality comes from the same arguments involving the bound (B.4). ∎

Lemma B.4.

$$Var\left(\frac{1}{\sqrt{n}} \sum_{i \in N} q_i \mid \mathcal{F}\right) = O_P\left(d_{av} d_{mx}^3\right).$$

Proof: Note that

(B.5) $$\text{Var}\left(\frac{1}{\sqrt{n}} \sum_{i \in N} q_i \mid \mathcal{F}\right) \leq 2\text{Var}\left(\frac{1}{\sqrt{n}} \sum_{i \in N} e_{i,1} a_i \mid \mathcal{F}\right) + 2\text{Var}\left(\frac{1}{\sqrt{n}} \sum_{i \in N} \psi_i \mid \mathcal{F}\right).$$

The leading term on the right-hand side is $O_P(d_{av} d_{mx}^3)$ by Lemma B.3(ii). Since G_{cau} is a DCG for (Y_0, Y_1) given $\mathcal{F}$, $Y_{j,0}$'s are conditionally independent across j's given $\mathcal{F}$. Therefore, by Assumption 3.1, ψ_{i_1} and ψ_{i_2} are conditionally independent given $\mathcal{F}$, whenever

$$\left(\overline{N}_{\text{cau}}(i_1) \cup \overline{N}_{\text{obs}}(i_1)\right) \cap \left(\overline{N}_{\text{cau}}(i_2) \cup \overline{N}_{\text{obs}}(i_2)\right) = \varnothing.$$

Hence, following the same arguments in the proof of Lemma B.3(i), we find that the last term on the right-hand side of (B.5) is $O_P(d_{av} d_{mx}^3)$. ∎

Lemma B.5. *Suppose that the contact network is formed according to (3.16) and the condition (3.17) is satisfied. Then*

$$d_{mx} = O_P\left((\log n)^{kt_1+1}\right),$$

where k is the constant that appears in (3.17).

Proof: Define

$$\pi_{n,c} = \max_{i \in N} \sum_{j \in N \setminus \{i\}} P\left\{\varphi_{n,ij}(X, \eta) \geq v_{ij} \mid X, \eta\right\}.$$

We let d_c be the geodesic distance on N in the contact network. Define

$$N_{\text{ctt}}^{\partial}(i; s) = \{j \in N : d_c(i, j) = s\}.$$

First, by (B.6) in the Supplemental Note of Kojevnikov, Marmer, and Song (2021), for each $s = 1, ..., (t_1 \wedge n)$,

$$(\text{B.6}) \qquad n^{-3.3} \geq P\left\{|N^{\partial}_{\text{ctt}}(i;s)| \geq 5.7 s^2 (\pi_{n,c} \vee 1)^s \log n\right\}$$
$$\geq P\left\{|N^{\partial}_{\text{ctt}}(i;s)| \geq 5.7 t_1^2 (\pi_{n,c} \vee 1)^{t_1} \log n\right\},$$

where $a \vee b$ denotes the maximum between a and b. For each $i \in N$, we have

$$N_{\text{cau}}(i) \subset \bigcup_{s=1}^{t_1} N^{\partial}_{\text{ctt}}(i;s).$$

(Note that for any two nodes j and i that are adjacent in the causal graph G_{cau}, the shortest path between the two nodes in the contact network G_{ctt} has a length less than or equal to t_1.) Furthermore, $N^{\partial}_{\text{ctt}}(i;s)$'s are djsjoint across s's. Hence

$$P\left\{|N_{\text{cau}}(i)| \geq 5.7 t_1^2 (\pi_{n,c} \vee 1)^{t_1} \log n\right\}$$
$$\leq P\left\{|N^{\partial}_{\text{ctt}}(i;s)| \geq 5.7 t_1^2 (\pi_{n,c} \vee 1)^{t_1} \log n, \text{ for some } s = 1, ..., t_1\right\} \leq t_1 n^{-3.3},$$

by (B.6). Therefore,

$$P\left\{\max_{i \in N} |N_{\text{cau}}(i)| \geq 5.7 t_1^2 (\pi_{n,c} \vee 1)^{t_1} \log n\right\}$$
$$\leq \sum_{i \in N} P\left\{|N_{\text{cau}}(i)| \geq 5.7 t_1^2 (\pi_{n,c} \vee 1)^{t_1} \log n\right\} \leq t_1 n^{-2.3}.$$

Using (3.16), we find that $\max_{i \in N} |N_{\text{cau}}(i)| = O_P\left((\log n)^{k t_1 + 1}\right)$. Using a similar argument, we find that $\max_{i \in N} |N_{\text{obs}}(i)| = O_P\left((\log n)^{k t_1 + 1}\right)$. Since $d_{mx} \leq \max_{i \in N} |N_{\text{cau}}(i)| + \max_{i \in N} |N_{\text{obs}}(i)|$, we obtain the desired result. ∎

B.3. Asymptotic Normality

Define

$$\hat{e}_{i,1} = Y_{i,1} - \hat{c}_i \text{ and } \hat{\varepsilon}_{j,0} = Y_{j,0} - \hat{\mu}_{j,0}.$$

Then we can write

$$\hat{C} = \frac{1}{n} \sum_{i \in N} \hat{e}_{i,1} \hat{a}_i,$$

where we recall $\hat{a}_i = \sum_{j \in N : ij \in E_{\text{obs}}} n w_j \hat{\varepsilon}_{j,0} / \hat{\sigma}^2_{j,0}$ and $\hat{\sigma}^2_{j,0} = \hat{\mu}_{j,0}(1 - \hat{\mu}_{j,0})$.

Lemma B.6.

$$\sqrt{n}(\hat{C} - C) = \frac{1}{\sqrt{n}} \sum_{i \in N} (q_i - \mathbf{E}[q_i \mid \mathcal{F}]) + o_P(1).$$

Proof: First, we show that

(B.7)
$$\max_{i \in N} |\hat{a}_i - a_i| = O_P\left(n^{-\kappa}(\log n)^C\right),$$

where κ is the constant in Assumption 3.2(ii) and C is the constant in Assumption 3.4. For $i \in N$, we write

$$\hat{a}_i - a_i = \sum_{j \in N: ij \in E_{obs}} n w_j \left(A_{1j} + A_{2j} + A_{3j}\right),$$

where

$$A_{1j} = \frac{\mu_{j,0} - \hat{\mu}_{j,0}}{\sigma_{j,0}^2}, \quad A_{2j} = (\mu_{j,0} - \hat{\mu}_{j,0})\left(\frac{1}{\hat{\sigma}_{j,0}^2} - \frac{1}{\sigma_{j,0}^2}\right), \text{ and } A_{3j} = \varepsilon_{j,0}\left(\frac{1}{\hat{\sigma}_{j,0}^2} - \frac{1}{\sigma_{j,0}^2}\right).$$

By Assumptions 2.2(ii) and 3.2(ii), we have

$$\max_{j \in N} |A_{1j}| = O_P(n^{-\kappa}).$$

As for A_{2j}, again, using Assumptions 2.2(ii) and 3.2(ii), we bound

$$\max_{j \in N} |A_{2j}| \leq \max_{j \in N} \left| (\mu_{j,0} - \hat{\mu}_{j,0})\left(\frac{\hat{\mu}_{j,0}(1 - \hat{\mu}_{j,0}) - \mu_{j,0}(1 - \mu_{j,0})}{c^4 + o_P(1)}\right) \right|$$

$$\leq \max_{j \in N} \frac{C(\mu_{j,0} - \hat{\mu}_{j,0})^2}{c^4 + o_P(1)} = O_P\left(n^{-2\kappa}\right).$$

Using the same argument, we also obtain that $\max_{j \in N} |A_{3j}| = O_P\left(n^{-\kappa}\right)$. Hence,

$$\max_{i \in N} |\hat{a}_i - a_i| = O_P\left(n^{-\kappa} d_{mx}\right) = O_P\left(n^{-\kappa}(\log n)^C\right),$$

by Assumption 3.4, obtaining (B.7).

By Assumption 3.2(ii),

$$\frac{1}{\sqrt{n}}\sum_{i \in N}(\hat{c}_i - c_i)(\hat{a}_i - a_i) \leq \sqrt{n}\max_{i \in N}|\hat{c}_i - c_i|\max_{i \in N}|\hat{a}_i - a_i| = O_P\left(n^{-2\kappa}(\log n)^C\right) = o_P(1).$$

Using this result and Assumption 3.1, we can write

$$\sqrt{n}\{\hat{C} - C\} = \frac{1}{\sqrt{n}}\sum_{i \in N}\left(\hat{e}_{i,1}\hat{a}_i - \mathbf{E}[e_{i,1}a_i \mid \mathcal{F}]\right) = \frac{1}{\sqrt{n}}\sum_{i \in N}\left((Y_{i,1} - \hat{c}_i)\hat{a}_i - \mathbf{E}[e_{i,1}a_i \mid \mathcal{F}]\right)$$

$$= \frac{1}{\sqrt{n}}\sum_{i \in N}(Y_{i,1} - c_i)(\hat{a}_i - a_i) - \frac{1}{\sqrt{n}}\sum_{i \in N}(\hat{c}_i - c_i)a_i$$

$$+ \frac{1}{\sqrt{n}}\sum_{i \in N}\left(e_{i,1}a_i - \mathbf{E}[e_{i,1}a_i \mid \mathcal{F}]\right) + o_P(1)$$

$$= \frac{1}{\sqrt{n}}\sum_{i \in N}(q_i - \mathbf{E}[q_i \mid \mathcal{F}]) + o_P(1).$$

∎

Lemma B.7. $\max_{i \in N} \mathbf{E}\left[q_i^4 \mid \mathcal{F}\right] = O_P\left(d_{mx}^4\right)$ and $\max_{i \in N} h_i^4 = O_P\left(d_{mx}^4\right)$.

Proof: For the first statement, note that

$$
\text{(B.8)} \qquad \max_{i \in N} \mathbf{E}\left[q_i^4 \mid G_{\text{obs}}, X\right] \le 8 \max_{i \in N} \mathbf{E}\left[e_{i,1}^4 a_i^4 \mid G_{\text{obs}}, X\right]
$$
$$
+ 8 \max_{i \in N} \mathbf{E}\left[\psi_i^4 \mid G_{\text{obs}}, X\right] = O_P\left(d_{mx}^4\right),
$$

because $a_i = \sum_{j \in N_{\text{obs}}(i)} n w_j \varepsilon_{j,0} / \sigma_{j,0}^2$, and due to (3.15).

