

# General Additive Network Effect Models

TRANG BUI, STEFAN H. STEINER, AND NATHANIEL T. STEVENS\*

---

## Abstract

In the interest of business innovation, social network companies often carry out experiments to test product changes and new ideas. In such experiments, users are typically assigned to one of two experimental conditions with some outcome of interest observed and compared. In this setting, the outcome of one user may be influenced by not only the condition to which they are assigned but also the conditions of other users via their network connections. This challenges classical experimental design and analysis methodologies and requires specialized methods. We introduce the general additive network effect (GANE) model, which encompasses many existing outcome models in the literature under a unified model-based framework. The model is both interpretable and flexible in modeling the treatment effect as well as the network influence. We show that (quasi) maximum likelihood estimators are consistent and asymptotically normal for a family of model specifications. Quantities of interest such as the global treatment effect are defined and expressed as functions of the GANE model parameters, and hence inference can be carried out using likelihood theory. We further propose the “power-degree” (POW-DEG) specification of the GANE model. The performance of POW-DEG and other specifications of the GANE model are investigated via simulations. Under model misspecification, the POW-DEG specification appears to work well. Finally, we study the characteristics of good experimental designs for the POW-DEG specification. We find that graph-cluster randomization and balanced designs are not necessarily optimal for precise estimation of the global treatment effect, indicating the need for alternative design strategies.

KEYWORDS AND PHRASES: Design of experiments, A/B testing, Social networks, SUTVA, Network effect modeling.

---

## 1. INTRODUCTION

### 1.1 Context

Network data encode not only information about individuals but also the connections among them. Such data exist in many settings: students who are friends with one another [4, 33]; intersecting streets [27]; interconnected brain regions [30]; and social network users who “follow” or “friend” one another [47]. Network data can be represented by a graph where nodes represent individuals and an edge between two nodes indicates that there is a connection between them. For example, Figure 1 visualizes a Caltech Facebook network, which was collected in September 2005 [42, 43]. The version of the data used in this paper contains information about 770 users and was retrieved from the Network Data Repository [38]. In Figure 1, the circles (nodes) represent Caltech Facebook users and a line between two nodes (an edge) signifies a Facebook friendship between two users.

The dependency structure in network data poses distinctive challenges and opportunities and thus motivates specialized theory and methodology. In this paper, we focus on network experimentation, a topic that has gained recent interest given the widespread experimentation happening in social network companies like LinkedIn, Facebook, and Twitter [17]. These businesses regularly develop updates and new

features to their platforms. To decide whether a new feature or update should be deployed for all users, these companies conduct controlled experiments on their user networks to test a new feature and evaluate its causal impact [34, 47].

Current literature on network experimentation generally focuses on the setting where the experiment is conducted on a given network of a fixed number of individuals. Connections among individuals on the network are further assumed to be known and unchanged over the duration of the experiment. In addition, the experiment usually involves only two experimental conditions: the proposed new condition (i.e., the *treatment*) and the existing or baseline condition (i.e., the *control*). This type of experiment is also known as an A/B test, since two conditions A (control) and B (treatment) are compared to determine which one performs better in terms of some business metric. For example, the outcome of interest may be a measure of engagement like time spent on the platform, and the goal of the experiment may be to determine which condition maximizes this metric. In general, the experimental outcome is measured on experimental units, who are assigned to the experimental conditions. In the social network setting, the experimental units are often the platform users.

There are two stages in running an experiment: designing the experiment and analyzing the results. In the design stage, the experimenter determines the treatment assignment for the experimental units. Then in the analysis stage,

---

\*Corresponding author.

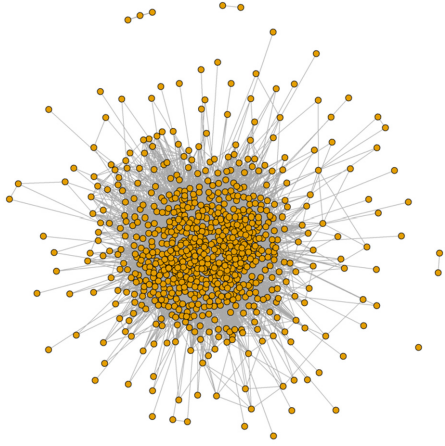


Figure 1: Network of Facebook connections among 770 Caltech Facebook users. The network was plotted using the `igraph` package with the Fruchterman-Reingold layout [14] generated from the `qgraph` package.

data collected from the experiment are used to estimate the effect of the treatment on the outcome of interest. A popular estimand in the literature is the *global treatment effect* (sometimes alternatively referred to as the average treatment effect), which is the average difference in experimental outcomes when everyone in the population is assigned to the treatment versus when everyone is assigned to the control [9, 12]. A significantly positive global treatment effect (in the case where higher outcome values are desired) would motivate a business decision to deploy the treatment instead of the control to all users and vice versa. As such, an accurate estimate of the global treatment effect is an important objective to consider when developing experimental design and analysis methodologies. Another relevant concern in network experimentation is the estimation of the *network effect*: the influence of the network on the outcome of interest.

Classical experimental design and analysis methodologies treat individuals (or experimental units) as independent entities, in the sense that the potential outcome of each unit is assumed to be independent of the experimental conditions the *other* units are assigned. This is known as the Stable Unit Treatment Value Assumption (SUTVA) [11]. However, when experimental units are connected on a network, SUTVA is often violated. For example, consider an experiment where the treatment is designed to increase the time a user spends on the social network. Suppose Connie is assigned to control and her friend, Tracy, is assigned to treatment. Due to the treatment, suppose Tracy spends more time on the network. Connie may observe Tracy's increased activity and interact with Tracy more, resulting in an increased amount of time spent on the platform despite being in the control group. Thus, the fact that Tracy is assigned to the treatment affects the potential outcome of Connie.

As such, SUTVA no longer holds and the potential outcome of an individual depends not only on their own treatment assignment but also the treatment assignment of other individuals in the network. Consequently, in a network setting, the global treatment effect depends not only on the experimental conditions but also on the structure of the network. Therefore, specialized methods are required for the design and analysis of network experiments.

## 1.2 Related Work

There are two major classes of approaches to the design and analysis of experiments on networks. The first base their methods on the exposure framework [2, 12, 16], in which experimental units (nodes) are classified into exposure groups where group members have the same level of exposure to the treatment. In a network experiment, even when two units are assigned to the same experimental condition, they may be exposed to different levels of the treatment (e.g., based on the treatment assignment of those around them) and thus belong to different exposure groups. For example, units  $i$  and  $j$  may belong to different exposure groups if all neighbors of unit  $i$  are assigned to treatment while all neighbors of unit  $j$  are assigned to control even when units  $i$  and  $j$  have the same treatment status. The literature often defines exposure groups based on local treatment assignment patterns of neighbors [2, 12, 39]. Based on the defined exposures, various causal effects, such as the global treatment effect and others [20, 41], can be defined as contrasts of the exposure groups' potential outcomes [2]. Accordingly, estimators can be derived using inverse probability weighting [2, 12]. As an example, Gui et al. [16] define the *treatment exposure* as containing units such that (i) they are assigned to the treatment and (ii) the majority of their neighbors are also assigned to the treatment. Similarly, units are said to belong to the *control exposure* if (i) they are assigned to the control and (ii) the majority of their neighbors are also assigned to the control. The potential outcome of the treatment exposure group serves as a proxy for the outcome observed if the whole network is assigned to treatment. Similarly, the control exposure group is assumed to behave as if the whole network is assigned to control. The global treatment effect can then be estimated by taking the (inverse probability weighted) difference between the average outcome of the treatment exposure group and that of the control exposure group.

To achieve a precise estimator of the global treatment effect, each exposure group needs to contain a large number of units. In other words, the exposure-based analysis strategy requires that experimental units are mostly surrounded by neighbors assigned to the same experimental conditions as themselves. This motivates a popular design strategy in the network experimentation literature based on *graph-cluster randomization* [8, 16, 23, 44]. Graph-cluster randomization designs first partition the network into clusters, i.e., subgraphs of nodes that are densely connected

within each cluster and sparsely connected between clusters. These clusters are then randomly assigned to treatment or control and all units in each cluster are assigned to the same experimental condition. This ensures that most units share the same experimental condition as their neighbors. Eckles et al. [12] illustrate via simulations that graph-cluster randomization, together with inverse-probability-weighted estimators, reduces bias in estimating the global treatment effect in experiments on networks.

The second class of approaches is based on assuming a statistical model for the experimental outcomes. There is a rich literature on network regression models for inference and prediction problems [1, 7, 22, 29, 31]. However, for network experiments, models need to characterize and quantify the influence of the treatment assignment to the outcome of interest, compared to the control, while taking into account the network influence. Parker et al. [35] and Koutra et al. [26] posit that an experimental condition can affect the outcome of a unit in two ways: (i) through the unit’s own treatment assignment and/or (ii) through the treatment assignment of the unit’s neighbors. These effects are assumed to be fixed and additive and then modeled via an ordinary linear regression model. Basse and Airolidi [5] and Pokhilko et al. [36] instead suggest that the experimental conditions only assert influence via the nodes’ own treatment assignment. Network interference is then modeled via the correlation of random errors based on the network structure, in a similar fashion to the conditional autoregressive (CAR) model from the spatial statistics literature [3, 10].

In the model-based framework, once the model is formalized, experimenters can select design criteria to reduce, for example, the variance of the model parameters and/or functions of them. Designs that optimize the posed design criteria can be found using exhaustive search [26, 35], random search [35] or other optimal design search methods [4, 36]. In the analysis stage, model parameters will be estimated by fitting the model to the observed data using methods such as least squares or maximum likelihood.

Of the two classes of approaches, the exposure framework seems more popular in the network experimentation literature thanks to its focus on the global treatment effect. When the main goal of the experiment is to accurately estimate the global treatment effect, the exposure framework only requires the experimenters to define the treatment and control exposures. For instance, in the Gui et al. [16] method the experimenter must determine what percentage of a unit’s neighbors must be assigned to treatment for it to be classified into the treatment exposure group. After that, graph-cluster randomization can be applied as a general design strategy to reduce the bias of the global treatment effect estimate. Nevertheless, there exist pitfalls associated with this class of approaches. Chin [9] points out that social networks are usually locally dense, making it hard for algorithms to partition the network into separate clusters. In addition, experimental outcomes on units not classified into the treatment or the control exposures will be wastefully discarded

in the analysis stage [16], resulting in a low effective sample size [39].

On the contrary, if the generating model for the experimental outcome is correctly specified, the model-based framework will have several advantages over the exposure framework. First, data from all units participating in the experiment can be utilized when fitting the model in the analysis stage. Second, additional inference beyond the global treatment effect can be achieved via model parameters and functions of them. Finally, design selection can be calibrated according to the experimenters’ specific interests, which may sometimes be something other than the global treatment effect.

### 1.3 Our Contribution

In this paper, we attempt to unify many of the model-based approaches by introducing the general additive network effect (GANE) model, in which the experimental outcome of a unit is modeled as an additive function of the effect of its own treatment assignment and the effects that come from other treated or controlled nodes in the network. The model not only encompasses (as special cases) many network experiment models from the literature, but it also flexibly extends the manner in which network effects are modeled. In particular, the network influence from the treatment and control groups can be modeled differently in terms of both size and functional form. Moreover, the model is specified such that its parameters are interpretable. We show that (quasi) maximum likelihood estimation is possible for a family of model specifications. The resulting (quasi) maximum likelihood estimators are then proven to be consistent and asymptotically normal, which facilitates familiar likelihood-based inference.

Existing work in the model-based direction concentrates on inference for the model parameters themselves, however, in many cases, the global treatment effect is the main quantity of interest. To expand the utility of the model-based framework, we illustrate how quantities such as the global treatment effect can be expressed as functions of the GANE model parameters, which permits inference via the delta method [45]. Similarly, via the model parameters, the experimenters can also build design criteria and select experimental designs that optimize these criteria. Overall, through the GANE model, we provide a generalized model-based framework for the design and analysis of experiments on networks.

We further propose a particular specification of the GANE model, the power-degree (POW-DEG) specification, where the network effects are modeled as a nonlinear function of the number of neighbors assigned to treatment or control. Using the power as an additional parameter, the POW-DEG model specification gains additional flexibility in capturing the network effect compared to the linear counterpart that has been proposed in the literature [35]. We investigate inferential properties of the POW-DEG specification as well as other existing specifications in the literature via simulations on real-life networks. We find that the

POW-DEG specification yields good estimates of the true global treatment effect, even under model misspecification. By examining designs in the context of the POW-DEG specification, we also find that cluster randomization and balanced treatment assignment are not necessarily optimal for this model specification and other design strategies should therefore be explored in the future.

As with any model-based approach, the GANE model framework faces the challenge of model selection. When the GANE model is estimated via maximum likelihood, experimenters can use popular model selection criteria such as AIC, BIC, etc. to choose the model specification that best fits the data. However, in the design phase, before data have been collected, experimenters must use their domain knowledge to select a suitable model specification, with which they can build design criteria and select a design. In the absence of information concerning which model specification is appropriate, we suggest the use of the POW-DEG specification because of its flexibility and good performance in terms of global treatment effect estimation and inference. Nevertheless, model selection for both the design and analysis of experiments on networks remains an open area for future study.

The rest of the paper is structured as follows. In Section 2, we introduce the general additive network effect (GANE) model framework. We present the theoretical results for maximum likelihood estimation in Section 3. In Section 4 we report the results of several simulation studies used to investigate the design and inferential properties of a few particular specifications of the GANE model. In Section 5 we conclude with a summary and discussion of several extensions to this work.

## 2. GENERAL ADDITIVE NETWORK EFFECT MODEL

### 2.1 Modeling Framework

Suppose that the experiment is on a network of  $n$  experimental units. Let  $\mathcal{G} = (\mathcal{V}, \mathcal{E})$  denote the network where  $\mathcal{V}$  denotes the set of experimental units (i.e., nodes) and  $\mathcal{E}$  denotes the set of connections among experimental units (i.e., edges). As in most of the network experiment literature, we assume that  $\mathcal{G}$  is known and fixed throughout the experiment. Furthermore,  $\mathcal{G}$  is assumed to be undirected and simple, and so  $\mathcal{G}$  can be represented by an  $n \times n$  adjacency matrix  $\mathbf{A}$  where for  $i, j = 1, 2, \dots, n$

$$A_{ij} = A_{ji} = \begin{cases} 1 & \text{if units } i \neq j \text{ are connected} \\ 0 & \text{if } i = j \text{ or if units } i \neq j \text{ are not connected} \end{cases}$$

If  $A_{ij} = 1$ , unit  $i$  and  $j$  are said to be *neighbors*. The number of neighbors of unit  $i$ , i.e., the sum of the  $i$ th row (or  $i$ th column) in the adjacency matrix, is called the *degree* of unit  $i$ , which we denote by  $k_i$ . We assume the treatment vs. control setting and let  $Z_i$  be the binary indicator denoting

the experimental condition of unit  $i$ , where  $Z_i = 1$  if unit  $i$  is assigned to treatment and  $Z_i = 0$  if unit  $i$  is assigned to control. Let  $Y_i$  denote the experimental outcome of unit  $i$  and further let  $\mathbf{D} = (\mathbf{A}, \mathbf{Z}, \mathbf{Y}, \mathbf{X})$  denote the experimental data where  $\mathbf{Z}$  and  $\mathbf{Y}$  are the treatment assignment and outcome vectors respectively,  $\mathbf{A}$  is the adjacency matrix defined above, and  $\mathbf{X}$  is a possible  $n \times p$  matrix containing the units' covariates.

Parker et al. [35] introduce the linear network effect (LNE) model

$$Y_i = \mu + \tau Z_i + \gamma_T \sum_{j=1}^n A_{ij} Z_j + \gamma_C \sum_{j=1}^n A_{ij} (1 - Z_j) + \epsilon_i, \quad (2.1)$$

where  $\epsilon_i \sim \mathcal{N}(0, 1)$  independently for all  $i = 1, 2, \dots, n$ . The model assumes that when a unit is assigned to the treatment group, the unit itself will experience an effect of size  $\tau$  while exerting an effect of size  $\gamma_T$  on each of its neighbors. If a unit is assigned to the control, it will exert an effect of size  $\gamma_C$  on each of its neighbors. The outcome of a unit  $i$  is then the sum of the baseline  $\mu$ , the effect of the treatment assignment, and the sum of the effects from its neighbors plus some random error  $\epsilon_i$ . In other words, the experimental outcome of unit  $i$  is a linear combination of which experimental condition it belongs to and the numbers of its neighbors that are assigned to treatment (*treatment degree*) and control (*control degree*). The model offers a straightforward parameter interpretation:  $\mu$  is the expected baseline outcome,  $\tau$  is the effect of the treatment assignment, and  $\gamma_T$  and  $\gamma_C$  quantify the effect of the neighbors' treatment assignments. The model is linear in the parameters and can therefore be fit using ordinary least squares.

