Identification and Estimation of Treatment Effects from Social Network Data

Licheng Liu*

This draft: September 12, 2023
First draft: August 30, 2023

Abstract

I introduce a generalized propensity score (GPS) based approach to the identification and estimation of treatment effects from observational social network data, where formation of social tie between pair of units depends on individual level characteristics. Ignoring the tie formation process, its interaction with the treatment assignment mechanism and interference induced by the social network can lead to biased estimation of treatment effects. I propose a unified framework that addresses these challenges by jointly modeling treatment assignment and network formation, incorporating their complex interactions in observational social network data. Generalized propensity score can be semi-parametrically estimated given probabilistic models for these two processes and functional form of exposure mapping (Aronow and Samii, 2017) for effective treatment. Average potential outcomes and treatment effects are estimated with inverse probability weighting estimators. I illustrate the proposed method in several Monte Carlo studies and an empirical analysis that investigates the effect of adopting a new political communication technology on political participation in Uganda.

Keywords: Causal Inference, Social Network Data, Interference

*Ph.D. Candidate, Department of Political Science, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139. Email: liulch@mit.edu.

1 Introduction

Causal inference with network data has emerged as a vibrant topic in both empirical social science studies and methodological research. For cross-country analyses, connections between countries can be assessed by geographical proximity or bilateral metrics such as trade volume and joint membership in international bodies. In behavioral studies, individuals might establish social ties, like friendships, based on shared characteristics like age or educational background. When the units of study are interconnected, a treatment assigned to one unit (referred to as the "ego") can influence the potential outcomes of other units (termed "peers"). The impact of the treatment on the ego's outcome is called the *direct effect*, while its influence on the peer's outcome is the *indirect* or *spillover effect*. Various mechanisms could underlie this indirect effect. For instance, a peer might adjust her actions based on information gleaned from the ego's treatment adoption (Sinclair et al., 2012). Alternatively, treatments might implicitly alter a peer's outcomes due to outcome interdependence, as discussed by (Simmons and Elkins, 2004). In observational studies, the adoption of treatment by one unit might also affect the likelihood that other units adopt the treatment. Identifying and quantifying these indirect effects is crucial. It not only helps in understanding the externalities of treatment assignment but also is essential for assessing societal welfare changes in cost-benefit analyses.

However, identification and estimation of treatment effect from network data can be challenging. While connections measured by geographical locations are fixed and exogenous, most social networks are random, and the formation of social ties between a pair of units depends on some unit-level features. If those features also affect the behavioral outcome, they confound the relationship between network formation and outcome (Goldsmith-Pinkham and Imbens, 2013), and the problem becomes more complex if some of those features also affect the adoption of treatment. In such scenarios, failing to account for the formation of social network may cause bias in estimation of indirect effect.

Consider an instance where we want to evaluate the impact of a new political communication technology (PCT) on political participation. If men are typically more engaged in adopting PCT and participating in political activities than women, and if individuals of the same gender are more likely to form social ties, then men tend to connect with

more adopters than women. Even after addressing the confounding influence of gender, neglecting the indirect effects could bias the estimated effect of PCT adoption on political participation. This stems from the fact that we are essentially estimating the combination of both direct and indirect effects. A positive indirect effect can lead to an overestimation of the direct effect, whereas a negative indirect effect can result in its underestimation.

The randomness of social networks complicates the identification and estimation of treatment effects. Even in experimental studies where treatments are randomized, their presence can pose challenges in identifying these effects. With observational data, the situation becomes even more complicated due to the potential interactions between treatment assignment and network formation. For example, treatment assignment mechanism and network formation can be simultaneously determined by exogenous covariates (Han et al., 2021), treatment assignment may affect the tie formation process (Comola and Prina, 2021), and social network may induce diffusion of treatment adoption (Leung and Loupos, 2022). In such settings, it is important to jointly model treatment assignment mechanism and network formation process. This paper aims at developing a unified framework to incorporate such complex interactions in social network data.

I propose a generalized propensity score (GPS) approach for identification and estimation of treatment effect. In network data, treatment effects are defined as differences in the average potential outcomes under different levels of "effective treatment" (Manski, 2013), which is a function, termed "exposure mapping" in Aronow and Samii (2017), of the treatment assignment vector and the social network. The chosen form of this exposure mapping reflects how researchers perceive interference between units. In this paper, I assume it is flexible but correctly specified [1]. Average potential outcomes, also known as average dose-response function (ADRF), can be identified with generalized propensity score for the adoption of (each level of) treatment (Imbens, 2000). If we properly specify probabilistic models for both the treatment assignment mechanism and network formation process, we can estimate the generalized propensity score for each unit, given that the exposure mapping depends on the treatment assignment vector and the social network. Researchers can semi-parametrically estimate generalized propensity scores via analytical

[1]For discussions on potential mis-specifications, readers can turn to studies like Sävje (2023).

expression if the exposure mapping has a simple functional form. For multi-valued or continuous treatments, or when the exposure mapping is complex, simulations based on the probabilistic models can be employed (Aronow and Samii, 2017).

Joint modeling of both the treatment assignment mechanism and the network formation process is central to the proposed method. Yet, estimating a combined probabilistic model for these elements can be challenging, particularly when we introduce additional constraints like row normalization on network entries. To circumvent this complexity, I develop assumptions to factorize these two processes so that researchers can model them separately or sequentially to simplify estimation. Although these assumptions seems be strong, I show that this approach is flexible to incorporate multiple settings.

Once the generalized propensity scores are estimated, average potential outcomes as well as treatment effects can be estimated with the inverse-probability weighting estimators like the Horvitz–Thompson estimator, Hajek estimator and doubly robust estimator. Uncertainty estimates can be obtained by implementing the network HAC estimator (Kojevnikov et al., 2021; Leung and Loupos, 2022) that accounts for correlations in the social network. Besides, I also propose a regression-based estimation procedure (Hirano and Imbens, 2004) to incorporate weighted networks and continuous treatment assignment.

This paper contributes to the burgeoning literature on causal inference with network data. It extends the framework of Aronow and Samii (2017) to observational settings and relaxes the assumption of exogenous network. The idea of modeling network formation process echoes the insights of Toulis et al. (2021). This studies the problem of network dynamics as treatment, while my approach distinctively evaluates the joint effect of a treatment assignment variable and a social network, which is static but random. Highlighting other relevant studies, Forastiere et al. (2021) and Sanchez-Becerra (2022) also propose propensity score based approach to casual inference with observational network data. While working under tighter assumptions, they contend that the exposure mapping is independent of potential outcomes when controlled for unit-level covariates and other network nuances. They further suggest parametric models for a direct estimation of generalized propensity scores. In contrast, the proposed method is flexible to accommodate multiple types of treatment variables, networks and their interactions. The trade-off is

computational demand, which may limit its feasibility on large scale network datasets. Additionally, Jackson et al. (2022) introduces a peer-influenced propensity score and Leung and Loupos (2022) proposed a graph neural network (GNN) based propensity score. Both methods can incorporate the diffusion of treatment adoption induced by the network, operating under the assumption that networks are either static or exogenous.

The rest of the paper is organized as follows. Section 2 outlines the contextual background of a motivating example that investigates the effect of adoption of a new political communication technology on political participation in Uganda. In Section 3, I set up the potential outcome framework, define the causal estimands, and develop key identification assumptions. Section 4 introduces the generalized propensity score based approach and details of joint modeling of treatment assignment and network formation process in different scenarios. Section 5 illustrates details of estimation and inference based on the generalized propensity score. Section 6 reports results of several Monte Carlo studies designed to investigate the finite sample properties of the proposed method. Section 7 provides a comparison of the estimation results from our motivating example, contrasting the proposed method with some existing estimation strategies. The last section concludes.

2 A Motivating Example

Political scientists are interested in whether the adoption of information and communication technology (ICT) promotes political participation in developing countries. However, existing literature provides mixed evidence on the effect of ICT adoption on political participation. For example, Harrison and List (2004) conducted a small-scale framed field experiment in Uganda. They found that, compared to existing political communication channels, marginalized populations utilized short message service (SMS)-based communication at relatively higher rates. They concluded that ICT adoption has a "flattening" effect on political participation. This conclusion is plausible as marginalized populations often have limited opportunities to communicate with politicians. As a result, they might be less inclined to bear the high costs of political participation. ICT innovations can reduce communication costs, potentially increasing political participation among marginalized groups. However, in a subsequent nation-wide field experiment in Uganda, Grossman

et al. (2020) found that introducing a new political communication system did not significantly affect political engagement. Most existing research assumes the independence of units under study. However, the presence of social networks might contaminate estimates of treatment effects. In this paper, I use replication data from Ferrali et al. (2020) and Eubank et al. (2021) to investigate the effect of adopting a new political communication technology, the U-Bridge program, on political participation in 16 Ugandan villages. In these villages, residents may share various types of social ties.

I begin by outlining the contextual background. The U-Bridge program was implemented in a district located in northwestern Uganda. It allows residents to contact district officials via text message, which is both free and anonymous. The program was implemented in a field experiment that encouraged usage in 131 randomly selected villages. Residents from treatment villages were invited to periodic community meetings about ways to communicate with local officials. The first round of meetings was held in late 2014. To investigate the pattern of adoption of U-Bridge, Ferrali et al. (2020) conducted in-person surveys in 16 treatment villages in 2016, two years after launch of the program. The surveys gathered multiple individual-level variables such as age, gender, attendance at meetings, U-Bridge adoption status, and various social ties between resident pairs.

In this motivating example, I regard the adoption of U-Bridge as a binary treatment indicator. The outcome variable is a continuous summary index of political participation that aggregates political actions in the last 12 months. For structure of social networks, I follow the methods of Ferrali et al. (2020) and Sanchez-Becerra (2022) to measure connection between resident pairs. I assume that two residents within the same village are connected if they share any of the four types of social ties: family relationships, friendships, lender relationships, or problem-solving connections. Consequently, connections between residents are undirected. I exclude social ties spanning different villages, which reduces the overall network connecting residents to 16 distinct components. Other individual level covariates include age, gender, levels of income, binary indicators representing whether a resident has attained secondary education, whether they occupy a formal leadership role in the village, and whether they own a phone, and a behavioral proxy measure of care for the community. A detailed variable description is available in the appendix of Ferrali et al.

(2020). The original survey data comprises information on 3,184 respondents across the 16 treatment villages. After excluding entries with missing values, this study's dataset covers 3,018 respondents, with 135 of them having adopted U-Bridge.

Identifying and estimating the effect of U-Bridge on political participation poses several challenges. First, individual-level covariates might influence both U-Bridge adoption and political participation. Second, the existence of social networks could lead to spillover effects. For instance, the adoption of U-Bridge by one resident might impact others' political activities due to the transmission of information along social ties. Moreover, the adoption of U-Bridge by one resident might also affect the likelihood of adoption by others (see Table 4 in Section 7), known as peer influence in treatment adoption (Jackson et al., 2022). Lastly, these social networks are not exogenously determined, and individual-level covariates could influence social tie formation. For example, Figure 1 suggests that residents of the same gender are more likely to form social ties and that village leaders generally have more connections. In Section 7, I employ the proposed method to estimate the treatment effect of U-Bridge adoption and contrast the findings with results from some other estimation strategies.

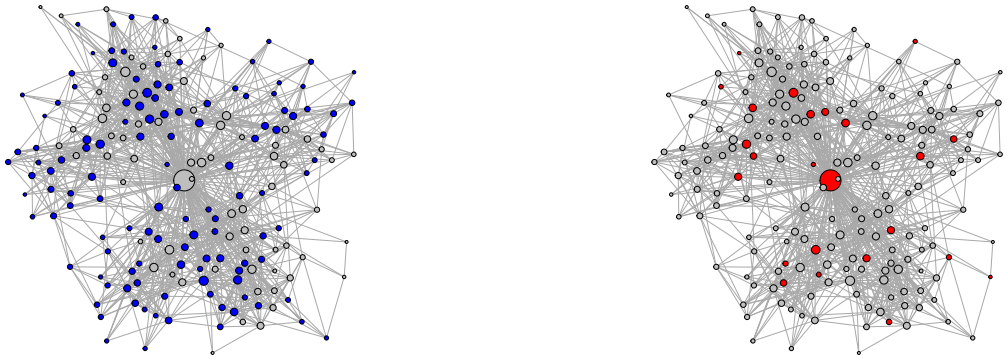

Figure 1: Network Visualization for Village Indexed 9

Note: Size of each vertex is proportional to the square root of degree. In the left panel, vertex in blue represents respondent whose gender is female. In the right panel, vertex in blue represents respondent who occupies formal leadership role within village.

3 Interference in the Presence of Social Network

3.1 Notation

Suppose we have a cross-sectional dataset that includes N units. We focus on the case of a single large network, while the results can be generated to datasets consisting of multiple networks or clusters like the settings in Hudgens and Halloran (2008) and Sanchez-Becerra (2022). Let us denote $\boldsymbol{W}$ an observed network, a $(N \times N)$ random adjacency matrix with entry w_{ij} specifying the connection between unit i and j. In the terminology of network analysis, we denote the associated random graph G for $\boldsymbol{W}$ as a pair: $(\mathcal{V}, \mathcal{E})$, where $\mathcal{V} = \{1, \ldots, N\}$ is the set of units [2] and $\mathcal{E}$ is the set of edges, i.e., $(i, j) \in \mathcal{E}$ if $w_{ij} > 0$. If G is undirected, $(i, j) = (j, i)$, otherwise (i, j) and (j, i) are different edges $\forall i, j$. $\boldsymbol{W}$ can be a weighted matrix, where the strength of connections between unit i, j and i, k are different if $w_{ij} \neq w_{ik}$. For notational convenience, in this paper we assume that W is unweighted. That is, $w_{ij} \in \{0, 1\} \quad \forall i, j$. We denote $\mathcal{W}$ the sample space of $\boldsymbol{W}$.

We denote D_i the treatment assigned to unit i, which can be binary, multi-valued or even continuous. Y_i is the observed outcome of interest. In vector form, we denote $\boldsymbol{D} = (D_1, \ldots, D_N)$ a $(N \times 1)$ the vector of treatment assignments and $\boldsymbol{Y} = (Y_1, \ldots, Y_N)$ the vector of observed outcomes for all units. We denote $\mathcal{D}$ the sample space of $\boldsymbol{D}$. When D_i is binary, we have $\mathcal{D} = \{0, 1\}^N$.

Finally, we denote X_i a $(p \times 1)$ vector of observed (pre-treatment) covariates for unit i. For network data, X_i can be decomposed into two parts: confounders that affect the adoption of the treatment, and covariates contributing to observed homophily that affects network formation. We denote the first part X_i^D and the second part X_i^W. It is worth noting that these two subsets, X_i^D and X_i^W, are not always mutually exclusive, meaning there may be instances of overlap where certain covariates simultaneously impact both the adoption of treatment and network formation. In addition, X_i^W could also incorporate dyadic covariates, such as geographical distance or the bilateral trade volume between country pairs. In matrix form, let $\boldsymbol{X} = (X_1', \ldots, X_N')$ denote the matrix that aggregates all

[2] They are also called nodes or vertices in network analysis. We use the term "units" through out this paper for consistency.

the covariate vectors for each unit. Similarly, $\boldsymbol{X}^D$ and $\boldsymbol{X}^W$ are the corresponding matrices for the aforementioned subsets of covariates. To maintain simplicity and consistency in our terminology, we will use the term "covariate" to refer to X_i, X_i^D, and X_i^W.

