Causal Inference for Social Network Data

ST 790 Project Report

Netra Prabhu, Ananya Roy

North Carolina State University

Date: December 7, 2024

Contents

1 Background

Social networks are not only a different type of data structure, but also the one where behavior, treatment assignment, belief or outcome of a given random effect depends on that of other related random effects. For the purpose of causal inference such data is complicated to analyze because the majority of established statistical methods, including the generalized linear models GLM or the generalized estimating equations GEE do not usually incorporate the network structure dependencies. This has in fact given rise to specific techniques that are designed specifically for usage in social networks. Causal inference in social networks

Figure 1: Social Network (Source: iStock)

addresses two main challenges: Interaction (the impact that one subject's treatment has on the other's results) and hidden dependence (unidentified common factors that affect the other connected subject's results). While previous work mainly tested theories about how treatments or outcomes are transferred through ties, recent developments also include dependence on potential 'overlapping variables', which was assumed to be present in pragmatic networks.

The basis for this paper dwells on works like Framingham Heart Study (FHS), which focused on causal relationships between social contacts and health statuses including obesity and happiness. These studies revealed that more refined approaches, which are able to address the dependency and confounding issues inherent to network data, were required.

The recent techniques employ causal structural equation models (SEM) and directed acyclic graphical models (DAG) to facilitate the definition as well as estimation of causal effects. These frameworks make it possible to work with individual connected networks and estimate causal effects, it is no longer necessary to have multiple independent networks, or always randomly applied treatments.

In this project, we review approaches to estimate identification of cross-sectional causal peer effects and pathways of treatment on networks. The goal is to help secure a comprehensive arsenal of techniques

to perform causal inference with the observational social network data, while increasing the breadth and certainty of conclusions derived from them.

2 Preliminary

Before we proceed, let's briefly introduce some notations and terminologies used throughout this report. They are as follows:

2.1 Network Structure

A network consists of *nodes* (individuals or units) and *edges* (ties) between them. Nodes, indexed as $i \in \{1, 2, \ldots, n\}$, are connected via an adjacency matrix A, where $A_{ij} = 1$ indicates a tie between nodes i and j, and $A_{ij} = 0$ otherwise. The *degree* of node i, $K_i = \sum_{j=1}^{n} A_{ij}$, is the total number of its connections. The *alters* of node i are its neighbors, defined as $\{j : A_{ij} = 1\}$. These elements collectively describe the network's structure.

2.2 Observed Data

For each node i, the observed data O_i includes three components: the *outcome* (Y_i), representing a variable of interest (e.g., health status or behavior); *covariates* (C_i), describing baseline characteristics of the node; and the *treatment* (X_i), indicating an exposure or intervention applied to node i. Together, these components characterize the observed data for each node in the network.

2.3 Potential Outcomes Framework

The *counterfactual outcome* for node i, $Y_i(x^*)$, represents the outcome under a hypothetical treatment $X = x^*$ for the entire network. The *network-wide potential outcome*, $\bar{Y}_n(x^*) = \frac{1}{n}\sum_{i=1}^{n} Y_i(x^*)$, is the average outcome across all nodes under treatment x^*.

2.4 Interventions

Interventions can be categorized as static, dynamic, or stochastic. In static interventions, $X_i = x^*$ is fixed for all i. Dynamic interventions assign X_i based on a deterministic function of i's covariates, $g(C_i)$. Stochastic interventions assign X_i according to a probability distribution dependent on C_i.

2.5 Summary Functions

The network structure is described by an adjacency matrix A, where A_{ij} indicates ties between nodes i and j. Summary measures capture individual and neighborhood characteristics. For node i, $W_i = s_{C,i}(\{C_j : A_{ij} = 1\})$ summarizes covariates of i and its neighbors, while $V_i = s_{X,i}(\{X_j : A_{ij} = 1\})$ summarizes treatments. For instance, $s_{C,i}(\{C_j : A_{ij} = 1\}) = (C_i, \sum_{j:A_{ij}=1} C_j)$ implies outcomes depend on i's covariates and the sum of neighbors' covariates. The conditional expectation $m(v, w)$ represents the expected outcome Y_i given $V_i = v$ and $W_i = w$.

2.6 Assumptions

The positivity assumption ensures all values of X_i have positive probability for each observed W_i. No unmeasured confounding implies that covariates C adequately control for confounding between X and Y. Latent variable dependence allows for unobserved similarities among nodes, with dependencies limited to a distance of two ties.