However, on a network, the outcome of a unit may not only be influenced by the *number* of neighbors it has. It may also be influenced by the *outcomes* of its neighbors [1, 40], or by neighbors of neighbors, etc. To enable experimenters to model the experimental outcomes more flexibly, we propose a more general model for the design and analysis of experiments on networks, which we call the *general additive network effect* (GANE) model:

$$Y_i = \mu + \tau Z_i + f_{T,i}(\mathbf{D}, \boldsymbol{\eta}) + f_{C,i}(\mathbf{D}, \boldsymbol{\eta}) + \epsilon_i, \quad (2.2)$$

where  $f_{T,i}$  models the effect of treated units on unit  $i$  and  $f_{C,i}$  models the effect of controlled units. Hence,  $f_T$  and  $f_C$  are functions of the network (via the adjacency matrix  $\mathbf{A}$ ) and the treatment assignment vector  $\mathbf{Z}$ . They may also be functions of the outcome vector  $\mathbf{Y}$ , a possible covariate matrix  $\mathbf{X}$ , and a parameter vector  $\boldsymbol{\eta}$ . By separating the network effect based on the source (treatment or control), the GANE model not only allows the network effect to come from both treated and controlled neighbors, but the network effect can also be of different forms and sizes, i.e.,  $f_T$  and  $f_C$  may have different forms and/or may depend on different parameters (which are all contained in the vector  $\boldsymbol{\eta}$ ). The

functional forms of  $f_T$  and  $f_C$  may be determined by the experimenters' domain knowledge or other considerations.

The GANE model has a structure similar to that of the LNE model (2.1) by Parker et al. [35], in which  $\mu$  parametrizes the baseline outcome,  $\tau$  parametrizes the effect of the treatment assignment to a unit and  $f_T$  and  $f_C$  respectively model the influence the unit receives from other treated and controlled units. Therefore, it retains the clear interpretability of Model (2.1) while increasing the flexibility with which the influence of other nodes is modeled.

## 2.2 Model Specifications

The GANE model encompasses several existing models in the literature, a few of which are identified below.

**The Linear Network Effect (LNE) Model:** Model (2.1) is a GANE model with  $f_T$  modeling the treatment degree and  $f_C$  modeling the control degree:

$$\begin{aligned} f_{T,i}(\mathbf{D}, \boldsymbol{\eta}) &= \gamma_T \sum_{j=1}^n A_{ij} Z_j \\ f_{C,i}(\mathbf{D}, \boldsymbol{\eta}) &= \gamma_C \sum_{j=1}^n A_{ij} (1 - Z_j), \end{aligned} \quad (2.3)$$

where  $\boldsymbol{\eta} = (\gamma_T, \gamma_C)^\top$ .

**The Local Aggregate (LAG) Model:** Advani and Malde [1] propose that the outcome of a unit may be influenced by the sum of its neighbors' outcomes in cases such as (i) a person's criminal behavior might depend on the number of crimes committed by their friends; (ii) the number of items purchased by a customer may depend on the number of items purchased by their friends; and (iii) the number of hours a student spends studying may depend on the number of hours their friends spend studying. Accordingly, under the GANE model, functions  $f_T$  and  $f_C$  can respectively model the sum of the treated and controlled neighbors' outcomes:

$$\begin{aligned} f_{T,i}(\mathbf{D}, \boldsymbol{\eta}) &= \rho_T \sum_{j=1}^n A_{ij} Z_j Y_j, \\ f_{C,i}(\mathbf{D}, \boldsymbol{\eta}) &= \rho_C \sum_{j=1}^n A_{ij} (1 - Z_j) Y_j, \end{aligned} \quad (2.4)$$

where  $\boldsymbol{\eta} = (\rho_T, \rho_C)^\top$ .

**The Homophily (HOM) Model:** In the social economics literature, a unit's outcome is often modeled as a function of the average outcome of its neighbors [7, 31]. This is often referred to as the *homophily effect* [40] which is summarized by the conjecture that "you are the average of the people around you", and a result of people's desire to conform to social norms [1]. Gui et al. [16] formalize a model with homophily effect for the analysis of a network A/B test as

follows:

$$Y_i = \mu + \tau Z_i + \gamma \sum_{j=1}^n A_{ij} Z_j + \rho \frac{1}{k_i} \sum_{j=1}^n A_{ij} Y_j + \epsilon_i. \quad (2.5)$$

That is, besides the homophily effect, each treated node will exert an effect of size  $\gamma$  on each of its neighbors. This model can be reparameterized and written in the GANE model framework with

$$\begin{aligned} f_{T,i}(\mathbf{D}, \boldsymbol{\eta}) &= \gamma_T \sum_{j=1}^n A_{ij} Z_j + \rho_T \frac{1}{k_i} \sum_{j=1}^n A_{ij} Z_j Y_j, \\ f_{C,i}(\mathbf{D}, \boldsymbol{\eta}) &= \rho_C \frac{1}{k_i} \sum_{j=1}^n A_{ij} (1 - Z_j) Y_j. \end{aligned}$$

In this case,  $\boldsymbol{\eta} = (\rho_T, \rho_C, \gamma_T)^\top$  with constraints  $\rho_T = \rho_C = \rho$  and  $\gamma_T = \gamma$ .

The above are existing models that have been proposed and discussed in the literature. We next propose a new specification under the GANE framework.

**The Power-Degree (POW-DEG) Model:** Let us continue with the example in Section 1.1 where the outcome of interest is the amount of time that users spend on a social network platform. Recall, since Tracy is assigned to treatment, Connie (in the control group) may spend more time on the platform simply because she is Tracy's friend. However, Connie's time increase due to her first treated friend may not necessarily be equal to the time increase due to her 100th treated friend. In particular, as the number of treated friends grows, the effect of an additional treated friend is likely to decrease. This intuition is similar to the law of diminishing marginal utility in economics [15].

When the number of neighbors is low, for example, in agricultural settings where units are plots placed on a lattice, it may be reasonable to model the network effects homogeneously as in the LNE specification (2.3). However, in the social network setting where a user can have hundreds to thousands of friends, this linearity assumption seems less likely. As a result, we suggest modeling the network effect as a *non-linear* function of the number of neighbors (i.e., the degree). In particular, we propose the following GANE model specification with

$$\begin{aligned} f_{T,i}(\mathbf{D}, \boldsymbol{\eta}) &= \gamma_T \left( \sum_{j=1}^n A_{ij} Z_j \right)^\lambda, \\ f_{C,i}(\mathbf{D}, \boldsymbol{\eta}) &= \gamma_C \left( \sum_{j=1}^n A_{ij} (1 - Z_j) \right)^\lambda, \end{aligned} \quad (2.6)$$

where  $\boldsymbol{\eta} = (\gamma_T, \gamma_C, \lambda)^\top$ . We call this the power-degree (POW-DEG) specification because the network effects are modeled as powers of the treatment and control degrees.

The power parameter  $\lambda$  serves to temper the growth of network effects as the treatment and control degrees increase, and so we expect that  $0 < \lambda < 1$ . However, in the interest of ample flexibility, we do not make this assumption. We allow for the possibility that  $\lambda > 1$  and also the possibility of the LNE specification (2.3) arising as a special case when  $\lambda = 1$ .

Which estimation method is appropriate for fitting the GANE model depends on the specification of the functions  $f_T$  and  $f_C$ . For instance, estimation and inference for the LNE specification (2.3) can be performed using ordinary least squares. However, when  $f_T$  and  $f_C$  are more complicated such as in specifications (2.4) or (2.5), the GANE model becomes (spatially) autoregressive, in which case maximum likelihood estimation is required.

### 2.3 Quantities of Interest

As discussed in Section 1.2, compared to the exposure framework, one of the advantages of the model-based framework is that more quantities can be easily estimated, assuming they can be expressed as functions of the model parameters. Here, we review several such quantities that may be of interest to experimenters.

**Global treatment effect (GTE):** As noted already, an important quantity for business decision-making is the *global treatment effect* (GTE), which is defined as the difference in average outcomes when everyone in the network is assigned to the treatment versus when everyone is assigned to the control. Note that GTE measures the treatment effect at the *global* level, instead of the individual level, taking into account the structure of the network and possible network effects. In the GANE framework, this quantity can be expressed as

$$\begin{aligned} \text{GTE} &= \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n Y_i \mid \mathbf{Z} = \mathbf{1}_n \right] - \mathbb{E} \left[ \frac{1}{n} \sum_{i=1}^n Y_i \mid \mathbf{Z} = \mathbf{0}_n \right] \\ &= \tau + \frac{1}{n} \sum_{i=1}^n \mathbb{E}[f_{T,i}(\mathbf{D}_{\mathbf{Z}=\mathbf{1}_n}, \boldsymbol{\eta})] \\ &\quad - \frac{1}{n} \sum_{i=1}^n \mathbb{E}[f_{C,i}(\mathbf{D}_{\mathbf{Z}=\mathbf{0}_n}, \boldsymbol{\eta})] \end{aligned} \tag{2.7}$$

where the subscripts on  $\mathbf{D}$  indicate that the functions are evaluated when all the experimental units are assigned to treatment or control, respectively. Note that the expectations in the second equivalence are only necessary when  $f_T$  and/or  $f_C$  are functions of the outcome  $\mathbf{Y}$ . As shown, the GTE can be expressed as a function of the model parameters. Moreover, GTE can be decomposed into two components, the direct and indirect treatment effects, which aid interpretation.

**Direct treatment effect (DTE):** Traditionally, the direct treatment effect with respect to a treatment assignment vector  $\mathbf{Z}$  is defined as the average difference in expected outcomes when the treatment assignment of one unit changes

while the others remain the same, i.e.,

$$\text{DTE}(\mathbf{Z}) = \frac{1}{n} \sum_{i=1}^n \left( \mathbb{E}[Y_i | Z_i = 1, \mathbf{Z}_{-i}] - \mathbb{E}[Y_i | Z_i = 0, \mathbf{Z}_{-i}] \right),$$

where  $\mathbf{Z}_{-i}$  denotes vector  $\mathbf{Z}$  without the  $i$ th element. This definition is difficult to interpret and calculate, especially when the network effect functions are complicated. So instead, we define the direct treatment effect as the expected difference in outcomes when a node is assigned to the treatment versus when a node is assigned to control, keeping the network effects fixed. In the GANE model framework, the direct treatment effect is simply

$$\text{DTE} = \tau. \tag{2.8}$$

Our definition is clear, easy to interpret, easy to calculate, and does not depend on any specific treatment assignment vector. Hence, it can be used across all specifications of the GANE model.

**Indirect treatment effect (ITE):** Interest may also lie in quantifying the amount of the global treatment effect due to the network. The ITE is therefore defined as the difference between the global treatment effect and the direct treatment effect. In the GANE framework, this is

$$\begin{aligned} \text{ITE} &= \text{GTE} - \tau \\ &= \frac{1}{n} \sum_{i=1}^n \mathbb{E}[f_{T,i}(\mathbf{D}_{\mathbf{Z}=\mathbf{1}_n}, \boldsymbol{\eta})] - \frac{1}{n} \sum_{i=1}^n \mathbb{E}[f_{C,i}(\mathbf{D}_{\mathbf{Z}=\mathbf{0}_n}, \boldsymbol{\eta})]. \end{aligned} \tag{2.9}$$

Hence, the indirect treatment effect can also be interpreted as the difference between the network effect induced by the treatment versus that induced by the control.

Each of these quantities can be expressed as functions of the model parameters  $\mu$ ,  $\tau$ , and  $\boldsymbol{\eta}$ . When the expectations of  $f_T$  and  $f_C$  are known, estimates and hypothesis tests for these quantities can be developed based on parametric inference associated with the model. With respect to hypothesis tests, the experimenters may be interested in one or more of the following hypotheses.

**Hypothesis 1** (Direct treatment effect).  $H_{01} : \text{DTE} = \tau = 0$  is the null hypothesis that the direct treatment effect is 0, i.e., keeping the network effect fixed, a node's outcome is the same no matter if it is assigned to treatment or to control.

**Hypothesis 2** (SUTVA).  $H_{02} : f_T = f_C = 0$  is the null hypothesis that there is no network effect and the SUTVA is satisfied.

**Hypothesis 3** (Indirect treatment effect).  $H_{03} : \text{ITE} = 0$  is the null hypothesis that the indirect treatment effect is 0, i.e., the network influence from treated and controlled neighbors is the same.

**Hypothesis 4** (Global treatment effect).  $H_{04} : \text{GTE} = 0$  is the null hypothesis that the global treatment effect is 0, i.e., on average, treatment does not have an effect on the nodes' outcomes.

When the GANE model can be estimated using maximum likelihood estimation, since DTE, ITE, and GTE are functions of the model parameters, Hypotheses 1, 3, and 4 can be tested using Wald-type tests, and Hypothesis 2 can be tested using a likelihood ratio test.

## 2.4 Experimental Design

To design an A/B test, the experimenter decides which units are assigned to the treatment group and which units are assigned to the control group. This corresponds to specifying the treatment assignment vector  $\mathbf{Z}$  where each unit  $i$  is assigned  $Z_i = 0$  or 1. Interest lies in determining an *optimal* treatment assignment.

Under the GANE framework, experimenters can define design criteria related to the variance of the parameters' estimators. For example, D-optimality [37] aims to minimize the determinant of the variance-covariance matrix and is therefore a popular choice to generally minimize the confidence region for the parameters. However, in the network experiment context, as discussed in Section 2.3, GTE (or possibly DTE or ITE) is the primary quantity of interest. Therefore, in the experimental design simulations in Section 4.4, we will focus on the variance of GTE estimates as our design criterion.

With the design criterion determined, experimenters can find good designs using methods such as exchange algorithms [26, 35], or random search [35]. Random search refers to the process in which the design is chosen by first randomly generating a large number of designs, and then choosing the design with the best-evaluated design criteria. Exchange algorithms [26, 35] instead take a greedy approach by iteratively changing the treatment assignment for each unit in the direction of optimizing the design criteria. Via simulation, Parker et al. [35] find that random search, despite its simplicity, yields nearly as good designs as the computationally less efficient exchange algorithm. We thus use random search as a design selection strategy in our simulations in Section 4.4.

## 3. MAXIMUM LIKELIHOOD INFERENCE

### 3.1 Estimation

We have discussed in Section 2.2 that different specifications of the GANE model may require different estimation techniques. In Section 2.3, we also mentioned that when maximum likelihood estimation is possible, different hypotheses can be tested using the maximum likelihood framework. Therefore, in this section, we discuss the maximum likelihood estimation for the GANE model.

In order to obtain the likelihood of the outcome vector  $\mathbf{Y}$ , we consider the family of GANE specifications where

the outcome  $Y_i$  either (i) does not depend on neighboring outcomes, or (ii) depends linearly on neighboring outcomes. That is,

$$\begin{aligned} f_{T,i}(\mathbf{D}, \boldsymbol{\eta}) &= \rho_T \sum_{j=1}^n W_{T,ij} Y_j + \gamma_T g_{T,i}(\boldsymbol{\varphi}), \\ f_{C,i}(\mathbf{D}, \boldsymbol{\eta}) &= \rho_C \sum_{j=1}^n W_{C,ij} Y_j + \gamma_C g_{C,i}(\boldsymbol{\varphi}), \end{aligned} \quad (3.1)$$

where  $\boldsymbol{\eta} = (\rho_T, \rho_C, \gamma_T, \gamma_C, \boldsymbol{\varphi}^\top)^\top$  and  $W_{T,ij}$  (or  $W_{C,ij}$ ) is the  $(i, j)^{\text{th}}$  element of pre-specified weight matrix  $\mathbf{W}_T$  (or  $\mathbf{W}_C$ ). The diagonals of these weight matrices are zero, i.e.,  $W_{l,ii} = 0$  for  $l \in \{T, C\}$  and  $i = 1, \dots, n$ . In addition,  $g_{T,i}(\boldsymbol{\varphi})$  and  $g_{C,i}(\boldsymbol{\varphi})$  are real-valued functions, possibly depending on the parameter vector  $\boldsymbol{\varphi}$ , the experimental data  $\mathbf{D}$ , but not the outcome vector  $\mathbf{Y}$ . We can see that Model (3.1) generalizes all model specifications discussed in Section 2.2, in which the outcome of an experiment may depend linearly on other unit's outcomes and/or possibly nonlinearly on other covariates. Model (3.1), however, excludes cases where the outcome of unit  $i$  is dependent on a *nonlinear* function of the outcome vector  $\mathbf{Y}$ , which complicates the maximum likelihood theory.