For the second statement, let

$$
M = \frac{1}{n} \sum_{i \in N} X_i X_i' \text{ and } b = \frac{1}{n} \sum_{i \in N} X_i \mathbf{E}\left[q_i \mid \mathcal{F}\right].
$$

Then, by Cauchy-Schwarz inequality,

$$
h_i^2 = ((M^{-1}b)'X_i)^2 \le (M^{-1}b)'(M^{-1}b)X_i'X_i
$$
$$
= \text{tr}\left(M^{-2}bb'\right)\|X_i\|^2 \le \frac{1}{c'^2}\max_{i \in N}\|X_i\|^4 \max_{i \in N}\mathbf{E}\left[q_i^2 \mid \mathcal{F}\right] = O(d_{mx}^2),
$$

by (B.8) and Assumption 3.3(ii). $\blacksquare$

Define a graph $G^* = (N, E^*)$, where for $i_1 \ne i_2$, $i_1 i_2 \in E^*$ if and only if $\overline{N}(i_1) \cap \overline{N}(i_2) \ne \varnothing$, and let $\overline{E}^* = E^* \cup \{ii : i \in N\}$. Then $G^* = (N, E^*)$ contains $G = (N, E)$ as a subgraph. Let d_{mx}^* and d_{av}^* be the maximum degree and the average degree of G^*, that is,

$$
d_{mx}^* = \max_{i \in N} |N^*(i)| \text{ and } d_{av}^* = \frac{1}{n} \sum_{i \in N} |N^*(i)|.
$$

By definition, we have

$$
\text{(B.9)} \qquad d_{mx}^* = O(d_{mx}^2), \text{ and } d_{av}^* = O(d_{mx} d_{av}),
$$

where d_{mx} and d_{av} are defined in (B.2). We also let E_{obs}^* be the collection of $i_1 i_2$ such that $\overline{N}_{\text{obs}}(i_1) \cap \overline{N}_{\text{obs}}(i_2) \ne \varnothing$, and similarly, let E_{cau}^* be the collection of $i_1 i_2$ such that $\overline{N}_{\text{cau}}(i_1) \cap \overline{N}_{\text{cau}}(i_2) \ne \varnothing$. We let $\overline{E}_{\text{obs}}^* = E_{\text{obs}}^* \cup \{ii : i \in N\}$ and $\overline{E}_{\text{cau}}^* = E_{\text{cau}}^* \cup \{ii : i \in N\}$. Then we define the graphs G_{obs}^* and G_{cau}^* as follows:

$$
\text{(B.10)} \qquad G_{\text{obs}}^* = (N, E_{\text{obs}}^*) \text{ and } G_{\text{cau}}^* = (N, E_{\text{cau}}^*).
$$

Recall the definition:

$$
\sigma^2 = \mathbf{E}\left[\left(\frac{1}{\sqrt{n}} \sum_{i \in N}(q_i - \mathbf{E}[q_i \mid \mathcal{F}])\right)^2 \mid \mathcal{F}\right].
$$

Lemma B.8.

$$\sup_{u\in\mathbf{R}}\left|P\left\{\frac{\sqrt{n}(\hat{C}-C)}{\sigma}\leq u\mid\mathcal{F}\right\}-\Phi(u)\right|\to 0,\ as\ n\to\infty,$$

where Φ is the distribution function of $N(0,1)$.

Proof: For each $i\in N$, q_i is a function of $(Y_{i,1},c_i,a_i,\psi_i)$ which is measurable with respect to the σ-field generated by $Y_{i,1}$, $(Y_{j,0})_{j\in\overline{N}_{\mathrm{cau}}(i)\cup\overline{N}_{\mathrm{obs}}(i)}$, and $\mathcal{F}$, by (2.4), and Assumption 3.1. By Assumption A.1, $\{q_i\}_{i\in N}$ has G^* as a conditional dependency graph given $\mathcal{F}$, which is a special case of conditional neighborhood dependency introduced in Lee and Song (2019). We apply their Corollary 3.1 to deduce that

$$\sup_{u\in\mathbf{R}}\left|P\left\{\frac{1}{\sigma\sqrt{n}}\sum_{i\in N}(q_i-\mathbf{E}[q_i\mid\mathcal{F}])\leq u\mid\mathcal{F}\right\}-\Phi(u)\right|$$
$$\leq C\left(\frac{\sqrt{d^*_{mx}d^*_{av}\mu^3_3}}{n^{1/4}}-\log\left(\frac{d^*_{mx}d^*_{av}\mu^3_3}{\sqrt{n}}\right)\frac{\sqrt{(d^*_{mx})^2 d^*_{av}\mu^4_4}}{\sqrt{n}}\right),$$

for some constant $C>0$ that does not depend on n, where

$$\mu^p_p=\max_{i\in N}\mathbf{E}\left[\left|\frac{q_i-\mathbf{E}[q_i\mid\mathcal{F}]}{\sigma}\right|^p\mid\mathcal{F}\right].$$

Thus, the desired result follows from this and Assumption 3.4 and Lemmas B.6 and B.7. ∎

Recall the definitions of $\tilde{\sigma}^2$ and $\tilde{\sigma}^2_{\mathrm{obs}}$ which we write as

$$\tilde{\sigma}^2=\frac{1}{n}\sum_{i_1 i_2\in\overline{E}^*}\mathbf{E}\left[(q_{i_1}-h_{i_1})(q_{i_2}-h_{i_2})\mid\mathcal{F}\right]\ \text{and}$$

$$\tilde{\sigma}^2_{\mathrm{obs}}=\frac{1}{n}\sum_{i_1 i_2\in\overline{E}^*_{\mathrm{obs}}}\mathbf{E}\left[(q_{i_1}-h_{i_1})(q_{i_2}-h_{i_2})\mid\mathcal{F}\right],$$

where h_i is as defined in (3.5). Note that when G_{obs} is a DCG, we have $\tilde{\sigma}^2=\tilde{\sigma}^2_{\mathrm{obs}}$.

Lemma B.9. $\tilde{\sigma}^2\geq\sigma^2$.

Proof: Let A be the $n\times n$ matrix whose (i,j)-th entry is $1/n$ if $ij\in\overline{E}^*$ and zero otherwise. Furthermore, A is positive semidefinite.[24] Define

$$q=[q_1,...,q_n]'\ \text{and}\ h=[h_1,...,h_n]',$$

---

[24]As shown in Kojevnikov (2021), this follows because $c'Ac=\frac{1}{n}\sum_{j\in N}b^2_j$ with $b_j=\sum_{i\in\overline{N}(j)}c_i$ for any vector $c=[c_1,...,c_n]'$.

where h_i's are as defined in (3.5). Then, note that

$$
\begin{aligned}
\sigma^2 &= \mathbf{E}\left[(q - \mathbf{E}[q \mid \mathcal{F}])'A(q - \mathbf{E}[q \mid \mathcal{F}]) \mid \mathcal{F}\right] \\
&= \mathbf{E}\left[(q - h)'A(q - h) \mid \mathcal{F}\right] - (\mathbf{E}[q \mid \mathcal{F}] - h)'A(\mathbf{E}[q \mid \mathcal{F}] - h) \\
&= \tilde{\sigma}^2 - (\mathbf{E}[q \mid \mathcal{F}] - h)'A(\mathbf{E}[q \mid \mathcal{F}] - h),
\end{aligned}
$$

because A and h are measurable with respect to $\mathcal{F}$. The last term is nonnegative because A is positive semidefinite. ∎

Lemma B.10. *For each $c > 0$,*

$$
P\left\{\left|\frac{\sqrt{n}(\hat{C} - C)}{\tilde{\sigma}}\right| \le c\right\} \ge P\{|\mathbb{Z}| \le c\} + o(1),
$$

where $\mathbb{Z} \in \mathbf{R}$ is a random variable distributed as $N(0,1)$.

Proof: First, by Lemma B.8, we find that for any Borel set B,

$$
P\left\{\frac{\sqrt{n}(\hat{C} - C)}{\tilde{\sigma}} \in B\right\} = P\left\{\frac{\sigma\mathbb{Z}}{\tilde{\sigma}} \in B\right\} + o(1).
$$

Note that

$$
\begin{aligned}
P\left\{\left|\frac{\sqrt{n}(\hat{C} - C)}{\tilde{\sigma}}\right| \le c\right\} &= P\left\{\left|\frac{\sqrt{n}(\hat{C} - C)}{\sigma}\right| \le \frac{\tilde{\sigma}c}{\sigma}\right\} \\
&= P\left\{|\mathbb{Z}| \le \frac{\tilde{\sigma}c}{\sigma}\right\} + o(1) \ge P\{|\mathbb{Z}| \le c\} + o(1),
\end{aligned}
$$

by Lemma B.9. ∎

B.4. Consistency of Variance Estimators

Lemma B.11. $\hat{C} = C + o_P(1)$.

Proof: The statement follows from Lemma B.4, Assumption 3.4, and Lemma B.6. ∎

Lemma B.12. $\max_{i \in N} |\hat{q}_i - q_i| = O_P\left(n^{-\kappa}(\log n)^C\right)$.

Proof: By Assumption 3.2, it suffices to show that

$$
\max_{i \in N} |\hat{e}_{i,1}\hat{a}_i - e_{i,1}a_i| = O_P\left(n^{-\kappa}(\log n)^C\right).
$$

Since c_i takes values from $[0,1]$,

$$
\max_{i \in N} |\hat{e}_{i,1}\hat{a}_i - e_{i,1}a_i| \le \max_{i \in N} |\hat{e}_{i,1}(\hat{a}_i - a_i)| + \max_{i \in N} |(\hat{e}_{i,1} - e_{i,1})a_i|
$$

$$
\le \max_{i \in N} |\hat{c}_i - c_i||\hat{a}_i - a_i| + 2\max_{i \in N} |\hat{a}_i - a_i| + \max_{i \in N} |\hat{c}_i - c_i||a_i| = O_P\left(n^{-\kappa}(\log n)^C\right),
$$

from Assumptions 3.2 and 3.4, and (B.7). ∎

Lemma B.13.

$$\frac{d_{mx}^6}{n}\sum_{i\in N}(\hat{q}_i - q_i)^2 = o_P(1) \text{ and } \frac{d_{mx}^6}{n}\sum_{i\in N}(\hat{h}_i - h_i)^2 = o_P(1).$$

Proof: The first result comes from Lemma B.12 and Assumption 3.4. As for the second result, $\hat{h} = [\hat{h}_1, ..., \hat{h}_n]'$ is the projection of $\hat{q} = [\hat{q}_1, ..., \hat{q}_n]'$ onto the range space of $X = [X_1, ..., X_n]'$. Let P_X be the projection matrix projecting from $\mathbf{R}^n$ onto the range space of X. Define $\tilde{h} = [\tilde{h}_1, ..., \tilde{h}_n]'$, where $\tilde{h}_i = \tilde{\lambda}'X_i$, with

$$\tilde{\lambda} = \left(\frac{1}{n}\sum_{i\in N}X_iX_i'\right)^{-1}\frac{1}{n}\sum_{i\in N}X_iq_i.$$

Then

$$\|\hat{h} - \tilde{h}\|^2 \leq \mathrm{tr}\left(P_X(\hat{q}-q)(\hat{q}-q)'P_X\right) = \mathrm{tr}\left(P_X(\hat{q}-q)(\hat{q}-q)'\right) \leq \|\hat{q}-q\|^2,$$

because the maximum eigenvalue of P_X is at most 1. Hence

(B.11) $$\frac{d_{mx}^6}{n}\sum_{i\in N}(\hat{h}_i - \tilde{h}_i)^2 \leq \frac{d_{mx}^6}{n}\sum_{i\in N}(\hat{q}_i - q_i)^2 = o_P(1).$$