2.7 Estimation Targets

Key estimation targets include causal estimands such as $\mathbb{E}[\bar{Y}_n(x^*)]$, the expected average outcome under treatment x^*. The influence function ϕ characterizes the first-order behavior of estimators, aiding in evaluating robustness and efficiency.

3 Setup and Assumptions

3.1 Setup

The basic setup of the observed data and the structural causal model forms the foundation of causal inference in social networks. These involve identifying the kind of network representation, the different observed variables and some key concepts concerning treatments and results.

3.2 Structural Equation Models (SEMs)

The relationships between covariates (C), treatments (X), and outcomes (Y) are represented using Structural Equation Models (SEMs):

3.2.1 Normal Case

The normal case models the direct and indirect effects of covariates and treatments without simplifying the network structure:

$$C_i = f_C(\epsilon_{C_i}),$$
$$X_i = f_X\left(\{C_j : A_{ij} = 1\}, \epsilon_{X_i}\right),$$
$$Y_i = f_Y\left(\{X_j : A_{ij} = 1\}, \{C_j : A_{ij} = 1\}, \epsilon_{Y_i}\right),$$

where: - f_C, f_X, f_Y are unknown functions describing covariate generation, treatment assignment, and outcome determination. $\epsilon_X, \epsilon_Y, \epsilon_Z$ represents unobserved random variables or noise terms.

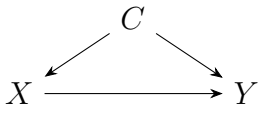


Figure 2: DAG representing the normal SEM case in a social network.

3.2.2 Case with Summary Functions

This structural model equations setup described although capable of identifying causal estimands under nonparametric setup. However, to avoid complexity under dynamic and stochastic interventions we simplify the structural equations. This is done by considering summary functions W_i and V_i (as mentioned earlier):

$$C_i = f_C(\epsilon_{C_i}),$$
$$X_i = f_X\left(W_i, \epsilon_{X_i}\right),$$
$$Y_i = f_Y\left(V_i, W_i, \epsilon_{Y_i}\right),$$

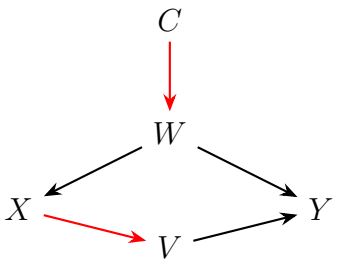


Figure 3: DAG representing the SEM case with summary functions.

This approach simplifies the SEM while preserving the network's causal structure.

3.3 Assumptions

Along with the assumptions of *Conditional Ignorability* and *No Unmeasured Confounding*, the Structural Equation Model also features the following assumptions -

- **(A1)** The vectors $(\epsilon_{X_1}, \ldots, \epsilon_{X_n})$, $(\epsilon_{Y_1}, \ldots, \epsilon_{Y_n})$, and $(\epsilon_{C_1}, \ldots, \epsilon_{C_n})$ are independent.
- **(A2a)** $\epsilon_{X_1}, \ldots, \epsilon_{X_n}$ are identically distributed, and $\epsilon_{Y_1}, \ldots, \epsilon_{Y_n}$ are identically distributed.
- **(A2b)** $\epsilon_{X_i} \perp \epsilon_{X_j}$ and $\epsilon_{Y_i} \perp \epsilon_{Y_j}$ for i, j such that $A_{ij} = 0$ and $\exists! k$ with $A_{ik} = A_{kj} = 1$.
- **(A3a)** $\epsilon_{C_1}, \ldots, \epsilon_{C_n}$ are identically distributed.
- **(A3b)** $\epsilon_{C_i} \perp \epsilon_{C_j}$ for i, j such that $A_{ij} = 0$ and $\exists! k$ with $A_{ik} = A_{kj} = 1$.

Here, the assumptions (A2b) and (A3b) help us incorporate the latent variable dependence up to a distance of two network ties relaxing the typical assumption of independent errors. Note that, the assumption (A3) can be omitted if our attention is restricted to causal effects conditional on C.

Although, some estimation strategies can only be implemented for stronger versions of assumptions (A2b) and (A3b), which can be written as follows -

- **(A4)** $\epsilon_{X_1}, \ldots, \epsilon_{X_n} \mid C$ are i.i.d., and $\epsilon_{Y_1}, \ldots, \epsilon_{Y_n} \mid C, X$ are i.i.d.
- **(A5)** $\epsilon_{C_i}, i = 1, \ldots, n$, are i.i.d.