To perform estimation, we consider the matrix form of Model (3.1) as follows

$$\begin{aligned} \mathbf{Y} &= \mu \mathbf{1}_n + \tau \mathbf{Z} + (\rho_T \mathbf{W}_T \mathbf{Y} + \gamma_T \mathbf{G}_T(\boldsymbol{\varphi})) \\ &\quad + (\rho_C \mathbf{W}_C \mathbf{Y} + \gamma_C \mathbf{G}_C(\boldsymbol{\varphi})) + \boldsymbol{\epsilon}, \end{aligned} \quad (3.2)$$

where  $\mathbf{G}_T(\boldsymbol{\varphi})$  denotes the  $n \times 1$  vector of  $g_{T,i}(\boldsymbol{\varphi})$  values and  $\mathbf{G}_C(\boldsymbol{\varphi})$  denotes the  $n \times 1$  vector of  $g_{C,i}(\boldsymbol{\varphi})$  values. Let  $\mathbf{M}(\boldsymbol{\varphi}) = [\mathbf{1}_n \ \mathbf{Z} \ \mathbf{G}_T(\boldsymbol{\varphi}) \ \mathbf{G}_C(\boldsymbol{\varphi})]$  be the model matrix, which depends on the parameter vector  $\boldsymbol{\varphi}$ . Further, let  $\boldsymbol{\beta} = (\mu, \tau, \gamma_T, \gamma_C)^\top$  and  $\boldsymbol{\rho} = (\rho_T, \rho_C)^\top$ . The model may be rewritten, isolating for  $\mathbf{Y}$  on the left-hand side, as follows

$$\begin{aligned} \mathbf{Y} &= (\rho_T \mathbf{W}_T + \rho_C \mathbf{W}_C) \mathbf{Y} + \mathbf{M}(\boldsymbol{\varphi}) \boldsymbol{\beta} + \boldsymbol{\epsilon}, \\ &= \mathbf{S}(\boldsymbol{\rho})^{-1} \left( \mathbf{M}(\boldsymbol{\varphi}) \boldsymbol{\beta} + \boldsymbol{\epsilon} \right), \end{aligned} \quad (3.3)$$

where  $\mathbf{S}(\boldsymbol{\rho}) = \mathbf{I}_n - \rho_T \mathbf{W}_T - \rho_C \mathbf{W}_C$ . The expression in (3.3) is possible if and only if  $\mathbf{S}(\boldsymbol{\rho})$  is invertible. Lemma 1 gives sufficient conditions on  $\boldsymbol{\rho}$  so that  $\mathbf{S}(\boldsymbol{\rho})$  is nonsingular. Although the condition is based on any matrix norm, in practice, we can use the popular spectral norm [19] to derive the constraints. The proof of Lemma 1 is given in Appendix A.1.

**Lemma 1.** *If*

$$\max(|\rho_T|, |\rho_C|) < \frac{1}{\|\mathbf{W}_T\| + \|\mathbf{W}_C\|},$$

where  $\|\cdot\|$  denotes a matrix norm [19], then  $\mathbf{S}(\boldsymbol{\rho})$  is invertible.

With  $\mathbf{Y}$  expressed as in Equation (3.3) and assuming  $\epsilon \sim \mathcal{N}(\mathbf{0}_n, \sigma^2 \mathbf{I}_n)$ , the log-likelihood function for  $\mathbf{Y}$  is

$$\begin{aligned} \log L(\boldsymbol{\theta}) &= -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log(\sigma^2) + \log |\mathbf{S}(\boldsymbol{\rho})| \\ &\quad - \frac{1}{2\sigma^2} \left( \mathbf{S}(\boldsymbol{\rho})\mathbf{Y} - \mathbf{M}(\boldsymbol{\varphi})\boldsymbol{\beta} \right)^\top \left( \mathbf{S}(\boldsymbol{\rho})\mathbf{Y} - \mathbf{M}(\boldsymbol{\varphi})\boldsymbol{\beta} \right), \end{aligned} \quad (3.4)$$

where  $\boldsymbol{\theta} = (\boldsymbol{\rho}^\top, \boldsymbol{\beta}^\top, \boldsymbol{\varphi}^\top, \sigma^2)^\top$  is the vector of all model parameters. If the normality assumption is not made, then (3.4) becomes the *quasi log-likelihood* [28] and the estimators  $\hat{\boldsymbol{\theta}}$  that maximize (3.4) are called the quasi maximum likelihood estimators.

To find the maximum likelihood estimates, we take the first order derivatives with respect to  $\boldsymbol{\beta}$  and  $\sigma^2$  and equate them to zero to obtain

$$\hat{\boldsymbol{\beta}}(\boldsymbol{\rho}, \boldsymbol{\varphi}) = \left( \mathbf{M}(\boldsymbol{\varphi})^\top \mathbf{M}(\boldsymbol{\varphi}) \right)^{-1} \mathbf{M}(\boldsymbol{\varphi})^\top \mathbf{S}(\boldsymbol{\rho})\mathbf{Y}; \quad (3.5)$$

$$\begin{aligned} \hat{\sigma}^2(\boldsymbol{\rho}, \boldsymbol{\varphi}) &= \frac{1}{n} \left( \mathbf{S}(\boldsymbol{\rho})\mathbf{Y} - \mathbf{M}(\boldsymbol{\varphi})\hat{\boldsymbol{\beta}}(\boldsymbol{\rho}, \boldsymbol{\varphi}) \right)^\top \\ &\quad \times \left( \mathbf{S}(\boldsymbol{\rho})\mathbf{Y} - \mathbf{M}(\boldsymbol{\varphi})\hat{\boldsymbol{\beta}}(\boldsymbol{\rho}, \boldsymbol{\varphi}) \right) > 0. \end{aligned} \quad (3.6)$$

Note that these are the solution of an ordinary least squares that regresses the transformed outcome variable  $\mathbf{S}(\boldsymbol{\rho})\mathbf{Y}$  on the covariate matrix  $\mathbf{M}(\boldsymbol{\varphi})$  when  $\boldsymbol{\rho}$  and  $\boldsymbol{\varphi}$  are known. Plugging this into the log-likelihood (3.4), we can obtain the profile log-likelihood

$$\ell_P(\boldsymbol{\rho}, \boldsymbol{\varphi}) = -\frac{n}{2} [\log(2\pi) + 1] + \log |\mathbf{S}(\boldsymbol{\rho})| - \frac{n}{2} \log \hat{\sigma}^2(\boldsymbol{\rho}, \boldsymbol{\varphi}). \quad (3.7)$$

Then, we can find  $\hat{\boldsymbol{\varphi}}$  and  $\hat{\boldsymbol{\rho}}$  by maximizing  $\ell_P$  using a grid search on their respective parameter space or using numerical algorithms such as the Nelder-Mead method [32]. The maximum likelihood estimates  $\hat{\boldsymbol{\beta}}$  and  $\hat{\sigma}^2$  can be obtained by plugging  $\hat{\boldsymbol{\varphi}}$  and  $\hat{\boldsymbol{\rho}}$  into Equations (3.5) and (3.6).

Note that in (3.5), it is implicitly required that  $\mathbf{M}(\boldsymbol{\varphi})^\top \mathbf{M}(\boldsymbol{\varphi})$  is invertible, i.e., that the columns of the model matrix  $\mathbf{M}(\boldsymbol{\varphi})$  are linearly independent. Although unlikely, multicollinearity may exist. For instance, in the POW-DEG specification (2.6), when the graph is fully connected (i.e., every node is connected with one another), or when the treatment and/or control degrees are the same for all nodes, the model matrix will have linearly dependent columns. It is thus important in the design stage to choose a design that ensures the model matrix has full rank.

## 3.2 Asymptotic Results

Here, we study the behavior of the maximum likelihood estimators as the network size increases to infinity. We use the subscript  $n$  to denote the data for a given network size  $n$ . Model (3.3) then becomes

$$\mathbf{Y}_n = \mathbf{S}_n(\boldsymbol{\rho})^{-1} \left( \mathbf{M}_n(\boldsymbol{\varphi})\boldsymbol{\beta} + \boldsymbol{\epsilon}_n \right),$$

where  $\mathbf{S}_n(\boldsymbol{\rho}) = \mathbf{I}_n - \rho_T \mathbf{W}_{T_n} - \rho_C \mathbf{W}_{C_n}$ . Let  $\boldsymbol{\theta}_0 = (\boldsymbol{\rho}_0^\top, \boldsymbol{\beta}_0^\top, \boldsymbol{\varphi}_0^\top, \sigma_0^2)^\top$  be the true parameter values. The consistency and asymptotic normality properties of the maximum likelihood estimators  $\hat{\boldsymbol{\theta}}_n$  are given in Theorem 1 below.

**Theorem 1.** *Under Assumption 1–6 (given in Appendix A.2), the (quasi) maximum likelihood estimator  $\hat{\boldsymbol{\theta}}_n$  obtained by maximizing the log-likelihood in (3.4) is consistent to  $\boldsymbol{\theta}_0$ . Further assuming that  $J_n(\boldsymbol{\theta}_0) = -\mathbb{E} \left[ \frac{\partial \log L_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} \right]$  and  $V_n(\boldsymbol{\theta}_0) = \mathbb{E} \left[ \left( \frac{\partial \log L_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} \right) \left( \frac{\partial \log L_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} \right)^\top \right]$  are positive definite,*

$$[V_n(\boldsymbol{\theta}_0)]^{-1/2} [J_n(\boldsymbol{\theta}_0)] (\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \xrightarrow{d} \mathcal{N}(\mathbf{0}_{\dim(\boldsymbol{\theta})}, \mathbf{I}_{\dim(\boldsymbol{\theta})}),$$

where  $\dim(\cdot)$  denotes the length of a vector.

The proof of Theorem 1 is given in Appendix A.3, following the ideas of Lee [28], treating  $\mathbf{S}_n(\boldsymbol{\rho})$  and  $\mathbf{M}_n(\boldsymbol{\varphi})$  as non-stochastic for any given  $\boldsymbol{\rho}$  and  $\boldsymbol{\varphi}$ . The random errors  $\epsilon_{i,n}$  are assumed to be independent and identically distributed with mean zero and variance  $\sigma_0^2$ . When  $\epsilon_{i,n}$  follow a normal distribution,  $\hat{\boldsymbol{\theta}}_n$  is the maximum likelihood estimator (instead of a quasi maximum likelihood estimator), and we have  $V_n(\boldsymbol{\theta}) = J_n(\boldsymbol{\theta})$  and

$$[J_n(\boldsymbol{\theta}_0)]^{1/2} (\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \xrightarrow{d} \mathcal{N}(\mathbf{0}_{\dim(\boldsymbol{\theta})}, \mathbf{I}_{\dim(\boldsymbol{\theta})}).$$

## 3.3 Inference for Causal Quantities

With the asymptotic normality result, inference for the parameters can be performed accordingly. The inference for the causal quantities given in Section 2.3 can then be carried out via the Delta method [45]. In particular, the global treatment effect for Model (3.1) is calculated as

$$\begin{aligned} \text{GTE}(\boldsymbol{\theta}) &= \frac{1}{n} \mathbf{1}_n^\top \left[ \left( \mu + \tau + \frac{1}{n} \sum_{i=1}^n g_{T,i}(\mathbf{D}_{\mathbf{z}=1_n}, \boldsymbol{\varphi}) \right) \right. \\ &\quad \times (\mathbf{I}_n - \rho_T \mathbf{W}_{T, \mathbf{z}=1_n})^{-1} \\ &\quad \left. - \left( \mu + \frac{1}{n} \sum_{i=1}^n g_{C,i}(\mathbf{D}_{\mathbf{z}=0_n}, \boldsymbol{\varphi}) \right) \right. \\ &\quad \left. \times (\mathbf{I}_n - \rho_C \mathbf{W}_{C, \mathbf{z}=0_n})^{-1} \right] \mathbf{1}_n. \end{aligned} \quad (3.8)$$

Using the Delta method, the variance of the GTE can be written as

$$\text{Var}[\text{GTE}(\boldsymbol{\theta})] = \mathbf{t}^\top \text{Var}(\boldsymbol{\theta}) \mathbf{t}, \quad (3.9)$$

where  $\mathbf{t} = \frac{\partial \text{GTE}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^\top}$ . As  $\text{DTE}(\boldsymbol{\theta}) = \tau$  and  $\text{ITE}(\boldsymbol{\theta}) = \text{GTE}(\boldsymbol{\theta}) - \tau$ , inference for DTE and ITE can be derived in a similar manner.



Table 1. Number of nodes and edges of the networks used in the simulations.

Networks	# of nodes ( $n$ )	# of edges
Caltech network	770	16,656
UMich network	3,749	81,903

## 4. SIMULATIONS

In this section, we use simulation to study the properties of different specifications of the GANE model. Specifically, we study our proposed POW-DEG specification (2.6) and the HOM specification (2.5) as an illustration of a spatially autoregressive specification. In order to study these model specifications on real-life networks, we use the Caltech Facebook network and the UMichigan Facebook network, both retrieved from the Network Repository [38]. The sizes of these networks are summarized in Table 1. In both cases, these networks provide realistic structure for the graph  $\mathcal{G}$ , but the experiment and outcomes are hypothetical and simulated using either the POW-DEG (2.6) or HOM (2.5) specifications.

### 4.1 The Distribution of the Estimates

We first investigate the asymptotic properties of the maximum likelihood estimates derived in Section 3. As the theoretical results concern the case where the treatment assignment vector  $\mathbf{Z}$  is known, in this simulation, we fix a particular design where half of the nodes are randomly assigned to treatment and the other half are assigned to control.

First, we investigate the results for the POW-DEG specification (2.6) by generating outcomes on the given network (either the Caltech or UMich Facebook network) with the following parameter settings:  $\beta = (0, 1, 0.5, 0.1)^\top$  and  $\sigma = 1$ . We further vary the power  $\lambda$  within the set  $\{0.5, 0.75, 1, 1.25\}$ , where  $\lambda = 1$  corresponds to the LNE specification (2.3). With each combination of parameters, 1,000 runs are conducted where the outcomes are generated and the maximum likelihood estimates are calculated accordingly.

The distribution of the parameter and GTE estimates for the POW-DEG specification (2.6) are plotted in Figure 2. We can see that the distributions of all estimates are reasonably bell-shaped and symmetric and centered around the true values (dashed vertical lines) as is expected given the asymptotic theory. While the distribution of  $\hat{\tau}$  remains the same under different values of  $\lambda$ , the variances of the other estimators decrease when  $\lambda$  increases. This is because the ranges of values within  $\mathbf{G}_T$  and  $\mathbf{G}_C$  in the model matrix  $\mathbf{M}$  increase as  $\lambda$  increases, which in turn decreases the variance of the parameter estimates.