3.2 Potential outcomes in social network data

In most social network data, the observed network $\boldsymbol{W}$ is random and endogenous. I extend the potential outcome framework proposed in Rubin (1974) to account for the potential interference induced by $\boldsymbol{W}$. In this scenario, the observable outcome for unit i is determined by the entire treatment assignment vector $\boldsymbol{D}$ and the network $\boldsymbol{W}$, which can be written as $Y_i(\boldsymbol{D}, \boldsymbol{W})$. Under the assumption of no multiple versions of treatment (consistency) (Rubin, 1986), the observed outcome $Y_i = Y_i(\boldsymbol{d}, \boldsymbol{w})$ if $\boldsymbol{D} = \boldsymbol{d}$ and $\boldsymbol{W} = \boldsymbol{w}$ [3]. Note that notation of potential outcome in the form of $Y_i = Y_i(\boldsymbol{d}, \boldsymbol{w})$ is quite general, in that we regard $\boldsymbol{D}$ and $\boldsymbol{W}$ as "joint" treatment assignments, and incorporates the following examples as special cases.

Example 1: No interference.

The consistency assumption implies that the mechanism used for treatment assignment and network formation does not matter for the potential outcomes. It is the first part of the stable unit treatment value assumption (SUTVA) (Rubin, 1986) often made in the potential outcomes approach to causal inference. In addition, SUTVA assumes away the interference among units, i.e., the treatment assigned to unit i does not affect the potential outcome of unit j: $Y_i(\boldsymbol{d}, \boldsymbol{w}) = Y_i(\boldsymbol{d}', \boldsymbol{w}')$ if $d_i = d_i'$. In this case, we can simplify the potential outcome as $Y_i(\boldsymbol{d}, \boldsymbol{w}) = Y_i(d_i)$.

Example 2: Exogenous networks.

The assumption of no interference is quite strong and is often violated when we make causal inference with network data. To relax this assumption, scholars have proposed design-based as well as model-based approaches to causal inference under interference with experimental and observational data (e.g., Aronow and Samii (2017); Forastiere et al. (2021)). In their work, they assume that the potential outcomes for each unit is determined

[3] We use bold uppercase letters to represent random vectors and matrices, bold lowercase letters their corresponding realizations.

by the whole treatment assignment vector, i.e., $Y_i(\boldsymbol{d}, \boldsymbol{w}) = Y_i(\boldsymbol{d}', \boldsymbol{w}')$ if $\boldsymbol{d} = \boldsymbol{d}'$. Therefore, the network $\boldsymbol{W}$ is often regarded as "fixed" and used to define the structure of interference among units. In this case, we can simplify the potential outcome as $Y_i(\boldsymbol{d}, \boldsymbol{w}) = Y_i(\boldsymbol{d})$.

Example 3: Networks as treatment assignment.

In some cases, an endogenous network itself can be regarded as treatment assignment. For example, changes in the network, like degree for each node, may affect potential outcomes (Toulis et al., 2021). In fact, network summary statistics are sometimes the key independent variables in political science studies. For example, Kinne (2012) studies the effect of a country's position in global trade network on conflicts. In this case, we can simplify the potential outcome as $Y_i(\boldsymbol{d}, \boldsymbol{w}) = Y_i(\boldsymbol{W})$. Since $\boldsymbol{W}$ is common for all units, therefore treatment assignments are interdependent [4] .

Given the notations above, a system of models for network formation, treatment assignment and behavioral outcomes is represented as follows:

$$\begin{aligned} W_{ij} &= g_w(i, j, \boldsymbol{X}, \epsilon_{ij}^w) \\ D_i &= g_d(i, \boldsymbol{X}, \epsilon_i^d) \\ Y_i(\boldsymbol{D}, \boldsymbol{W}) &= g_y(i, \boldsymbol{D}, \boldsymbol{W}, \boldsymbol{X}, \epsilon_i^y) \end{aligned} \tag{1}$$

where ϵ_{ij}^w, ϵ_i^d and ϵ_i^y are vector (or matrix) of random errors. The correlation between ϵ_{ij}^w and ϵ_i^d characterizes interactions between network formation and treatment assignment. When the network is undirected and link formation between pair of units depends only on pairwise covariates, the network formation model can be simplified as $W_{ij} = g_w(X_i, X_j, \epsilon_{ij}^w)$. It is worth noting that the covariate matrix $\boldsymbol{X}$ for all units may determine treatment adoption and behavioral outcome for each unit i. In Equation (1), $\boldsymbol{X}$ can be written as $(X_i, \boldsymbol{X}_{-i})$, where $\boldsymbol{X}_{-i}$ is named *contextual variable* in the literature of network analysis (Jackson et al., 2022). By incorporating such variables in g_d and g_y, researchers consider the influence of *social norm* on each unit's action on treatment adoption and behavioral outcome. If researchers have strong prior belief that such influence does not exist, the models can be simplified as $D_i = g_d(X_i, \epsilon_i^d)$ and $Y_i(\boldsymbol{D}, \boldsymbol{W}) = g_y(\boldsymbol{D}, \boldsymbol{W}, X_i, \epsilon_i^y)$.

When the treatment vector and network jointly determine the potential outcomes, the "essential" treatment assignment is high-dimensional, which makes the identification and

[4] Problems of using network measures as covariates in linear regression are discussed in Cai (2022).

estimation of treatment effects a challenging task. Following existing literature, I assume that there exists a low-dimensional, and possibly vector-valued, function of the original treatment vector and network that represents the potential outcomes.

Assumption 1: Exposure mapping: There exists a known function $g : \mathcal{N} \times \mathcal{D} \times \mathcal{W} \to \mathcal{T}$, such that

$$Y_i(\boldsymbol{d}, \boldsymbol{w}) = Y_i(t_i)$$

if $g(i, \boldsymbol{d}, \boldsymbol{w}) = t_i$ [5]. Therefore, we have $Y_i(\boldsymbol{d}, \boldsymbol{w}) = Y_i(\boldsymbol{d}', \boldsymbol{w}')$ if $g(i, \boldsymbol{d}, \boldsymbol{w}) = g(i, \boldsymbol{d}', \boldsymbol{w}')$. In the literature of casual inference under interference, $T_i = g(i, \boldsymbol{D}, \boldsymbol{W})$ is called "exposure mapping" (Aronow and Samii, 2017) or "effective treatment" (Manski, 2013). Exposure mapping has been adopted by methodological research on interference (Forastiere et al., 2021) as well as empirical research on spillover effect (Arpino and Mattei, 2016). In these papers, the authors assume that $\boldsymbol{W}$ is fixed and part of the exposure mapping function that defines the interference structure. In this paper, I regard $\boldsymbol{W}$ as an input like $\boldsymbol{D}$ for g_i. Therefore, $\boldsymbol{W}$ and $\boldsymbol{D}$ jointly determine the effective treatment for each unit. Specifically, the treatment assignment vector $\boldsymbol{D}$ and the network $\boldsymbol{W}$ are common inputs for all units under study, while the output of the exposure mapping or effective treatment is unit-heterogeneous. Note that the specification of exposure mapping needs substantial knowledge on the treatment of interest. Here are some examples. When there is no interference, we have $T_i = g(i, \boldsymbol{D}, \boldsymbol{W}) = D_i$. For a binary treatment indicator, the range of T_i is $\mathcal{T} = \{0, 1\}$. For exogenous network, if the potential outcome for each unit is determined by its own treatment status as well as the number of treated neighbors (defined by connections on the network) , we have a vector-valued exposure mapping $T_i = g(i, \boldsymbol{D}, \boldsymbol{W}) = (D_i, \sum_j W_{ij} D_j)$. For endogenous network, if treatment is defined as out-degree, we have $T_i = g(i, \boldsymbol{D}, \boldsymbol{W}) = \sum_j W_{ij}$. While its functional form can be flexible, that it is known is a rather strong assumption. Mis-specification of exposure mapping has been an emerging research topic (Sävje, 2023), which is beyond the scope of this paper.

[5] The potential outcome with exposure mapping is denoted as $Y_i(\boldsymbol{d}, \boldsymbol{w}) = \tilde{Y}_i(t_i)$ in Leung and Loupos (2022). For notational consistency, I keep using $Y_i(\cdot)$ for potential outcomes throughout this papaer.

3.3 Causal quantities of interest

For binary treatment indicator, there are two potential outcomes, and the causal estimands are well-defined. For example, the average treatment effect (ATE) is defined as $ATE = \mathbb{E}(Y_i(1) - Y_i(0))$, and we have average treatment effect on the treated (ATT) or control (ATC) defined on subgroups. In network data, the exposure mapping is multi-valued, and can be continuous and high-dimensional. We define the average potential outcome at a given level of exposure mapping t as:

$$\mu(t) = \mathbb{E}(Y_i(t))$$

The average potential outcome is also called the average dose-response function (ADRF), a concept in medical statistics and recently adopted in the casual inference literature to describe the relation between an outcome of interest and a continuous treatment. In the rest of this paper, we use ADRF to represent the average potential outcomes. For binary treatment, we have $\mu(1) = \mathbb{E}(Y_i(1))$ and $\mu(0) = \mathbb{E}(Y_i(0))$, and it is straightforward that ATE $= \mu(1) - \mu(0)$. For network data, we define the treatment effect of exposure mapping at level t compared to level t' as:

$$\tau(t, t') = \mathbb{E}(Y_i(t) - Y_i(t')) = \mu(t) - \mu(t') \tag{2}$$

While simple in its form, the treatment effect in Equation (2) may have rich meanings given researchers' choices of the exposure mapping. For example, a researcher is interested in estimating direct and spillover effects of a treatment, and defines $g(i, \boldsymbol{d}, \boldsymbol{w}) = (d_i, \sum_j w_{ij} D_j) = (d_i, z_i)$. For binary treatment and weighted matrix ($0 \leq w_{ij} \leq 1$ and $\sum_j w_{ij} = 1$), we have $d_i \in \{0, 1\}$ and $z_i \in [0, 1]$. In this case, the ADRF becomes

$$\mu(d, z) = \mathbb{E}[Y_i(d, z)], \quad d \in \{0, 1\}, z \in [0, 1]$$

Following existing literature on interference and spillover effect, researchers may be interested in multiple causal estimands. First, the *conditional direct effect* is defined as:

$$\tau(z) = \mathbb{E}[Y_i(1, z) - Y_i(0, z)] = \mu(1, z) - \mu(0, z)$$

and the *marginal direct effect* is defined by averaging $\tau(z)$ over the probability distribution of z. that is

$$\tau = \int \tau(z) f_Z(z) dz$$

where $f_Z(z)$ is the probability density of Z. Next, the *conditional spillover effect* is defined as:

$$\delta(z, z', d) = \mathbb{E}[Y_i(d, z) - Y_i(d, z')] = \mu(d, z) - \mu(d, z')$$

which is the difference between ADRF when we fix the value of individual treatment (d) and compare two alternative values z and z'. When z' is some benchmark value, e.g., $z' = 0$, we can simplify the notation by dropping z' in the equation and denote $\delta(z, d) = \mu(d, z) - \mu(d, 0)$. The *marginal spillover effect* is defined as

$$\delta(d) = \int \delta(z, d) f_Z(z) dz$$

Finally, the *overall (total) effect* is defined as the sum of direct and spillover effects:

$$\begin{aligned} TE(z) &= \mathbb{E}[Y_i(1, z) - Y_i(0, 0)] \\ &= \mu(1, z) - \mu(0, 0) \\ &= (\mu(1, z) - \mu(0, z)) + (\mu(0, z) - \mu(0, 0)) \\ &= \tau(z) + \delta(z, 0) \end{aligned}$$

Similarly, we define the *marginal total effect* as follows:

$$TE = \int TE(z) f_Z(z) dz = \tau + \delta(0)$$

3.4 Unconfoundedness of the treatment assignment and network formation

Since casual estimands are differences in the ADRF (average potential outcomes), their identification relies on the identification of ARDF. When SUTVA holds, identification results are based on the conditional unconfoundedness (ignorability) assumption:
$\{Y_i(0), Y_i(1)\} \perp\!\!\!\perp D_i | X_i$. It states that, conditional on the observed confounders X_i, the potential outcomes are independent of the observed treatment assignment. This assumption focuses on the unconfoundedness of individual treatment. For network data, potential outcomes are determined by the exposure mapping, a function of both the whole treatment assignment vector and the social network, and the assumption must be restated to account for confounders that affect treatment assignment or (and) network formation, and potential outcomes.

Existing literature has made several alternative assumptions for identification of treatment effects from network data. When the network is fixed and exogenous, Forastiere et al. (2021) show that under the following condition:

$$\{Y_i(\cdot)\} \perp\!\!\!\perp T_i | S_i \tag{3}$$

treatment effect can be identified, provided that there is no diffusion of treatment adoption. In Equation (3), S_i includes individual level covariates X_i as well as "network controls" like the contextual effects (or exogenous network influence) $\sum_j W_{ij}X_j / \sum_j W_{ij}$ and network measures like degree and centrality. If this condition holds, researchers can adopt the conventional propensity-score based methods by first specifying a probabilistic model for treatment assignment. While the assumption of unconfoundedness is weaker and easy to implement in practice, it ignores peer effect in treatment adoption. To incorporate treatment diffusion, Leung and Loupos (2022) consider a high-dimensional condition of unconfoundedness that incorporates Equation (3) as a special case [6]:

$$\{Y_i(\cdot)\} \perp\!\!\!\perp \boldsymbol{D} | \boldsymbol{X}, \boldsymbol{W} \tag{4}$$

In Equation (7), the effect of $\boldsymbol{X}_{-i}$ on D_i is called *exogenous network influence*. Since covariates of one unit may affect treatment adoption of others, treatment assignments are interdependent. Furthermore, in network data, it is also reasonable to assume that units under study are strategic, i.e., their adoption of treatment depends on whether their peers adopt treatment. To account for such equilibrial behavior, Jackson et al. (2022) make the following *Societal Conditional Unconfoundedness* assumption [7]:

$$\{Y_i(\cdot)\} \perp\!\!\!\perp D_i | \boldsymbol{D}_{-i}, \boldsymbol{X}, \boldsymbol{W} \tag{5}$$

Equation (5) is weaker than Equation (7) in that it directly incorporate the influence of other units' treatment adoption. In Equation (5), the effect of $\boldsymbol{D}_{-i}$ on D_i is called *endogenous network influence*, which characterizes the equilibrium of treatment assignments in the social network.

[6] The original assumption made in Leung and Loupos (2022) focuses on treatment assignment: $\{Y_i(\cdot)\} \perp\!\!\!\perp \boldsymbol{D} | \boldsymbol{X}, \boldsymbol{W}$. Given that the network is exogenous, I replace the treatment assignment vector with the exposure mapping.

[7] In Jackson et al. (2022), the authors assume that potential outcomes for each unit only depend on individual treatment status.