4 Estimation Procedures

In estimation and inference for causal effects in social networks, we have a statistical model $M = \mathrm{P}(O)$ that aims to approximate the true data generating process $P(O)$. The estimation procedures build on some structural assumptions and factorization conditions for causal inference when network dependencies exist.

4.1 Modeling the Observed Data

Thus the statistical model M defined a set of distributions over the observed data O. These distributions satisfy the following factorization, as derived from the structural assumptions:

$$P(O = o) = p_C(c) \cdot g(x|w) \cdot p_Y(y|v, w),$$

where:

- $p_C(c)$: Marginal distribution of covariates (C).

- $g(x|w)$: Propensity score model, describing the conditional probability of treatment assignment given covariates (W).
- $p_Y(y|v,w)$: Outcome model, representing the conditional probability of outcomes given treatments (V) and covariates (W).

4.2 Decomposition of the Model

The statistical model is decomposed into two independent components:

- M_h: Collection of conditional distributions $h(v|w)$ for treatments given covariates.
- M_m: Collection of conditional expectations $m(v,w)$ for outcomes given treatments and covariates.

This decomposition ensures flexibility by allowing the estimation of treatment and outcome models independently without inducing restrictions on the marginal distribution $p_C(c)$.

4.3 Regularity and Estimation Challenges

- Nonparametric estimation of $h(v|w)$ and $m(v,w)$ is feasible under iid data but may not achieve sufficient convergence rates for dependent network data.
- Parametric models are often employed to estimate h and m, ensuring the regularity conditions required for asymptotic normality of the estimators.

4.4 Influence Function-Based Estimation

The identifying functional ψ_n for causal inference can be expressed using influence functions ([4]). The sample average of the influence function provides a robust and doubly robust estimator:

$$Dn(o) = \frac{1}{n}\sum_{i=1}^{n}\left(m(v_i,w_i) + \frac{h^*(v_i,w_i)}{h(v_i,w_i)}\left(y_i - m(v_i,w_i)\right) - \psi_n\right),$$

where $h^*(v,w)$ is the induced density under the specified intervention, and $h(v,w)$ is the observed data density.

4.5 Targeted Maximum Likelihood Estimation (TMLE)

Targeted Maximum Likelihood Estimation (TMLE) is a general template for estimating smooth parameters in semi- and nonparametric models. The estimation algorithm is designed to solve an influence function estimating equation, achieving equivalence with the standard estimating equation approach. TMLE integrates three main components:

- **A loss function L:** Used to model the outcome regression m.
- **Initial working estimators:** $\hat{m}$ for m and $\hat{h}$ for h, which are treated as nuisance models.
- **Parametric submodel m_ϵ:** Derived from the primary model m, with the score of m_ϵ tied to the influence function $D_n(o)$. The submodel satisfies $m_{\epsilon=0} = m(\cdot)$.

The TMLE algorithm iteratively updates the outcome regression model $\hat{m}$ until the influence function estimating equation is solved. A parametric submodel adjusts the predictions of $\hat{m}$ through weighted optimization of the loss function. At convergence, the final estimator is computed as the solution to the estimating equation.

4.5.1 The TMLE Algorithm

The TMLE of ψ_n is computed as follows:

- **Step 1: Auxiliary Weights**

  Define auxiliary weights H_i as the ratio of estimated densities:

  $$H_i = \frac{\bar{h}^*(V_i^*, W_i)}{\bar{h}(V_i, W_i)},$$

  where $\bar{h}(V_i, W_i)$ is the conditional mixture density of observed covariates V_i and W_i, while $\bar{h}^*(V_i^*, W_i)$ is the density under the counterfactual distribution V_i^*. These weights reweight the observed data to mimic a counterfactual intervention.

- **Step 2: Initial Predictions**

  Compute the initial predicted outcome values $\hat{Y}_i = \hat{m}(V_i, W_i)$ and predicted counterfactual outcomes $\hat{Y}_i^* = \hat{m}(V_i^*, W_i)$, where V_i^* is the degenerate random variable corresponding to the intervention $s_{X,i}(x^*)$.