Now, consider

$$\frac{1}{n}\|\tilde{h} - h\|^2 = \left(\frac{1}{n}\sum_{i\in N}(q_i - \mathbf{E}[q_i \mid \mathcal{F}])X_i'\right)\left(\frac{1}{n}\sum_{i\in N}X_iX_i'\right)^{-1}\left(\frac{1}{n}\sum_{i\in N}(q_i - \mathbf{E}[q_i \mid \mathcal{F}])X_i\right).$$

Let X_{ik} be the k-th entry of X_i. Then,

$$\mathbf{E}\left[\left(\frac{1}{n}\sum_{i\in N}(q_i - \mathbf{E}[q_i \mid \mathcal{F}])X_{ik}\right)^2 \mid \mathcal{F}\right] = \frac{1}{n^2}\sum_{i_1i_2\in\overline{E}^*}\mathrm{Cov}\left(q_{i_1}X_{i_1k}, q_{i_2}X_{i_2k} \mid \mathcal{F}\right)$$

$$= O_P\left(\frac{d_{av}^*}{n}\right) = O_P\left(\frac{d_{av}d_{mx}}{n}\right),$$

where the second to the last equality uses the assumption that $\max_{i\in N}\|X_i\|^2 < C$ in Assumption 3.3 (ii) and the last equality uses (B.9). Hence we have

$$\frac{d_{mx}^6}{n}\sum_{i\in N}(\tilde{h}_i - h_i)^2 = O_P\left(\frac{d_{mx}^7 d_{av}}{n}\right) = o_P(1),$$

by Assumption 3.4. Combining this with (B.11), we obtain the second result of this lemma. ∎

Define $\eta_i = q_i - h_i$, $\hat{\eta}_i = \hat{q}_i - \hat{h}_i$, and

$$V = \frac{1}{n} \sum_{i_1 i_2 \in \overline{E}_{\text{obs}}^*} \left(\hat{\eta}_{i_1} \hat{\eta}_{i_2} - \eta_{i_1} \eta_{i_2} \right),$$

where we recall $\overline{E}_{\text{obs}}^* = E_{\text{obs}}^* \cup \{ii : i \in N\}$.

Lemma B.14. $V = o_P(1)$.

Proof: We write

$$V = \frac{1}{n} \sum_{i_1 i_2 \in \overline{E}_{\text{obs}}^*} \left(\hat{\eta}_{i_1} - \eta_{i_1} \right) \hat{\eta}_{i_2} + \frac{1}{n} \sum_{i_1 i_2 \in \overline{E}_{\text{obs}}^*} \left(\hat{\eta}_{i_2} - \eta_{i_2} \right) \eta_{i_1} = A_{1n} + A_{2n}, \text{ say.}$$

We consider first A_{1n}. We write it as $A_{1n}' + A_{1n}''$, where

$$A_{1n}' = \frac{1}{n} \sum_{i_1 i_2 \in \overline{E}_{\text{obs}}^*} \left(\hat{\eta}_{i_1} - \eta_{i_1} \right) \eta_{i_2}, \text{ and}$$

$$A_{1n}'' = \frac{1}{n} \sum_{i_1 i_2 \in \overline{E}_{\text{obs}}^*} \left(\hat{\eta}_{i_1} - \eta_{i_1} \right) \left(\hat{\eta}_{i_2} - \eta_{i_2} \right).$$

Write A_{1n}' as

$$\frac{1}{n} \sum_{i_1 i_2 \in \overline{E}_{\text{obs}}^*} (\hat{\eta}_{i_1} - \eta_{i_1}) \eta_{i_2} = \frac{1}{n} \sum_{i_1 i_2 \in \overline{E}_{\text{obs}}^*} (\hat{q}_{i_1} - q_{i_1})(q_{i_2} - h_{i_2})$$

$$- \frac{1}{n} \sum_{i_1 i_2 \in \overline{E}_{\text{obs}}^*} (\hat{h}_{i_1} - h_{i_1})(q_{i_2} - h_{i_2}) = A_{11n}' - A_{12n}', \text{ say.}$$

We bound A_{11n}' by

$$\sqrt{\frac{1}{n} \sum_{i_1 \in N} |\overline{N}_{\text{obs}}^*(i_1)| \left(\hat{q}_{i_1} - q_{i_1} \right)^2} \sqrt{\frac{1}{n} \sum_{i_1 \in N} \sum_{i_2 \in \overline{N}_{\text{obs}}^*(i_1)} (q_{i_2} - h_{i_2})^2}.$$

Defining P_X, q and h as in the proof of Lemma B.13,

$$\sum_{i \in N} \mathbf{E}\left[h_i^2 \mid G_{\text{obs}}, X \right] = \mathbf{E}\left[\|h\|^2 \mid G_{\text{obs}}, X \right] \le \mathbf{E}\left[\|P_X \mathbf{E}[q \mid \mathcal{F}]\|^2 \mid G_{\text{obs}}, X \right]$$

$$\le \mathbf{E}\left[\mathbf{E}\left[\|P_X q\|^2 \mid \mathcal{F} \right] \mid G_{\text{obs}}, X \right] \le \sum_{i \in N} \mathbf{E}\left[q_i^2 \mid G_{\text{obs}}, X \right].$$

Hence we have

$$\frac{1}{n} \sum_{i_1 \in N} \sum_{i_2 \in \overline{N}_{\text{obs}}^*(i_1)} (q_{i_2} - h_{i_2})^2 \le \frac{2}{n} \sum_{i_1 \in N} \sum_{i_2 \in \overline{N}_{\text{obs}}^*(i_1)} q_{i_2}^2 + \frac{2}{n} \sum_{i_1 \in N} \sum_{i_2 \in \overline{N}_{\text{obs}}^*(i_1)} h_{i_2}^2$$

$$\le \frac{2d_{mx}^*}{n} \left(\|q\|^2 + \|h\|^2 \right) \le \frac{4d_{mx}^* \|q\|^2}{n} = O_P(d_{mx}^4),$$

because $\|h\|^2 \leq \|q\|^2$, and by (B.8). By Lemma B.13, $A'_{11n} = o_P(1)$. Using the same argument and Lemma B.13, we also find that $A'_{12n} = o_P(1)$.

Let us turn to A''_{1n} which is bounded by

$$\sqrt{\frac{1}{n} \sum_{i_1 \in N} |\overline{N}^*_{obs}(i_1)| \left(\hat{\eta}_{i_1} - \eta_{i_1}\right)^2} \sqrt{\frac{1}{n} \sum_{i_1 \in N} \sum_{i_2 \in \overline{N}^*_{obs}(i_1)} \left(\hat{\eta}_{i_2} - \eta_{i_2}\right)^2}$$

$$\leq \frac{d^*_{mx} + 1}{n} \sum_{i \in N} (\hat{\eta}_i - \eta_i)^2 .$$

The last term is bounded by

$$\frac{2(d^*_{mx} + 1)}{n} \sum_{i \in N} (\hat{q}_i - q_i)^2 + \frac{2(d^*_{mx} + 1)}{n} \sum_{i \in N} \left(\hat{h}_i - h_i\right)^2 = o_P(1),$$

by Lemma B.13. Finally, note that A_{2n} is the same as A'_{1n}. Thus the proof is complete. ∎

Let

$$W = \frac{1}{n} \sum_{i_1 i_2 \in \overline{E}^*_{obs}} \left(\eta_{i_1} \eta_{i_2} - \mathbf{E}[\eta_{i_1} \eta_{i_2} \mid \mathcal{F}]\right).$$

Lemma B.15. $W = o_P(1)$.

Proof: First, observe that by Lemma B.7,

(B.12) $$\max_{i_1, i_2 \in N} \mathbf{E}\left[\eta_{i_1}^2 \eta_{i_2}^2 \mid \mathcal{F}\right] = O_P(d^4_{mx}).$$

Write

$$\mathbf{E}\left[W^2 \mid \mathcal{F}\right] = \frac{1}{n^2} \sum_{i_1 i_2 \in \overline{E}^*_{obs}} \sum_{i'_1 i'_2 \in \overline{E}^*_{obs}} \mathrm{Cov}\left(\eta_{i_1} \eta_{i_2}, \eta_{i'_1} \eta_{i'_2} \mid \mathcal{F}\right).$$

Since G_{cau} is a DCG, for any $i_1 i_2, i'_1 i'_2 \in \overline{E}^*_{obs}$, $\mathrm{Cov}\left(\eta_{i_1} \eta_{i_2}, \eta_{i'_1} \eta_{i'_2} \mid \mathcal{F}\right)$ is not zero only if $\{i_1, i_2\}$ and $\{i'_1, i'_2\}$ are adjacent in G^*. Thus the number of the terms in the double sum above is of the same order as the number of ways one chooses paths $i_1 k_1 i_2 k_2 i'_1 k_3 i'_2$ in G, which is $O(n d_{av} d^5_{mx})$. With (B.12), this implies that by Assumption 3.4,

$$\mathbf{E}\left[W^2 \mid \mathcal{F}\right] = O\left(\frac{d_{av} d^9_{mx}}{n}\right) = o_P(1).$$

∎

Lemma B.16. $\hat{\sigma}^2 = \tilde{\sigma}^2_{obs} + o_P(1)$.

Proof: Note that $\hat{\sigma}^2 - \tilde{\sigma}^2_{obs} = V + W = o_P(1)$, by Lemmas B.14 and B.15. ∎

B.5. Proofs of Theorems 3.1 and 4.1

Proof of Theorem 3.1: Since G_{obs} is a DCG, we have $\text{ADM} = C$ and $\tilde{\sigma}^2 = \tilde{\sigma}^2_{\text{obs}}$. Hence

$$\hat{\sigma}^2 = \tilde{\sigma}^2_{\text{obs}} + o_P(1) \geq \sigma^2 + o_P(1),$$

by Lemmas B.9 and B.16. We find from Lemma B.10 that

$$P\{C \in \mathbb{C}_{1-\alpha}\} \geq P\{z_{\alpha/2} \leq \mathbb{Z} \leq z_{1-\alpha/2}\} + o(1) = 1 - \alpha + o(1).$$

∎

We define for any $\delta > 0$,

$$\overline{N}_{\text{ctt}}(i; \delta) = \{j \in N : \text{Cov}(Y_{i,1}, Y_{j,0} \mid \mathcal{F}) \geq \delta\} \cap \overline{N}_{\text{ctt}}(i),$$

and

$$d^{\Delta}_{\text{ctt}}(\delta) = \frac{1}{n}\sum_{i \in N}|\overline{N}_{\text{ctt}}(i; \delta) \setminus \overline{N}_{\text{obs}}(i)|, \text{ and}$$

$$d^{\Delta}_{\text{cau}} = \frac{1}{n}\sum_{i \in N}|\overline{N}_{\text{cau}}(i) \setminus \overline{N}_{\text{obs}}(i)|.$$

We let

$$e_n(\delta) = \begin{cases} d^{\Delta}_{\text{ctt}}(\delta)/d^{\Delta}_{\text{cau}}, & \text{if } d^{\Delta}_{\text{cau}} > 0, \\ \infty, & \text{if } d^{\Delta}_{\text{cau}} = 0. \end{cases}$$

The quantity $e_n(\delta)$ decreases in δ. If the observed graph G_{obs} contains the causal graph G_{cau} as a subgraph (so that G_{obs} is a DCG), then $e_n(\delta) = \infty$. The following assumption requires that $e_n(\delta_n)$ does not shrink to zero too quickly as δ_n decreases to zero.