There are a few papers that consider formation of social networks. In a recent paper, Sanchez-Becerra (2022) derives the condition of unconfoundedenss under relatively strict assumptions. If the triple $(Y_i(\cdot), X_i, D_i)$ are drawn from i.i.d. distributions and the formation of link W_{ij} between unit i and j is undirected and depends on X_i and X_j, then the condition of unconfoundedenss is[8]:

$$\{Y_i(\cdot)\} \perp\!\!\!\perp T_i | X_i \tag{6}$$

This is because, *a priori*, link formation between any pair of units is randomized and there is no treatment diffusion. While easy to implement, in observational data, there can be rich dynamics between treatment assignment and network formation. For example, treatment may cause the formation of network ($\boldsymbol{D} \to \boldsymbol{W}$), network may induce the diffusion of treatment ($\boldsymbol{W} \to \boldsymbol{D}$), and treatment assignment and network formation can be interdependent ($\boldsymbol{D} \leftrightarrow \boldsymbol{W}$). In these scenarios, the unconfoundedness condition in Equation (6) is less sufficient. To account for network formation, treatment diffusion and their interaction, I make the following assumption:

Assumption 2: Unconfoundedness of treatment assignment and network formation:

$$Y_i(g(i, \boldsymbol{d}', W')) \perp\!\!\!\perp \boldsymbol{D}, W | \boldsymbol{X} \quad \forall i \in \mathcal{N}, \boldsymbol{d} \in \mathcal{D}, W' \in \mathcal{W}$$

This assumption states that, conditional on the observed covariates matrix $\boldsymbol{X}$, the potential outcomes for each unit are independent of the treatment assignment vector and the observed social network. It differs from the classical unconfoundedness assumption in that we condition on the observed covariates for all units $\boldsymbol{X}$ rather than the individual covariates X_i. This is because the covariate matrix $\boldsymbol{X}^D$ determines assignment of the treatment vector $\boldsymbol{D}$ while $\boldsymbol{X}^W$ determines the formation of the network $\boldsymbol{W}$. If they also affects the potential outcomes, then $\boldsymbol{X}^D$ and $\boldsymbol{X}^W$ confound the relation between $\boldsymbol{D}$, $\boldsymbol{W}$ and the potential outcomes. Given our definition of exposure mapping, equivalently, we have

$$\{Y_i(\cdot)\} \perp\!\!\!\perp T_i | \boldsymbol{X}, \quad \forall i \tag{7}$$

Under this assumption, the assignment mechanism can be represented by a probabilistic model $p(\boldsymbol{D}, \boldsymbol{W}; \boldsymbol{X})$. It is rather general since we don't consider the relation between $\boldsymbol{D}$

[8]The original setting in Sanchez-Becerra (2022) contains multiple separate networks. Here we consider only one network for notaional consistency.

and W. Instead, we regard them together as "joint assignment". To simplify the modeling of the assignment mechanism, I make an additional assumption to model the assignment of treatment vector $p(\boldsymbol{D};\boldsymbol{X})$ and the formation of network $p(\boldsymbol{W};\boldsymbol{X})$ separately. Formally, I assume that the treatment vector and network formation are conditionally independent given the observed covariates $\boldsymbol{X}$.

Remark 1: the assumptions above explicitly account for the endogeneity of social network that induces neighborhood treatments for the identification of ADRF. While they are weaker than SUTVA and assumptions made in some existing approaches to causal inference under interference that regard the network as fixed, they can be quite strong and are violated when there exists unobserved confounder. In section 5, I extend the estimation strategy to incorporate unobserved confounders by directly modeling it. Alternatively, researchers can conduct a sensitivity analysis for the existence of unobserved confounders.

4 A Generalized Propensity Score Approach

The propensity score is defined as the conditional probability of receiving a treatment level given observed confounders (Rosenbaum and Rubin, 1983). Under SUTVA, treatment assignments are individualized, and propensity score can be estimated with parametric models. For binary treatment indicator, the propensity score is usually modeled as a Logit or Probit model. More generally, we can have multinomial logit or normal linear models for the estimation of generalized propensity scores for multi-valued and continuous treatment (Hirano and Imbens, 2004).

While the condition of unconfoundedness above is weaker than those made in most existing research on interference and allows for rich interactions between treatment assignment and network formation, in practice, it can be difficult to estimate the ARDF by conditioning on the observed covariate matrix $\boldsymbol{X}$ for all units especially when the number of covariates is large. This paper contributes to the literature by extending the propensity score based method to identify and estimate treatment effects in the presence of endogenous networks. Unlike existing approaches (Sanchez-Becerra, 2022; Forastiere et al., 2021) that assume conditional independence of (or exchangeable) treatment assignment, it is more plausible that exposure mapping assigned to each unit is interdependent, because inputs

of the function are treatment assignment vector and the network, which are common to all units. Therefore, we can not directly assume a parametric model for exposure mapping and then estimate the propensity score. Rather, we fit a model for treatment assignment given $\boldsymbol{X}^D$ and a network formation model given $\boldsymbol{X}^W$. Given these two probabilistic models and the exposure mapping function, we implicitly estimate the propensity score for each level of exposure mapping for each unit. Since the exposure mapping $T_i = g(i, \boldsymbol{D}, \boldsymbol{W})$ is multi-valued or continuous, we follow the existing literature (Imbens, 2000) and name the propensity score for each level "generalized propensity score". Note that my approach is semi-parametric in that I do not assume an explicit model for the exposure mapping.

4.1 Generalized propensity score for exposure mapping

I define the *generalized propensity score* (gps) for unit i as the probability distribution of exposure mapping for unit i, given the observed covariate matrix $\boldsymbol{x}$:

$$r(i, t; \boldsymbol{x}) = Pr(T_i = t | \boldsymbol{X} = \boldsymbol{x}) = Pr(g(i, \boldsymbol{D}, \boldsymbol{W}) = t | \boldsymbol{X} = \boldsymbol{x}) \tag{8}$$

where $r(i, t; \boldsymbol{x})$ is the probability that the exposure mapping for unit i takes value of t given the observed covariate matrix. The definition is similar to the (ordinal) propensity score for a binary treatment indicator, which is formulated as:

$$\psi(\boldsymbol{x}_i) = Pr(D_i = 1 | \boldsymbol{X}_i = \boldsymbol{x}_i)$$

which represents the probability that the treatment assigned to unit i equal to 1. Note that the propensity score $\psi(\boldsymbol{x}_i)$ for binary treatment indicator is individualized in that it only condition on unit i's own covariates (or confounders), and the differences in propensity scores across units come from the differences in individual covariates $\boldsymbol{x}_i$. For the generalized propensity score $r(i, t; \boldsymbol{x})$ in network setting, the input $\boldsymbol{x}$ is common for all units, and we use the label $r(i, \cdot; \cdot)$ to denote unit-level heterogeneity. Like the propensity score for binary treatment indicator, we assume positivity of the generalized propensity score for exposure mapping.

Assumption 3: Positivity of the generalized propensity score for exposure mapping:

$$0 < r(i, t; \boldsymbol{x}) < 1, \quad \forall i \in \mathcal{N}, t \in \mathcal{T} \tag{9}$$

where $\mathcal{T}$ denotes all possible values that the exposure mapping may take.

4.2 Joint modeling of treatment assignment and network formation

As in our discussion of the unconfoundedness condition in Equation (7), it is weaker than the unconfoundedness condition at individual level in Equation (3) and can incorporate rich interactions between treatment assignment and network formation, but conditioning on the whole covariate matrix makes the estimation of propensity score a challenging task, especially when the number of covariate is large. For Equation (3) it is possible to assume an explicit probabilistic model for exposure mapping, but for Equation (7) the estimation problem is high-dimensional and much trickier. Leung and Loupos (2022) propose a graph neural network (GNN) based estimator for such propensity scores conditional on the whole covariate matrix. In their paper, the network is assumed to be exogeneous, and they estimate the probability for exposure mapping conditional on the covariate matrix and the observed exogeneous network, i.e., $Pr(T_i = t|\boldsymbol{X}, \boldsymbol{W})$. Becasue the estimation problem is high-dimensional, they assume that $Pr(T_i = t|\boldsymbol{X}, \boldsymbol{W}) = Pr(T_i = t|\boldsymbol{X}_{\mathcal{N}(i,L)}, \boldsymbol{W}_{\mathcal{N}(i,L)})$, where L is the depth, or number of layers, of the graph neural network, a key parameter that determines the *receptive field* $(\boldsymbol{X}_{\mathcal{N}(i,L)}, \boldsymbol{W}_{\mathcal{N}(i,L)}))$ used to predict unit i's outcome. In the terminology of network analysis, $\mathcal{N}(i, L)$ is the set of unit i's up to L^{th} order neighbors. If $L = 1$, only units who share a common link with unit i are included, i.e., $\mathcal{N}(i, L) = \{j : W_{ij} \neq 0\}$. When $\boldsymbol{W}$ is endogenous, the network as an input of the exposure mapping itself should be regarded as a "treatment", and conditioning on it may cause selection bias. Besides, since we assume $\boldsymbol{X}^D$ determines the assignment of treatment $\boldsymbol{D}$ and $\boldsymbol{X}^W$ determines the formation of network $\boldsymbol{W}$, and neither $\boldsymbol{X}^D$ nor $\boldsymbol{X}^W$ *directly* determines the exposure mapping T_i. Therefore, I propose an alternative approach to estimating generalized propensity score $r(i, t; \boldsymbol{x})$ by modeling the assignment mechanism for treatment $\boldsymbol{D}$ and the formation of network $\boldsymbol{W}$. Suppose $f_{\boldsymbol{D},\boldsymbol{W}|\boldsymbol{X}}(\boldsymbol{d}, \boldsymbol{w}|\boldsymbol{x})$ is the joint distribution of $\boldsymbol{D}$ and $\boldsymbol{W}$ conditional on the covariate matrix $\boldsymbol{X}$, the generalized propensity score of exposure mapping for unit i is formulated as:

$$\begin{aligned} r(i, t; \boldsymbol{x}) &= Pr(g(i, \boldsymbol{D}, \boldsymbol{W}) = t|\boldsymbol{X} = \boldsymbol{x}) \\ &= \int_{\boldsymbol{D}\in\mathcal{D}} \int_{\boldsymbol{W}\in\mathcal{W}} \mathbb{1}\{g(i, \boldsymbol{d}, \boldsymbol{w}) = t\} f_{\boldsymbol{D},\boldsymbol{W}|\boldsymbol{X}}(\boldsymbol{d}, \boldsymbol{w}|\boldsymbol{x}) \, \mathrm{d}\boldsymbol{d} \, \mathrm{d}\boldsymbol{w} \end{aligned} \tag{10}$$

Therefore, if we can properly model the joint distribution of the treatment assignment $\boldsymbol{D}$ and the social network $\boldsymbol{W}$ given the covariate matrix $\boldsymbol{X}$, the generalized propensity score of the exposure mapping equals the integral over the sample spaces $\mathcal{D}$ and $\mathcal{W}$ such that a pair $(\boldsymbol{d}', \boldsymbol{w}')$, $\boldsymbol{d}' \in \mathcal{D}$ and $\boldsymbol{w}' \in \mathcal{W}$, satisfies $g(i, \boldsymbol{d}', \boldsymbol{w}') = t$.

4.3 Factorization of treatment assignment and network formation

we face two challenges when estimating propensity scores of exposure mapping for social network data. First, the exposure mapping for each unit depends on the whole treatment vector and the network, i.e., treatment values are entangled. Second, the network itself is endogenous. As we discussed above, we need to condition on the whole covariate matrix $\boldsymbol{X}$ rather than individual covariate vector X_i, and direct estimation of the generalized propensity score can be difficult. I consider modeling the joint distribution of $\boldsymbol{D}$ and $\boldsymbol{W}$ and the generalized propensity score of the exposure mapping equals the integral over the sample spaces that satisfy $g(i, \boldsymbol{d}, \boldsymbol{w}) = t$ for each i. However, modeling the joint distribution of $\boldsymbol{D}$ and $\boldsymbol{W}$ is also challenging, especially for observational data where there may be rich interactions between treatment assignment and network formation. Treatment assignment and network formation are two processes. The simplest case is that these two processes are independent given the observed covariate $\boldsymbol{X}$, and there may be more complex dynamics. In the long term, they may mutually affect each other and achieve an equilibrium, which is called *simultaneity* in econometrics literature. On the other hand, if we observe the data generating process for a relatively short period, it is possible that network formation precedes treatment assignment, and even induces diffusion of treatment adoption. It is also possible that treatments assigned to a pair of units affect the tie formation process between them.

Jointly modeling the distribution of $\boldsymbol{D}$ and $\boldsymbol{W}$ incorporates complex dynamics between them, but in practice it is not straightforward to specify a probabilistic model for treatment assignment together with network formation. We consider a factorization of their joint distribution to simplify modeling without ignoring the interaction between $\boldsymbol{D}$ and $\boldsymbol{W}$. I

start with the baseline case, i.e., $\boldsymbol{D}$ and $\boldsymbol{W}$ are independent conditional on $\boldsymbol{X}$:

$$\boldsymbol{D} \perp\!\!\!\perp \boldsymbol{W} | \boldsymbol{X} \tag{11}$$

Under this assumption, the joint distribution $f_{\boldsymbol{D},\boldsymbol{W}|\boldsymbol{X}}(\boldsymbol{d}, \boldsymbol{w}|\boldsymbol{x})$ can be factorized as:

$$f_{\boldsymbol{D},\boldsymbol{W}|\boldsymbol{X}}(\boldsymbol{d}, \boldsymbol{w}|\boldsymbol{x}) = f_{\boldsymbol{D}|\boldsymbol{X}}(\boldsymbol{d}|\boldsymbol{x}) f_{\boldsymbol{W}|\boldsymbol{X}}(\boldsymbol{w}|\boldsymbol{x}) = f_{\boldsymbol{D}|\boldsymbol{X}^D}(\boldsymbol{d}|\boldsymbol{x}^D) f_{\boldsymbol{W}|\boldsymbol{X}^W}(\boldsymbol{w}|\boldsymbol{x}^W) \tag{12}$$

when treatment assignment is individualized, treatment assigned to each unit is independent, and we have $f_{\boldsymbol{D}|\boldsymbol{X}^D}(\boldsymbol{d}|\boldsymbol{x}^D) = \prod_{i=1}^{N} f_{D_i|\boldsymbol{X}_i^D}(d_i|\boldsymbol{x}_i)$. Under this factorization, researchers can separately fit probabilistic models for treatment assignment and network formation. However, it assumes that the observed covariate matrix $\boldsymbol{X}$ contains all information, and there are no unobserved confounders or homophily. This assumption is quite strong and implausible when unobserved covaraite affect both treatment assignment and network formation, i.e., *simultaneity*. When researchers observed data for long periods, it is possible that treatment assignment and network formation are mutually interdependent, and we consider the following factorization:

$$\begin{aligned} f_{\boldsymbol{D},\boldsymbol{W}|\boldsymbol{X}}(\boldsymbol{d}, \boldsymbol{w}|\boldsymbol{x}) &= \int f_{\boldsymbol{D},\boldsymbol{W}|\boldsymbol{X},\boldsymbol{U}}(\boldsymbol{d}, \boldsymbol{w}|\boldsymbol{x}, \boldsymbol{u}) f_{\boldsymbol{U}|\boldsymbol{X}}(\boldsymbol{u}|\boldsymbol{x}) \, \mathrm{d}\boldsymbol{u} \\ &= \int f_{\boldsymbol{D}|\boldsymbol{X},\boldsymbol{U}}(\boldsymbol{d}|\boldsymbol{x}, \boldsymbol{u}) f_{\boldsymbol{W}|\boldsymbol{X},\boldsymbol{U}}(\boldsymbol{w}|\boldsymbol{x}, \boldsymbol{u}) f_{\boldsymbol{U}|\boldsymbol{X}}(\boldsymbol{u}|\boldsymbol{x}) \, \mathrm{d}\boldsymbol{u} \\ &= \int f_{\boldsymbol{D}|\boldsymbol{X}^D,\boldsymbol{U}}(\boldsymbol{d}|\boldsymbol{x}^D, \boldsymbol{u}) f_{\boldsymbol{W}|\boldsymbol{X}^W,\boldsymbol{U}}(\boldsymbol{w}|\boldsymbol{x}^W, \boldsymbol{u}) f_{\boldsymbol{U}|\boldsymbol{X}}(\boldsymbol{u}|\boldsymbol{x}) \, \mathrm{d}\boldsymbol{u} \end{aligned} \tag{13}$$

where $\boldsymbol{U} = (\boldsymbol{U}_1, \ldots, \boldsymbol{U}_N)$ is a vector of unobserved covariates and $f_{\boldsymbol{U}|\boldsymbol{X}}(\boldsymbol{u}|\boldsymbol{x})$ is its distribution conditional on $\boldsymbol{X}$. When $\boldsymbol{U}$ and $\boldsymbol{X}$ are independent, we have $f_{\boldsymbol{U}|\boldsymbol{X}}(\boldsymbol{u}|\boldsymbol{x}) = f_{\boldsymbol{U}}(\boldsymbol{u})$. Given this factorization, we essentially assume:

$$\boldsymbol{D} \perp\!\!\!\perp \boldsymbol{W} | \boldsymbol{X}, \boldsymbol{U} \tag{14}$$

like the latent ignorability assumption in causal inference literature (Frangakis and Rubin, 1999). It states that, conditional on observed covariates $\boldsymbol{X}$ and unobserved covariates $\boldsymbol{U}$, treatment assignment and network formation are independent. In structural econometrics, the control function approach is proposed to jointly model the distribution of $\boldsymbol{D}$ and $\boldsymbol{W}$ accounting for unobserved $\boldsymbol{U}$.