The coverage of 95% asymptotic confidence intervals and variances of the parameter estimates are given in Figure 3, where left axes correspond to variances and right axes correspond to coverage. The blue lines depict the asymptotic variances derived from  $\mathbf{J}(\theta_0)$  and the red lines depict the sample

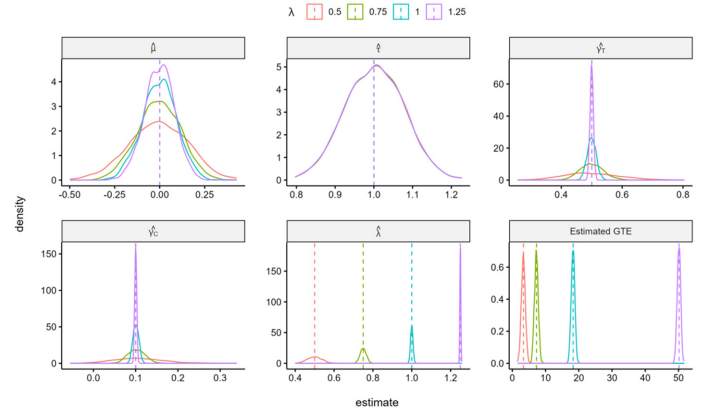


Figure 2: The distribution of parameter estimates of the POW-DEG specification on the Caltech Facebook network with  $\beta = (0, 1, 0.5, 0.1)^\top$  and  $\lambda \in \{0.05, 0.75, 1.00, 1.25\}$  over 1,000 simulation runs.

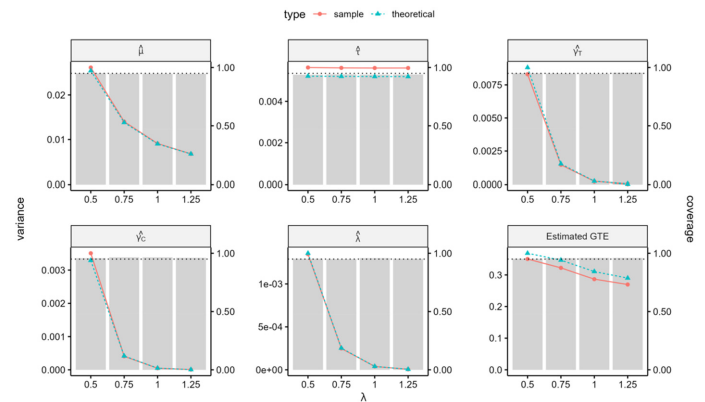


Figure 3: The variances of the estimates (left axes, lines) and coverage rates (right axes, bars) of POW-DEG specification on the Caltech Facebook network with  $\beta = (0, 1, 0.5, 0.1)^\top$  and  $\lambda \in \{0.5, 0.75, 1.00, 1.25\}$  over 1,000 simulation runs.

variance of the 1,000 parameter estimates. The agreement between these lines suggests that the asymptotic variances may be used reliably for inference. With respect to coverage, the coverage rates for the 95% confidence intervals are plotted as grey bars on the right axes and the dotted lines serve as a reference at 0.95. We can see that the obtained confidence intervals have the correct coverage. To summarize, the simulation corroborates the asymptotic theory and indicates that maximum likelihood procedures work as expected for the POW-DEG specification (2.6). Simulation results for a different set of parameter values are included in Section S1 of the Supplementary Material. These results suggest that when the network effect is small and the network size is moderate, consistent estimation of  $\gamma_T$ ,  $\gamma_C$ , and  $\lambda$  is more difficult. However, the battery of simulations was also run on the UMich Facebook network, whose size ( $n = 3,749$ ) is almost 5 times that of the Caltech network, and we find

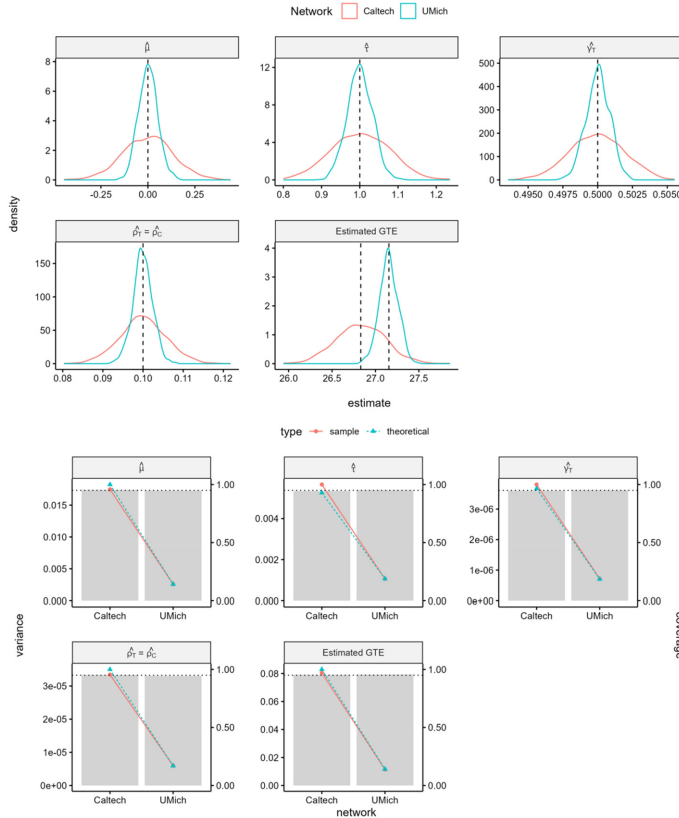


Figure 4: (upper) The distribution of parameter estimates of the HOM specification with  $\mu = 0$ ,  $\tau = 1$ ,  $\gamma_T = 0.5$ ,  $\rho_T = \rho_C = 0.1$  over 1,000 simulation runs. (lower) The corresponding variances of the estimates (left axes, lines) and coverage rates (right axes, bars).

that estimation of all parameters, whether the network and treatment effects are large or small, agrees with the asymptotic theory. These results are also available in Section S1 of the Supplementary Material.

We conducted a similar simulation study on the HOM specification (2.5) with  $\mu = 0$ ,  $\tau = 1$ ,  $\gamma_T = 0.5$ ,  $\rho_T = \rho_C = 0.1$  and  $\sigma^2 = 1$ . The results for both the Caltech Facebook network and the UMich Facebook network are shown in Figure 4. As with the POW-DEG (2.6) estimates, and in agreement with the likelihood theory, the distributions of these parameter estimates are bell-shaped and centered at the true values. Moreover, since the UMich Facebook network is larger, the variation in the estimates decreases, as expected. Notice that the true values of GTE are different for the two networks, even though all parameters used are the same. This illustrates how the true value of GTE depends not only on the parameters but also on the structure of the graph. Variances and confidence interval coverage are also plotted in Figure 4. As we would expect, the asymptotic variances are suitable for inference and the asymptotic confidence intervals have acceptable coverage. To demonstrate

the generality of these findings we present additional simulation results for another set of parameter values in Section S1 of the Supplementary Material. The theory developed in Section 3 and the simulations presented here (for multiple GANE specifications, parameter values, and networks) demonstrate the general utility of maximum likelihood inference with GANE models.

## 4.2 Hypothesis Testing

As discussed in Section 2.3, under the GANE framework, we can test hypotheses about the DTE, SUTVA, the ITE and the GTE. In particular, testing  $DTE = 0$  is equivalent to testing  $H_{01} : \tau = 0$ ; testing whether SUTVA is satisfied is equivalent to testing  $H_{02} : f_T = f_C = 0$ ; the null hypothesis for testing the indirect treatment effect is  $H_{03} : ITE = 0$ ; and the null hypothesis for testing the global treatment effect is  $H_{04} : GTE = 0$ . In the maximum likelihood framework, Hypotheses 1, 3, and 4 can be tested using Wald-type tests and Hypothesis 2 can be tested with a likelihood ratio test.

We study the characteristics of these tests via simulation. Again, as the design  $\mathbf{Z}$  is treated as fixed in our analysis, we randomly pick a design where half of the nodes are assigned to treatment and the other half are assigned to control. The parameters of the POW-DEG specification (2.6) are set at  $\beta = (0, 1, 0.5, 0.1)^\top$ ,  $\sigma = 1$  and  $\lambda \in \{0.5, 0.75, 1.00, 1.25\}$ . Separate simulations are conducted to investigate each of the four hypothesis tests. For each simulation, values of certain parameters in  $\beta$  vary while the others stay as stated. In particular, in the simulation for Hypothesis 1,  $\tau$  varies in the range  $[0, 1]$ ; in the simulation for Hypothesis 2,  $\gamma_T = \gamma_C$  and their values vary in the range  $[0, 0.05]$ ; for Hypothesis 3,  $\gamma_C$  is fixed at 0.1 and  $\gamma_T - \gamma_C$  varies in the range  $[0, 0.5]$ . Hypothesis 4 with  $H_{04} : GTE = 0$  is also tested within each of the three simulations (with different  $\lambda$ ) and the results are aggregated over different values of GTE, which corresponds to different parameter combinations. All tests are done at a 5% significance level and 1,000 runs are conducted for each parameter combination. The results are presented in Figure 5. The dotted horizontal line serves as a reference at the 5% level.

As expected, the rejection rates for each test increase as the respective parameter values depart from their null values. Moreover, tests for  $H_{01} : \tau = 0$  seem to behave similarly over different values of  $\lambda$ . This is consistent with the model estimation results in Figure 2 where the variances for  $\hat{\tau}$  and  $\widehat{GTE}$  look similar over different values of  $\lambda$  while the variances for  $\hat{\gamma}$ 's decrease as  $\lambda$  increases. We also remark that the results at null values deviate slightly from the nominal 5% level. This can be attributed to the use of asymptotic (inexact) variances in these tests. Although we do not include the results for the UMich Facebook network, given its size and given the results from Section 4.1, we expect similar results to those presented here for the Caltech Facebook network.

Table 2. Parameters for the simulation in Section 4.3.

Specification	$\mu$	$\tau$	$\rho_T$	$\rho_C$	$\gamma_T$	$\gamma_C$	$\lambda$	$\sigma$
SUTVA	0	2	0	0	0	0		1
LNE	0	1	0	0	0.1231	0.1		1
POW-DEG	0	1	0	0	0.2691	0.1	0.5	1
LAG	0	1	0.008492	0.001	0	0		0.9977
HOM	0	1	0.1	0.1	0.01728	0		0.9999

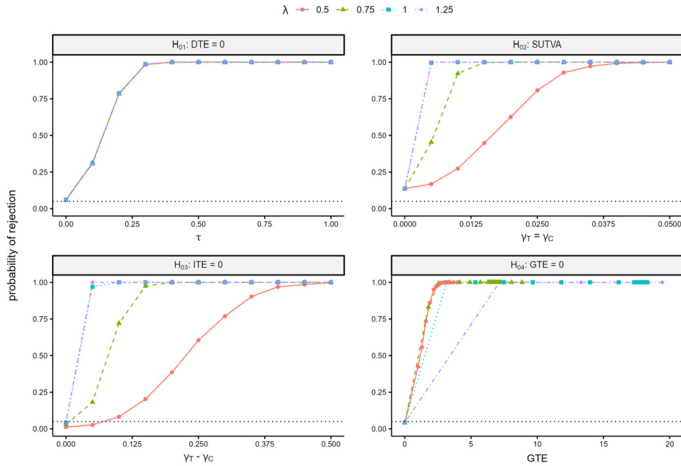


Figure 5: Rejection rates of hypothesis tests for POW-DEG specification on the Caltech Facebook network with varying parameters.

We conducted similar simulations for the HOM specification (2.5) with  $\mu = 0$ ,  $\tau = 1$ ,  $\gamma_T = 0.5$ ,  $\rho_T = \rho_C = 0.1$ , and varying  $\tau$ ,  $\gamma_T = \rho_C$  and  $\gamma_T - \rho_C$  in different simulations for different hypothesis tests. The results are included in Section S2 of the Supplementary Material. It can be noted that the results are similar in both networks, except for different values of  $\tau$ , signifying that this is an important parameter for the HOM specification (2.5). We also note that the rejection rates for Hypothesis 3 always stay at 100% even at  $\gamma_T = \rho_C$  (i.e., when the scaling coefficients for  $f_T$  and  $f_C$  are equal). This shows that the indirect effect is not only affected by the sizes of the network effects, but also by the functional forms of  $f_T$  and  $f_C$ . These are different in the HOM specification (2.5).

### 4.3 Model Misspecification

The simulations in Sections 4.1 and 4.2 explore properties of maximum likelihood inference for different GANE specifications when they are *correctly specified*. In this section, we further investigate properties of these specifications under model misspecification. The specifications considered here are (i) the SUTVA specification, in which network effects do not exist and  $f_T = f_C = 0$ ; (ii) the linear network effect (LNE) specification in (2.3); (iii) the POW-DEG specification in (2.6); (iv) the local aggregate (LAG) specification in (2.4); and (v) the homophily (HOM) specification in (2.5).

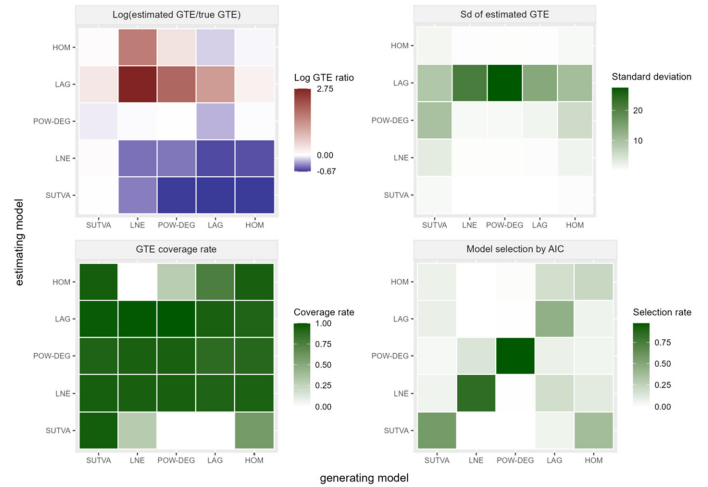


Figure 6: Model misspecification simulation results.

In this simulation, on the Caltech Facebook network, outcomes are generated 1,000 times for each of the listed model specifications. The data are then fitted using each of the five model specifications. We use global treatment effect (GTE) estimation and its inference results to compare performance among specifications because the GTE is generally of primary interest. To make the comparison fair, parameters for each model specification are chosen such that the true global treatment effect (GTE) is fixed at 2.0 and the average outcome variance is 1.0 in all data-generating scenarios. The exact parameter values for each specification are provided in Table 2.

Results of the simulation are plotted as heatmaps in Figure 6. The columns correspond to outcome-generating models and the rows correspond to estimating models. The top left panel shows the log ratio of the average estimated GTE to the true GTE. The desired value is 0, which is colored white. Red represents overestimation and blue represents underestimation. We see that all specifications can estimate the SUTVA specification well because it is nested within all GANE specifications. The POW-DEG specification seems to provide estimates with the lowest bias, even under model misspecification.

The top right panel shows the variances of the GTE estimates where white represents low variances and dark green represents high variances. We can see that the SUTVA and HOM (2.5) specifications provide the lowest variances while

the highest variances come from the LAG specification (2.4). Both the POW-DEG (2.6) and the LNE (2.3) specifications provide reasonably low variances.

The bottom left panel shows the coverage rate of 95% confidence intervals for the GTE constructed by each estimating model. The results show that LNE (2.3), POW-DEG (2.6) and LAG (2.4) specifications have high coverage rates while the HOM specification (2.5) has lower coverage rates and the SUTVA specification has the worst. This is because the SUTVA specification does not capture the network effects introduced by other specifications.

Finally, on the bottom right, the model selection results by AIC are presented. Green represents high selection rates while white represents low selection rates. AIC works well as it selects the correct model specification most of the time, which is shown by the green diagonal. This supports the use of likelihood-based model selection criteria such as AIC for the GANE framework. Furthermore, it can be seen that POW-DEG specification (2.6) is selected fairly often no matter the data-generating model, which suggests that it fits the data reasonably well even under model misspecification.

As illustrated in Figure 6, the POW-DEG specification (2.6) is the only one that performs well in each dimension. This illustrates the flexibility of the POW-DEG model to capture a variety of network effects. Hence, we advocate its use generally, especially when there is no prior information or preference for another specification.

### 4.4 Experimental Design

Given the suggestion for the general use of the POW-DEG specification (2.6) in Section 4.3, in this section, we examine the characteristics of good designs in the context of this specification. Identifying such characteristics will provide useful insights to consider when designing the experiment. To do so, we randomly generate 10,000 designs for the Caltech Facebook network as follows: (i) take a random draw  $m$  from the discrete uniform distribution on  $[1, n - 1]$ ; (ii) randomly select  $m$  nodes and assign them to treatment, and assign the remaining  $n - m$  nodes to control.