Assumption B.2 (Nontrivially Effective Dependency Causal Graph)**.** There exists a positive sequence δ_n such that for any constant $C > 0$,

(B.13) $$P\left\{\left(d^{\text{cau}}_{mx} d_{mx}\right)^2 \leq C\sqrt{n}e_n(\delta_n)\delta_n\right\} \to 1, \text{ as } n \to \infty,$$

where d^{cau}_{mx} denotes the maximum degree of G_{cau}.

Suppose that Assumption 4.1 holds, and we take δ_n to be such that $\delta_n \leq c_1 n^{-1/2+v}$, where $c_1 > 0$ is the constant that appears in Assumption 4.1. Then,

$$e_n(\delta_n) \geq d^{-(t_1-1)}_{mx}.$$

because, for all $i \in N$,

$$|\overline{N}_{\text{cau}}(i) \setminus \overline{N}_{\text{obs}}(i)| \leq |\overline{N}_{\text{ctt}}(i) \setminus \overline{N}_{\text{obs}}(i)|\left(d^{\text{ctt}}_{mx}\right)^{t_1-1}.$$

Hence, we have (B.13) by Assumption 3.4. Therefore, Assumption B.2 is weaker than Assumption 4.1.

Lemma B.17. *Suppose that the conditions of Lemma 2.1(ii) and Assumption B.2 hold. Then,*

(B.14)
$$P\left\{z_{1-\alpha}(\tilde{\sigma} - \tilde{\sigma}_{\text{obs}}) \le \sqrt{n}(\text{ADM} - \text{C})\right\} \to 1,$$

as $n \to \infty$.

Proof: Let

$$\overline{N}^*_{\text{cau}}(i) = \left\{i' \in N : \overline{N}_{\text{cau}}(i) \cap \overline{N}_{\text{cau}}(i') \neq \varnothing\right\}, \text{ and}$$
$$\overline{N}^*_{\text{obs}}(i) = \left\{i' \in N : \overline{N}_{\text{obs}}(i) \cap \overline{N}_{\text{obs}}(i') \neq \varnothing\right\}.$$

For any sets A, A', B, B', we have

$$(A \cap A') \setminus (B \cap B') = ((A \cap A') \setminus B) \cup ((A \cap A') \setminus B')$$
$$= (A' \cap (A \setminus B)) \cup (A \cap (A' \setminus B')).$$

Hence

(B.15)
$$\frac{1}{n}\sum_{i \in N}|\overline{N}^*_{\text{cau}}(i) \setminus \overline{N}^*_{\text{obs}}(i)| \le \frac{1}{n}\sum_{i \in N}\sum_{i' \in N}\left|(\overline{N}_{\text{cau}}(i) \cap \overline{N}_{\text{cau}}(i')) \setminus (\overline{N}_{\text{obs}}(i) \cap \overline{N}_{\text{obs}}(i'))\right|$$
$$\le \frac{1}{n}\sum_{i \in N}\sum_{i' \in N}\left|\overline{N}_{\text{cau}}(i') \cap (\overline{N}_{\text{cau}}(i) \setminus \overline{N}_{\text{obs}}(i))\right|$$
$$+ \frac{1}{n}\sum_{i \in N}\sum_{i' \in N}\left|\overline{N}_{\text{cau}}(i) \cap (\overline{N}_{\text{cau}}(i') \setminus \overline{N}_{\text{obs}}(i'))\right|.$$

The last two double summations are identical, and hence

$$\frac{1}{n}\sum_{i \in N}|\overline{N}^*_{\text{cau}}(i) \setminus \overline{N}^*_{\text{obs}}(i)| \le \frac{2}{n}\sum_{i \in N}\sum_{i' \in N}\left|\overline{N}_{\text{cau}}(i') \cap (\overline{N}_{\text{cau}}(i) \setminus \overline{N}_{\text{obs}}(i))\right|.$$

As for the last term, we write it as

$$\frac{2}{n}\sum_{i \in N}\sum_{i':\overline{N}_{\text{cau}}(i') \cap (\overline{N}_{\text{cau}}(i) \setminus \overline{N}_{\text{obs}}(i)) \neq \varnothing}\left|\overline{N}_{\text{cau}}(i') \cap (\overline{N}_{\text{cau}}(i) \setminus \overline{N}_{\text{obs}}(i))\right|$$
$$\le \frac{2}{n}\sum_{i \in N}\sum_{i':\overline{N}_{\text{cau}}(i') \cap (\overline{N}_{\text{cau}}(i) \setminus \overline{N}_{\text{obs}}(i)) \neq \varnothing}\left|\overline{N}_{\text{cau}}(i) \setminus \overline{N}_{\text{obs}}(i)\right|$$
$$\le \frac{2\left(d^{\text{cau}}_{mx}\right)^2}{n}\sum_{i \in N}\left|\overline{N}_{\text{cau}}(i) \setminus \overline{N}_{\text{obs}}(i)\right| = 2\left(d^{\text{cau}}_{mx}\right)^2 d^{\Delta}_{mx}.$$

Let us first find an upper bound for $\tilde{\sigma}^2 - \tilde{\sigma}^2_{\text{obs}}$. Now, observe that

(B.16)
$$\tilde{\sigma}^2 - \tilde{\sigma}^2_{\text{obs}} = \frac{1}{n} \sum_{i_1 i_2 \in \overline{E}^* \setminus \overline{E}^*_{\text{obs}}} \mathbf{E}[\eta_{i_1} \eta_{i_2} \mid \mathcal{F}]$$

$$\leq O_P(d_{mx}^2) \times \frac{1}{n} \sum_{i_1 \in N} \sum_{i_2 \in N} \mathbf{1}\left\{ i_2 \in \overline{N}^*_{\text{cau}}(i_1), i_2 \notin \overline{N}^*_{\text{obs}}(i_1) \right\}$$

$$= O_P(d_{mx}^2) \times \frac{1}{n} \sum_{i \in N} \left| \overline{N}^*_{\text{cau}}(i) \setminus \overline{N}^*_{\text{obs}}(i) \right| \leq O_P(d_{mx}^2) \times 2 \left(d_{mx}^{\text{cau}} \right)^2 \cdot d_{\text{cau}}^{\Delta},$$

by (B.15), and because $\max_{i \in N} \mathbf{E}[\eta_i^4 \mid \mathcal{F}] = O_P\left(d_{mx}^4 \right)$, which is due to Lemma B.7. Let δ_n be the sequence in Assumption B.2. Recalling $\varepsilon_{i,1} = Y_{i,1} - \mathbf{E}[Y_{i,1} \mid \mathcal{F}]$, we note that

(B.17)
$$\text{ADM} - \text{C} = \frac{1}{n} \sum_{i \in N} \sum_{j \in N_{\text{cau}}(i) \setminus N_{\text{obs}}(i)} \frac{n w_j \mathbf{E}\left[\varepsilon_{i,1} \varepsilon_{j,0} \mid \mathcal{F} \right]}{\sigma_{j,0}^2}$$

$$\geq \frac{1}{n} \sum_{i \in N} \sum_{j \in N_{\text{ctt}}(i; \delta_n) \setminus N_{\text{obs}}(i)} \frac{n w_j \mathbf{E}\left[\varepsilon_{i,1} \varepsilon_{j,0} \mid \mathcal{F} \right]}{\sigma_{j,0}^2} \geq \frac{\overline{w} d_{\text{ctt}}^{\Delta}(\delta_n) \delta_n}{c^2},$$

for the constant $c > 0$ in Assumption 2.2(ii). The first inequality follows by the condition (2.7) assumed for $(Y_{ij}^*(1), Y_{ij}^*(0))$ and by Lemma B.2.

The probability on the left-hand side of (B.14) is equal to

$$P\left\{ \frac{z_{1-\alpha}(\tilde{\sigma}^2 - \tilde{\sigma}^2_{\text{obs}})}{\tilde{\sigma} + \tilde{\sigma}_{\text{obs}}} \leq \sqrt{n}(\text{ADM} - \text{C}) \right\} \geq P\left\{ \frac{z_{1-\alpha}(\tilde{\sigma}^2 - \tilde{\sigma}^2_{\text{obs}})}{\sqrt{c'}} \leq \sqrt{n}(\text{ADM} - \text{C}) \right\},$$

by Assumption 3.3 and Lemma B.9. By (B.16) and (B.17), for any $\epsilon > 0$, there exist $n_0 \geq 1$ and $C_\epsilon > 0$ such that for all $n \geq n_0$, the last probability is bounded from below by

$$P\left\{ z_{1-\alpha} d_{\text{cau}}^{\Delta} \left(d_{mx}^{\text{cau}} d_{mx} \right)^2 \leq C_\epsilon \sqrt{n} d_{\text{ctt}}^{\Delta}(\delta_n) \delta_n \right\} - \epsilon$$

$$= P\left\{ z_{1-\alpha} \left(d_{mx}^{\text{cau}} d_{mx} \right)^2 \leq C_\epsilon \sqrt{n} e_n(\delta_n) \delta_n \right\} - \epsilon \to 1 - \epsilon,$$

as $n \to \infty$, by Assumption B.2. Since the choice of $\epsilon > 0$ was arbitrary, we obtain the desired result. ∎

Proof of Theorem 4.1: By Lemma B.6, we have

$$\sqrt{n}\{\hat{\text{C}} - \text{C}\} = \frac{1}{\sqrt{n}} \sum_{i \in N} (q_i - \mathbf{E}[q_i \mid \mathcal{F}]) + o_P(1).$$

We write

(B.18)
$$P\{\text{ADM} \geq \hat{L}_{1-\alpha}\} = P\left\{ \frac{\sqrt{n}(\text{C} - \hat{\text{C}})}{\tilde{\sigma}} \geq -z_{1-\alpha} + R_{1n} + R_{2n} \right\},$$

where

(B.19)
$$R_{1n} = \frac{z_{1-\alpha}(\tilde{\sigma} - \tilde{\sigma}_{\mathrm{obs}}) - \sqrt{n}(\mathrm{ADM} - \mathrm{C})}{\tilde{\sigma}}, \text{ and}$$

$$R_{2n} = -\frac{z_{1-\alpha}(\hat{\sigma} - \tilde{\sigma}_{\mathrm{obs}})}{\tilde{\sigma}}.$$

By Lemmas B.9 and B.16, and Assumption 3.3, we have $R_{2n} = o_P(1)$. On the other hand, by Lemma B.17, $P\{R_{1n} \le 0\} \to 1$ as $n \to \infty$. The probability on the right-hand side of (B.18) is bounded from below by

(B.20)
$$P\left\{\frac{\sqrt{n}(\mathrm{C} - \hat{\mathrm{C}})}{\tilde{\sigma}} \ge -z_{1-\alpha}\right\} + o(1) = P\left\{\frac{\sqrt{n}(\mathrm{C} - \hat{\mathrm{C}})}{\sigma} \ge -\frac{\tilde{\sigma}}{\sigma}z_{1-\alpha}\right\} + o(1)$$

$$\ge P\left\{\frac{\sqrt{n}(\mathrm{C} - \hat{\mathrm{C}})}{\sigma} \ge -z_{1-\alpha}\right\} + o(1),$$

since $\tilde{\sigma}^2 \ge \sigma^2$ by Lemma B.9. The last probability converges to $1 - \alpha$ by Lemma B.8. ∎

C. Low Level Conditions for Assumptions 3.1 and 3.2

In this section, we focus on the specifications in (3.6) and (3.9), and their estimators constructed using (3.8) and (3.11), as explained in Section 3.3. We provide low level conditions for Assumptions 3.1 and 3.2. For the results below, we assume that Assumptions A.1, 2.2, 3.3 and 3.4 hold.