Finally, since treatment assignment and network formation are two processes, it is possible that one precedes the other and intervenes the latter. While determining the sequential

order depends on the substantial question of interest and researchers' prior knowledge, in general we have two types of factorization. If treatment assignment precedes network formation, we have

$$f_{\boldsymbol{D},\boldsymbol{W}|\boldsymbol{X}}(\boldsymbol{d},\boldsymbol{w}|\boldsymbol{x}) = f_{\boldsymbol{D}|\boldsymbol{X}^D}(\boldsymbol{d}|\boldsymbol{x}^D) f_{\boldsymbol{W}|\boldsymbol{X}^W,\boldsymbol{D}}(\boldsymbol{w}|\boldsymbol{x}^W,\boldsymbol{d})$$

If network formation induces diffusion of treatment adoption, we have

$$f_{\boldsymbol{D},\boldsymbol{W}|\boldsymbol{X}}(\boldsymbol{d},\boldsymbol{w}|\boldsymbol{x}) = f_{\boldsymbol{W}|\boldsymbol{X}^W}(\boldsymbol{w}|\boldsymbol{x}^W) f_{\boldsymbol{D}|\boldsymbol{X}^D,\boldsymbol{W}}(\boldsymbol{d}|\boldsymbol{x}^D,\boldsymbol{w})$$

and we can model the diffusion process with the spatial autoregressive (SAR) model accounting for network endogeneity (Goldsmith-Pinkham and Imbens, 2013).

4.4 Back to the generalized propensity score

Once we model the joint distribution of treatment assignment and network formation, given the functional form of exposure mapping, we can estimate the generalized propensity score for each unit. I show that for simple functional forms of exposure mapping, some special types of network like unweighted networks, and binary treatment indicators, it is possible to get closed-form expressions for the generalized propensity score. We can also adopt simulation-based methods for estimation. I show that the generalized propensity score has the desirable balancing property, i.e., given the propensity score, exposure mapping is independent of the observed covariate matrix $\boldsymbol{X}$ for each unit.

Suppose we have a binary treatment $D_i \in \{0,1\} \quad \forall i$, and an unweighted network $W_{ij} \in \{0,1\} \quad \forall i,j$. The exposure mapping is defined as $T_i = g(i,\boldsymbol{D},\boldsymbol{W}) = (D_i, \sum_j W_{ij}D_{ij})$, i.e., the potential outcome is determined by the unit's own treatment status and the number of units under treatment that it shares a common link. For notational convenience, we denote $K_i = \sum_j W_{ij}D_{ij}$ and write $T_i = (D_i, K_i)$. Suppose treatment assignment is individualized and $\boldsymbol{D} \perp\!\!\!\perp \boldsymbol{W}|\boldsymbol{X}$, we can formulate the generalized propensity score $r(i,t;\boldsymbol{x})$ as:

$$\begin{aligned} r(i,t;\boldsymbol{x}) &= Pr(g(i,\boldsymbol{D},\boldsymbol{W}) = (d,k)|\boldsymbol{x}) \\ &= Pr(D_i = d|x_i^D) \sum_{\sum_j w_{ij}d_j = k} Pr(\boldsymbol{D}_{-i} = \boldsymbol{d}_{-i}|\boldsymbol{x}^D) Pr(\boldsymbol{W} = \boldsymbol{w}|\boldsymbol{x}^W) \\ &= Pr(D_i = d|x_i^D) \sum_{\sum_j w_{ij}d_j = k} \prod_{j \neq i} Pr(D_j = d_j|x_i^D) Pr(\boldsymbol{W} = \boldsymbol{w}|\boldsymbol{x}^W) \end{aligned} \tag{15}$$

The factorization above shows a closed-form formula for $r(i,t;\boldsymbol{x})$ under constraints on treatment indicator, types of network and functional form of exposure mapping. It essentially is a synthesis of the probabilistic models of treatment assignment $f_{\boldsymbol{D}|\boldsymbol{X}^D}(\boldsymbol{d}|\boldsymbol{x}^D)$ and network formation $f_{\boldsymbol{W}|\boldsymbol{X}^W}(\boldsymbol{w}|\boldsymbol{x}^W)$. For complicated scenarios, like continuous treatment, weighted networks and complex functional form of exposure mapping incorporating influences from higher order neighbors, it is difficult to get close-form expression for generalized propensity scores, and we need to adopt simulation-based methods (Aronow and Samii, 2017; Toulis et al., 2021). Details of simulation-based methods for estimating generalized propensity scores are introduced in the next section.

Given the definitions of the generalized propensity score $r(i,t;\boldsymbol{x})$, under the assumptions above, we have the following propositions:

Proposition 1: Balancing property of the generalized propensity score:

$$Pr(T_i = t|\boldsymbol{x}, r(i,t;\boldsymbol{x})) = Pr(T_i = t|r(i,t;\boldsymbol{x}))$$

Proposition 2: Conditional unconfoundedness given the generalized propensity score:

$$\{Y_i(\cdot)\} \perp\!\!\!\perp T|r(i,t;\boldsymbol{x})$$

In the next section, I introduce how to use the generalized propensity score derived above to identify and estimate the average dose-response function (ADRF) as well as the treatment effects.

5 Estimation and Inference

In the previous section, I introduce the generalized propensity score $r(i,t;\boldsymbol{x})$ of the exposure mapping and its estimation based on modeling the joint distribution of treatment assignment vector $\boldsymbol{D}$ and the social network $\boldsymbol{W}$, and we show the balancing property of $r(i,t;\boldsymbol{x})$. In this section, I study the identification of average dose-response function (ADRF) and treatment effects and propose propensity score based estimators. Note that the balancing property of generalized propensity score implies the identification of ADRF:

Proposition 3: Identification of ADRF:

$$\mathbb{E}\{\frac{Y_i\mathbb{1}\{T_i = g(i,\boldsymbol{D},\boldsymbol{W}) = t\}}{r(i,t;\boldsymbol{x})}\} = \mathbb{E}\{Y_i(t)\} = \mu(t) \tag{16}$$

Since we assume positivity of $r(i,t;\boldsymbol{x})$, given the estimated generalized propensity score $\hat{r}(i,t;\boldsymbol{x})$, we can estimate the ADRF $\mu(t)$ without bias with the inverse probability (Horvitz-Thompson) estimator:

$$\hat{\mu}(t)^{HT} = \sum_{i=1}^{N} \mathbb{1}\{T_i = t\} \frac{Y_i}{\hat{r}(i,t;\boldsymbol{x})} \tag{17}$$

It is possible that we get extreme values of the Horvitz-Thompson estimates when $\hat{r}(i,t;\boldsymbol{x})$ is close to 0, which implies a large value of weight $1/\hat{r}(i,t;\boldsymbol{x})$ for that observation. Therefore, the Horvitz-Thompson estimator is of high variance. Alternatively, we can estimate ADRF with the Hajek estimator, which is proved to improve efficiency at the cost of finite sample bias.

$$\hat{\mu}(t)^{H} = \frac{\sum_{i=1}^{N} \mathbb{1}\{T_i = t\} \frac{Y_i}{\hat{r}(i,t;\boldsymbol{x})}}{\sum_{i=1}^{N} \mathbb{1}\{T_i = t\} \frac{1}{\hat{r}(i,t;\boldsymbol{x})}} \tag{18}$$

The Hajek refinement allows the denominator to vary according to the sum of the weights ($\frac{1}{\hat{r}(i,t;\boldsymbol{x})}$), and it improves estimation efficiency by shrinking the magnitude of the estimator when its value is large, and increasing the magnitude of the estimator when its value is small.

With a properly-specified outcome model $\mu(i,T,\boldsymbol{X},\boldsymbol{W})$, we can also estimate $\mu(t)$ with the doubly-robust estimator:

$$\hat{\mu}(t)^{DR} = \sum_{i=1}^{N} \mathbb{1}\{T_i = t\} \frac{Y_i - \hat{\mu}(i,t,\boldsymbol{x},\boldsymbol{w})}{\hat{r}(i,t;\boldsymbol{x})} + \hat{\mu}(i,t,\boldsymbol{x},\boldsymbol{w}) \tag{19}$$

since $\boldsymbol{x}$ is of high dimension, we can fit $\hat{\mu}(i,t,\boldsymbol{x},\boldsymbol{w})$ with machine learning models like random forests. An estimator for treatment effect at exposure mapping level t compared to level t' is just the difference between estimated ADRF. Taking the doubly-robust estimator as an example, we have:

$$\hat{\tau}(t,t') = \hat{\mu}(t)^{DR} - \hat{\mu}(t')^{DR} \tag{20}$$

For inference, provided that the generalized propensity score is consistently estimated, i.e., $\hat{r}(i,t;\boldsymbol{x}) \xrightarrow{\text{plim}} r(i,t;\boldsymbol{x})$, we can estimate the asymptotic variance with the network HAC estimator (Kojevnikov et al., 2021). Leung and Loupos (2022) consider a uniform kernel, and the HAC estimator for $\hat{\mu}(t)^{HT}$ is written as:

$$\hat{\sigma}^2_{\hat{\mu}(t)^{HT}} = \frac{1}{n} \sum_{i=1}^{N} \sum_{j=1}^{N} (\hat{\mu}(t)_i^{HT} - \hat{\mu}(t)^{HT})(\hat{\mu}(t)_j^{HT} - \hat{\mu}(t)^{HT}) \mathbb{1}\{\ell_{\boldsymbol{W}}(i,j) \leq b\} \tag{21}$$

where $\hat{\mu}(t)_i^{HT} = \mathbb{1}\{T_i = t\}\frac{Y_i}{\hat{r}(i,t;\boldsymbol{x})}$, b is a pre-specified bandwidth, and $\ell_{\boldsymbol{W}}(i,j)$ is the shortest path length from i to j (and $\ell_{\boldsymbol{W}}(i,j) = \ell_{\boldsymbol{W}}(j,i)$ for undirected network). Note that the HAC estimator can also incorporate grouped data as a special case. In this case, $\ell_{\boldsymbol{W}}(i,j) = \infty$ if unit i and j belong to different groups. If b is greater than or equal to the largest group size, then the HAC estimator equals clustered robust variance estimator.

5.1 Regression-based estimation and inference for continuous exposure mapping

While we can directly use the inverse probability weighting type estimators (e.g. Horvitz-Thompson estimator) to estimate the average dose-response function given the estimated generalized propensity score, it is possible that for a given level of exposure mapping t, we do not have any observed outcomes. For example, when the treatment assignment $\boldsymbol{D}$ is continuous or the network $\boldsymbol{W}$ is weighted, i.e., $0 \leq W_{ij} \leq 1$. Therefore, I follow Forastiere et al. (2021) and propose a regression-based estimation procedure for the practical implementation of generalized propensity score based methods to estimate ADRF as well as treatment effects. The regression-based estimation procedure is proposed by Hirano and Imbens (2004) to estimate the effect of a single continuous treatment. I briefly introduce the two-step regression based estimator. Suppose X_i is a vector of observed confounders for unit i, T_i is a continuous treatment variable, and Y_i is the outcome of interest. In the first step, a parametric model for treatment assignment given observed confounder is estimated. For continuous T_i, we can assume a linear model:

$$T_i|X_i \sim N(\beta_0 + \beta_1' X_i, \sigma^2)$$

And the estimated generalized propensity score equals:

$$\hat{R}_i(T_i; X_i) = \frac{1}{\sqrt{2\pi}\hat{\sigma}} \exp(-\frac{1}{2\hat{\sigma}^2}(T_i - \hat{\beta}_0 - \hat{\beta}_1' X_i)^2)$$

Given the estimated generalized propensity score, we estimate a parametric model for the outcome given the generalized propensity score and the treatment. Hirano and Imbens (2004) consider a quadratic approximation:

$$\mathbb{E}[Y_i|X_i, R_i] = \alpha_0 + \alpha_1 T_i + \alpha_2 T_i^2 + \alpha_3 R_i + \alpha_4 R_i^2 + \alpha_5 T_i R_i$$

An estimator for the average dose-response function $\hat{\mathbb{E}}(Y_i(t))$ is:

$$\hat{\mathbb{E}}(Y_i(t)) = \frac{1}{N}\sum_{i=1}^{N}(\hat{\alpha}_0 + \hat{\alpha}_1 t + \hat{\alpha}_2 t^2 + \hat{\alpha}_3 \hat{r}_i(t; x_i) + \hat{\alpha}_4 \hat{r}_i(t; x_i)^2 + \hat{\alpha}_5 t \hat{r}_i(t; x_i))$$

My approach is similar to the two-step estimator above, except that in the first step, I estimate the generalized propensity score in a semi-parametric way rather than assume a parametric model. I use the baseline case, i.e., independence of treatment assignment and network formation given observed covariate matrix, to describe the estimation procedure. Modifications can be made to incorporate simultaneity or sequential orders. First, I estimate a model for treatment adoption at individual level, which is similar to the conventional propensity score analysis, and a model for network formation. Next, I combine the two probabilistic models and the functional form of the exposure mapping to estimate the generalized propensity score for each level of exposure mapping. Finally, an outcome regression model is estimated with adjustment of the estimated propensity score. Average dose-response functions are estimated based on the estimated outcome regression model and estimated generalized propensity score. Detailed steps are illustrated as follows:

Step 1 Fit probabilistic models for individual treatment adoption and network formation given observed covariates $\boldsymbol{X}^D$ and $\boldsymbol{X}^W$:

(1) Estimate a probabilistic model $D \sim f^D(D; \boldsymbol{X}^D, \theta^D)$ for treatment assignment. For binary treatment indicator, we can estimate $D_i \sim Bern(\phi(1; X_i^D))$ with a Probit or Logit model for D_i conditional on observed covariates X_i^D.