- **Step 3: TMLE Model Update**

  Construct a parametric submodel update for $\hat{m}$ by performing a weighted intercept-only logistic regression. The weights H_i are used to adjust the fit of $\hat{m}$, with $\hat{Y}_i$ as the offset:

  $$\text{logit } \hat{m}_\epsilon(V, W) = \text{logit } \hat{m}(V, W) + \epsilon,$$

  where ϵ is the intercept parameter estimated from the weighted regression. $\text{logit}(x) = \log\left(\frac{x}{1-x}\right)$ is the logit function.

- **Step 4: Updated Predictions**

  Using the updated submodel, compute the updated predicted outcomes:

$$\hat{Y}_i^* = \text{expit}\left(\text{logit}\ \hat{Y}_i^* + \epsilon\right),$$

  where $\text{expit}(x) = \frac{1}{1+e^{-x}}$ is the inverse logit function.

- **Step 5: Final TMLE Estimate**

  The TMLE estimate of ψ_n is obtained as:

$$\hat{\psi}_n = \frac{1}{n}\sum_{i=1}^{n} \hat{Y}_i^*.$$

At the end of the iterative procedure, the influence function estimating equation is solved, ensuring consistency and asymptotic efficiency of the TMLE.

4.5.2 Properties of TMLE

Doubly Robust: The TMLE estimator is doubly robust, meaning it remains consistent for ψ_n if either the working model $\hat{h}$ for h (e.g., $\bar{h}$ and $\bar{h}^*$) is correctly specified, or the working model $\hat{m}$ for m is correctly specified. This robustness ensures reliability even when one of the nuisance models is misspecified.

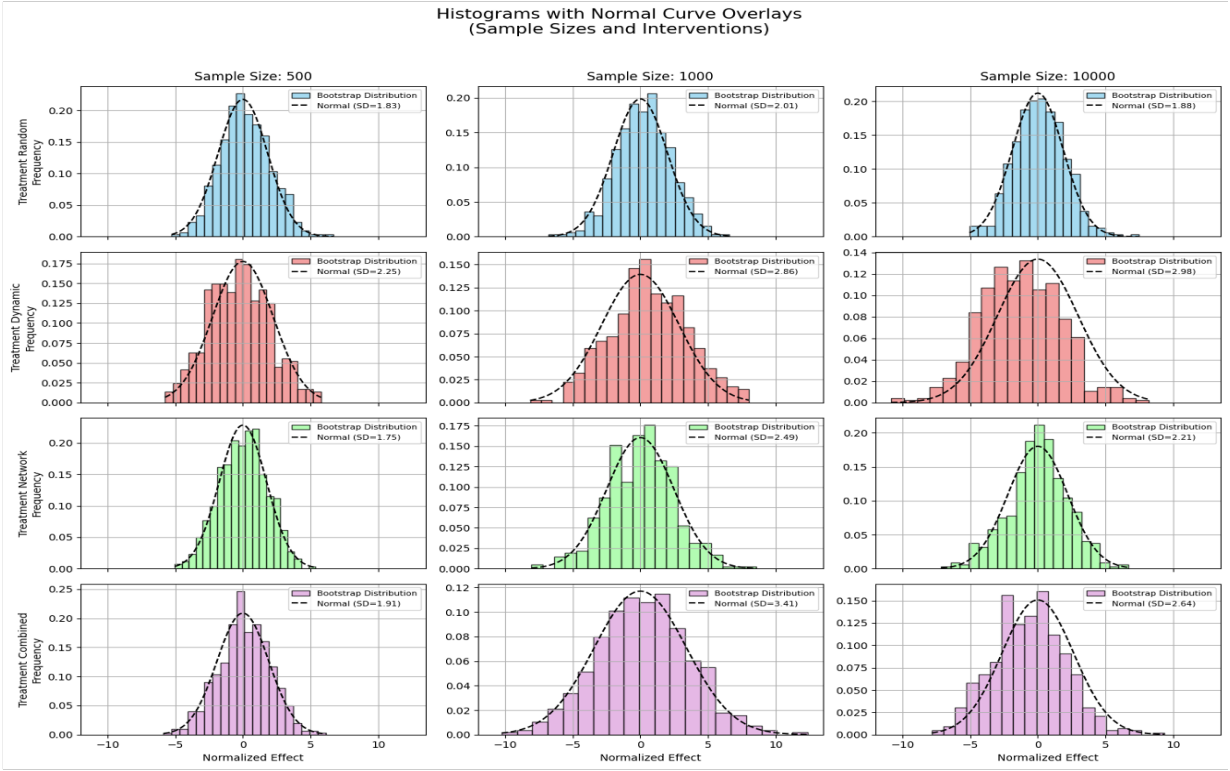
Asymptotic Properties: TMLE is CAN (consistent and asymptotically normal) for ψ_n. $\hat{\psi}_n$ converges to a normal limiting distribution under assumptions A2, A3, A4 and A5, provided at least one of the nuisance models m or h is correctly specified and converges at the rate $\sqrt{C_n}$, or both models converge at slower but compatible rates.