As GTE is the primary focus for many experiments on networks, in this simulation we use the variance of the estimated global treatment effect  $\text{Var}[\widehat{\text{GTE}}]$  as the design criterion to evaluate the designs. Good designs are ones that give lower values of the design criterion. The criterion is evaluated based on the Fisher information matrix of the POW-DEG specification (2.6) with parameters  $\mu = 0, \tau = 1, \lambda \in \{0.50, 0.75, 1.00, 1.25\}$ . We also vary  $\gamma_T$  and  $\gamma_C$  in three settings: (i)  $\gamma_T = 0.5 > \gamma_C = 0.1$ ; (ii)  $\gamma_T = 0.5 = \gamma_C$ ; (iii)  $\gamma_T = 0.1 < \gamma_C = 0.5$ .

For each parameter combination, the best and worst 500 (i.e., 5%) designs with respect to  $\text{Var}[\widehat{\text{GTE}}]$  are recorded. We investigate three characteristics of these designs using three metrics. First, as Bowers et al. [6] point out, in the presence of network interference, balanced allocation may not be optimal. As such, we examine the percentage of treated

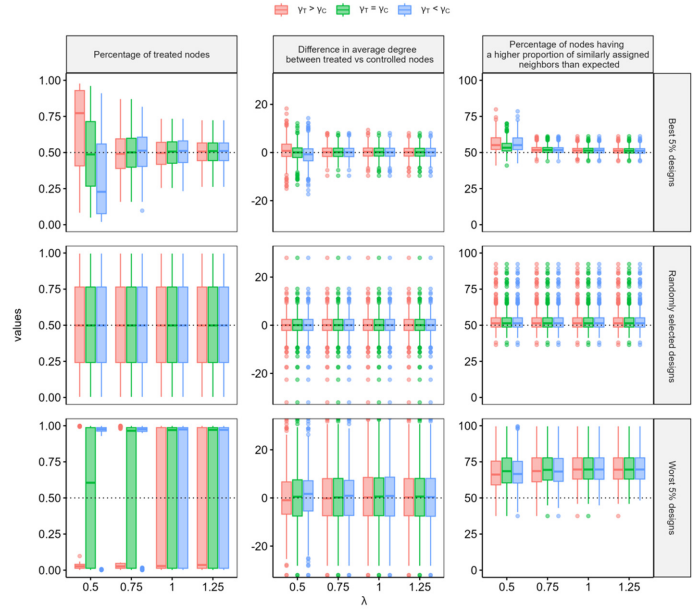


Figure 7: Characteristics of “best” 5%, randomly selected, and “worst” 5% designs with respect to the variance of estimated global treatment effect on the Caltech Facebook network.

nodes in each of the designs. Second, Parker et al. [35] observe that for certain design criteria, it is better to assign nodes with high degrees to treatment and nodes with low degrees to control. Hence we calculate the difference in average degrees of treated and controlled nodes to verify this observation. Third, if the best designs are clustered as graph-cluster randomization algorithms suggest, then nodes will be more likely to be surrounded by neighbors who share the same treatment assignment as themselves. This means that for each node, the percentage of neighbors having the same treatment assignment will be high. However, with our two-step design-generating procedure, we are considering both balanced and unbalanced designs. Suppose a design has  $b\%$  of treated nodes, then the nodes will be expected to have  $b\%$  of neighbors treated and  $(100 - b)\%$  of neighbors controlled. Nodes in a clustered design will have a high percentage of similarly treated neighbors regardless of such expectation. Therefore, we use the percentage of treated neighbors in each design as a threshold and calculate the percentage of nodes in the design that have the proportion of neighbors sharing the same treatment with themselves higher than this threshold.

Figure 7 shows characteristics of the best 500, the worst 500, and 500 randomly selected designs from the 10,000 that were generated. First, let us look at the best 5% designs. We see that the designs are more balanced at higher values of  $\lambda$ . At smaller values of  $\lambda$ , when  $\gamma_T$  is greater, we need more treated nodes and when  $\gamma_C$  is lower, we need more controlled nodes. Thus, in order to decide which treatment

Table 3. Efficiency in terms of  $\text{Var}[\widehat{GTE}]$  of the best and worst 500 designs selected from 10,000 randomly generated designs compared to 500 randomly generated designs on the Caltech Facebook network.

Designs	$\lambda$	0.5	0.75	1.00	1.75
Best 500 designs	$\gamma_T > \gamma_C$	1.7521	2.3841	5.5776	23.7801
	$\gamma_T = \gamma_C$	1.3875	2.0251	5.1557	23.2936
	$\gamma_T < \gamma_C$	1.6177	2.3152	5.5369	23.7562
Worst 500 designs	$\gamma_T > \gamma_C$	0.2469	0.16627	0.0807	0.0486
	$\gamma_T = \gamma_C$	0.4772	0.2032	0.0792	0.04791
	$\gamma_T < \gamma_C$	0.2291	0.1588	0.0797	0.0484

allocation to use, we need to know which of  $\gamma_T$  or  $\gamma_C$  is greater. This information may or may not be available in practice. Second, the differences in average degrees between treated nodes and controlled nodes distribute evenly around zero, suggesting that there is no preference in terms of degree when assigning treatments to the nodes. Last, in the third panel, we only observe a slight deviance of the boxplots from the 0.5 reference line at low values of  $\lambda$ . This suggests that graph cluster randomization is not very effective, at least in this specific setting. In contrast, the random designs have all the characteristics randomly distributed around the reference lines, while the worst designs will assign almost all of the nodes to either treatment or control.

Table 3 shows the efficiency gained (lost) by selecting the best (worst) 500 designs out of 10,000 random designs relative to a randomly selected design. We can see that the efficiency gained by the best 500 designs increases when network effects are present, i.e., when  $ITE \neq 0$  or  $\gamma_T \neq \gamma_C$  for the POW-DEG specification (2.6). It also increases when the sizes of the network effects increase, and when  $\lambda$  increases. Hence, even though a randomly generated design may be suitable when there is no network effect, when some network effect is present, other design selection techniques are necessary. We perform an analogous investigation with the UMich Facebook network, the results of which are provided in Section S3 of the Supplementary Material. The findings are similar to those presented here for the Caltech Facebook network, but we also notice that with a larger network, the efficiency of this random search procedure decreased. This suggests that when there are many experimental units, a balanced randomized design may still be efficient.

## 5. DISCUSSION

We introduce the general additive network effect model for network A/B tests, which unifies many existing models in the literature and enhances the modeling flexibility. We further bridge the model-based framework and the exposure framework by defining causal quantities of interest: the global treatment effect, the direct treatment effect, and the indirect treatment effect as functions of the model parameters. Inference for all three quantities may be carried out via the maximum likelihood framework.

Although the model is studied under the A/B testing setting where there are just two experimental conditions

(treatment and control) and under the normal independent error assumptions, the GANE model framework can be extended for use in other settings. First, by expanding the model equation, the GANE model can be used to analyze experiments with more than two experimental conditions. Second, by introducing link functions and other distributional and functional assumptions, the framework can be extended to deal with non-normal distributions and discrete outcomes in manners similar to generalized linear models.

Despite the GANE framework’s wide modeling possibilities, we specifically propose the POW-DEG specification (2.6), which models the network effect as powers of the treatment and control degrees. Via simulation, we found that the specification is robust against model misspecification in terms of inference for the global treatment effect. Thus we suggest the use of this specification, especially in the design stage, when there is no prior information or modeling preference. Characteristics of good designs with respect to different criteria on the POW-DEG specification (2.6) are also investigated via simulation. We find that balanced randomization to treatment and control and graph cluster randomization are not necessarily optimal for the estimation of the global treatment effect in this specification.

The design simulation in Section 4.4 suggests that different parameter settings lead to a different desired percentage of treated nodes. In addition, the optimal design search requires the formula of the Fisher information matrix, which further demands the knowledge of both the model and parameters. This is often unknown in the design stage, thus posing challenges for the model-based framework. Future work could consider methods such as Bayesian design to alleviate the dependence on unknown parameters [13]. Moreover, although AIC appears to work in the model misspecification simulation in Section 4.3, the use of AIC is only possible in the analysis stage once the data is observed, or when preliminary data are available. Model selection for the design and analysis of experiments on networks thus remains an open problem for future research.

## SUPPLEMENTARY MATERIAL

The Supplementary Material to “General Additive Network Effect Models” contains additional simulation results for Section 4.

## APPENDIX A. MATHEMATICAL DETAILS FOR SECTION 3

### A.1 Proof of Lemma 1

We use the following two lemmas.

**Lemma 2** (Theorem 18.2.16 of Harville [18]). *Let  $\mathbf{A}$  represent an  $n \times n$  matrix. Then, the infinite series  $\mathbf{I} + \mathbf{A} + \mathbf{A}^2 + \mathbf{A}^3 + \dots$  converges if and only if  $\lim_{k \rightarrow \infty} \mathbf{A}^k = \mathbf{0}$ , in which case  $\mathbf{I} - \mathbf{A}$  is nonsingular and*

$$(\mathbf{I} - \mathbf{A})^{-1} = \sum_{k=0}^{\infty} \mathbf{A}^k = \mathbf{I} + \mathbf{A} + \mathbf{A}^2 + \mathbf{A}^3 + \dots,$$

where  $\mathbf{A}^0 = \mathbf{I}$ .

**Lemma 3** (Lemma 5.6.11 of Horn and Johnson [19]). *Let  $\mathbf{A}$  be an  $n \times n$  given matrix. If there is a matrix norm  $\|\cdot\|$  such that  $\|\mathbf{A}\| < 1$ , then  $\lim_{k \rightarrow \infty} \mathbf{A}^k = \mathbf{0}$ , that is, each entry of  $\mathbf{A}^k$  tends to zero as  $k \rightarrow \infty$ .*

From the two lemmas, if we have  $\|\rho_T \mathbf{W}_T + \rho_C \mathbf{W}_C\| < 1$  for any matrix norm  $\|\cdot\|$ , then  $\mathbf{S}(\boldsymbol{\rho})$  will be invertible. Now, if the condition of Lemma 1 is satisfied, using triangle inequality, we can derive

$$\begin{aligned} \|\rho_T \mathbf{W}_T + \rho_C \mathbf{W}_C\| &\leq |\rho_T| \|\mathbf{W}_T\| + |\rho_C| \|\mathbf{W}_C\|, \\ &\leq \max(|\rho_T|, |\rho_C|) [\|\mathbf{W}_T\| + \|\mathbf{W}_C\|], \\ &< 1, \end{aligned}$$

i.e.,  $\mathbf{S}(\boldsymbol{\rho})$  is invertible.

### A.2 Assumptions Needed for Asymptotic Results

In order to achieve the asymptotic results in Theorem 1 in Appendix A.3, we make the following assumptions.

**Assumption 1.**  $\boldsymbol{\epsilon}_n = (\epsilon_{1n}, \epsilon_{2n}, \dots, \epsilon_{nn})^\top$  are independently and identically distributed with mean 0 and variance  $\sigma_0^2 > 0$ . In addition, the moment  $\mathbb{E}(|\epsilon_{i,n}|^{4+\eta})$  exists for some  $\eta > 0$ .

**Assumption 2.** The true parameters  $\boldsymbol{\rho}_0$  and  $\boldsymbol{\varphi}_0$  lie in the interior of a compact parameter space  $\mathbf{P} \times \boldsymbol{\Phi}$ . The parameters are uniquely identifiable, in the sense that  $\mathbb{P}(L_n(\boldsymbol{\theta}_1) = L_n(\boldsymbol{\theta}_2)) = 0$  for  $\boldsymbol{\theta}_1 \neq \boldsymbol{\theta}_2$ .

**Assumption 3.** The elements of  $\mathbf{W}_{Tn}$  and  $\mathbf{W}_{Cn}$  are at most of order  $h_n$  uniformly, i.e.,  $W_{ln,ij} = O(1/h_n) \forall i, j$  and  $l \in \{T, C\}$ . The sequence  $h_n$  can be bounded or divergent. Furthermore,  $\lim_{n \rightarrow \infty} h_n/n = 0$ .

**Assumption 4.** The matrix  $\mathbf{S}_n(\boldsymbol{\rho}_0)$  is nonsingular.

**Assumption 5.** The weight matrices  $\mathbf{W}_{Tn}$ ,  $\mathbf{W}_{Cn}$  and the matrix  $\mathbf{S}_n(\boldsymbol{\rho}_0)^{-1}$  are uniformly bounded in both row and column sums. Moreover,  $\mathbf{S}_n(\boldsymbol{\rho})^{-1}$  is uniformly bounded in either row or column sums.

**Assumption 6.** For each  $i$ , the functions  $g_{T,i}$  and  $g_{C,i}$  are twice continuously differentiable with respect to  $\boldsymbol{\varphi}$ .

The values of these functions and their derivatives are uniformly bounded  $\forall \boldsymbol{\varphi} \in \boldsymbol{\Phi}$ . Furthermore,  $\forall \boldsymbol{\varphi} \in \boldsymbol{\Phi}$ ,  $\lim_{n \rightarrow \infty} \mathbf{M}_n(\boldsymbol{\varphi})^\top \mathbf{M}_n(\boldsymbol{\varphi})/n$  exists and nonsingular.

Assumptions 1, 2 and the differentiability requirement for  $g_{T,i}$  and  $g_{C,i}$  in Assumption 6 are usual regularity conditions for the consistency and asymptotic normality of nonlinear least squares regression [21]. The identifiability requirement in Assumption 2 contains the requirement that the columns of the model matrix  $\mathbf{M}_n(\boldsymbol{\varphi})$  are linearly independent as discussed in Section 3.1.

Note that we cannot use the usual central limit theorems to derive the asymptotic behavior of Model (3.3) because as the sample size  $n$  changes, the weight matrices may also change, leading to changes in the outcomes  $\mathbf{Y}_n$ . For example, when a new unit is added to the network, it can be connected to other existing units, which in turn changes the degree  $k_i$  for each existing unit  $i$ ,  $i = 1, 2, \dots, n$ . This results in a different set of weight matrices for autoregressive specifications such as the LAG (2.4) or the HOM (2.5) specifications. Therefore, we need to use the Central Limit Theorem for linear-quadratic forms of triangular arrays [25]. Assumptions 3, 5, and 6 are introduced to satisfy the assumptions of this theorem. Essentially, the bounds in these assumptions control the spatial correlations to a manageable degree so that they do not diverge as  $n$  goes to infinity [28]. For example, suppose  $\mathbf{W}_{Tn} = \mathbf{A}_n$ , Assumption 3 is satisfied as all elements of  $\mathbf{A}$  are either 1 or 0, i.e., bounded. However, to satisfy Assumption 5 in this case, we need to further require that the degree of each unit  $i$ , i.e., the number of connections, is bounded as  $n$  goes to infinity. This is reasonable in social network settings as one will not have infinitely many friends.

Finally, Assumption 4 makes sure that  $\mathbf{Y}_n$  can be expressed in the reduced form as in (3.3).

### A.3 Proof of Theorem 1

To prove consistency, we use the following lemma.