(2) Estimate $W \sim f^W(W; \boldsymbol{X}^W, \theta^W)$ with a probabilistic model like the latent space model.

Step 2 Estimate the generalized propensity score $r(i, t; \boldsymbol{X})$ of each exposure mapping level t for each unit given the estimated probabilistic models $f^D(D; \boldsymbol{X}^D, \theta^D)$ and $f^W(W; \boldsymbol{X}^W, \theta^W)$. When it is difficult to get closed-form expression, we consider the following simulation based methods:

(1) For a given large number N_s, we simulate N_s treatment vectors $\tilde{D}^k$ from

$f^D(D; \boldsymbol{X}^D, \theta^D)$ and networks $\tilde{W}^k$ from $f^W(W; \boldsymbol{X}^W, \theta^W)$ $(1 \leq k \leq N_s)$.

(2) For $1 \leq k \leq N_s$, we get a simulated exposure mapping $\tilde{T}_i^k = g(i, \tilde{D}^k, \tilde{W}^k)$ for each unit. Therefore, we get an empirical distribution of exposure mapping for each unit: $\{\tilde{T}_i^1, \ldots, \tilde{T}_i^k, \ldots, \tilde{T}_i^{N_s}\}$.

(3) We estimate the probability for a given level t of the exposure mapping for each unit, based on the empirical distribution $\{\tilde{T}_i^1, \ldots, \tilde{T}_i^k, \ldots, \tilde{T}_i^{N_s}\}$, as the estimated generalized propensity score $\hat{r}(i, t; \boldsymbol{x})$). When T takes continuous values, densities can be estimated with kernel density estimators.

Step 3 Given the estimated generalized propensity score, we fit a regression model for observed outcomes:

$$Y_i(T_i)|r(i, T_i, \boldsymbol{X}) \sim f^Y(Y; T_i, r(i, T_i, \boldsymbol{X}), \theta^Y)$$

Step 4 For a given level t of exposure mapping, an estimator for the average dose-response function is:

$$\hat{\mu}(t) = \hat{\mathbb{E}}(Y_i(t)) = \sum_{i=1}^{N} \hat{Y}_i(t)$$

where $\hat{Y}_i(t)$ is the imputed counterfactual outcome for unit i under exposure mapping t. The counterfactual outcomes $\hat{Y}_i(t)$ are estimated based on the outcome regression using estimated propensity score $\hat{r}(i, t, \boldsymbol{X})$.

For uncertainty estimates, we can also apply the network HAC variance estimator as long as the generalized propensity score is consistently estimated. To account for uncertainty in propensity score estimation, it is also plausible to adopt modified bootstrap procedure (Forastiere et al., 2021) or Bayesian inference (Forastiere et al., 2022).

5.2 Extension: incorporating unobserved unit-level heterogeneity

When there exists unobserved covariates that affect both treatment assignment and network formation, we can not separately model the distributions $f_{\boldsymbol{D}|\boldsymbol{X}^D}(\boldsymbol{d}|\boldsymbol{x}^D)$ and $f_{\boldsymbol{W}|\boldsymbol{X}^W}(\boldsymbol{w}|\boldsymbol{x}^W)$.

In this subsection, I illustrate how to model the joint distribution of $\boldsymbol{D}$ and $\boldsymbol{W}$ when there exists unobserved unit-level heterogeneity. Suppose the assumption of latent independence holds, i.e., $\boldsymbol{D} \perp\!\!\!\perp \boldsymbol{W}|\boldsymbol{X}, \boldsymbol{U}$. We consider a simple example with binary treatment indicator and unweighted social network without directions. Therefore, $D_i \in \{0,1\}$, $W_{ij} \in \{0,1\}$ and $W_{ij} = W_{ji}$. We assume a Probit model for treatment adoption and latent space model for network formation:

$$\begin{aligned} D_i &= \mathbb{1}\{X_i'\beta_d + \alpha_d U_i + \varepsilon_i\} \\ W_{ij} &= \mathbb{1}\{X_{ij}'\beta_w + \alpha_w|U_i - U_j| + e_{ij}\} \end{aligned} \tag{22}$$

where X_{ij} is a vector of dyadic covariates for dyad (i,j), which includes observed homophily, like $|X_{ip} - X_{jp}|$ for some covariate p, and other measures like geographical approximation. U_i and U_j represent individual-level unobserved heterogeneity, which are parameters to be estimated. $|U_i - U_j|$ represents unobserved homophily, with α_w the corresponding effect on the probability of link formation between unit i and j. The following constraints are added for identification of parameters.

$$\varepsilon_i \sim N(0,1), \quad e_{ij} \sim N(0,1), \quad U_i \sim N(0,1) \tag{23}$$

which implies

$$\begin{aligned} Pr(D_i = 1) &= \Phi(X_i'\beta_d + \alpha_d U_i) \\ Pr(W_{ij} = 1) &= \Phi(X_{ij}'\beta_w + \alpha_w|U_i - U_j|) \end{aligned}$$

where $\Phi(\cdot)$ is the cumulative distribution function (CDF) for standard normal distribution. The proposed model is under-identified because U_i is unobserved and we have $\alpha_d U_i = (-\alpha_d)(-U_i)$ and $|U_i - U_j| = |(-U_i) - (-U_j)|$. We consider a Bayesian approach to the system of models in Equation (22) for estimation and inference. Details of the Markov Chain Monte Carlo (MCMC) algorithm can be found in the appendix.

6 Monte Carlo Studies

In this section, I conduct a series of Monte Carlo studies to investigate finite sample property of the proposed generalized propensity score based estimator for estimating treatment effects from observational network data. I consider bias, root mean squared error

(RMSE) and coverage rate of 95% confidnece interval as key performance metrics. Difference between average estimated standard error (SE) and the sampling variation (SD) is also reported.

I use clustered data for simulation studies, where each simulated dataset consists of multiple independent clusters. Units belonging to the same cluster may form link between each other, but units belonging to different clusters will not be connected. For a given simulated dataset, I use $c(\cdot)$ to denote the cluster that each unit belongs to. Thus $c(i) = 1$ if unit i belongs to cluster 1. Suppose $\boldsymbol{W}$ is the observed network that specifies connections between units for this simulated dataset, then if $c(i) \neq c(j)$, $W_{ij} \equiv 0$, otherwise $W_{ij} \in \{0, 1\}$. For notational convenience, we assume that sizes of clusters are equal. We consider binary treatment indicator and unweighted network with no directions. The data generating process (DGP) for treatment assignment and network formation is as follows:

$$
\begin{aligned}
X_i &\sim \begin{cases} N(0, 0.5) & \text{if } i \text{ is odd} \\ N(0, -0.5) & \text{if } i \text{ is even} \end{cases} \\
W_{ij} &= \begin{cases} 0 & \text{if} \quad c(i) \neq c(j) \\ \mathbb{1}\{0.1 - 0.25|X_i - X_j| + \nu_i \geq 0\}, \quad \nu_i \sim N(0,1) & \text{if} \quad c(i) = c(j) \end{cases} \\
D_i &= \mathbb{1}\{0.1 + 0.5X_i + e_i \geq 0\}, \quad e_i \sim N(0,1)
\end{aligned}
\tag{24}
$$

where cluster $c(\cdot)$ is pre-specified. For each cluster, observed covariates X_i are drawn from independent but not identical normal distributions. For simplicity, I assume that cluster size equals 4 and cluster is determined by unit's id, i.e., unit 1, 2, 3, and 4 belongs to cluster 1, unit 5, 6, 7, and 8 belongs to cluster 2, and so on. The exposure mapping is defined as $T_i = g(i, \boldsymbol{D}, \boldsymbol{W}) = (D_i, \sum_j W_{ij} D_j) = (D_i, Z_i)$, and obviously $Z_i \in \{0, 1, 2, 3\}$. Therefore, the exposure mapping T_i may take 8 different levels: $(0, 0)$, $(0, 1)$, $(0, 2)$, $(0, 3)$, $(1, 0)$, $(1, 1)$, $(1, 2)$, $(1, 3)$. DGP for the potential outcome is:

$$
Y_i(d, z) = 1 + X_i + d + 0.5z + dz + \varepsilon_i, \quad \varepsilon_i \sim N(0, 1) \tag{25}
$$

Note that there are multiple quantities of interest (QoI), like the average dose-response function (ADRF) and treatment effect as differences in ADRF. Therefore, I focus on a direct treatment effect, $\tau((1, 1), (0, 1)) = \mu(1, 1) - \mu(0, 1)$, and a spillover treatment effect,

$\tau((1,1),(1,0)) = \mu(1,1) - \mu(1,0)$, as quantities of interest. Equation (25) implies constant treatment effects, i.e., the individual level treatment effect $Y_i(d,z) - Y_i(\tilde{d},\tilde{z})$ is constant across units and does not depend on covariate X_i. I also consider an alternative DGP for potential outcome with heterogeneous treatment effect. I generated simulated datasets from the DGP above with varying sample sizes. Since the size of cluster is fixed, I generate datasets with differnet number of clusters. Specifically, I consider the number of clusters equals $N_c = 20$, 50, 100, 500 and 1000. For each number of clusters, I conduct 2,000 simulations.

In the simulation setting, since we only have 8 levels of exposure mapping and the generalized propensity score has a closed-form expression, I use the Horvitz-Thompson (HT) estimator in Equation (17) to estimate treatment effect. For the proposed model, I use the network HAC variance estimator in Equation (21) to construct 95% confidence interval. Since I only consider correlation for units within the same cluster, the HAC variance estimator is equal to the clustered robust variance estimator. Results of Monte Carlo studies for the two treatment effects are summarized in Table 1 and Table 2.

Table 1: Finite Sample Properties: $\tau((1,1),(0,1))$

N_c	Bias	RMSE	Coverage Rate	SE - SD
20	0.041	1.119	0.972	0.123
50	0.011	0.695	0.978	0.115
100	0.006	0.480	0.982	0.091
500	0.004	0.210	0.981	0.049
1000	0.002	0.155	0.983	0.029

Table 2: Finite Sample Properties: $\tau((1,1),(1,0))$

N_c	Bias	RMSE	Coverage Rate	SE - SD
20	0.003	1.157	0.961	0.105
50	0.018	0.693	0.974	0.110
100	0.004	0.507	0.973	0.060
500	−0.001	0.222	0.976	0.032
1000	0.001	0.157	0.978	0.023

I find that, the proposed model produces small bias even when sample sizes are small. Meanwhile, bias decreases fast as the number of clusters increases. The decrease is not

monotone for the spillover effect though. For root mean squared error, it also decreases as the number of clusters increases, as both bias and sampling variation decrease.

Next, I investigate the properties of the network HAC variance estimator. I find that, the coverage rate of 95% confidence interval is slightly larger than the nominal rate. Therefore, the variance estimator is conservative. However, the difference between average estimated standard error and sampling variation of the proposed estimator decreases as the number of clusters increases, indicating that the variance estimator is consistent.

Additional Monte Carlo studies indicate that properties of the proposed estimator are robust to heterogeneous treatment effect. Details of the results can be found in appendix.

7 Empirical Analysis

To demonstrate its utility in empirical social science research, I apply the proposed method to investigate the effect of U-Bridge on political participation in Uganda. Additionally, I compare this method with several existing identification strategies, discussing the plausibility of their underlying assumptions in the empirical setting described in Section 2.

First, consider the scenario where we disregard the network structure, assuming that the Stable Unit Treatment Value Assumption (SUTVA) holds. Given that resident-level covariates might influence both the adoption of U-Bridge and political participation, I employ linear regression models to account for such confounders. Estimation results are presented in Column (1) of Table 3. I observe a positive and statistically significant effect, suggesting that adopting U-Bridge increases political participation.

Next, recognizing that residents are inter-connected through various forms of social ties within each village, we are interested in whether U-Bridge adoption by one resident affects political participation of other residents. Such spillover effects are conceivable due to the transmission of information across social networks. Indeed, Sanchez-Becerra (2022) has identified a notable spillover effect from attending U-Bridge meetings on political participation. To identify spillover effects, we need to specify the exposure mapping. I assume $g(i, \boldsymbol{D}, \boldsymbol{W}) = (D_i, \sum_j W_{ij} D_j)$, meaning that political participation hinges both on a resident's own adoption of U-Bridge and the number of her connections who have adopted it. The term $\sum_j W_{ij} D_j$ is then incorporated as an additional treatment variable in the

Table 3: Regression Models for the Effect of U-Bridge on Political Participation

Outcome Variable	Political Participation Index (Y_i)			
	Model (1)	Model (2)	Model (3)	Model (4)
Adoption (D_i)	0.340***	0.261***	0.256***	0.369***
	(0.068)	(0.068)	(0.064)	(0.089)
Spillover ($\sum_j W_{ij} D_j$)		0.048***	0.051***	0.056***
		(0.007)	(0.007)	(0.007)
Interaction ($D_i \cdot \sum_j W_{ij} D_j$)				-0.032**
				(0.014)
Control Variable (X_i)	✓	✓	✓	✓
Contextual Variable ($\sum_j W_{ij} X_j$)			✓	✓
Village Fixed Effects	✓	✓	✓	✓
Observations	3,018	3,018	3,018	3,018

Notes: Robust standard errors clustered at village level are in the parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

regression model, with its coefficient representing the spillover effect. Estimations are detailed in Column (2) of Table 3, revealing a positive and substantial spillover effect. The coefficient of D_i, representing the effect of direct treatment, diminishes, though remains positive and significant. This diminution is expected if individuals who adopted U-Bridge tend to have more connections with other adopters, a phenomenon which can be attributed to homophily, given that spillover effects are positive.

In social networks, covariates of one resident might affect potential outcome of others due to peer influence (Jackson et al., 2022). Consequently, I introduce contextual variables, which are average of linked residents' covariates, to mitigate confounding influence. Lastly, an interaction term between D_i and $\sum_j W_{ij} D_j$ is added to capture treatment effect heterogeneity. The results of these estimations are showcased in Columns (3) and (4) of Table 3. When incorporating the contextual variables, the coefficient of D_i decreases slightly, while the coefficient of $\sum_j W_{ij} D_j$ sees a marginal increase. The coefficient of the interaction term is negative, indicating decreasing effect of D_i as $\sum_j W_{ij} D_j$ increases.

While regression models are flexible and easy to implement, the estimates may be biased due to model mis-specification. Moreover, it is hard to interpret the treatment effect when the social network itself is part of the exposure mapping. As a result, I employ

the generalized propensity score based methods, design-based alternatives that focus on treatment assignment mechanism and network formation process, to estimate treatment effects. First, suppose the social network is exogenously given, the propensity score $Pr(T_i = t)$ only depends on individual level covariates X_i as well as "network controls" according to Equation (3) (Forastiere et al., 2021). Here I use average value $\sum_j W_{ij}X_j / \sum_j W_{ij}$ as the network controls. On the other hand, if the assumption of exchangeability in sampling is satisfied (Sanchez-Becerra, 2022), the propensity score only depends on individual level covariates. To account for the correlation between D_i and $\sum_j W_{ij}D_j$, I regard the exposure mapping $T_i = (D_i, \sum_j W_{ij}D_j)$ as a multi-valued treatment, and fit parametric models to estimate generalized propensity scores. In this paper, I fit multinomial logit models (Feng et al., 2012; McCaffrey et al., 2013) with village fixed effects to account for unobserved village level heterogeneity [9]. I refer to this approach as "MN PS" for *Multinomial propensity score*. If network controls are included, it is denoted by "MN Net PS".