Theorem 1. *Suppose that* $\frac{K_{\max,n}^2}{n} \to 0$ *as* $n \to \infty$*, where* $K_{\max,n} = \max_i\{K_i\}$ *for network size* n*. Under assumptions* $(A2)$*,* $(A3)$*,* $(A6)$*,* $(A7)$*, and regularity conditions,*

$$\sqrt{C_n}(\hat{\psi}_n - \psi_n) \xrightarrow{d} N(0, \sigma^2),$$

for some finite σ^2 *and for some* C_n *such that* $\frac{n}{K_{\max,n}^2} \leq C_n \leq n$*.*

It achieves the efficiency bound by leveraging the influence function to minimize estimation bias and variance. We verify this theorem with simulation. The results are as follows:



4.6 Direct Estimation of $\bar{h}$ and $\bar{h}^*$

A direct estimation approach for $\bar{h}$ optimizes the log-likelihood $\sum_{i=1}^{n} \log \bar{h}(V_i|W_i)$, as if (V_i, W_i) were independent and identically distributed (i.i.d.). Similarly, $\bar{h}^*$ can be estimated by creating a counterfactual sample (V_i^*, W_i) and optimizing $\sum_{i=1}^{n} \log \bar{h}^*(V_i^*|W_i)$.

For conditional mixture densities, methods such as conditional histograms or logistic regression are employed to estimate $\bar{h}$. Despite challenges posed by network dependencies, direct estimation remains computationally efficient and practically viable.

4.7 Intervention

Under dynamic and stochastic interventions, the intervention distribution $h^*(v \mid w)$ replaces the original $h(v \mid w)$, and potential outcomes are identified under no unmeasured confounding and positivity assumptions. For stochastic interventions, the potential outcomes are defined as:

$$E[\bar{Y}_n^*] = \frac{1}{n} \sum_{i=1}^{n} E[Y_i^*] = \frac{1}{n} \sum_{i=1}^{n} \int_c E[Y \mid V_i^*, W_i = s_{C,i}(c)] p_C(c).$$

This expression can be further expanded using the intervention distribution $h_i^*(v, w)$ as:

$$E[\bar{Y}_n^*] = \frac{1}{n} \sum_{i=1}^{n} \int_w \int_v m(v, w) h_i^*(v, w) \, dv \, dw,$$

where $m(v, w) = E[Y \mid V = v, W = w]$ is the outcome regression.

For static interventions, $h_i^*(v, w)$ assigns mass to a single value $v = v_i^*$, reducing to $E[\bar{Y}_n^*] = \frac{1}{n}\sum_{i=1}^n m(v_i^*, W_i)$. The TMLE algorithm uses these potential outcomes, with the influence function $D_n(o)$ modified only by the form of h^*. In the stochastic case, the weights for the intervention are defined as $H_i = \frac{h^*(V_i^*, W_i)}{h(V_i, W_i)}$. The estimation remains consistent across static, dynamic, and stochastic interventions as long as the conditional support of V^* lies within that of V, ensuring positivity.

Stochastic Intervention SEM

The intervention SEM modifies the structural equations as:

$$C_i = f_C(\epsilon_{C_i}), \quad X_i^* = f_X(W_{X,i}^*, \epsilon_{X_i}), \quad Y_i^* = f_Y(W_i, V_i^*, \epsilon_{Y_i}),$$

where $W_{X,i}^* = s_{C,X,i}^*(C), \quad V_i^* = s_{X,i}^*(X_i^*)$. The probabilities are defined as $P(X = x^* \mid W = s_{C,X,i}^*(C))$, and $P(V = v \mid W = s_{C,X,i}^*(C)) = P(X \in \{x^* : s_{X,i}^*(x^*) = v\} \mid W = s_{C,X^*}(C))$. Under these adjustments, potential outcomes are derived as:

$$E[\bar{Y}_n^*] = \frac{1}{n}\sum_{i=1}^n \int_{w,v} m(v, w) h_i^*(v, w)\, dv\, dw.$$

The intervention-induced distribution h^* is crucial for ensuring the validity of estimation results, and its specification directly impacts the resulting estimators.