**Lemma 4** (Theorem 3.4 of White [46]). *Let  $(\Omega, \mathcal{F}, \mathbb{P})$  be a complete probability space, let  $\Theta$  be a compact subset of  $\mathbb{R}^p$ ,  $p \in \mathbb{N}$  and let  $\{\Theta_n\}$  be a sequence of compact subsets of  $\Theta$ . Let  $\{Q_n\}$  be a sequence of random functions continuous on  $\Theta$  a.s. and let  $\hat{\boldsymbol{\theta}}_n = \arg \max_{\Theta_n} Q_n(\cdot, \boldsymbol{\theta})$  a.s. If  $Q_n(\cdot, \boldsymbol{\theta}) - \bar{Q}_n(\boldsymbol{\theta}) \rightarrow 0$  as  $n \rightarrow \infty$  a.s. uniformly on  $\Theta$  and if  $\{Q_n : \Theta \rightarrow \mathbb{R}\}$  has identifiably unique maximizers  $\{\boldsymbol{\theta}_n^*\}$  on  $\{\Theta_n\}$  then  $\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_n^* \rightarrow 0$  as  $n \rightarrow \infty$  a.s.*

From the reduced form (3.3), consider

$$\begin{aligned} Q_n(\boldsymbol{\rho}, \boldsymbol{\varphi}) &= \max_{\boldsymbol{\beta}, \sigma^2} \mathbb{E}(\log L_n(\boldsymbol{\theta})), \\ &= \max_{\boldsymbol{\beta}, \sigma^2} \left[ -\frac{n}{2} \log 2\pi - \frac{n}{2} \log(\sigma^2) + \log |\mathbf{S}_n(\boldsymbol{\rho})| \right. \\ &\quad \left. - \frac{1}{2\sigma^2} \boldsymbol{\beta}^\top \mathbf{M}_n(\boldsymbol{\varphi})^\top \mathbf{M}_n(\boldsymbol{\varphi}) \boldsymbol{\beta} - \frac{\sigma_0^2}{2\sigma^2} \text{tr}(\mathbf{B}_n(\boldsymbol{\rho})) \right] \end{aligned}$$

$$\begin{aligned}
& + \frac{1}{\sigma^2} \boldsymbol{\beta}^\top \mathbf{M}_n(\boldsymbol{\varphi})^\top \mathbf{S}_n(\boldsymbol{\rho}) \mathbf{S}_n(\boldsymbol{\rho}_0)^{-1} \mathbf{M}_n(\boldsymbol{\varphi}_0) \boldsymbol{\beta}_0 \\
& - \frac{1}{2\sigma^2} \boldsymbol{\beta}_0^\top \mathbf{M}_n(\boldsymbol{\varphi}_0)^\top \mathbf{S}_n(\boldsymbol{\rho}_0)^{-\top} \mathbf{S}_n(\boldsymbol{\rho})^\top \\
& \times \mathbf{S}_n(\boldsymbol{\rho}) \mathbf{S}_n(\boldsymbol{\rho}_0)^{-1} \mathbf{M}_n(\boldsymbol{\varphi}_0) \boldsymbol{\beta}_0 \Big], \\
& - \frac{1}{n} \left[ \frac{1}{\sqrt{n}} \mathbf{M}_n(\boldsymbol{\varphi})^\top \mathbf{S}_n(\boldsymbol{\rho}) \mathbf{S}_n(\boldsymbol{\rho}_0)^{-1} \boldsymbol{\epsilon}_n \right]^\top \\
& \times \left[ \frac{1}{n} [\mathbf{M}_n(\boldsymbol{\varphi})^\top \mathbf{M}_n(\boldsymbol{\varphi})]^{-1} \right] \\
& \times \left[ \frac{1}{\sqrt{n}} \mathbf{M}_n(\boldsymbol{\varphi})^\top \mathbf{S}_n(\boldsymbol{\rho}) \mathbf{S}_n(\boldsymbol{\rho}_0)^{-1} \boldsymbol{\epsilon}_n \right].
\end{aligned}$$

where  $\mathbf{B}_n(\boldsymbol{\rho}) = \mathbf{S}_n(\boldsymbol{\rho}_0)^{-\top} \mathbf{S}_n(\boldsymbol{\rho})^\top \mathbf{S}_n(\boldsymbol{\rho}) \mathbf{S}_n(\boldsymbol{\rho}_0)^{-1}$ . Taking the first derivative with respect to  $\boldsymbol{\beta}$  and  $\sigma^2$ , we obtain the maximizers of  $Q_n(\boldsymbol{\rho}, \boldsymbol{\varphi})$  as follows:

$$\begin{aligned}
\boldsymbol{\beta}_n^*(\boldsymbol{\rho}, \boldsymbol{\varphi}) &= [\mathbf{M}_n(\boldsymbol{\varphi})^\top \mathbf{M}_n(\boldsymbol{\varphi})]^{-1} \mathbf{M}_n(\boldsymbol{\varphi})^\top \mathbf{S}_n(\boldsymbol{\rho}) \\
&\quad \times \mathbf{S}_n(\boldsymbol{\rho}_0)^{-1} \mathbf{M}_n(\boldsymbol{\varphi}_0) \boldsymbol{\beta}_0; \\
\sigma_n^{*2}(\boldsymbol{\rho}, \boldsymbol{\varphi}) &= \frac{1}{n} \left\{ \boldsymbol{\beta}_0^\top \mathbf{M}_n(\boldsymbol{\varphi}_0)^\top \mathbf{S}_n^{-\top}(\boldsymbol{\rho}_0) \mathbf{S}_n(\boldsymbol{\rho})^\top [\mathbf{I}_n - \mathbf{H}_n(\boldsymbol{\varphi})] \right. \\
&\quad \left. \times \mathbf{S}_n(\boldsymbol{\rho}) \mathbf{S}_n(\boldsymbol{\rho}_0)^{-1} \mathbf{M}_n(\boldsymbol{\varphi}_0) \boldsymbol{\beta}_0 + \sigma_0^2 \text{tr}(\mathbf{B}_n(\boldsymbol{\rho})) \right\}. \tag{A.1}
\end{aligned}$$

where  $\mathbf{H}_n(\boldsymbol{\varphi}) = \mathbf{M}_n(\boldsymbol{\varphi}) [\mathbf{M}_n(\boldsymbol{\varphi})^\top \mathbf{M}_n(\boldsymbol{\varphi})]^{-1} \mathbf{M}_n(\boldsymbol{\varphi})^\top$ . Now,

$$\begin{aligned}
\ell_{P,n}(\boldsymbol{\rho}, \boldsymbol{\varphi}) &= -\frac{n}{2} \log 2\pi - \frac{n}{2} \log \hat{\sigma}_n^2(\boldsymbol{\rho}, \boldsymbol{\varphi}) \\
&\quad + \log |\mathbf{S}_n(\boldsymbol{\rho})| - \frac{n}{2}, \\
Q_n(\boldsymbol{\rho}, \boldsymbol{\varphi}) &= -\frac{n}{2} \log 2\pi - \frac{n}{2} \log \sigma_n^{*2}(\boldsymbol{\rho}, \boldsymbol{\varphi}) \\
&\quad + \log |\mathbf{S}_n(\boldsymbol{\rho})| - \frac{n}{2},
\end{aligned}$$

where  $\hat{\sigma}_n^2(\boldsymbol{\rho}, \boldsymbol{\varphi})$  was given in (3.6). To use Lemma A4, we first need to show that

$$\begin{aligned}
& \frac{1}{n} \{ \ell_{P,n}(\boldsymbol{\rho}, \boldsymbol{\varphi}) - Q_n(\boldsymbol{\rho}, \boldsymbol{\varphi}) \} \\
&= -\frac{1}{2} \{ \log \hat{\sigma}_n^2(\boldsymbol{\rho}, \boldsymbol{\varphi}) - \log \sigma_n^{*2}(\boldsymbol{\rho}, \boldsymbol{\varphi}) \} = o_P(1).
\end{aligned}$$

Note that

$$\begin{aligned}
\hat{\sigma}_n^2(\boldsymbol{\rho}, \boldsymbol{\varphi}) - \sigma_n^{*2}(\boldsymbol{\rho}, \boldsymbol{\varphi}) &= 2R_{1n}(\boldsymbol{\rho}, \boldsymbol{\varphi}) + R_{2n}(\boldsymbol{\rho}, \boldsymbol{\varphi}) \\
&\quad - \frac{1}{n} \sigma_0^2 \text{tr}(\mathbf{B}_n(\boldsymbol{\rho})),
\end{aligned}$$

where

$$\begin{aligned}
R_{1n}(\boldsymbol{\rho}, \boldsymbol{\varphi}) &= \frac{1}{n} \boldsymbol{\beta}_0^\top \mathbf{M}_n(\boldsymbol{\rho}_0)^\top \mathbf{S}_n^{-\top}(\boldsymbol{\rho}_0) \mathbf{S}_n(\boldsymbol{\rho})^\top [\mathbf{I}_n - \mathbf{H}_n(\boldsymbol{\varphi})] \\
&\quad \times \mathbf{S}_n(\boldsymbol{\rho}) \mathbf{S}_n(\boldsymbol{\rho}_0)^{-1} \boldsymbol{\epsilon}_n,
\end{aligned}$$

and

$$\begin{aligned}
R_{2n}(\boldsymbol{\rho}, \boldsymbol{\varphi}) &= \frac{1}{n} \boldsymbol{\epsilon}_n^\top \mathbf{S}_n^{-\top}(\boldsymbol{\rho}_0) \mathbf{S}_n(\boldsymbol{\rho})^\top [\mathbf{I}_n - \mathbf{H}_n(\boldsymbol{\varphi})] \mathbf{S}_n(\boldsymbol{\rho}) \mathbf{S}_n(\boldsymbol{\rho}_0)^{-1} \boldsymbol{\epsilon}_n \\
&= \frac{1}{n} \boldsymbol{\epsilon}_n^\top \mathbf{S}_n^{-\top}(\boldsymbol{\rho}_0) \mathbf{S}_n(\boldsymbol{\rho})^\top \mathbf{S}_n(\boldsymbol{\rho}) \mathbf{S}_n(\boldsymbol{\rho}_0)^{-1} \boldsymbol{\epsilon}_n
\end{aligned}$$

It can be shown that  $R_{1n}(\boldsymbol{\rho}, \boldsymbol{\varphi}) = o_P(1)$  and  $R_{2n}(\boldsymbol{\rho}, \boldsymbol{\varphi}) - \frac{1}{n} \sigma_0^2 \text{tr}(\mathbf{B}_n(\boldsymbol{\rho})) = o_P(1)$  using the following three lemmas given in Lee [28] and assumptions on the bounds of matrices.

**Lemma 5.** *Suppose the elements  $A_{n,ij}$  of  $n \times n$  matrices  $\mathbf{A}_n$  are  $O(1/h_n)$  uniformly for all  $i, j$ . If  $n \times n$  matrices  $\{\mathbf{B}_n\}$  are uniformly bounded in column sums (respectively, row sums), then the elements of  $\mathbf{A}_n \mathbf{B}_n$  (respectively,  $\mathbf{B}_n \mathbf{A}_n$ ) have the uniform order  $O(1/h_n)$ . For these cases,  $\text{tr}(\mathbf{A}_n \mathbf{B}_n) = \text{tr}(\mathbf{B}_n \mathbf{A}_n) = O(n/h_n)$ .*

**Lemma 6.** *Suppose  $\{\mathbf{A}_n\}$  are uniformly bounded either in row or column sums and their elements  $A_{n,ij}$  have order  $O(1/h_n)$  uniformly in  $i$  and  $j$ . Then  $\mathbb{E}(\boldsymbol{\epsilon}_n^\top \mathbf{A}_n \boldsymbol{\epsilon}_n) = \sigma_0^2 \text{tr}(\mathbf{A}_n) = O(n/h_n)$  and  $\text{Var}(\boldsymbol{\epsilon}_n^\top \mathbf{A}_n \boldsymbol{\epsilon}_n) = O(n/h_n)$ . If  $\lim_{n \rightarrow \infty} h_n/n = 0$ , then  $(h_n/n) [\boldsymbol{\epsilon}_n^\top \mathbf{A}_n \boldsymbol{\epsilon}_n - \mathbb{E}(\boldsymbol{\epsilon}_n^\top \mathbf{A}_n \boldsymbol{\epsilon}_n)] = o_P(1)$ , where  $\boldsymbol{\epsilon}_n$  satisfies Assumption 1 (possibly without normality but with  $\mathbb{E}(|\epsilon_n|^{4+\gamma}) < \infty$  for some  $\gamma > 0$ ).*

**Lemma 7.** *Suppose that  $\mathbf{A}_n$  is a square matrix with its column sums being uniformly bounded and elements of the  $n \times k$  matrix  $\mathbf{Z}_n$  are uniformly bounded. Then  $(1/\sqrt{n}) \mathbf{Z}_n^\top \mathbf{A}_n \boldsymbol{\epsilon}_n = o_P(1)$ . Furthermore, if the limit of  $\mathbf{Z}_n^\top \mathbf{A}_n \mathbf{A}_n^\top \mathbf{Z}_n/n$  exists and is positive definite, then  $(1/\sqrt{n}) \mathbf{Z}_n^\top \mathbf{A}_n \boldsymbol{\epsilon}_n \xrightarrow{d} \mathcal{N}(0, \sigma_0^2 \lim_{n \rightarrow \infty} \mathbf{Z}_n^\top \mathbf{A}_n \mathbf{A}_n^\top \mathbf{Z}_n/n)$ .*

Therefore,  $\hat{\sigma}_n^2(\boldsymbol{\rho}, \boldsymbol{\varphi}) - \sigma_n^{*2}(\boldsymbol{\rho}, \boldsymbol{\varphi}) = o_P(1)$  uniformly on  $\mathbf{P} \times \boldsymbol{\Phi}$ . Hence,  $\sup_{(\boldsymbol{\rho}, \boldsymbol{\varphi}) \in \mathbf{P} \times \boldsymbol{\Phi}} \frac{1}{n} \{ \ell_{P,n}(\boldsymbol{\rho}, \boldsymbol{\varphi}) - Q_n(\boldsymbol{\rho}, \boldsymbol{\varphi}) \} = o_P(1)$ . Second, we need to prove the identification uniqueness condition that, for any  $\epsilon_1, \epsilon_2 > 0$ ,

$$\limsup_{n \rightarrow \infty} \max_{\boldsymbol{\rho} \in \bar{N}_{\epsilon_1}(\boldsymbol{\rho}_0), \boldsymbol{\varphi} \in \bar{N}_{\epsilon_2}(\boldsymbol{\varphi}_0)} \frac{1}{n} [Q_n(\boldsymbol{\rho}, \boldsymbol{\varphi}) - Q_n(\boldsymbol{\rho}_0, \boldsymbol{\varphi}_0)] < 0,$$

where  $\bar{N}_{\epsilon_1}(\boldsymbol{\rho}_0)$  denotes the complement of an open neighborhood of  $\boldsymbol{\rho}_0$  of diameter  $\epsilon_1$  and likewise for  $\boldsymbol{\varphi}$ . To see this, we can write

$$\begin{aligned}
& \frac{1}{n} [Q_n(\boldsymbol{\rho}, \boldsymbol{\varphi}) - Q_n(\boldsymbol{\rho}_0, \boldsymbol{\varphi}_0)] \\
&= \frac{1}{n} \left\{ \mathbb{E}[\log L_n(\boldsymbol{\rho}, \boldsymbol{\beta}_0, \boldsymbol{\varphi}_0)] - \mathbb{E}[\log L_n(\boldsymbol{\rho}_0, \boldsymbol{\beta}_0, \boldsymbol{\varphi}_0)] \right\} \\
&\quad - \frac{1}{2} \left\{ \log \sigma_n^{*2}(\boldsymbol{\rho}, \boldsymbol{\varphi}) - \log \left( \frac{\sigma_0^2}{n} \text{tr}(\mathbf{B}_n(\boldsymbol{\rho})) \right) \right\}.
\end{aligned}$$

The first term is less than 0 by Jensen's inequality and the identifiability condition of Assumption 2. Furthermore,  $\sigma_n^{*2}(\boldsymbol{\rho}, \boldsymbol{\varphi}) \geq \frac{\sigma_0^2}{n} \text{tr}(\mathbf{B}_n(\boldsymbol{\rho}))$  from (A.1) by the positive semi-definiteness of the annihilator matrix  $\mathbf{I} - \mathbf{H}_n(\boldsymbol{\varphi})$ . Putting all of these together, we proved  $\hat{\boldsymbol{\theta}}_n = \boldsymbol{\theta}_0 + o_P(1)$ .