While the approaches above consider either network formation process or possible diffusion of treatment adoption by including network controls as confounders, neither can incorporate these two processes at the same time. Furthermore, they falls short in addressing the interaction between treatment assignment and network formation. This oversight can lead to biased estimations of the generalized propensity score. Besides, probability of tie formation between pairs of residents may not remain equal after the realization of resident-level covariates. To address these concerns, I implement the proposed method, termed by "JM PS" standing for "joint modeling". This approach explicitly models both processes, effectively integrating their interactions. In this empirical setting, it is plausible to believe that the social network formation process precedes treatment assignment. I fit network formation models separately for each village to account for village-level heterogeneity like trust and informal institutions. I fit a Logit model, first line in Equation (26), for the presence of social tie between a pair of respondents, and control variables are ab-

[9]Note that the estimation strategies are different from those in Forastiere et al. (2021) and Sanchez-Becerra (2022). Forastiere et al. (2021) consider the factorization of D_i and $\sum_j W_{ij}D_j$ and fit two separate models to estimate propensity scores. Sanchez-Becerra (2022) assumes a functional form for the potential outcomes. Here I follow the assumptions like sampling and network formation in Sanchez-Becerra (2022) to construct the generalized propensity score and apply the inverse probability weighing estimators.

solute difference in age, gender, levels of income, secondary education, and proxy measure of care for the community. These variables represent observed homophily. I also include binary indicators for whether at least one of them own a phone and whether at least one of them occupy a leadership role within the village. While the estimated coefficients are heterogeneous across villages, for most villages the coefficients of absolute differences in respondent-level covaraites are significantly negative, and the effects of occupying formal leadership role and owning a phone are significantly positive on social tie formation. Details of estimation results can be found in the appendix.

For treatment assignment mechanism, the presence of social network could induce the diffusion of treatment adoption. In fact, Table 4 indicates that treatment assignments come in clusters. Adopters tend to have links with adopters compared to non-adopters. Thus we need to incorporate peer effects in adoption of U-Bridge. For model fitting, I fit a Logit model, second line in Equation (26), for adoption of U-Bridge that include resident-level covariates, contextual variables for peer effects, and village-level fixed effects ($\gamma_{w,g(i)}$ and $\gamma_{d,g(i)}$) to control for unobserved village-level heterogeneity. The estimation results are summarized in column (2) in Table 7. I find that women are less likely to adopt U-Bridge than men. Additionally, secondary education, formal leadership and owning a phone have significant positive effects on adoption. The coefficients for some contextual variables are statistically significant, indicating the existence of peer effects. I simulate 20,000 networks from the network formation model to construct generalized propensity scores for the exposure mapping $(D_i, \sum_j W_{ij} D_j)$. Clustered robust standard errors at village level are calculated for each estimation strategy to account for correlation within each village.

$$\begin{aligned} Pr(W_{ij} = 1) &= \text{logit}^{-1}(|X_i - X_j|'\beta_{w1,g(i)} + X_{ij}'\beta_{w2,g(i)} + \gamma_{w,g(i)}) \\ Pr(D_i = 1) &= \text{logit}^{-1}(X_i'\beta_{d1} + (\sum_j W_{ij} X_j)'\beta_{d2} + \gamma_{d,g(i)}) \end{aligned} \tag{26}$$

I evaluate treatment effect estimates based on these different estimation strategies. Given that only a small proportion of respondents have adopted the U-Bridge program and that adoption might be interdependent, certain levels of exposure mapping contain limited observations. As a result, my analysis primarily concentrates on two direct treat-

Table 4: Contingency Table for Exposure mapping

	$\sum_j W_{ij}D_j = 0$	$\sum_j W_{ij}D_j > 0$
$D_i = 0$	990	1893
$D_i = 1$	5	130

ment effects defined as $\tau_{d1} = \mu(1,1) - \mu(0,1)$ and $\tau_{d2} = \mu(1,2) - \mu(0,2)$ and two direct effects $\tau_{i1} = \mu(0,1) - \mu(0,0)$ and $\tau_{i2} = \mu(0,2) - \mu(0,0)$. For generalized propensity score based approaches, since the number of observations under these exposure mapping levels can be small, I fit a random forest model for outcomes with 2,000 trees and implement the doubly robust estimator. The estimates are similar if I adopt a linear model with village fixed effects. I derive the estimates using estimated coefficients for regression based approach. Estimated treatment effects with 95% confidence intervals based on different model specifications are displayed in Figure 2.

For the direct effect $\hat{\tau}_{d1}$, both propensity score based approaches and regression models yield positive and statistically significant estimates, and the magnitudes of estimated effects are similar, though magnitudes are smaller for Model (2) and (3). We may conclude that, when residents have connection to one resident that adopted U-Bridge, the adoption of U-Bridge increased her political participation. Next, we are interested in whether the direct effect is constant or is it moderated by residents' connections.

Regarding the direct effect $\hat{\tau}_{d2}$, regression models still yield significant estimates, but with the interaction term $D_i \cdot \sum_j W_{ij}D_j$ included, the estimated effect attenuates. For generalized propensity score based approaches, Multinomial propensity score based estimates are marginally significant and their magnitudes are smaller. The proposed method produces insignificant estimate, though its magnitude is slightly larger. As previously discussed, regression models might either presuppose a constant effect or misrepresent the structure of treatment effect heterogeneity.

Based on these estimation results, we can infer that adopting U-Bridge boosts political participation of a resident, but such effect is not constant and is moderated by the number of connected adopters. This may be because that residents can learn from connected adopters about the efficacy of U-Bridge. After acquiring information, they may adjust their own

behaviors, and such adjustment is not monotone increasing with respect to the number of connected adopters.

For both indirect effects, estimates from regression models are all positive and statistically significant. However, the estimated may be biased due to model mis-specification. For the proposed method, the indirect effect $\hat{\tau}_{i1}$ is positive and significant, though its magnitude is smaller than those from regression models. For indirect effect $\hat{\tau}_{i2}$, it has larger magnitude but is marginally insignificant. Given that proposed estimator has relatively large sampling variation, we may conclude that the adoption of U-Bridge has positive spillover effect. If a resident is connected to other adopters, she may also be more active in political participation. However, the spillover effect may not be constant or monotone increasing with respect to the number of adopters. For propensity scores estimated with multinomial logit models, $\hat{\tau}_{i1}$ is marginally insignificant while $\hat{\tau}_{i2}$ is marginally significant. The discrepancy between propensity score based approaches might arise from neglecting the diffusion of treatment adoption induced by social network and heterogeneity in probability of tie formation *a posteriri*.

In summary, when estimating treatment effects from social network data, researchers should consider the network formation process, treatment assignment mechanisms, and potential interactions between these processes. Only then can they select the most appropriate estimation strategy for their specific empirical setting.

8 Conclusion

In this paper, I propose a generalized propensity score based approach to identification and estimation of treatment effects from observational social network data. In such data, treatment assigned to one unit may affect potential outcome of other units. Meanwhile, there may be some covariates that determines adoption of the treatment, formation of social ties and observed behavioral outcomes, making identification and estimation a challenging task. To incorporate the rich interactions between treatment assignment and network formation process, I propose to jointly model these two processes. Given a known functional form of exposure mapping that determines the effective treatment level, generalized propensity score for each treatment level is semi-parametrically estimated based on the probabilistic

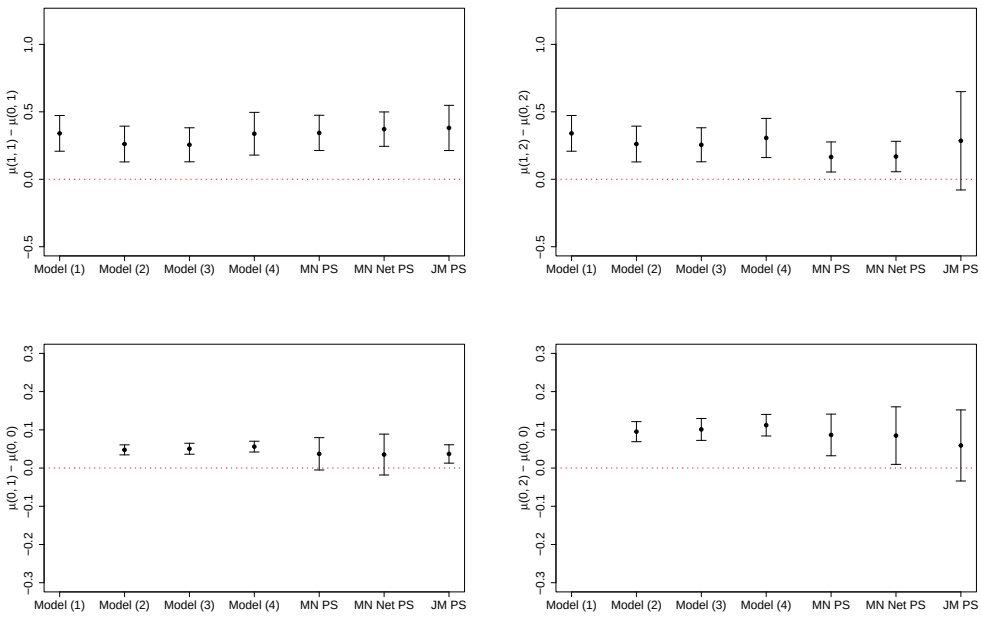

Figure 2: Estimated Effects of U-Bridge Adoption on Political Participation

Note: The upper panel shows estimated direct effects and the lower panel shows estimated indirect effects based on different model specification. Abbreviations "Model (·)" represents regression models, "MN PS" represents "Multinomial propensity score", and "JM PS" represents "joint modeling propensity score" (the proposed model).

models for treatment assignment and network formation. An estimate of average potential outcome and treatment effect is obtained by implementing inverse probability weighting estimators. I investigate its performance and finite sample properties in several Monte Carlo studies and illustrate its applicability in an empirical application on the effect of adoption of a new political communication technology on political participation.

While the proposed method can incorporate different types of treatment variables and networks and complex interactions between them in observational data, it has several limitations too. First, I still make strong assumption on the parametric form of treatment assignment mechanism, network formation process, and exposure mapping. Mis-specification of them may cause bias in estimation. Second, the approach is more computationally intensive than some existing approaches that directly estimate the generalized propensity scores via parametric models, which limits its practical applicability to relatively small datasets. Finally, the proposed method aims at cross-sectional data where the network is random but

static. Addressing the interactions and feedback among treatment assignment, network dynamics and behavioral outcomes is important for causal inference with longitudinal social network data and worth exploring in future research.

References

Aronow, P. M. and C. Samii (2017). Estimating average causal effects under general interference, with application to a social network experiment. *Annals of Applied Statistics 11*(4), 1912–1947. 1, 2, 3, 8, 10, 21

Arpino, B. and A. Mattei (2016). Assessing the causal effects of financial aids to firms in tuscany allowing for interference. *The Annals of Applied Statistics*, 1170–1194. 10

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

Cai, Y. (2022). Linear regression with centrality measures. *arXiv preprint arXiv:2210.10024*. 9

Comola, M. and S. Prina (2021). Treatment effect accounting for network changes. *The Review of Economics and Statistics 103*(3), 597–604. 2

Eubank, N., G. Grossman, M. Platas, J. Rodden, et al. (2021). Viral voting: Social networks and political participation. *Quarterly Journal of Political Science 16*(3), 265–284. 5

Feng, P., X.-H. Zhou, Q.-M. Zou, M.-Y. Fan, and X.-S. Li (2012). Generalized propensity score for estimating the average treatment effect of multiple treatments. *Statistics in medicine 31*(7), 681–697. 31

Ferrali, R., G. Grossman, M. R. Platas, and J. Rodden (2020). It takes a village: Peer effects and externalities in technology adoption. *American Journal of Political Science 64*(3), 536–553. 5

Forastiere, L., E. M. Airoldi, and F. Mealli (2021). Identification and estimation of treatment and interference effects in observational studies on networks. *Journal of the American Statistical Association 116*(534), 901–918. 3, 8, 10, 13, 15, 23, 25, 31

Forastiere, L., F. Mealli, A. Wu, and E. M. Airoldi (2022). Estimating causal effects under network interference with bayesian generalized propensity scores. *Journal of Machine Learning Research 23*(289), 1–61. 25

Frangakis, C. E. and D. B. Rubin (1999). Addressing complications of intention-to-treat analysis in the combined presence of all-or-none treatment-noncompliance and subsequent missing outcomes. *Biometrika 86*(2), 365–379. 19

Goldsmith-Pinkham, P. and G. W. Imbens (2013). Social networks and the identification of peer effects. *Journal of Business & Economic Statistics 31*(3), 253–264. 1, 20

Griffith, A. (2022). Random assignment with non-random peers: A structural approach to counterfactual treatment assessment. *Review of Economics and Statistics*, 1–40.

Grossman, G., M. Humphreys, and G. Sacramone-Lutz (2020). Information technology and political engagement: Mixed evidence from uganda. *The Journal of Politics 82*(4), 1321–1336. 4

Han, X., C.-S. Hsieh, and S. I. Ko (2021). Spatial modeling approach for dynamic network formation and interactions. *Journal of Business & Economic Statistics 39*(1), 120–135. 2

Harrison, G. W. and J. A. List (2004). Field experiments. *Journal of Economic literature 42*(4), 1009–1055. 4

Hirano, K. and G. W. Imbens (2004). The propensity score with continuous treatments. *Applied Bayesian modeling and causal inference from incomplete-data perspectives 226164*, 73–84. 3, 15, 23

Hoff, P. (2021). Additive and multiplicative effects network models. *Statistical Science 36*(1), 34–50.

Horvitz, D. G. and D. J. Thompson (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association 47*(260), 663–685.

Hudgens, M. G. and M. E. Halloran (2008). Toward causal inference with interference. *Journal of the American Statistical Association 103*(482), 832–842. 7

Imbens, G. W. (2000). The role of the propensity score in estimating dose-response functions. *Biometrika 87*(3), 706–710. 2, 16

Jackson, M. O., Z. Lin, and N. N. Yu (2022). Adjusting for peer-influence in propensity scoring when estimating treatment effects. *Available at SSRN 3522256*. 4, 6, 9, 13, 30

Jung, J., R. Shroff, A. Feller, and S. Goel (2020). Bayesian sensitivity analysis for offline policy evaluation. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pp. 64–70.