While we follow the standard arguments (e.g., White (1982), Amemiya (1985), and Newey and McFadden (1994)), some modifications are necessary, as the setting here is different from the standard situation for two reasons. First, the "population" objective functions, $Q_0(\gamma)$ and $Q_1(\beta)$, depend on the sample size n and are random, due to the conditioning on $\mathcal{F}$. Second, the second period outcomes, $Y_{i,1}$, are cross-sectionally dependent along a large network, conditional on $\mathcal{F}$. In dealing with this dependence, we rely on the results on empirical processes with conditional dependency graphs in Lee and Song (2019).

Define

$$F_{j,0}(\gamma) = F_0(X'_j\gamma), \quad f_{j,0}(\gamma) = f_0(X'_j\gamma), \quad f'_{j,0}(\gamma) = f'_0(X'_j\gamma), \text{ and}$$

$$F_{i,1}(\beta) = F_1(X'_i\beta), \quad f_{i,1}(\beta) = f_1(X'_i\beta), \quad f'_{i,1}(\beta) = f'_1(X'_i\beta),$$

where f_0 and f_1 are the first order derivatives of F_0 and F_1, and f_0' and f_1' are the first order derivatives of f_0 and f_1. We define the sample objective functions:

$$\hat{Q}_0(\gamma) = \frac{1}{n}\sum_{j \in N}\left(Y_{j,0}\log F_{j,0}(\gamma) + (1 - Y_{j,0})\log(1 - F_{j,0}(\gamma))\right) \text{ and }$$

$$\hat{Q}_1(\beta) = \frac{1}{n}\sum_{i \in N}\left(Y_{i,1}\log F_{i,1}(\beta) + (1 - Y_{i,1})\log(1 - F_{i,1}(\beta))\right).$$

We take the following as their "population" objective functions:

$$Q_0(\gamma) = \mathbf{E}\left[\hat{Q}_0(\gamma) \mid \mathcal{F}\right] \text{ and } Q_1(\beta) = \mathbf{E}\left[\hat{Q}_1(\beta) \mid \mathcal{F}\right].$$

We let

$$H_0(\gamma) = \frac{\partial^2 Q_0(\gamma)}{\partial \gamma \partial \gamma'} \text{ and } H_1(\beta) = \frac{\partial^2 Q_1(\beta)}{\partial \beta \partial \beta'},$$

and

$$\hat{H}_0(\gamma) = \frac{\partial^2 \hat{Q}_0(\gamma)}{\partial \gamma \partial \gamma'} \text{ and } \hat{H}_1(\beta) = \frac{\partial^2 \hat{Q}_1(\beta)}{\partial \beta \partial \beta'}.$$

Let β^* be a maximizer of $Q_1(\beta)$ over $\beta \in B$. We introduce a set of conditions, and later show that under these conditions, Assumptions 3.1 and 3.2 hold.

Assumption C.3. (i) The parameter spaces Γ and B for γ_0 and β^* are compact, and γ_0 and β^* are in the interior of the parameter spaces Γ and B respectively.

(ii) The densities f_0 and f_1 of F_0 and F_1 are log-concave.

(iii) F_0 and F_1 are three times continuously differentiable with bounded derivatives, and for any compact set $K \subset \mathbf{R}$, there exists a constant $c_K > 0$ that depends only on K such that

$$\inf_{z \in K} F_0(z)(1 - F_0(z)) > c_K \text{ and } \inf_{z \in K} F_1(z)(1 - F_1(z)) > c_K.$$

(iv) There exists a constant $\bar{c} > 0$ such that for all $n \geq 1$,

(C.1) $$\lambda_{\min}(-H_0(\gamma_0)) \geq \bar{c} \text{ and } \lambda_{\min}(-H_1(\beta^*)) \geq \bar{c},$$

with probability one.

Assumption C.3 ensures that the unique maximizers, γ_0 and β^*, are well separated from other points in the parameter spaces in a way that is uniform over the sample size $n \geq 1$. The negative definiteness of the hessian matrix $H_1(\beta^*)$ is known to be a necessary condition for the identifiability of β^*. (See Theorem 3.1 of White (1982).) Assumption C.3(iv) strengthens this negative definiteness to be uniform over $n \geq 1$. This condition prevents the hessian matrix from being nearly singular as $n \to \infty$. The log-concavity condition is made to ensure that the log likelihood function is globally concave in γ and β. (See Pratt (1981).) This condition is

satisfied, for example, when F_0 and F_1 are distribution functions of a normal distribution or a logistic distribution. In Assumption C.3(iii), we require the functions F_0 and F_1 to be three times continuously differentiable, instead of twice continuously differentiable. This ensures that the hessian matrices $H_0(\gamma)$ and $H_1(\beta)$ behave smoothly so that the first order derivatives of the matrices with respect to γ and β around γ_0 and β^* are bounded uniformly over $n \geq 1$.

C.1. Asymptotic Linear Representation of $\hat{\gamma}$ and $\hat{\beta}$

Let us first establish the asymptotic linear representation of estimators $\hat{\gamma}$ and $\hat{\beta}$. From here on, we assume that Assumption C.3 holds. Our first focus is on the consistency of the estimators $\hat{\gamma}$ and $\hat{\beta}$. For this, we establish the following uniform convergence result.

Lemma C.1. *(i)*

$$\sup_{\gamma \in \Gamma} \|\hat{Q}_0(\gamma) - Q_0(\gamma)\| = O_P(n^{-1/2}) \text{ and } \sup_{\beta \in B} \|\hat{Q}_1(\beta) - Q_1(\beta)\| = O_P(n^{-1/2}d_{mx}^2).$$

(ii)

$$\sup_{\gamma \in \Gamma} \|\hat{H}_0(\gamma) - H_0(\gamma)\| = O_P(n^{-1/2}) \text{ and } \sup_{\beta \in B} \|\hat{H}_1(\beta) - H_1(\beta)\| = O_P(n^{-1/2}d_{mx}^2).$$

Proof: We focus on proving (i). The statement (ii) is proved similarly. We also prove only the second statement of (i). The first statement of (i) is simpler to prove because $Y_{j,0}$'s are conditionally independent given $\mathcal{F}$. We write

$$\hat{Q}_1(\beta) - Q_1(\beta) = \frac{1}{n} \sum_{i \in N} (Y_{i,1} - \mathbf{E}[Y_{i,1} \mid \mathcal{F}]) \left(\log(F_{i,1}(\beta)) - \log(1 - F_{i,1}(\beta)) \right).$$

Define $\mathcal{H} = \{h(\cdot, \cdot; \beta) : \beta \in B\}$, where $h(x, y; \beta) = y \left(\log(F_1(x'\beta)) - \log(1 - F_1(x'\beta)) \right)$, $x \in \mathbf{R}^p, y \in \{0, 1\}$. Define the norm for any measurable function $h : \mathbf{R}^p \times \{0, 1\} \to \mathbf{R}$,

$$\overline{\rho}_n(h) = \sqrt{\frac{1}{n} \sum_{i \in N} \mathbf{E}\left[h^2(X_i, Y_{i,1}) \mid \mathcal{F} \right]}.$$

For any $\beta, \tilde{\beta} \in B$, we have $|h(x, y; \beta) - h(x, y; \tilde{\beta})| \leq \Psi(x) \|\tilde{\beta} - \beta\|$, where

$$\Psi(x) = \sup_{\beta \in B} \left\| \left(\frac{f_1(x'\beta)}{F_1(x'\beta)} + \frac{f_1(x'\beta)}{1 - F_1(x'\beta)} \right) x \right\|.$$

By Theorem 2.7.11 of van der Vaart and Wellner (1996), p.164, for any $\epsilon > 0$,

(C.2) $$N_{[]}(2\epsilon\overline{\rho}_n(\Psi), \mathcal{H}, \overline{\rho}_n) \leq N(\epsilon, B, \|\cdot\|) \leq C\epsilon^{-p},$$

for some constant C that does not depend on n, where $N_{[]}(\epsilon, \mathcal{H}, \overline{\rho}_n)$ denotes the ϵ-bracketing number of $\mathcal{H}$ with respect to the norm $\overline{\rho}_n$ and $N(\epsilon, B, \|\cdot\|)$ denotes the ϵ-covering number of B with respect to the Euclidean norm $\|\cdot\|$. (See Sections 2.6 and 2.7 of van der Vaart

and Wellner (1996).) The last inequality follows because B is a compact set in $\mathbf{R}^p$. Note that $(Y_{i,1}, X_i)$'s have G_{cau}^* (defined in (B.10)) as a conditional dependency graph given $\mathcal{F}$, because (Y_1, Y_0) has G_{cau} as a DCG given $\mathcal{F}$. By Assumption 3.3(ii) and Assumption C.3(iii), there exists a constant $C > 0$ that does not depend on n such that $\overline{\rho}_n(\Psi) \le C$. By Lemma 3.4 of Lee and Song (2019), there exist $C, C' > 0$ such that for all $n \ge 1$,

$$
\begin{aligned}
\mathbf{E}\left[\sqrt{n}\sup_{\beta \in B}|\hat{Q}_1(\beta) - Q_1(\beta)| \mid \mathcal{F}\right] &\le C(d_{mx}^* + 1)\int_0^{\overline{\rho}_n(\Psi)}\sqrt{1 + \log N_{[]}\left(\epsilon, \mathcal{H}, \overline{\rho}_n\right)}d\epsilon \\
&= C(d_{mx}^* + 1)\int_0^{C'}\sqrt{1 + \log N_{[]}\left(\epsilon\overline{\rho}_n(\Psi), \mathcal{H}, \overline{\rho}_n\right)}d\epsilon\overline{\rho}_n(\Psi).
\end{aligned}
$$

By (C.2), we obtain the desired result. ∎

The lemma below establishes the consistency of the estimators, $\hat{\gamma}$ and $\hat{\beta}$.