Now, to prove the asymptotic normality, we apply the mean-value theorem on the first order derivative of  $\log L_n(\boldsymbol{\theta})$  at  $\hat{\boldsymbol{\theta}}_n$  yielding

$$\frac{\partial \log L_n(\hat{\boldsymbol{\theta}}_n)}{\partial \boldsymbol{\theta}} = \mathbf{0} = \frac{\partial \log L_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} + (\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \frac{\partial^2 \log L_n(\tilde{\boldsymbol{\theta}}_n)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top},$$

where  $\tilde{\boldsymbol{\theta}}_n$  lies between  $\hat{\boldsymbol{\theta}}_n$  and  $\boldsymbol{\theta}_0$ . Therefore

$$\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0 = - \left[ \frac{\partial^2 \log L_n(\tilde{\boldsymbol{\theta}}_n)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} \right]^{-1} \left( \frac{\partial \log L_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} \right).$$

We can write down the first derivatives of  $\log L_n(\boldsymbol{\theta})$  with respect to  $\boldsymbol{\theta}$  as follows:

$$\begin{aligned} \frac{\partial \log L_n(\boldsymbol{\theta})}{\partial \boldsymbol{\beta}} &= \frac{1}{\sigma^2} \boldsymbol{\epsilon}_n^\top \mathbf{M}_n(\boldsymbol{\varphi}), \\ \frac{\partial \log L_n(\boldsymbol{\theta})}{\partial \sigma^2} &= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \boldsymbol{\epsilon}_n^\top \boldsymbol{\epsilon}_n, \\ \frac{\partial \log L_n(\boldsymbol{\theta})}{\partial \boldsymbol{\rho}_j} &= \frac{1}{\sigma^2} \boldsymbol{\epsilon}_n^\top \mathbf{W}_{jn} \mathbf{S}_n(\boldsymbol{\rho})^{-1} \mathbf{M}_n(\boldsymbol{\varphi}) \boldsymbol{\beta} \\ &\quad + \left( \frac{1}{\sigma^2} \boldsymbol{\epsilon}_n^\top \mathbf{W}_{jn} \mathbf{S}_n(\boldsymbol{\rho})^{-1} \boldsymbol{\epsilon}_n - \text{tr}(\mathbf{W}_{jn} \mathbf{S}_n^{-1}(\boldsymbol{\rho})) \right), \\ \frac{\partial \log L_n(\boldsymbol{\theta})}{\partial \boldsymbol{\varphi}} &= \frac{1}{\sigma^2} \boldsymbol{\epsilon}_n^\top \frac{\partial \mathbf{M}_n(\boldsymbol{\varphi})}{\partial \boldsymbol{\varphi}} \boldsymbol{\beta}, \end{aligned}$$

for  $j \in \{T, C\}$ . Note that these are linear and quadratic functions of  $\boldsymbol{\epsilon}_n$ . Therefore we can apply the Central Limit Theorem for linear-quadratic functions [25] given as Lemma A8 below.

**Lemma 8** (Theorem A.1 of Kelejian and Prucha [25]). *Consider the linear quadratic forms ( $r = 1, \dots, m$ )*

$$\mathbf{Q}_{r,n} = \boldsymbol{\epsilon}_n^\top \mathbf{A}_{r,n} \boldsymbol{\epsilon}_n + \mathbf{b}_{r,n}^\top \boldsymbol{\epsilon}_n,$$

where  $\boldsymbol{\epsilon}_n = (\epsilon_{1,n}, \dots, \epsilon_{n,n})^\top$  is an  $n \times 1$  random vector, and  $\mathbf{A}_{r,n} = (a_{ij,r,n})_{i,j=1,\dots,n}$  is an  $n \times n$  non-stochastic real matrix, and  $\mathbf{b}_{r,n} = (b_{1,r,n}, \dots, b_{n,r,n})^\top$  is an  $n \times 1$  non-stochastic real vector. We make the following assumptions:

1. The real-valued random variables of the array  $\{\epsilon_{i,n} : 1 \leq i \leq n, n \geq 1\}$  satisfy  $\mathbb{E}[\epsilon_{i,n}] = 0$ . Furthermore, for each  $n \geq 1$ , the random variables  $\epsilon_{1,n}, \dots, \epsilon_{n,n}$  are totally independent.
2. For  $r = 1, \dots, m$ , the elements of the array of real numbers  $\{a_{ij,r,n} : 1 \leq i, j \leq n, n \geq 1\}$  satisfy  $a_{ij,r,n} = a_{ji,r,n}$  and  $\sup_{1 \leq j \leq n, n \geq 1} \sum_{i=1}^n |a_{ij,r,n}| < \infty$ . The elements of the array of real numbers  $\{b_{i,r,n} : 1 \leq i \leq n, n \geq 1\}$  satisfy  $\sup_n n^{-1} \sum_{i=1}^n |b_{i,r,n}|^{2+\eta_1} < \infty$  for some  $\eta_1 > 0$ .
3. For  $r = 1, \dots, m$ , one of the following two conditions holds
  - (a)  $\sup_{1 \leq i \leq n, n \geq 1} \mathbb{E}|\epsilon_{i,n}|^{2+\eta_2} < \infty$  for some  $\eta_2 > 0$  and  $a_{ii,r,n} = 0$ .
  - (b)  $\sup_{1 \leq i \leq n, n \geq 1} \mathbb{E}|\epsilon_{i,n}|^{4+\eta_2} < \infty$  for some  $\eta_2 > 0$  (but possibly  $a_{ii,r,n} \neq 0$ )

Let

$$\mathbf{U}_n = [\mathbf{Q}_{1,n}, \dots, \mathbf{Q}_{m,n}]^\top,$$

and  $\boldsymbol{\mu}_{\mathbf{U}_n} = \mathbb{E}[\mathbf{U}_n]$  and  $\boldsymbol{\Sigma}_{\mathbf{U}_n}$  denote the mean and variance-covariance matrix of  $\mathbf{U}_n$ , respectively. Suppose the assumptions hold and  $n^{-1} \lambda_{\min}(\boldsymbol{\Sigma}_{\mathbf{U}_n}) \geq c$  for some  $c > 0$ . Let  $\boldsymbol{\Sigma}_{\mathbf{U}_n} = (\boldsymbol{\Sigma}_{\mathbf{U}_n}^{1/2})(\boldsymbol{\Sigma}_{\mathbf{U}_n}^{1/2})^\top$ , then

$$\boldsymbol{\Sigma}_{\mathbf{U}_n}^{-1/2}(\mathbf{U}_n - \boldsymbol{\mu}_{\mathbf{U}_n}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbf{I}_m).$$

Therefore, we can apply the above theorem to  $\frac{\partial \log L_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\beta}}$  with  $m = \dim(\boldsymbol{\theta})$  since all the multipliers to  $\boldsymbol{\epsilon}_n$  are bounded. Note that the assumption on the minimum eigenvalue of the variance-covariance matrix is to ensure that matrices  $\boldsymbol{\Sigma}_{\mathbf{V}_n}$  stay invertible as  $n \rightarrow \infty$ , to which we have an equivalent condition in Theorem (1). The assumption of symmetry is W.L.O.G since  $\boldsymbol{\epsilon}_n \mathbf{A}_n \boldsymbol{\epsilon}_n = \boldsymbol{\epsilon}_n^\top [(\mathbf{A}_n + \mathbf{A}_n^\top)/2] \boldsymbol{\epsilon}_n$  [24]. Hence we have

$$[\mathbf{V}_n(\boldsymbol{\theta}_0)]^{-1/2} \frac{\partial \log L_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} \xrightarrow{d} \mathcal{N}(\mathbf{0}_{\dim(\boldsymbol{\theta})}, \mathbf{I}_{\dim(\boldsymbol{\theta})}) \quad (\text{A.2})$$

Now, what is left to be proved is

$$\begin{aligned} \frac{1}{n} \frac{\partial \log L_n(\tilde{\boldsymbol{\theta}}_n)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} &= \frac{1}{n} \frac{\partial \log L_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} + o_P(1) \\ &= \frac{1}{n} \mathbb{E} \left[ \frac{\partial \log L_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} \right] + o_P(1). \end{aligned} \quad (\text{A.3})$$

The second derivatives of  $\log L_n(\boldsymbol{\theta})$  with respect to  $\boldsymbol{\theta}$  are

$$\begin{aligned} \frac{\partial^2 \log L_n(\boldsymbol{\theta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\top} &= -\frac{1}{\sigma^2} \mathbf{M}_n(\boldsymbol{\varphi})^\top \mathbf{M}_n(\boldsymbol{\varphi}), \\ \frac{\partial^2 \log L_n(\boldsymbol{\theta})}{(\partial \sigma)^2} &= \frac{n}{2\sigma^4} - \frac{1}{\sigma^6} \boldsymbol{\epsilon}_n^\top \boldsymbol{\epsilon}_n, \\ \frac{\partial^2 \log L_n(\boldsymbol{\theta})}{\partial \boldsymbol{\rho}_j \partial \boldsymbol{\rho}_l} &= -\frac{1}{\sigma^2} \mathbf{Y}_n^\top \mathbf{W}_{ln}^\top \mathbf{W}_{jn} \mathbf{Y}_n \\ &\quad - \text{tr}(\mathbf{W}_{ln} \mathbf{S}_n(\boldsymbol{\rho})^{-1} \mathbf{W}_{jn} \mathbf{S}_n(\boldsymbol{\rho})^{-1}), \\ \frac{\partial^2 \log L_n(\boldsymbol{\theta})}{\partial \boldsymbol{\varphi} \partial \boldsymbol{\varphi}^\top} &= \frac{1}{\sigma^2} \left[ -\boldsymbol{\beta}^\top \left( \frac{\partial \mathbf{M}_n(\boldsymbol{\varphi})}{\partial \boldsymbol{\varphi}} \right)^\top \left( \frac{\partial \mathbf{M}_n(\boldsymbol{\varphi})}{\partial \boldsymbol{\varphi}} \right) \boldsymbol{\beta} \right. \\ &\quad \left. + \boldsymbol{\beta}^\top \left( \frac{\partial^2 \mathbf{M}_n(\boldsymbol{\varphi})}{\partial \boldsymbol{\varphi} \partial \boldsymbol{\varphi}^\top} \right)^\top \boldsymbol{\epsilon}_n \right], \\ \frac{\partial^2 \log L_n(\boldsymbol{\theta})}{\partial \sigma^2 \partial \boldsymbol{\beta}^\top} &= -\frac{1}{\sigma^4} \mathbf{M}_n(\boldsymbol{\varphi})^\top \boldsymbol{\epsilon}_n, \\ \frac{\partial^2 \log L_n(\boldsymbol{\theta})}{\partial \sigma^2 \partial \boldsymbol{\rho}_j} &= -\frac{1}{\sigma^4} \mathbf{Y}_n^\top \mathbf{W}_{jn}^\top \boldsymbol{\epsilon}_n, \\ \frac{\partial^2 \log L_n(\boldsymbol{\theta})}{\partial \sigma^2 \partial \boldsymbol{\varphi}^\top} &= -\frac{1}{\sigma^4} \boldsymbol{\beta}^\top \left( \frac{\partial \mathbf{M}_n(\boldsymbol{\varphi})}{\partial \boldsymbol{\varphi}} \right)^\top \boldsymbol{\epsilon}_n, \\ \frac{\partial^2 \log L_n(\boldsymbol{\theta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\rho}_j} &= -\frac{1}{\sigma^2} \mathbf{Y}_n^\top \mathbf{W}_{jn}^\top \mathbf{M}_n(\boldsymbol{\varphi}), \\ \frac{\partial^2 \log L_n(\boldsymbol{\theta})}{\partial \boldsymbol{\varphi} \partial \boldsymbol{\rho}_j} &= -\frac{1}{\sigma^2} \mathbf{Y}_n^\top \mathbf{W}_{jn}^\top \left( \frac{\partial \mathbf{M}_n(\boldsymbol{\varphi})}{\partial \boldsymbol{\varphi}} \right) \boldsymbol{\beta}, \end{aligned}$$



$$\frac{\partial^2 \log L_n(\boldsymbol{\theta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\varphi}^\top} = \frac{1}{\sigma^2} \left[ -\boldsymbol{\beta}^\top \left( \frac{\partial \mathbf{M}_n(\boldsymbol{\varphi})}{\partial \boldsymbol{\varphi}} \right)^\top \mathbf{M}_n(\boldsymbol{\varphi}) + \boldsymbol{\epsilon}_n^\top \left( \frac{\partial \mathbf{M}_n(\boldsymbol{\varphi})}{\partial \boldsymbol{\varphi}} \right) \right],$$

for  $l, j \in \{T, C\}$ . Let  $\mathbf{C}_{kn}(\boldsymbol{\rho}) = \mathbf{W}_{kn} \mathbf{S}_n(\boldsymbol{\rho})^{-1}$  for  $k \in \{T, C\}$ . Using the mean-value theorem, we have

$$\begin{aligned} & \frac{1}{n} \text{tr}(\mathbf{W}_{ln} \mathbf{S}_n(\tilde{\boldsymbol{\rho}}_n)^{-1} \mathbf{W}_{jn} \mathbf{S}_n(\tilde{\boldsymbol{\rho}}_n)^{-1}) \\ &= \frac{1}{n} \text{tr}(\mathbf{C}_{ln}(\tilde{\boldsymbol{\rho}}_n) \mathbf{C}_{jn}(\tilde{\boldsymbol{\rho}}_n)) \\ &= \frac{1}{n} \text{tr}(\mathbf{C}_{ln}(\boldsymbol{\rho}_0) \mathbf{C}_{jn}(\boldsymbol{\rho}_0)) + \frac{1}{n} (\tilde{\boldsymbol{\rho}}_n - \boldsymbol{\rho}_0)^\top \\ & \times \left[ \begin{array}{c} \dots \\ \mathbf{C}_{kn}(\bar{\boldsymbol{\rho}}_n) \left( \mathbf{C}_{ln}(\bar{\boldsymbol{\rho}}_n) \mathbf{C}_{jn}(\bar{\boldsymbol{\rho}}_n) + \mathbf{C}_{jn}(\bar{\boldsymbol{\rho}}_n) \mathbf{C}_{ln}(\bar{\boldsymbol{\rho}}_n) \right) \\ \dots \end{array} \right] \\ &= \frac{1}{n} \text{tr}(\mathbf{C}_{ln}(\boldsymbol{\rho}_0) \mathbf{C}_{jn}(\boldsymbol{\rho}_0)) + o_P(1), \end{aligned}$$

where  $\bar{\boldsymbol{\rho}}_n$  lies between  $\tilde{\boldsymbol{\rho}}_n$  and  $\boldsymbol{\rho}_0$ . By consistency,  $\hat{\boldsymbol{\theta}}_n = \boldsymbol{\theta}_0 + o_P(1)$ , thus  $\hat{\boldsymbol{\theta}}_n = \boldsymbol{\theta}_0 + o_P(1)$  and  $\bar{\boldsymbol{\rho}}_n = \boldsymbol{\theta}_0 + o_P(1)$ . The elements of  $\mathbf{C}_{kn}(\bar{\boldsymbol{\rho}}_n) \mathbf{C}_{ln}(\bar{\boldsymbol{\rho}}_n) \mathbf{C}_{jn}(\bar{\boldsymbol{\rho}}_n)$  and  $\mathbf{C}_{kn}(\bar{\boldsymbol{\rho}}_n) \mathbf{C}_{jn}(\bar{\boldsymbol{\rho}}_n) \mathbf{C}_{ln}(\bar{\boldsymbol{\rho}}_n)$  are  $O(1)$  and the elements of  $\mathbf{Y}_n^\top \mathbf{W}_{ln}^\top \mathbf{W}_{jn} \mathbf{Y}_n$  are  $O_P(n/h_n)$  by Lemma A7. Therefore

$$\frac{1}{n} \frac{\partial^2 \log L_n(\hat{\boldsymbol{\theta}}_n)}{\partial \boldsymbol{\rho}_j \partial \boldsymbol{\rho}_l} = \frac{1}{n} \frac{\partial^2 \log L_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\rho}_j \partial \boldsymbol{\rho}_l} + o_P(1).$$