Kinne, B. J. (2012). Multilateral trade and militarized conflict: Centrality, openness, and asymmetry in the global trade network. *The Journal of Politics 74*(1), 308–322. 9

Kojevnikov, D., V. Marmer, and K. Song (2021). Limit theorems for network dependent random variables. *Journal of Econometrics 222*(2), 882–908. 3, 22

Leung, M. P. and P. Loupos (2022). Unconfoundedness with network interference. *arXiv preprint arXiv:2211.07823*. 2, 3, 4, 10, 13, 17, 22

Manski, C. F. (2013). Identification of treatment response with social interactions. *The Econometrics Journal 16*(1), S1–S23. 2, 10

McCaffrey, D. F., B. A. Griffin, D. Almirall, M. E. Slaughter, R. Ramchand, and L. F. Burgette (2013). A tutorial on propensity score estimation for multiple treatments using generalized boosted models. *Statistics in medicine 32*(19), 3388–3414. 31

Rosenbaum, P. R. and D. B. Rubin (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika 70*(1), 41–55. 15

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology 66*(5), 688. 8

Rubin, D. B. (1986). Which ifs have causal answers; comment on holland (1986). *Journal of the American Statististical Association 81*, 961–962. 8

Sanchez-Becerra, A. (2022). The network propensity score: Spillovers, homophily, and selection into treatment. *arXiv preprint arXiv:2209.14391*. 3, 5, 7, 14, 15, 29, 31

Sävje, F. (2023). Causal inference with misspecified exposure mappings: separating definitions and assumptions. *Biometrika*, asad019. 2, 10

Simmons, B. A. and Z. Elkins (2004). The globalization of liberalization: Policy diffusion in the international political economy. *American political science review 98*(1), 171–189. 1

Sinclair, B., M. McConnell, and D. P. Green (2012). Detecting spillover effects: Design and analysis of multilevel experiments. *American Journal of Political Science 56*(4), 1055–1069. 1

Steinberg, D. A., S. C. Nelson, and C. Nguyen (2018). Does democracy promote capital account liberalization? *Review of International Political Economy 25*(6), 854–883.

Tomz, M., J. L. Goldstein, and D. Rivers (2007). Do we really know that the wto increases trade? comment. *American Economic Review 97*(5), 2005–2018.

Tomz, M., J. Wittenberg, and G. King (2001). Clarify: Software for interpreting and presenting statistical results.

Toulis, P., A. Volfovsky, and E. M. Airoldi (2021). Estimating causal effects when treatments are entangled by network dynamics. 3, 9, 21

Xu, X., S. N. MacEachern, and B. Lu (2023). Bridging the design and modeling of causal inference: A bayesian nonparametric perspective. *Observational Studies 9*(1), 119–124.

SUPPLEMENTARY MATERIAL

A Proofs

Proof for proposition 1:

$$\begin{aligned} & Pr(T_i = t|\boldsymbol{X}, r(i,t;\boldsymbol{X})) \\ =&\mathbb{E}[\mathbb{1}\{g(i,\boldsymbol{D},\boldsymbol{W}) = t\}|\boldsymbol{X}, r(i,t;\boldsymbol{X})] \\ =&\mathbb{E}[\mathbb{1}\{g(i,\boldsymbol{D},\boldsymbol{W}) = t\}|\boldsymbol{X}] \\ =&Pr(T_i = t|\boldsymbol{X}) \\ =&r(i,t;\boldsymbol{X}) \end{aligned} \tag{27}$$

We also have

$$\begin{aligned} & Pr(T_i = t|r(i,t;\boldsymbol{X})) \\ =&\mathbb{E}[\mathbb{1}\{g(i,\boldsymbol{D},\boldsymbol{W}) = t\}|r(i,t;\boldsymbol{X})] \\ =&\mathbb{E}[\mathbb{E}[\mathbb{1}\{g(i,\boldsymbol{D},\boldsymbol{W}) = t\}|\boldsymbol{X}]r(i,t;\boldsymbol{X})] \\ =&\mathbb{E}[r(i,t;\boldsymbol{X})|\boldsymbol{X}] \\ =&r(i,t;\boldsymbol{X}) \end{aligned} \tag{28}$$

Therefore, we have $Pr(T_i = t|\boldsymbol{X}, r(i,t;\boldsymbol{X})) = Pr(T_i = t|r(i,t;\boldsymbol{X}))$.

Proof for proposition 2:

$$\begin{aligned} & Pr(T_i = t|Y_i(\cdot)r(i,t;\boldsymbol{X})) \\ =&\mathbb{E}[\mathbb{1}\{g(i,\boldsymbol{D},\boldsymbol{W}) = t\}|Y_i(\cdot), r(i,t;\boldsymbol{X})] \\ =&\mathbb{E}[\mathbb{E}[\mathbb{1}\{g(i,\boldsymbol{D},\boldsymbol{W}) = t\}|\boldsymbol{X}, Y_i(\cdot), r(i,t;\boldsymbol{X})]Y_i(\cdot), r(i,t;\boldsymbol{X})] \\ =&\mathbb{E}[\mathbb{E}[\mathbb{1}\{g(i,\boldsymbol{D},\boldsymbol{W}) = t\}|\boldsymbol{X}, Y_i(\cdot)]Y_i(\cdot), r(i,t;\boldsymbol{X})] \\ =&\mathbb{E}[\mathbb{E}[\mathbb{1}\{g(i,\boldsymbol{D},\boldsymbol{W}) = t\}|\boldsymbol{X}]Y_i(\cdot), r(i,t;\boldsymbol{X})] \\ =&\mathbb{E}[r(i,t;\boldsymbol{X})|Y_i(\cdot), r(i,t;\boldsymbol{X})] \\ =&r(i,t;\boldsymbol{X}) \end{aligned} \tag{29}$$

Proof for proposition 3:

$$
\begin{aligned}
&\mathbb{E}\{\frac{Y_i \mathbb{1}\{T_i = g(i, \boldsymbol{D}, \boldsymbol{W}) = t\}}{r(i,t;\boldsymbol{X})}\} \\
=&\mathbb{E}\{\mathbb{E}\{\frac{Y_i \mathbb{1}\{T_i = g(i, \boldsymbol{D}, \boldsymbol{W}) = t\}}{r(i,t;\boldsymbol{X})}|\boldsymbol{X}\}\} \\
=&\mathbb{E}\{\mathbb{E}\{\frac{Y_i}{r(i,t;\boldsymbol{X})}|T_i = t, \boldsymbol{X}\} Pr(T_i = t|\boldsymbol{X})\} \\
=&\mathbb{E}\{\mathbb{E}\{\frac{Y_i}{r(i,t;\boldsymbol{X})}|T_i = t, \boldsymbol{X}\} r(i,t;\boldsymbol{X})\} \\
=&\mathbb{E}\{\mathbb{E}\{Y_i|T_i = t, \boldsymbol{X}\}\} \\
=&\mathbb{E}\{\mathbb{E}\{Y_i(t)|\boldsymbol{X}\}\} \\
=&\mathbb{E}\{Y_i(t)\} \\
=&\mu(t)
\end{aligned}
\tag{30}
$$

B MCMC Algorithm for Modeling Unobserved Heterogeneity

Denote $\Xi = (\xi_1, \ldots, \xi_N)$, the parameters to be estimated are $\Theta = \{\beta_d, \beta_w, \alpha_d, \alpha_w, \Xi\}$. The likelihood function is written as:

$$
\begin{aligned}
L(\boldsymbol{D}, \boldsymbol{W}|\Theta) &= \prod_{i=1}^{N} Pr(D_i = 1)^{D_i} (1 - Pr(D_i = 1))^{1-D_i} \prod_{1 \le i < j \le N} Pr(W_{ij} = 1)^{W_{ij}} (1 - Pr(W_{ij}) = 1)^{1-W_{ij}} \\
&= \prod_{i=1}^{N} \Phi(X_i'\beta_d + \alpha_d \xi_i)^{D_i} (1 - \Phi(X_i'\beta_d + \alpha_d \xi_i))^{1-D_i} \\
&\times \prod_{1 \le i < j \le N} \Phi(X_{ij}'\beta_w + \alpha_w |\xi_i - \xi_j|)^{W_{ij}} (1 - \Phi(X_{ij}'\beta_w + \alpha_w |\xi_i - \xi_j|))^{1-W_{ij}}
\end{aligned}
$$

We assign priors to the remaining parameters:

$$
\beta_d \sim N(\beta_d^0, B_0), \quad \beta_w \sim N(\beta_w^0, C_0), \quad \alpha_d \sim N(\alpha_d^0, \sigma_d^2), \quad \alpha_w \sim N(\alpha_w^0, \sigma_w^2)
$$

By Bayes rule, we have

$$
p(\Theta|\boldsymbol{D}, \boldsymbol{W}) \propto L(\boldsymbol{D}, \boldsymbol{W}|\Theta) P(\Theta)
$$

For simulation, we use Gibbs sampler for the conditional posteriors for (β_d, α_d) and (β_w, α_w) given the data and other parameters. Besides, we use Metropolis-Hastings steps (random walk Metropolis) to update each ξ_i separately.

The MCMC steps are summarized as follows:

1. Draw latent outcomes D_i^* and W_{ij}^* given the parameters:

$$D_i^* \sim \begin{cases} TN_{[0,\infty)}(\mu_i^d, 1) & if D_i = 1 \\ TN_{(-\infty,0)}(\mu_i^d, 1) & if D_i = 0 \end{cases}$$

$$W_{ij}^* \sim \begin{cases} TN_{[0,\infty)}(\mu_{ij}^w, 1) & if W_{ij} = 1 \\ TN_{(-\infty,0)}(\mu_{ij}^w, 1) & if W_{ij} = 0 \end{cases}$$

   where $u_i^d = X_i'\beta_d + \alpha_d\xi_i$ and $u_{ij}^w = X_{ij}'\beta_w + \alpha_w|\xi_i - \xi_j|$. $TN_{(a,b)}(\mu, \nu)$ means normal distribution with mean μ and variance ν truncated to the interval (a, b).

2. Jointly draw (β_d, α_d): denote $\tilde{X}_i = (X_i, \xi_i)$

$$\tilde{\beta}_d \sim N(\tilde{\beta}_1, \tilde{B}_1)$$
$$\tilde{B}_1 = [\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{X}} + \tilde{B}_0^{-1}]^{-1}$$
$$\tilde{\beta}_d^1 = \tilde{B}_1[\tilde{\boldsymbol{X}}'\boldsymbol{D}^* + \tilde{B}_0^{-1}\tilde{\beta}_d^0]$$

3. Jointly draw (β_w, α_q): denote $\tilde{X}_{ij} = (X_{ij}, |\xi_i - \xi_j|)$

$$\tilde{\beta}_w \sim N(\tilde{\beta}_w^1, \tilde{C}_1)$$
$$\tilde{C}_1 = [\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{X}} + \tilde{C}_0^{-1}]^{-1}$$
$$\tilde{\beta}_w^1 = \tilde{C}_1[\tilde{\boldsymbol{X}}'\boldsymbol{W}^* + \tilde{C}_0^{-1}\tilde{\beta}_w^0]$$

4. Separately draw ξ_i: We first draw a proposal from

$$\xi_i^* \sim N(\xi_i^{(h-1)}, \sigma_\alpha^2)$$

   and then accept $\xi_i^h = \xi_i^*$ with probability $\min\{1, \psi\}$ where

$$\begin{aligned}\psi = & \frac{\Phi(X_i'\beta_d + \alpha_d\xi_i^*)^{D_i}(1 - \Phi(X_i'\beta_d + \alpha_d\xi_i^*))^{1-D_i}}{\Phi(X_i'\beta_d + \alpha_d\xi_i^{(h-1)})^{D_i}(1 - \Phi(X_i'\beta_d + \alpha_d\xi_i^{(h-1)}))^{1-D_i}} \\ & \times \frac{\prod_{1\le i<j\le N}\Phi(X_{ij}'\beta_w + \alpha_w|\xi_i^* - \xi_j|)^{W_{ij}}(1 - \Phi(X_{ij}'\beta_w + \alpha_w|\xi_i^* - \xi_j|))^{1-W_{ij}}}{\prod_{1\le i<j\le N}\Phi(X_{ij}'\beta_w + \alpha_w|\xi_i^{(h-1)} - \xi_j|)^{W_{ij}}(1 - \Phi(X_{ij}'\beta_w + \alpha_w|\xi_i^{(h-1)} - \xi_j|))^{1-W_{ij}}} \\ & \times \frac{\phi(\xi_i^*)}{\phi(\xi_i^{(h-1)})}\end{aligned}$$

C Additional results for Monte Carlo Studies

To investigate whether the proposed method is robust to heterogeneous treatment effects (HTE), I consider an alternative data generating process for the potential outcome:

$$Y_i(d,z) = 1 + X_i + d + 0.5z + dz + X_i dz + \varepsilon_i, \quad \varepsilon_i \sim N(0,1) \tag{31}$$

Therefore, the individual level treatment effect $Y_i(d,z) - Y_i(\tilde{d},\tilde{z})$ depends on the covariate X_i. Results for both treatment effects are summarized in Table 5 and Table 6. I find that the results are quite similar to the case of constant treatment effect. Therefore, the proposed method performs well even when treatment effects are heterogeneous.

Table 5: Finite Sample Properties (HTE): $\tau((1,1),(0,1))$

N_c	Bias	RMSE	Coverage Rate	SE - SD
20	0.048	1.213	0.972	0.082
50	−0.003	0.726	0.978	0.119
100	0.010	0.514	0.982	0.082
500	−0.001	0.223	0.981	0.048
1000	0.002	0.164	0.983	0.030

Table 6: Finite Sample Properties (HTE): $\tau((1,1),(1,0))$

N_c	Bias	RMSE	Coverage Rate	SE - SD
20	0.037	1.213	0.947	0.079
50	0.001	0.732	0.970	0.091
100	0.009	0.536	0.964	0.047
500	−0.007	0.235	0.973	0.026
1000	0.003	0.163	0.975	0.022

D Additional Results and Plots for the Empirical Application

This section includes additional results and plots for the empirical application. Figure 3 and Figure 4 display network structures for the rest 15 villages in the data sample. Residents who adopted the U-Bridge program are highlighted in blue and Residents who occupy

a formal leader role within village are highlighted in red. Figure 5 and Figure 6 show estimated coefficients with 95% confidence intervals in the network formation models across the 16 villages under study.

Table 7: Estimation Results for Treatment Assignment Mechanism

Treatment Variable	Individual Adoption of U-Bridge Program (D_i)	#Linked Adopters ($\sum_j W_{ij}D_j$)
	JM PS	Net PS
Age	-0.002	-0.007***
	(0.010)	(0.001)
Female	-1.524***	-0.072
	(0.347)	(0.061)
Income	0.018	0.001
	(0.073)	(0.010)
Education	1.911***	0.218***
	(0.191)	(0.051)
Leader	0.775***	-0.130***
	(0.244)	(0.045)
Care Community	-1.030*	-0.214**
	(0.528)	(0.094)
Has Phone	1.230***	0.054
	(0.352)	(0.050)
Contextual Age	-0.029***	
	(0.015)	
Contextual Female	1.439***	
	(0.466)	
Contextual Income	-0.419	
	(0.287)	
Contextual Education	1.177**	
	(0.471)	
Contextual Leader	-0.205	
	(0.533)	
Contextual Care Community	-4.611**	
	(1.798)	
Contextual Has Phone	0.098	
	(0.570)	
Village Fixed Effects	✓	✓

Notes: Robust standard errors clustered at village level are in the parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.
Contextual variables mean average value of villagers with common links, i.e., $\sum_j W_{ij}X_j / \sum_j W_{ij}$ for each villager i.