5 Results and Simulation Study

This simulation study enables the estimation of treatment effects under a range of intervention approaches in large network contexts. It compares the mean treatment effect, its standard deviation, and the corresponding confidence intervals, for various sample sizes, types of network, and methods of treatment assignment. The aim is to investigate how network structure and treatments affect causal treatment effect estimation under TMLE and various variance estimation techniques.

Two network models are considered to represent realistic connectivity patterns. First, preferential attachment networks are generated using the Barabási–Albert model, which produces scale-free networks characterized by the presence of hubs with a high degree of connectivity. These networks are representative of social and biological systems. Second, small-world networks are created using the Newman–Watts–Strogatz algorithm, which captures high clustering and short average path lengths, often observed in transportation or neural networks.

For each network, node-level covariates are simulated to reflect individual characteristics, including continuous, uniform, and binary features. A latent variable representing unobserved node traits is also introduced. Treatment assignments are performed using six strategies: random assignment, dynamic intervention based on covariate thresholds, network-based intervention influenced by neighboring nodes, combinations of dynamic and network-based strategies, stochastic assignment based on baseline covariate probabilities, and a combined stochastic network intervention. The outcome is modeled as a linear combination of treatments, covariates, and network influence, with added Gaussian noise to introduce variability.

TMLE is used to estimate the causal effects of treatments since it handles confounding and minimizes bias. The TMLE framework involves estimating treatment probabilities (propensity scores) and outcome, followed by a targeting step to improve the effect estimate. Variance is estimated using three methods: the dependent influence curve (IC) approach, bootstrap and an IID based method.

Simulations are conducted on networks with 500, 1000, and 10,000 nodes to assess the impact of sample size. Both preferential attachment and small-world networks are evaluated for all six treatment assignment strategies. For each scenario, the TMLE effect estimate, variance, and 95% confidence intervals are computed using the three variance estimation methods.

We can observe from the following figure that TMLE procedure (dependent_ic) estimates the variances more accurately by taking into account the network dependence while other methods have tighter CIs as they ignore the network dependence. As the network size increases, the variance in case of dependent_ic decreases which is expected. IID method completely ignores network dependence.