For other partial derivative terms, we can use the fact that

$$\begin{aligned} \mathbf{M}_n(\tilde{\boldsymbol{\varphi}}_n) &= \mathbf{M}_n(\boldsymbol{\varphi}_0) + o(1), \\ \frac{\partial \mathbf{M}_n(\tilde{\boldsymbol{\varphi}}_n)}{\partial \boldsymbol{\varphi}} &= \frac{\partial \mathbf{M}_n(\boldsymbol{\varphi}_0)}{\partial \boldsymbol{\varphi}} + o(1), \\ \frac{\partial^2 \mathbf{M}_n(\tilde{\boldsymbol{\varphi}}_n)}{\partial \boldsymbol{\varphi} \partial \boldsymbol{\varphi}^\top} &= \frac{\partial^2 \mathbf{M}_n(\boldsymbol{\varphi}_0)}{\partial \boldsymbol{\varphi} \partial \boldsymbol{\varphi}^\top} + o(1), \end{aligned}$$

by Assumption 6 and the continuous mapping theorem [45]; the fact that elements of  $\mathbf{M}_n(\boldsymbol{\varphi})$ ,  $\partial \mathbf{M}_n(\boldsymbol{\varphi})/\partial \boldsymbol{\varphi}$  and  $\partial^2 \mathbf{M}_n(\boldsymbol{\varphi})/\partial \boldsymbol{\varphi} \partial \boldsymbol{\varphi}^\top$  are all  $O(1)$ ; and Lemma A5, A6 and A7. Similarly, we also can prove

$$\frac{1}{n} \frac{\partial \log L_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} + o_P(1) = \frac{1}{n} \mathbb{E} \left[ \frac{\partial \log L_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} \right] + o_P(1),$$

with

$$\begin{aligned} \mathbb{E} \left[ \frac{\partial^2 \log L_n(\boldsymbol{\theta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\top} \right] &= -\frac{1}{\sigma^2} \mathbf{M}_n(\boldsymbol{\varphi})^\top \mathbf{M}_n(\boldsymbol{\varphi}), \\ \mathbb{E} \left[ \frac{\partial^2 \log L_n(\boldsymbol{\theta})}{(\partial \sigma)^2} \right] &= -\frac{n}{2\sigma^4}, \\ \mathbb{E} \left[ \frac{\partial^2 \log L_n(\boldsymbol{\theta})}{\partial \boldsymbol{\rho}_j \partial \boldsymbol{\rho}_l} \right] &= -\frac{1}{\sigma^2} \boldsymbol{\beta}^\top \mathbf{M}_n(\boldsymbol{\varphi})^\top \mathbf{C}_{ln}(\boldsymbol{\rho})^\top \mathbf{C}_{jn}(\boldsymbol{\rho}) \mathbf{M}_n(\boldsymbol{\varphi}) \boldsymbol{\beta} \\ &\quad - \text{tr}(\mathbf{C}_{ln}(\boldsymbol{\rho})^\top \mathbf{C}_{jn}(\boldsymbol{\rho})) - \text{tr}(\mathbf{C}_{ln}(\boldsymbol{\rho}) \mathbf{C}_{jn}(\boldsymbol{\rho})), \end{aligned}$$

$$\mathbb{E} \left[ \frac{\partial^2 \log L_n(\boldsymbol{\theta})}{\partial \boldsymbol{\varphi} \partial \boldsymbol{\varphi}^\top} \right] = -\frac{1}{\sigma^2} \boldsymbol{\beta}^\top \left( \frac{\partial \mathbf{M}_n(\boldsymbol{\varphi})}{\partial \boldsymbol{\varphi}} \right)^\top \left( \frac{\partial \mathbf{M}_n(\boldsymbol{\varphi})}{\partial \boldsymbol{\varphi}} \right) \boldsymbol{\beta},$$

$$\mathbb{E} \left[ \frac{\partial^2 \log L_n(\boldsymbol{\theta})}{\partial \sigma^2 \partial \boldsymbol{\beta}^\top} \right] = \mathbf{0}_4,$$

$$\mathbb{E} \left[ \frac{\partial^2 \log L_n(\boldsymbol{\theta})}{\partial \sigma^2 \partial \boldsymbol{\rho}_j} \right] = -\frac{1}{\sigma^2} \text{tr}(\mathbf{C}_{jn}),$$

$$\mathbb{E} \left[ \frac{\partial^2 \log L_n(\boldsymbol{\theta})}{\partial \sigma^2 \partial \boldsymbol{\varphi}^\top} \right] = \mathbf{0}_p,$$

$$\mathbb{E} \left[ \frac{\partial^2 \log L_n(\boldsymbol{\theta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\rho}_j} \right] = -\frac{1}{\sigma^2} \boldsymbol{\beta}^\top \mathbf{M}_n(\boldsymbol{\varphi})^\top \mathbf{C}_{jn}(\boldsymbol{\rho})^\top \mathbf{M}_n(\boldsymbol{\varphi}),$$

$$\mathbb{E} \left[ \frac{\partial^2 \log L_n(\boldsymbol{\theta})}{\partial \boldsymbol{\varphi} \partial \boldsymbol{\rho}_j} \right] = -\frac{1}{\sigma^2} \boldsymbol{\beta}^\top \mathbf{M}_n(\boldsymbol{\varphi})^\top \mathbf{C}_{jn}(\boldsymbol{\rho})^\top \left( \frac{\partial \mathbf{M}_n(\boldsymbol{\varphi})}{\partial \boldsymbol{\varphi}} \right) \boldsymbol{\beta},$$

$$\mathbb{E} \left[ \frac{\partial^2 \log L_n(\boldsymbol{\theta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\varphi}^\top} \right] = -\frac{1}{\sigma^2} \boldsymbol{\beta}^\top \left( \frac{\partial \mathbf{M}_n(\boldsymbol{\varphi})}{\partial \boldsymbol{\varphi}} \right)^\top \mathbf{M}_n(\boldsymbol{\varphi}),$$

where  $p$  is the number of parameters in  $\boldsymbol{\varphi}$ . This completes the proof for (A.3). Finally, from (A.2), (A.3), and Slutsky's theorem, Theorem 1 is proved.

## FUNDING

This work was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) by way of Grants RGPIN-2019-04212 and RGPIN-2023-03245.

*Accepted 6 April 2023*

## REFERENCES

- [1] ADVANI, A. and MALDE, B. (2018). Methods to identify linear network models: a review. *Swiss Journal of Economics and Statistics* **154**(1) 12.
- [2] ARONOW, P. M., SAMII, C. et al. (2017). Estimating average causal effects under general interference, with application to a social network experiment. *The Annals of Applied Statistics* **11**(4) 1912–1947. <https://doi.org/10.1214/16-AOAS1005>. MR3743283
- [3] BANERJEE, S., CARLIN, B. P. and GELFAND, A. E. (2003). *Hierarchical modeling and analysis for spatial data*. Chapman and Hall/CRC. MR3362184
- [4] BASSE, G. and FELLER, A. (2018). Analyzing two-stage experiments in the presence of interference. *Journal of the American Statistical Association* **113**(521) 41–55. <https://doi.org/10.1080/01621459.2017.1323641>. MR3803438
- [5] BASSE, G. W. and AIROLDI, E. M. (2018). Model-assisted design of experiments in the presence of network-correlated outcomes. *Biometrika* **105**(4) 849–858. <https://doi.org/10.1093/biomet/asy036>. MR3877869
- [6] BOWERS, J., DESMARAIS, B. A., FREDERICKSON, M., ICHINO, N., LEE, H.-W. and WANG, S. (2018). Models, methods and network topology: experimental design for the study of interference. *Social Networks* **54** 196–208.
- [7] BRAMOULLÉ, Y., DJEBBARI, H. and FORTIN, B. (2009). Identification of peer effects through social networks. *Journal of Econometrics* **150**(1) 41–55. <https://doi.org/10.1016/j.jeconom.2008.12.021>. MR2525993
- [8] CANDOGAN, O., CHEN, C. and NIAZADEH, R. (2021). Correlated cluster-based randomized Experiments: robust Variance Minimization. *Chicago Booth Research Paper No. 21-17*.
- [9] CHIN, A. (2019). Regression adjustments for estimating the global treatment effect in experiments with interference. *Journal*

- of *Causal Inference* **7**(2). <https://doi.org/10.1515/jci-2018-0026>. MR4350071
- [10] CLIFF, A. D. (1981). *Spatial processes: models & applications*. Pion, London. MR0632256
- [11] COX, D. R. (1958). *Planning of experiments*. Wiley. MR0095561
- [12] ECKLES, D., KARRER, B. and UGANDER, J. (2016). Design and analysis of experiments in networks: reducing bias from interference. *Journal of Causal Inference* **5**(1). <https://doi.org/10.1515/jci-2015-0021>. MR4323809
- [13] FEDOROV, V. V. and LEONOV, S. L. (2013). *Optimal design for nonlinear response models*. CRC Press. MR3113978
- [14] FRUCHTERMAN, T. M. and REINGOLD, E. M. (1991). Graph drawing by force-directed placement. *Software: Practice and Experience* **21**(11) 1129–1164.
- [15] GOSSEN, H. H. (1983). *The laws of human relations and the rules of human action derived therefrom*. MIT Press.
- [16] GUI, H., XU, Y., BHASIN, A. and HAN, J. (2015). Network A/B testing: from sampling to estimation. In *Proceedings of the 24th International Conference on World Wide Web* 399–409.
- [17] GUPTA, S., KOHAVI, R., TANG, D., XU, Y., ANDERSEN, R., BAKSHY, E., CARDIN, N., CHANDRAN, S., CHEN, N., COEY, D., CURTIS, M., DENG, A., DUAN, W., FORBES, P., FRASCA, B., GUY, T., IMBENS, G. W., SAINT JACQUES, G., KANTAWALA, P., KATSEV, I., KATZWER, M., KONUTGAN, M., KUNAKOVA, E., LEE, M., LEE, M., LIU, J., MCQUEEN, J., NAJMI, A., SMITH, B., TREHAN, V., VERMEER, L., WALKER, T., WONG, J. and YASHKOV, I. (2019). Top challenges from the first practical online controlled experiments summit. *ACM SIGKDD Explorations Newsletter* **21**(1) 20–35. <https://doi.org/10.1145/3331651.3331655>
- [18] HARVILLE, D. A. (1998). *Matrix algebra from a statistician's perspective*. Taylor & Francis. <https://doi.org/10.1007/b98818>. MR1467237
- [19] HORN, R. A. and JOHNSON, C. R. (2012). *Matrix analysis*. Cambridge University Press. MR2978290
- [20] HUDGENS, M. G. and HALLORAN, M. E. (2008). Toward causal inference with interference. *Journal of the American Statistical Association* **103**(482) 832–842. <https://doi.org/10.1198/016214508000000292>. MR2435472
- [21] JENNRICH, R. I. (1969). Asymptotic properties of non-linear least squares estimators. *The Annals of Mathematical Statistics* **40**(2) 633–643. <https://doi.org/10.1214/aoms/1177697731>. MR0238419
- [22] JOCHMANS, K. and WEIDNER, M. (2019). Fixed-Effect Regressions on Network Data. *Econometrica* **87**(5) 1543–1560. <https://doi.org/10.3982/ECTA14605>. MR4021456
- [23] KARRER, B., SHI, L., BHOLE, M., GOLDMAN, M., PALMER, T., GELMAN, C., KONUTGAN, M. and SUN, F. (2021). Network experimentation at scale. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining* 3106–3116.
- [24] KELEJIAN, H. H. and PRUCHA, I. R. (2001). On the asymptotic distribution of the Moran I test statistic with applications. *Journal of Econometrics* **104**(2) 219–257. [https://doi.org/10.1016/S0304-4076\(01\)00064-1](https://doi.org/10.1016/S0304-4076(01)00064-1). MR1864417
- [25] KELEJIAN, H. H. and PRUCHA, I. R. (2010). Specification and estimation of spatial autoregressive models with autoregressive and heteroskedastic disturbances. *Journal of Econometrics* **157**(1) 53–67. <https://doi.org/10.1016/j.jeconom.2009.10.025>. MR2652278
- [26] KOUTRA, V., GILMOUR, S. G. and PARKER, B. M. (2021). Optimal block designs for experiments on networks. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **70**(3) 596–618. <https://doi.org/10.1111/rssc.12473>. MR4275838
- [27] LA VIGNE, N. G., LOWRY, S. S., MARKMAN, J. A. and DWYER, A. M. (2011). *Evaluating the use of public surveillance cameras for crime control and prevention*. Washington, DC: US Department of Justice, Office of Community Oriented Policing Services. Urban Institute, Justice Policy Center.
- [28] LEE, L.-F. (2004). Asymptotic distributions of quasi-maximum likelihood estimators for spatial autoregressive models. *Econometrica* **72**(6) 1899–1925. <https://doi.org/10.1111/j.1468-0262.2004.00558.x>. MR2095537
- [29] LI, T., LEVINA, E. and ZHU, J. (2019). Prediction models for network-linked data. *The Annals of Applied Statistics* **13**(1) 132–164. <https://doi.org/10.1214/18-AOAS1205>. MR3937424
- [30] LYNN, C. W. and BASSETT, D. S. (2019). The physics of brain network structure, function and control. *Nature Reviews Physics* **1**(5) 318–332.
- [31] MANSKI, C. F. (1993). Identification of endogenous social effects: the reflection problem. *The Review of Economic Studies* **60**(3) 531–542. <https://doi.org/10.2307/2298123>. MR1236836
- [32] NELDER, J. A. and MEAD, R. (1965). A simplex method for function minimization. *The Computer Journal* **7**(4) 308–313. <https://doi.org/10.1093/comjnl/7.4.308>. MR3363409
- [33] PALUCK, E. L., SHEPHERD, H. and ARONOW, P. M. (2016). Changing climates of conflict: a social network experiment in 56 schools. *Proceedings of the National Academy of Sciences* **113**(3) 566–571.
- [34] PANGER, G. (2016). Reassessing the Facebook experiment: critical thinking about the validity of big data research. *Information, Communication & Society* **19**(8) 1108–1126.
- [35] PARKER, B. M., GILMOUR, S. G. and SCHORMANS, J. (2017). Optimal design of experiments on connected units with application to social networks. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* **66**(3) 455–480. <https://doi.org/10.1111/rssc.12170>. MR3632337
- [36] POKHILKO, V., ZHANG, Q., KANG, L. and MAYS, D. P. (2019). D-optimal design for network A/B testing. *Journal of Statistical Theory and Practice* **13**(4) 1–23. <https://doi.org/10.1007/s42519-019-0058-3>. MR4021855
- [37] PUKELSHEIM, F. (2006). *Optimal design of experiments*. SIAM. <https://doi.org/10.1137/1.9780898719109>. MR2224698
- [38] ROSSI, R. A. and AHMED, N. K. (2015). The network data repository with interactive graph analytics and visualization. In *AAAI*. <https://networkrepository.com>.
- [39] SAINT-JACQUES, G., VARSHNEY, M., SIMPSON, J. and XU, Y. (2019). Using Ego-Clusters to Measure Network Effects at LinkedIn. *arXiv preprint arXiv:1903.08755*.
- [40] SHALIZI, C. R. and THOMAS, A. C. (2011). Homophily and contagion are generically confounded in observational social network studies. *Sociological Methods & Research* **40**(2) 211–239. <https://doi.org/10.1177/0049124111404820>. MR2767833
- [41] TCHETGEN, E. J. T. and VANDERWEELE, T. J. (2012). On causal inference in the presence of interference. *Statistical Methods in Medical Research* **21**(1) 55–75. <https://doi.org/10.1177/0962280210386779>. MR2867538
- [42] TRAUD, A. L., MUCHA, P. J. and PORTER, M. A. (2012). Social structure of Facebook networks. *Journal of Physics A* **391**(16) 4165–4180.
- [43] TRAUD, A. L., KELSIC, E. D., MUCHA, P. J. and PORTER, M. A. (2011). Comparing community structure to characteristics in online collegiate social networks. *SIAM Review* **53**(3) 526–543. <https://doi.org/10.1137/080734315>. MR2834086
- [44] UGANDER, J., KARRER, B., BACKSTROM, L. and KLEINBERG, J. (2013). Graph cluster randomization: network exposure to multiple universes. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 329–337.
- [45] VAN DER VAART, A. W. (2000). *Asymptotic statistics* **3**. Cambridge University Press. <https://doi.org/10.1017/CBO9780511802256>. MR1652247
- [46] WHITE, H. (1996). *Estimation, inference and specification analysis*. Cambridge University Press. <https://doi.org/10.1017/COL0521252806>. MR1292251
- [47] XU, Y., CHEN, N., FERNANDEZ, A., SINNO, O. and BHASIN, A. (2015). From infrastructure to culture: A/B testing challenges in large scale social networks. In *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 2227–2236.

Trang Bui. Department of Statistics and Actuarial Science, University of Waterloo, Canada.

E-mail address: [tqtbui@uwaterloo.ca](mailto:tqtbui@uwaterloo.ca)

Stefan H. Steiner. Department of Statistics and Actuarial Science, University of Waterloo, Canada.

E-mail address: [shsteiner@uwaterloo.ca](mailto:shsteiner@uwaterloo.ca)

Nathaniel T. Stevens. Department of Statistics and Actuarial Science, University of Waterloo, Canada.

E-mail address: [nstevens@uwaterloo.ca](mailto:nstevens@uwaterloo.ca)