Lemma C.2. *As $n \to \infty$, $\|\hat{\gamma} - \gamma_0\| \to_P 0$, and $\|\hat{\beta} - \beta^*\| \to_P 0$.*

Proof: We prove the second statement only. The proof of the first statement is similar. The proof follows the proof of Theorem 2.1 in Newey and McFadden (1994) with minor modifications. We fix $\epsilon > 0$. Assumption C.3(ii) implies that $Q_1(\beta)$ is globally concave, due to the result from Pratt (1981). By Assumption C.3(iv), there exists $c_\epsilon > 0$ such that

(C.3) $$Q_1(\beta^*) \ge \sup_{\beta \in B: \|\beta - \beta^*\| > \epsilon} Q_1(\beta) + c_\epsilon,$$

with probability one for all $n \ge 1$. Since $\hat{\beta}$ is a maximizer of $\hat{Q}_1(\beta)$ over $\beta \in B$, by Lemma C.1, we have with probability approaching one,

$$Q_1(\hat{\beta}) > \hat{Q}_1(\hat{\beta}) - \frac{c_\epsilon}{2} \ge \hat{Q}_1(\beta^*) - \frac{c_\epsilon}{2} > Q_1(\beta^*) - c_\epsilon \ge \sup_{\beta \in B: \|\beta - \beta^*\| > \epsilon} Q_1(\beta).$$

This implies that $\|\hat{\beta} - \beta^*\| \le \epsilon$ with probability approaching one. ∎

Lemma C.3. *There exist constants $C > 0$ and $\bar{\epsilon} > 0$ that do not depend on n such that, for each $\epsilon \in (0, \bar{\epsilon})$, and for all $n \ge 1$,*

$$
\begin{aligned}
Q_0(\gamma_0) &\ge \sup_{\beta \in B: \epsilon < \|\gamma - \gamma_0\| \le \bar{\epsilon}} Q_0(\gamma) + C\epsilon^2, \text{ and} \\
Q_1(\beta^*) &\ge \sup_{\beta \in B: \epsilon < \|\beta - \beta^*\| \le \bar{\epsilon}} Q_1(\beta) + C\epsilon^2.
\end{aligned}
$$

Proof: We prove the second statement only. Using the same arguments, we can arrive at the first statement as well. First, we write

$$H_1(\beta^*) = \frac{1}{n}\sum_{i \in N}\xi_i(\beta^*)X_iX_i',$$

where

$$\xi_i(\beta^*) = \frac{\mathbf{E}[Y_{i,1} \mid \mathcal{F}]\left(f'_{i,1}(\beta^*)F_{i,1}(\beta^*) - f^2_{i,1}(\beta^*)\right)}{F^2_{i,1}(\beta^*)}$$

$$- \frac{(1 - \mathbf{E}[Y_{i,1} \mid \mathcal{F}])\left(f'_{i,1}(\beta^*)(1 - F_{i,1}(\beta^*)) + f^2_{i,1}(\beta^*)\right)}{\left(1 - F_{i,1}(\beta^*)\right)^2}.$$

Therefore, by Assumption 3.3(ii), we find that

(C.4) $$|\lambda_{\min}(-H_1(\beta)) - \lambda_{\min}(-H_1(\beta^*))| \leq \|H_1(\beta) - H_1(\beta^*)\| \leq \tilde{C}\|\beta - \beta^*\|,$$

where the first inequality follows by the result in Hoffman and Wielandt (1953) and the constant $\tilde{C}$ depends only on the supremum of the derivatives of F_1 and c_K in Assumption C.3(iii), with $K = \{x'\beta : \beta \in B, x \in \mathbf{R}^p, \|x\| \leq C\}$ for $C > 0$ in Assumption 3.3.

Now, take a small enough $\bar{\epsilon} > 0$ such that $\bar{c} - \tilde{C}\bar{\epsilon} > \bar{c}/2$, where $\bar{c}$ is the constant in Assumption C.3(iv). Take any $\epsilon \in (0, \bar{\epsilon})$ and $\beta \in B$ such that $\epsilon < \|\beta - \beta^*\| \leq \bar{\epsilon}$. For any $\tau \in [0,1]$ and $\tilde{\beta} = \tau\beta + (1-\tau)\beta^*$, we have

$$\lambda_{\min}(-H_1(\tilde{\beta})) = \lambda_{\min}(-H_1(\beta^*)) + (\lambda_{\min}(-H_1(\tilde{\beta})) - \lambda_{\min}(-H_1(\beta^*)))$$

$$\geq \lambda_{\min}(-H_1(\beta^*)) - \tilde{C}\|\tilde{\beta} - \beta^*\|$$

$$\geq \lambda_{\min}(-H_1(\beta^*)) - \tilde{C}\|\beta - \beta^*\| \geq \bar{c} - \tilde{C}\bar{\epsilon},$$

by (C.4) and Assumption C.3(iv). Since β^* is a maximizer of $Q_1(\beta)$,

$$Q_1(\beta^*) - Q_1(\beta) = -\frac{1}{2}(\beta - \beta^*)'H_1(\tilde{\beta})(\beta - \beta^*)$$

$$\geq \frac{1}{2}\lambda_{\min}(-H_1(\tilde{\beta}))\|\beta - \beta^*\|^2 \geq \frac{\bar{c} - \tilde{C}\bar{\epsilon}}{2}\|\beta - \beta^*\|^2 > \frac{\bar{c}\epsilon^2}{4},$$

where $\tilde{\beta}$ is a point on the line segment between β and β^*. ∎

Lemma C.4. *For any vanishing sequences $\epsilon_{n,0}$ and $\epsilon_{n,1}$ such that $n^{1/4}\epsilon_{n,0} \to \infty$ and $n^{1/4}\epsilon_{n,1}/d_{mx} \to \infty$ as $n \to \infty$,*

$$\hat{\gamma} - \gamma_0 = O_P(\epsilon_{n,0}), \text{ and } \hat{\beta} - \beta^* = O_P(\epsilon_{n,1}).$$

Proof: We prove the second statement only. Let $\mathcal{N}(\epsilon_{n,1}) = \{\beta \in B : \|\beta - \beta^*\| \leq \epsilon_{n,1}\}$ for the sequence $\epsilon_{n,1}$ chosen to satisfy the condition of the lemma. Let $C > 0$ and $\bar{\epsilon} > 0$ be the constants in Lemma C.3. Let n_0 be such that, for all $n \geq n_0$, $\epsilon_{n,1} < \bar{\epsilon}$. Then, for all $n \geq n_0$,

$$P\left\{\hat{\beta} \in B \setminus \mathcal{N}(\epsilon_{n,1})\right\} = P\left\{\hat{\beta} \in B \setminus \mathcal{N}(\bar{\epsilon})\right\} + P\left\{\hat{\beta} \in \mathcal{N}(\bar{\epsilon}) \setminus \mathcal{N}(\epsilon_{n,1})\right\}.$$

The first probability on the right-hand side vanishes by Lemma C.2. The last probability is bounded by

$$P\left\{Q_1(\hat{\beta}) \leq \sup_{\beta \in \mathcal{N}(\bar{\epsilon}) \setminus \mathcal{N}(\epsilon_{n,1})} Q_1(\beta)\right\} \leq P\left\{Q_1(\hat{\beta}) \leq Q_1(\beta^*) - C\epsilon_{n,1}^2\right\},$$

by Lemma C.3. Since $\hat{Q}_1(\hat{\beta}) \geq \hat{Q}_1(\beta^*)$, by Lemma C.1, the last probability is bounded by

$$P\left\{Q_1(\beta^*) \leq Q_1(\beta^*) - C\epsilon_{n,1}^2 + O_P\left(n^{-1/2}d_{mx}^2\right)\right\} \to 0,$$

as $n \to \infty$, because $n^{1/4}\epsilon_{n,1}/d_{mx} \to \infty$. Hence, $P\left\{\hat{\beta} \in B \setminus \mathcal{N}(\epsilon_{n,1})\right\} \to 0$ as $n \to \infty$. ∎

Lemma C.5. *For any sequences $\epsilon_{n,0}$ and $\epsilon_{n,1}$ that satisfy the conditions of Lemma C.4,*

$$\sqrt{n}(\hat{\gamma} - \gamma_0) = -H_0^{-1}(\gamma_0)\frac{1}{\sqrt{n}}\sum_{j \in N}\pi_{j,0}f_{j,0}(\gamma_0)X_j + O_P(\epsilon_{n,0}), \text{ and}$$

$$\sqrt{n}(\hat{\beta} - \beta^*) = -H_1^{-1}(\beta^*)\frac{1}{\sqrt{n}}\sum_{i \in N}(\pi_{i,1} - \mathbf{E}[\pi_{i,1} \mid \mathcal{F}])f_{i,1}(\beta^*)X_i + O_P\left(\sqrt{d_{av}d_{mx}}\epsilon_{n,1}\right),$$

where

$$\pi_{j,0} = \frac{Y_{j,0} - \mu_{j,0}}{\mu_{j,0}(1 - \mu_{j,0})} \text{ and } \pi_{i,1} = \frac{Y_{i,1} - F_{i,1}(\beta^*)}{F_{i,1}(\beta^*)(1 - F_{i,1}(\beta^*))}.$$

Proof: By the definition of the estimators, we can write

$$(C.5) \qquad \sqrt{n}(\hat{\gamma} - \gamma_0) = \hat{H}_0^{-1}(\bar{\gamma})\left(-\sqrt{n}\frac{\partial \hat{Q}_0(\gamma_0)}{\partial \gamma}\right),$$

where $\bar{\gamma}$ is a point in the line segment between $\hat{\gamma}$ and γ_0. By Lemma C.4,

$$(C.6) \qquad \hat{\gamma} = \gamma_0 + O_P(\epsilon_{n,0}).$$

Since

$$\sqrt{n}\frac{\partial \hat{Q}_0(\gamma_0)}{\partial \gamma} = \frac{1}{\sqrt{n}}\sum_{j \in N}\pi_{j,0}f_{j,0}(\gamma_0)X_j,$$

and $Y_{j,0}$'s are independent across j's conditional on $\mathcal{F}$, we find that

$$(C.7) \qquad \mathbf{E}\left[\left\|\sqrt{n}\frac{\partial \hat{Q}_0(\gamma_0)}{\partial \gamma}\right\|^2 \mid \mathcal{F}\right] = \frac{1}{n}\sum_{j \in N}\mathbf{E}[\pi_{j,0}^2 \mid \mathcal{F}]f_{j,0}^2(\gamma_0)\|X_j\|^2 = O_P(1),$$

by Assumptions C.3(iii) and 3.3(ii). On the other hand,

$$\left\|\hat{H}_0^{-1}(\bar{\gamma}) - H_0^{-1}(\gamma_0)\right\| \leq \left\|\hat{H}_0^{-1}(\bar{\gamma})\right\|\left\|H_0^{-1}(\gamma_0)\right\|\left\|\hat{H}_0(\bar{\gamma}) - H_0(\gamma_0)\right\|$$

$$\leq \left(\left\|H_0^{-1}(\gamma_0)\right\|^2 + o_P(1)\right)\left\|\hat{H}_0(\bar{\gamma}) - H_0(\gamma_0)\right\|,$$

by Lemmas C.1 and C.2. As for the last term,

$$\left\|\hat{H}_0(\overline{\gamma}) - H_0(\gamma_0)\right\| \le \left\|\hat{H}_0(\overline{\gamma}) - H_0(\overline{\gamma})\right\| + \|H_0(\overline{\gamma}) - H_0(\gamma_0)\|$$
$$\le O_P(n^{-1/2}) + \tilde{C}\|\overline{\gamma} - \gamma_0\| = O_P(\epsilon_{n,0}),$$

by Lemmas C.1 and C.4 and (C.4). Therefore, by Assumption C.3(iv), we find from (C.5) that

$$\sqrt{n}(\hat{\gamma} - \gamma_0) = H_0^{-1}(\gamma_0)\left(-\sqrt{n}\frac{\partial \hat{Q}_0(\gamma_0)}{\partial \gamma}\right) + O_P(\epsilon_{n,0}).$$