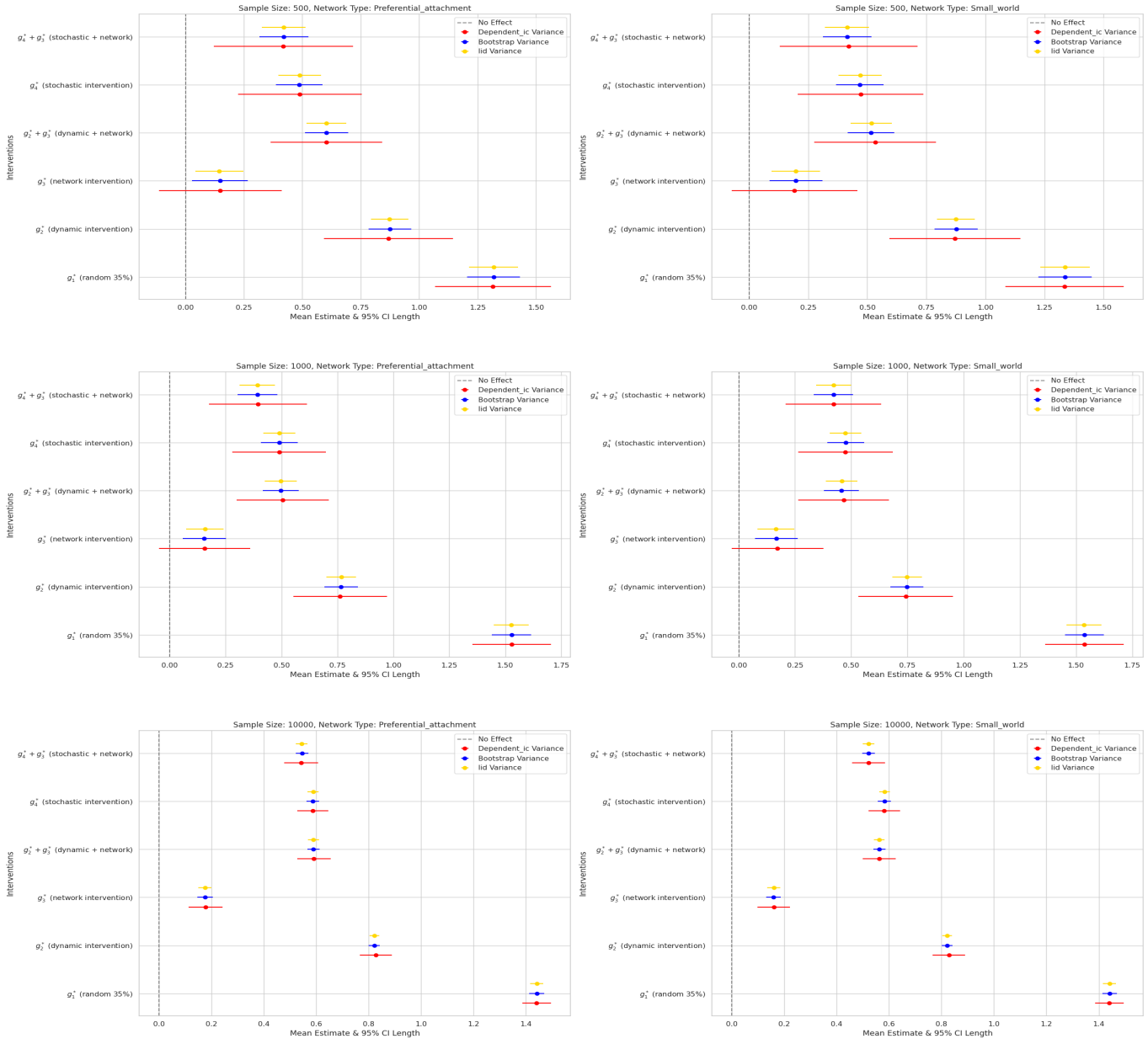


Figure 4: Mean 95% CI length and coverage for TMLE in preferential attachment network and small world network

6 Too Many Friends, Too Much Influence

We tried to analyze "Too Many Friends, Too Much Influence" to evaluate the effects of various interventions on network dynamics in Preferential Attachment and Small-World networks generated incorporating hub influence. The figures illustrate the mean estimates and 95% confidence intervals (CI) for different intervention strategies—random (g_1^*), dynamic (g_2^*), network-based (g_3^*), combined dynamic and network-based ($g_2^* + g_3^*$), and high-degree (g_5^*)—under three variance estimation methods (Dependent_ic, Bootstrap, and iid).

As seen in the Preferential Attachment networks, all the high-degree interventions (g_5^*) hit harder because of the nature of graph. These hubs escalate the influence application, resulting in massive ripple impact. The interventions diffuse influences more rapidly in Small-World networks because the network is tightly connected and has rather small distances between two nodes, wherein both dynamic ($g_2^* + g_3^*$) network-based interventions were also significant. Variance of Dependent_ic is higher, giving wider bands of confidence, while revealing dependencies networks underestimated by iid models. These results point toward the necessity of incorporating variance estimation techniques that generalize from patterns of dependencies and implementing interventions based on the specific structures of the defined networks. However, if the network size increases rapidly, the procedure may not estimate these causal effects correctly.

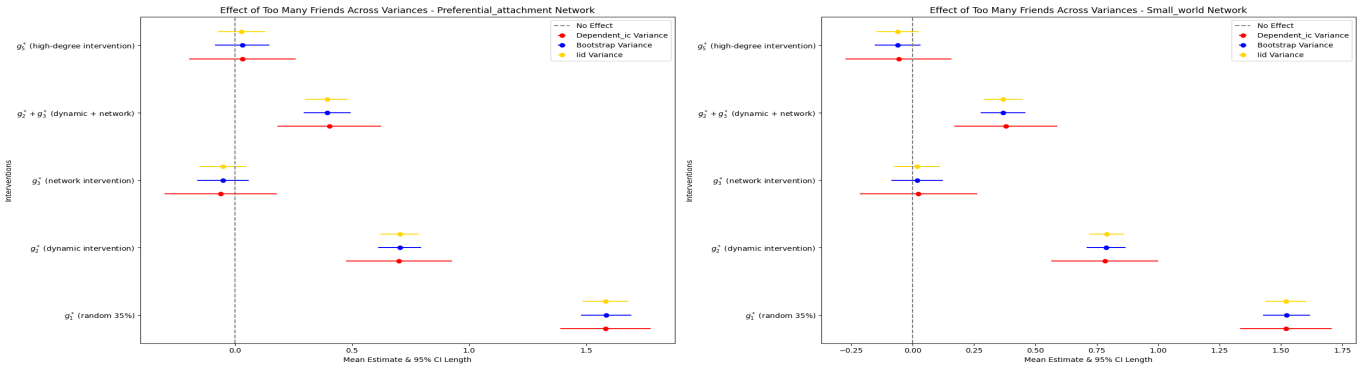


Figure 5: Mean 95% CI length and coverage for TMLE in preferential attachment network and small world network with hub influence