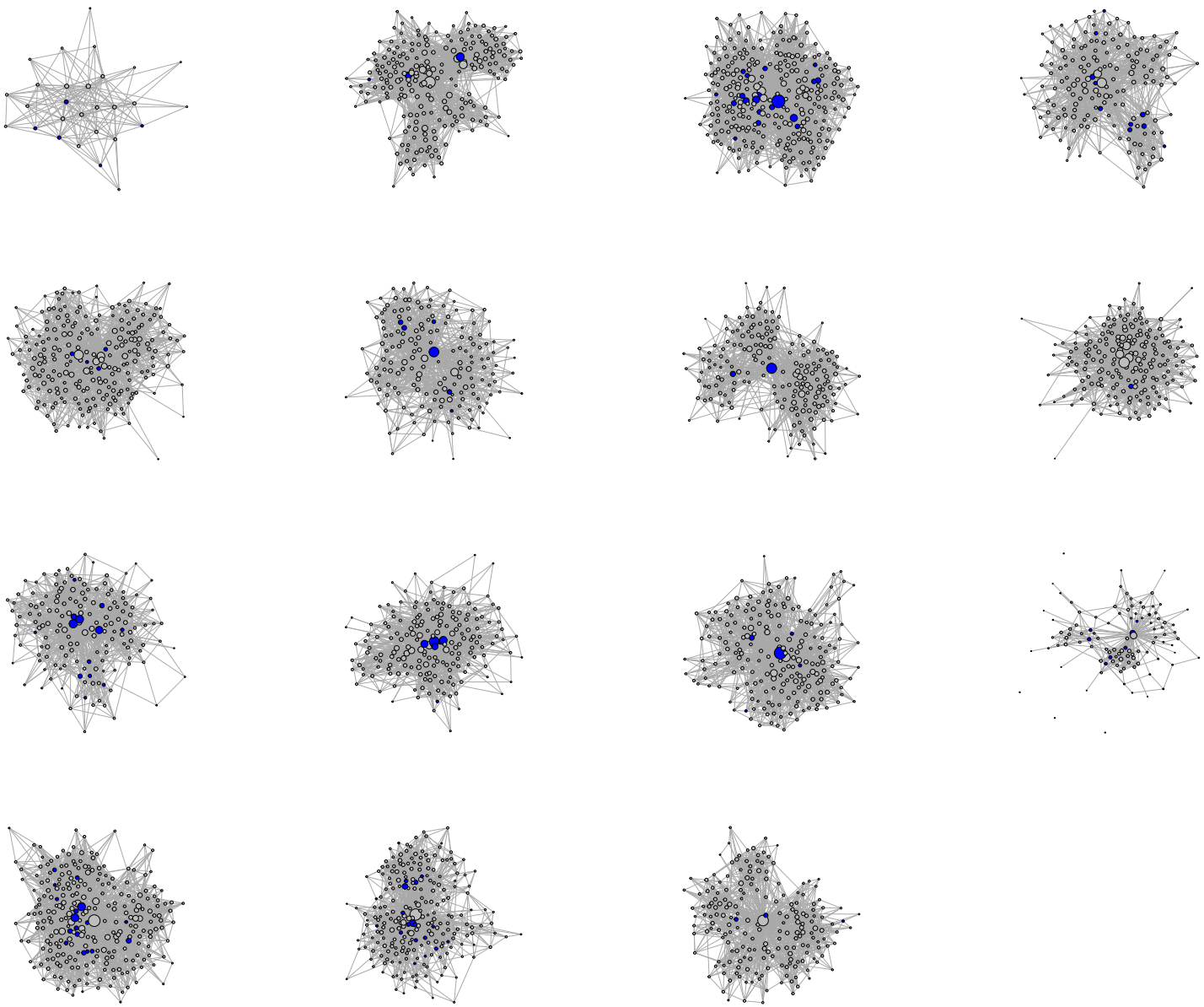

Figure 3: Additional Network Visualizations for Villages (Treatment)

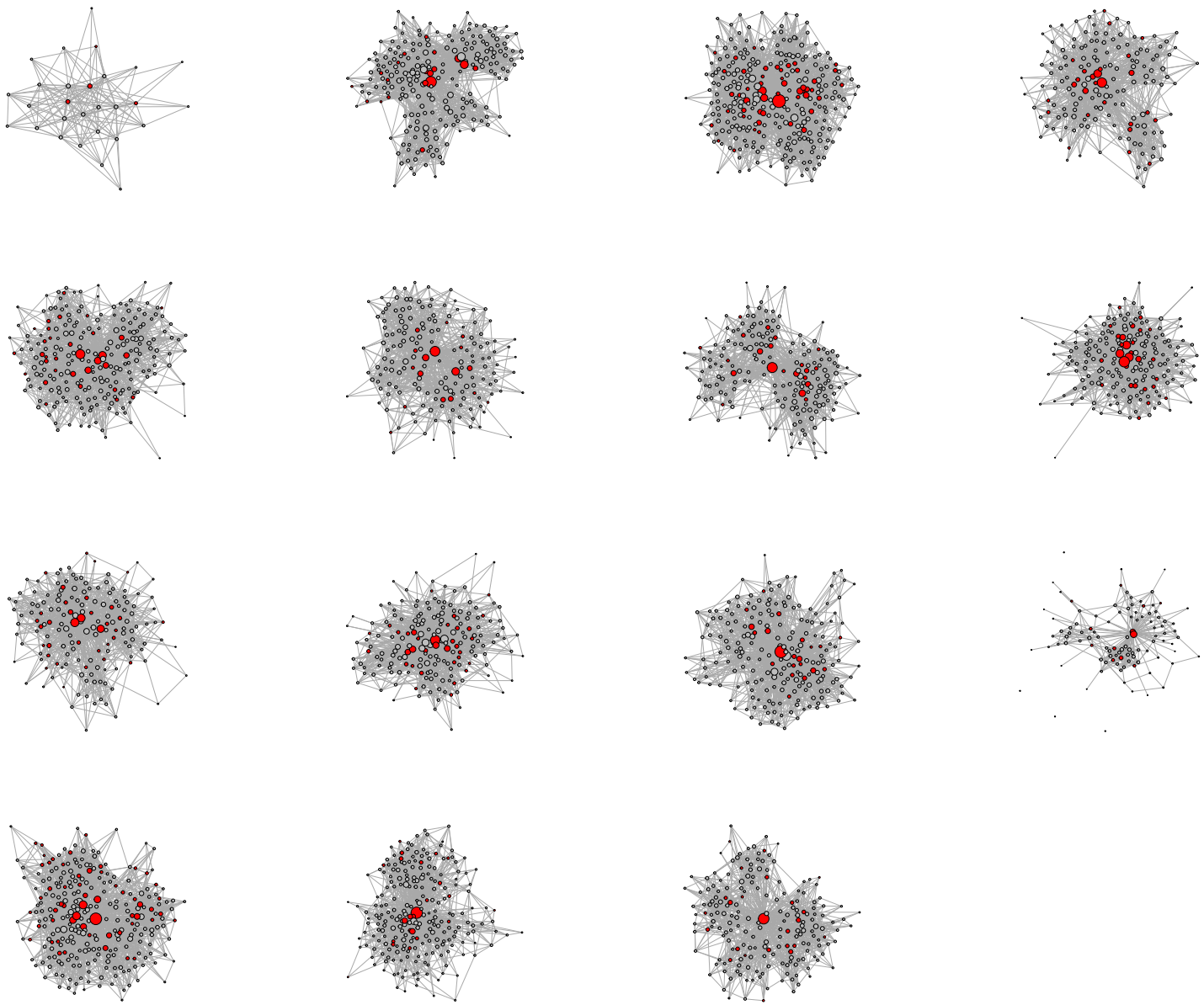

Figure 4: Additional Network Visualizations for Villages (Leadership)

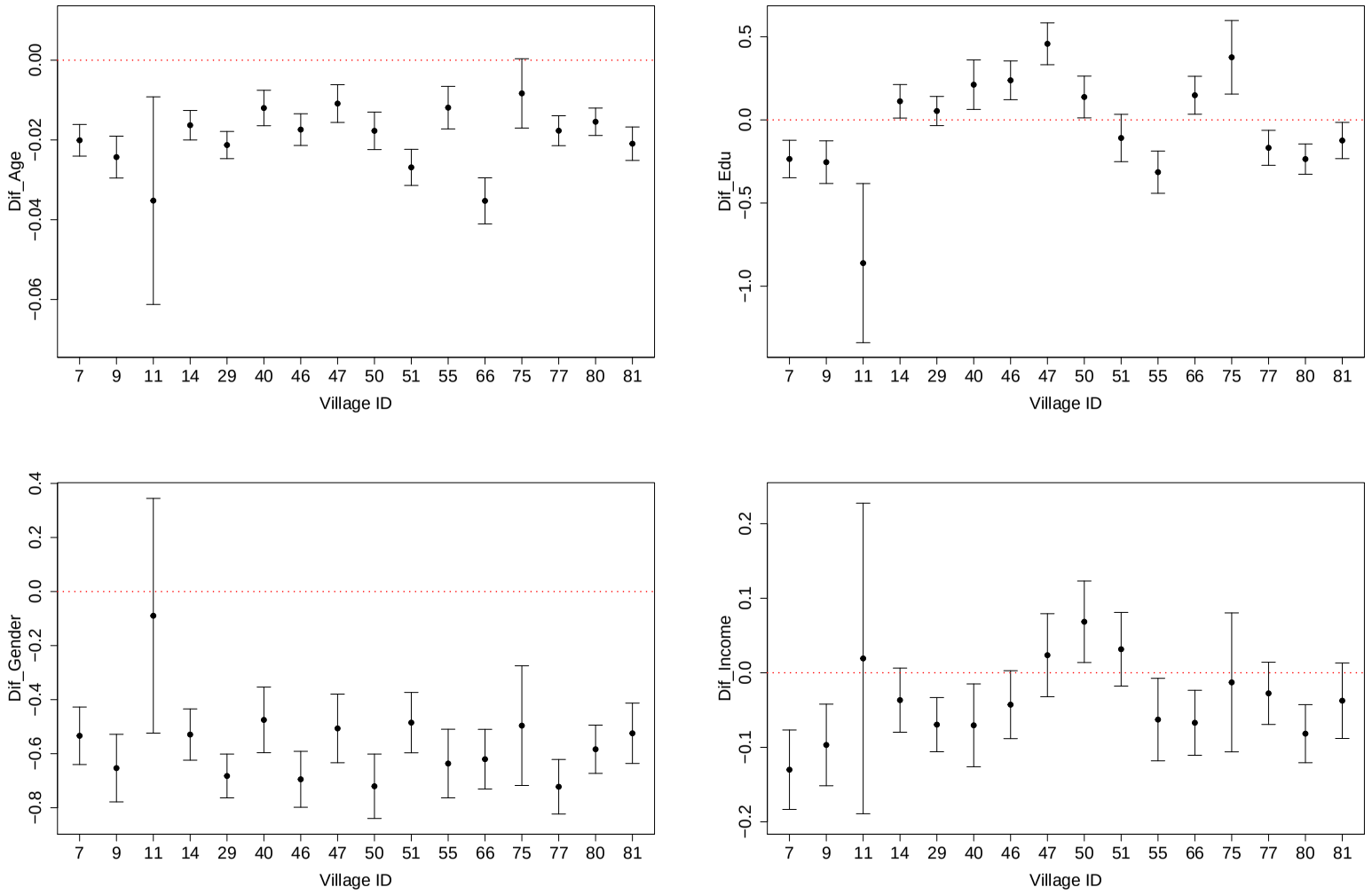

Figure 5: Heterogeneous Coefficients in Network Formation Models

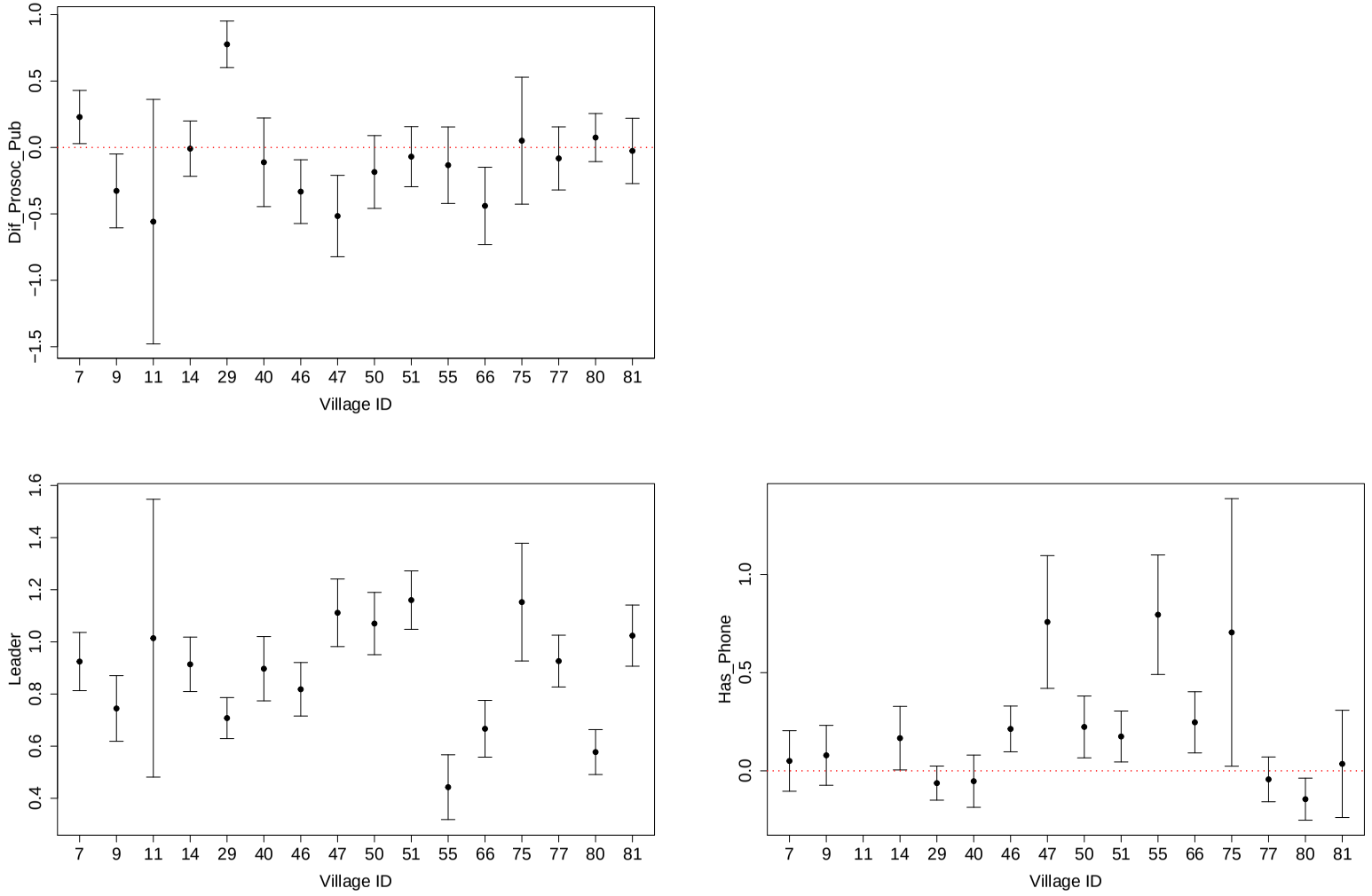

Figure 6: Heterogeneous Coefficients in Network Formation Models Cont'