Let us turn to the second statement. We write

$$\sqrt{n}\frac{\partial \hat{Q}_1(\beta^*)}{\partial \beta} = \sqrt{n}\left(\frac{\partial \hat{Q}_1(\beta^*)}{\partial \beta} - \mathbf{E}\left[\frac{\partial \hat{Q}_1(\beta^*)}{\partial \beta} \mid \mathcal{F}\right]\right)$$
$$= \frac{1}{\sqrt{n}}\sum_{i \in N}\frac{Y_{i,1} - \mathbf{E}[Y_{i,1} \mid \mathcal{F}]}{F_{i,1}(\beta^*)(1 - F_{i,1}(\beta^*))}f_{i,1}(\beta^*)X_i = \frac{1}{\sqrt{n}}\sum_{i \in N}(\pi_{i,1} - \mathbf{E}[\pi_{i,1} \mid \mathcal{F}])f_{i,1}(\beta^*)X_i.$$

The first equality follows because β^* is in the interior of B and the unique maximizer of $Q_1(\beta) = \mathbf{E}[\hat{Q}_1(\beta) \mid \mathcal{F}]$ over $\beta \in B$. We can show that

$$(C.8) \qquad \mathbf{E}\left[\left\|\sqrt{n}\frac{\partial \hat{Q}_1(\beta^*)}{\partial \beta}\right\|^2 \mid \mathcal{F}\right] = \mathbf{E}\left[\left\|\frac{1}{\sqrt{n}}\sum_{i \in N}(\pi_{i,1} - \mathbf{E}[\pi_{i,1} \mid \mathcal{F}])f_{i,1}(\beta^*)X_i\right\|^2 \mid \mathcal{F}\right]$$
$$\le \frac{C}{n}\sum_{i_1, i_2 \in N}\left|\mathrm{Cov}\left(\pi_{i_1,1}, \pi_{i_2,1} \mid \mathcal{F}\right)\right| = O_P(d_{av}d_{mx}),$$

for some constant C that does not depend on n. The last equality follows because $\pi_{i,1}$'s have G^* as a conditional dependency graph given $\mathcal{F}$. The rest of the proof is the same as in the first statement and is omitted. ∎

C.2. Verifying Assumptions 3.1 and 3.2

Let $\hat{\mu}_{j,0}$, $\hat{c}_i$, and $\hat{\psi}_j$ be constructed as in (3.7), (3.10), and (3.14).

Lemma C.6.

$$\max_{j \in N}|\hat{\mu}_{j,0} - \mu_{j,0}| = O_P\left(n^{-1/2}\right),$$
$$\max_{i \in N}|\hat{c}_i - c_i| = O_P\left(n^{-1/2}\sqrt{d_{av}d_{mx}}\right), \text{ and}$$
$$\max_{j \in N}|\hat{\psi}_j - \psi_j| = O_P\left(n^{-1/2}\sqrt{d_{av}d_{mx}}\right).$$

Proof: By Lemma C.5, (C.1), and (C.7), $\|\hat{\gamma} - \gamma_0\| = O_P(n^{-1/2})$. Hence, the first result of Lemma C.6 follows from Assumption C.3(iii). As for the second result, the result follows similarly from $\|\hat{\beta} - \beta^*\| = O_P\left(n^{-1/2}\sqrt{d_{av}d_{mx}}\right)$, which is due to (C.8) and Assumption C.3(iv).

The third statement comes from the first two statements and expanding summands in $\hat{\psi}_j$. Since the arguments are standard, details are omitted. ∎

Lemma C.7. *Assumptions 3.1 and 3.2 hold.*

Proof: Let us first show Assumption 3.1. Recall the notation $e_{i,1} = Y_{i,1} - c_i$, and write

$$\frac{1}{\sqrt{n}} \sum_{i \in N} e_{i,1}(\hat{a}_i - a_i) = A_{n,1} + A_{n,2},$$

where

$$A_{n,1} = \frac{1}{\sqrt{n}} \sum_{i \in N} e_{i,1} \sum_{j \in N_{\text{obs}}(i)} n w_j \varepsilon_{j,0} \left(\frac{1}{\hat{\sigma}_{j,0}^2} - \frac{1}{\sigma_{j,0}^2} \right), \quad \text{and}$$

$$A_{n,2} = \frac{1}{\sqrt{n}} \sum_{i \in N} e_{i,1} \sum_{j \in N_{\text{obs}}(i)} \frac{n w_j (\hat{\varepsilon}_{j,0} - \varepsilon_{j,0})}{\hat{\sigma}_{j,0}^2}.$$

Let us focus on $A_{n,1}$. We write it as

$$\frac{1}{\sqrt{n}} \sum_{i \in N} e_{i,1} \sum_{j \in N_{\text{obs}}(i)} n w_j \varepsilon_{j,0} \frac{\sigma_{j,0}^2 - \hat{\sigma}_{j,0}^2}{\sigma_{j,0}^4 (1 + o_P(1))}$$

$$= \frac{1}{\sqrt{n}} \sum_{i \in N} e_{i,1} \sum_{j \in N_{\text{obs}}(i)} n w_j \varepsilon_{j,0} \frac{(\hat{\mu}_{j,0} - \mu_{j,0})(2\mu_{j,0} - 1)}{\sigma_{j,0}^4 (1 + o_P(1))} + R_n,$$

where

$$R_n = \frac{1}{\sqrt{n}} \sum_{i \in N} e_{i,1} \sum_{j \in N_{\text{obs}}(i)} n w_j \varepsilon_{j,0} \frac{(\hat{\mu}_{j,0} - \mu_{j,0})^2}{\sigma_{j,0}^4 (1 + o_P(1))}.$$

By Lemma C.6 and Assumption 2.2,

$$R_n \leq O_P(n^{-1}) \times \frac{1}{\sqrt{n}} \sum_{i \in N} \sum_{j \in N_{\text{obs}}(i)} n w_j |\varepsilon_{j,0}| = O_P(n^{-1/2} d_{av}) = o_P(1).$$

Expanding $F_0(X_j' \hat{\gamma}) - F_0(X_j' \gamma_0)$, we find that

$$\frac{1}{\sqrt{n}} \sum_{i \in N} e_{i,1} \sum_{j \in N_{\text{obs}}(i)} n w_j \varepsilon_{j,0} \frac{(\hat{\mu}_{j,0} - \mu_{j,0})(2\mu_{j,0} - 1)}{\sigma_{j,0}^4 (1 + o_P(1))}$$

$$= \left(\frac{1}{\sqrt{n}} \sum_{i \in N} e_{i,1} \sum_{j \in N_{\text{obs}}(i)} n w_j \varepsilon_{j,0} f_{j,0}(\gamma_0) \frac{2\mu_{j,0} - 1}{\sigma_{j,0}^4} X_j' \right)(\hat{\gamma} - \gamma_0) + o_P(1).$$

By Lemma C.5 and Assumption 3.4, the leading term on the right-hand side is equal to

$$\left(\frac{1}{n} \sum_{i \in N} e_{i,1} \sum_{j \in N_{\text{obs}}(i)} n w_j \varepsilon_{j,0} f_{j,0}(\gamma_0) \frac{2\mu_{j,0} - 1}{\sigma_{j,0}^4} X_j' \right) \left(-H_0^{-1}(\gamma_0) \frac{1}{\sqrt{n}} \sum_{j \in N} \pi_{j,0} f_{j,0}(\gamma_0) X_j \right) + o_P(1).$$

Note that

$$\frac{1}{n}\sum_{i\in N}e_{i,1}\sum_{j\in N_{\text{obs}}(i)}nw_j\varepsilon_{j,0}f_{j,0}(\gamma_0)\frac{2\mu_{j,0}-1}{\sigma_{j,0}^4}X_j'$$

$$=\frac{1}{n}\sum_{i\in N}\mathbf{E}\left[e_{i,1}\sum_{j\in N_{\text{obs}}(i)}nw_j\varepsilon_{j,0}f_{j,0}(\gamma_0)\frac{2\mu_{j,0}-1}{\sigma_{j,0}^4}X_j'\mid\mathcal{F}\right]+o_P(1)$$

$$=\frac{1}{n}\sum_{i\in N}\mathbf{E}\left[e_{i,1}\sum_{j\in N_{\text{obs}}(i)}\frac{nw_jf_{j,0}(\gamma_0)X_j'}{\sigma_{j,0}^2}\pi_{j,0}(2\mu_{j,0}-1)\mid\mathcal{F}\right]+o_P(1),$$

using similar arguments as before. We find that

$$A_{n,1}=\left(\frac{1}{n}\sum_{i\in N}\mathbf{E}\left[e_{i,1}\sum_{j\in N_{\text{obs}}(i)}\frac{nw_jf_{j,0}(\gamma_0)X_j'}{\sigma_{j,0}^2}\pi_{j,0}(2\mu_{j,0}-1)\mid\mathcal{F}\right]\right)$$

$$\times\left(-H_0^{-1}(\gamma_0)\frac{1}{\sqrt{n}}\sum_{j\in N}\pi_{j,0}f_{j,0}(\gamma_0)X_j\right)+o_P(1).$$

Let us deal with $A_{n,2}$. We write it as

$$\frac{1}{\sqrt{n}}\sum_{i\in N}e_{i,1}\sum_{j\in N_{\text{obs}}(i)}\frac{nw_j(\hat{\varepsilon}_{j,0}-\varepsilon_{j,0})}{\sigma_{j,0}^2}+o_P(1)=-\frac{1}{\sqrt{n}}\sum_{i\in N}e_{i,1}\sum_{j\in N_{\text{obs}}(i)}\frac{nw_j(\hat{\mu}_{j,0}-\mu_{j,0})}{\sigma_{j,0}^2}+o_P(1).$$

Again, expanding $F_0(X_j'\hat{\gamma})-F_0(X_j'\gamma_0)$, we can rewrite the last leading term as

$$-\left(\frac{1}{\sqrt{n}}\sum_{i\in N}e_{i,1}\sum_{j\in N_{\text{obs}}(i)}\frac{nw_jf_{j,0}(\gamma_0)X_j'}{\sigma_{j,0}^2}\right)(\hat{\gamma}-\gamma_0)+o_P(1)$$

$$=-\left(\frac{1}{n}\sum_{i\in N}e_{i,1}\sum_{j\in N_{\text{obs}}(i)}\frac{nw_jf_{j,0}(\gamma_0)X_j'}{\sigma_{j,0}^2}\right)\times\left(-H_0^{-1}(\gamma_0)\frac{1}{\sqrt{n}}\sum_{j\in N}\pi_{j,0}f_{j,0}(\gamma_0)X_j\right)+o_P(1)$$

$$=-\left(\frac{1}{n}\sum_{i\in N}\mathbf{E}\left[e_{i,1}\sum_{j\in N_{\text{obs}}(i)}\frac{nw_jf_{j,0}(\gamma_0)X_j'}{\sigma_{j,0}^2}\mid\mathcal{F}\right]\right)$$

$$\times\left(-H_0^{-1}(\gamma_0)\frac{1}{\sqrt{n}}\sum_{j\in N}\pi_{j,0}f_{j,0}(\gamma_0)X_j\right)+o_P(1).$$

Recall the definition of ψ_j in (3.12). Since $H_0(\gamma_0) = -\frac{1}{n}\sum_{i\in N}(f_{i,0}^2(\gamma_0)/\sigma_{i,0}^2)X_iX_i'$, we find that

$$\frac{1}{\sqrt{n}}\sum_{i\in N}e_{i,1}(\hat{a}_i - a_i) = \Gamma_n\frac{1}{\sqrt{n}}\sum_{j\in N}\pi_{j,0}f_{j,0}(\gamma_0)X_j + o_P(1)$$

$$= \frac{1}{\sqrt{n}}\sum_{j\in N}\left(\psi_j - \mathbf{E}[\psi_j \mid \mathcal{F}]\right) + o_P(1).$$

The condition (3.15) for ψ_j are satisfied under Assumptions 2.2, 3.3 and C.3.