7 Data Analysis

The paper [2] does reanalysis of the Christakis and Fowler (CF) study on peer effects for obesity focuses on addressing limitations in the original approach by incorporating the entire social network structure and accounting for causal and statistical dependencies among participants.

Original CF Study

CF ([1]) modeled obesity status over multiple exams using longitudinal logistic regression, focusing on individual social ties (e.g., siblings, spouses) while assuming independence across individuals. They reported increased obesity risks (27%–171%) linked to social connections. The pairwise analysis, treating each network tie as independent, leads to incoherent models for the network. Network dependence across observations invalidates the analysis under statistical assumptions.

Reanalysis Methodology

The new method simultaneously analyzes all social connections for 3,766 participants using network-wide structural modeling. It replaces binary indicators of a single obese alter with the proportion of obese alters as the exposure of interest. Additionally, it controls for confounders such as previous obesity status, alters' obesity at the previous visit, and individual covariates, including age, sex, and education.

Predicted obesity probability under intervention (to increase the number of obese alters) closely matched observed data (0.137 vs. 0.147). Confidence intervals were similar to CF's, supporting the peer effects hypothesis but also reflecting potential bias from unaccounted dependencies.
While peer effects for obesity appear supported, the new analysis suggests a more robust approach, highlighting issues with unconfoundedness assumptions and pairwise model limitations.

We attempted to replicate the analysis using FHS data but due to access issues we couldn't complete it. We also tried a similar dataset but encountered several challenges, including issues with the `tmlenet` package. While coding simulations from scratch in `Python` was feasible, handling the entire dataset proved to be significantly difficult.

8 Conclusion

In this project, we explored various methodologies for causal inference in social network data, focusing on the identification and estimation of peer effects and treatment pathways within networked systems. Our work highlights the challenges inherent in analyzing network data, particularly the issues of interaction and unmeasured dependence between connected units. Traditional methods, such as generalized linear models (GLMs) and generalized estimating equations (GEEs), often fall short in capturing the complex dependencies in social networks, necessitating specialized approaches like Structural Equation Models (SEMs) and Targeted Maximum Likelihood Estimation (TMLE).

Through the integration of causal SEMs, we have outlined a robust framework that allows for the identification of causal effects in the presence of network dependencies. We discussed how dynamic, stochastic,

and static interventions can be modeled within this framework, and how these interventions influence the estimation of potential outcomes. Specifically, the stochastic and dynamic interventions, which introduce randomness or depend on covariates, allow for a more nuanced understanding of treatment effects across the network.

The TMLE method, particularly its doubly robust nature, provides a consistent and efficient estimator of causal effects, even when one of the models for treatment assignment or outcome prediction is misspecified. Our discussion of the direct estimation methods and the intervention SEM further enriches this estimation process, providing flexibility in dealing with the complexities of network data. Additionally, we provided an extensive simulation study comparing different intervention strategies and highlighting the importance of variance estimation methods tailored to the network structure.

Finally, our analysis of the "Too Many Friends, Too Much Influence" scenario demonstrated the practical applications of these techniques, revealing how high-degree nodes (hubs) in network structures amplify the effect of interventions. These findings underscore the importance of understanding network topology and treatment assignment strategies in designing effective interventions.

References

[1] Nicholas A Christakis and James H Fowler. The spread of obesity in a large social network over 32 years. *The New England journal of medicine*, 357(4):370—379, July 2007.

[2] Iván Díaz Elizabeth L. Ogburn, Oleg Sofrygin and Mark J. van der Laan. Causal inference for social network data. *Journal of the American Statistical Association*, 119(545):597–611, 2024.

[3] OpenAI. Chatgpt: Language models are few-shot learners, 2023. Accessed: 2024-12-06.

[4] Mark J van der Laan. Causal inference for a population of causally connected units. *Journal of Causal Inference*, 2:13 – 74, 2014.

Declaration: This paper utilizes ChatGPT for light editing ([3]).