As for the second statement in Assumption 3.1, we write

(C.9) $$\frac{1}{\sqrt{n}}\sum_{i\in N}(\hat{c}_i - c_i)a_i = \frac{1}{n}\sum_{i\in N}f_{i,1}(\beta^*)a_iX_i'\sqrt{n}(\hat{\beta} - \beta^*) + o_P(1).$$

By Lemma C.5 and Assumption 3.4, the leading term on the right-hand side is equal to

$$\left(\frac{1}{n}\sum_{i\in N}f_{i,1}(\beta^*)a_iX_i'\right)\left(-H_1^{-1}(\beta^*)\frac{1}{\sqrt{n}}\sum_{i\in N}\pi_{i,1}f_{i,1}(\beta^*)X_i\right) + o_P(1)$$

$$= \left(\frac{1}{n}\sum_{i\in N}f_{i,1}(\beta^*)\mathbf{E}[a_i \mid \mathcal{F}]X_i'\right)\left(-H_1^{-1}(\beta^*)\frac{1}{\sqrt{n}}\sum_{i\in N}\left(\pi_{i,1} - \mathbf{E}[\pi_{i,1}\mid\mathcal{F}]\right)f_{i,1}(\beta^*)X_i\right) + o_P(1).$$

To see the equality, note that

$$\mathrm{Var}\left(\frac{1}{n}\sum_{i\in N}f_{i,1}(\beta^*)a_iX_i' \mid \mathcal{F}\right) = O_P\left(\frac{d_{av}d_{mx}}{n}\right),$$

following the same arguments as in the proof of Lemma B.3(i), and

$$\frac{1}{\sqrt{n}}\sum_{i\in N}\pi_{i,1}f_{i,1}(\beta^*)X_i = O_P\left(\sqrt{d_{av}d_{mx}}\right),$$

because $\{\pi_{i,1}f_{i,1}(\beta^*)X_i\}_{i\in N}$ has G^* as a conditional dependency graph and $\mathbf{E}[\pi_{i,1}\mid\mathcal{F}] = 0$. Thus the equality below (C.9) follows by Assumption C.3(iv) and Assumption 3.4. However, $\mathbf{E}[a_i \mid \mathcal{F}] = 0$, and hence we obtain the second condition in Assumption 3.1.

Assumption 3.2 follows by Lemma C.6 and Assumption 3.4. ∎

D. Further Details on Assumption 4.1

In this section, we provide further details that support the discussion after Assumption 4.1. First, let us recall the linear threshold diffusion model considered in the discussion:

$$\rho_{i,t}\left((A_{j,t-1})_{j\in N_{ctt}(i)}, U_{i,t}; G_{ctt}\right) = 1\left\{\frac{\beta}{|N_{ctt}(i)|}\sum_{j\in N_{ctt}(i)}A_{j,t-1} \geq U_{i,t}\right\},$$

where $U_{i,t}$'s are i.i.d. across $t \geq 1$ and i's, following the uniform distribution on $[0,1]$, and conditionally independent of $A_{i,0}$'s given $\mathcal{F}$, and β does not depend on n. Our primary goal here is to verify the claim that when $\beta > 0$, Assumption 4.1 follows.

Let us assume that $\beta > 0$. This implies that for all $t = 1, ..., t_1$, $A_{i,t}$ is a non-decreasing function of $A_{j,0}$, $j \in N_{\text{cau}}(i)$, which are conditionally independent of $U_{k,s}$'s, $k \in N$ and $s = 1, ..., t_1$, given $\mathcal{F}$. Hence, for all $i \in N$ and $j \in N_{\text{cau}}(i)$, and all $t = 1, ..., t_1$,

$$\text{Cov}\left(A_{i,t}, A_{j,0} \mid \mathcal{F}\right) \geq 0.$$

Thus for $j \in N_{\text{ctt}}(i)$, $\text{Cov}\left(Y_{j,0}, Y_{i,1} \mid \mathcal{F}\right) = \sum_{t=0}^{t_1} \text{Cov}\left(A_{j,0}, A_{i,t} \mid \mathcal{F}\right) \geq \text{Cov}\left(A_{j,0}, A_{i,1} \mid \mathcal{F}\right)$. Note that

$$\mathbf{E}\left[A_{i,1} \mid \mathcal{F}\right] = \mathbf{E}\left[1\left\{\frac{\beta}{|N_{\text{ctt}}(i)|} \sum_{k \in N_{\text{ctt}}(i)} A_{k,0} \geq U_{i,1}\right\}(1 - A_{i,0}) \mid \mathcal{F}\right]$$

$$= \mathbf{E}\left[\left(\frac{\beta}{|N_{\text{ctt}}(i)|} \sum_{k \in N_{\text{ctt}}(i)} A_{k,0}\right)(1 - A_{i,0}) \mid \mathcal{F}\right]$$

$$= \frac{\beta}{|N_{\text{ctt}}(i)|} \sum_{k \in N_{\text{ctt}}(i)} \mu_{k,0}\left(1 - \mu_{i,0}\right) \equiv \xi(i), \text{ say.}$$

The first equality follows by the definition of $A_{i,1}$. The second equality follows because $U_{i,1}$ is independent of $A_{k,0}$'s and $A_{i,0}$, following the uniform distribution on $[0,1]$. The last equality follows because $A_{j,0}$'s are conditionally independent given $\mathcal{F}$ and $\mu_{j,0} = \mathbf{E}[A_{j,0} \mid \mathcal{F}]$.

Write

$$\text{Cov}\left(A_{j,0}, A_{i,1} \mid \mathcal{F}\right) = \mathbf{E}\left[A_{j,0} 1\left\{\frac{\beta}{|N_{\text{ctt}}(i)|} \sum_{k \in N_{\text{ctt}}(i)} A_{k,0} \geq U_{i,1}\right\}(1 - A_{i,0}) \mid \mathcal{F}\right] - \mu_{j,0}\xi(i)$$

$$= \mathbf{E}\left[A_{j,0} 1\left\{\frac{\beta}{|N_{\text{ctt}}(i)|} \sum_{k \in N_{\text{ctt}}(i)} A_{k,0} \geq U_{i,1}\right\} \mid \mathcal{F}\right](1 - \mu_{i,0}) - \mu_{j,0}\xi(i).$$

For the leading term on the right-hand side, since $i \neq j$, we can write it as

$$\mathbf{E}\left[A_{j,0}\left(\frac{\beta}{|N_{\text{ctt}}(i)|} \sum_{k \in N_{\text{ctt}}(i)} A_{k,0}\right) \mid \mathcal{F}\right](1 - \mu_{i,0}) = \frac{\beta \mu_{j,0}}{|N_{\text{ctt}}(i)|}\left(1 + \sum_{k \in N_{\text{ctt}}(i)\backslash\{j\}} \mu_{k,0}\right)(1 - \mu_{i,0}).$$

The above equality follows because $j \in N_{\text{cau}}(i)$. The last term is equal to

$$\frac{\beta \mu_{j,0}}{|N_{\text{ctt}}(i)|}\left(1 + \sum_{k \in N_{\text{ctt}}(i)\backslash\{j\}} \mu_{k,0}\right)(1 - \mu_{i,0}) = \mu_{j,0}\xi(i) + \frac{\beta \mu_{j,0}(1 - \mu_{j,0})(1 - \mu_{i,0})}{|N_{\text{ctt}}(i)|}.$$

Hence

$$\text{Cov}\big(A_{j,0}, A_{i,1} \mid \mathcal{F}\big) = \frac{\beta \mu_{j,0}(1-\mu_{j,0})(1-\mu_{i,0})}{|N_{\text{ctt}}(i)|} \geq \frac{\beta c^3}{d_{mx}},$$

by Assumption 2.2(ii). Therefore, we have

$$\text{Cov}\big(Y_{j,0}, Y_{i,1} \mid \mathcal{F}\big) \geq \frac{\beta c^3}{d_{mx}}.$$

By Assumption 3.4, we find that Assumption 4.1 is satisfied.

References

AMEMIYA, T. (1985): *Advanced Econometrics*. Harvard University Press, Cambridge, Massachusetts.

HE, X., AND K. SONG (2022): "Measuing Diffusion Over a Large Network," *arXiv:1812.04195v4 [stat.ME]*.

HOFFMAN, A. J., AND H. W. WIELANDT (1953): "The Variation of the Spectrum of a Normal Matrix," *Duke Mathematical Journal*, 20, 37–39.

KOJEVNIKOV, D. (2021): "The Bootstrap for Network Dependent Processes," *arXiv preprint arXiv:2101.12312*.

KOJEVNIKOV, D., V. MARMER, AND K. SONG (2021): "Limit Theorems for Network Dependent Random Variables," *Journal of Econometrics*, 222(2), 882–908.

KOLLER, D., AND N. FRIEDMAN (2009): *Probabilistic Graphical Models: Principles and Techniques*. The MIT Press, Cambridge, Massachusetts.

LEE, J. H., AND K. SONG (2019): "Stable Limit Theorems for Empirical Processes Under Conditional Neighborhood Dependence," *Bernoulli*, 25, 1189–1224.

NEWEY, W. K., AND D. MCFADDEN (1994): "Large Sample Estimation and Hypothesis Testing," in *Handbook of Econometrics*, vol. 4, pp. 193–281. North-Holland, Ambsterdam.

PRATT, J. W. (1981): "Concavity of the Log Likelihood," *Journal of the American Statistical Association*, 76, 103–106.

VAN DER VAART, A. W., AND J. A. WELLNER (1996): *Weak Convergence and Empirical Processes*. Springer, New York, USA.

WHITE, H. (1982): "Maximum Likelihood Estimation of Misspecified Models," *Econometrica*, 50, 1 – 25.