

# Causal Inference for Social Network Data

Elizabeth L. Ogburn, Oleg Sofrygin, Iván Díaz, Mark J. van der Laan

Journal of the American Statistical Association

March 27, 2025

# Motivating Example: Framingham Heart Study

## Background:

- ▶ **Framingham Heart Study (FHS):** A longitudinal cohort study with over 15,000 participants from Framingham, Massachusetts, initially focused on cardiovascular epidemiology.

## Christakis and Fowler's Findings:

- ▶ **Obesity as Socially Contagious (2007):** Evidence suggested that obesity spreads through social ties (connections or relationships between individuals within a social network)

## Motivating Example Assumption:

- ▶ Assume Framingham is a **closed community** where all social connections are known.
- ▶ Peer effects are hypothesized for various outcomes, such as **happiness**.
  - ▶ Peer effects refer to the influence that individuals have on each other's behaviors, attitudes, or outcomes due to social interactions within a group or network. Individual's happiness may be impacted by the happiness levels of their peers (e.g., friends, family, coworkers, or neighbors).

# Motivation and Research Statement

- ▶ Challenges: Standard methods often fail due to dependencies across network nodes.
- ▶ Christakis Fowler (2007): Used longitudinal logistic regression models to examine the ego's obesity status at time  $t+1$  as a function of their age, sex, educational level, and obesity status at time  $t$  and the alter's obesity status at times  $t$  and  $t+1$ .
- ▶ Network phenomena appear to be relevant to the biologic and behavioral trait of obesity, and obesity appears to spread through social ties.
- ▶ Objective: Develop methods to estimate causal effects within single interconnected social networks (all members are directly/indirectly connected).

# Motivation and Research Statement

- ▶ Many real-world interactions are non-independent (e.g., friend or family influence).
- ▶ Need to distinguish between:
  - ▶ Causal effects transmitted across network ties.
    - ▶ Refers to the direct influence one individual (or "node") exerts on another through their connection (or "tie") in the network.
    - ▶ Example: If a friend adopts a healthy habit, their influence might directly cause you to adopt it too.
  - ▶ Dependence due to unobserved, shared characteristics.
    - ▶ Refers to apparent similarities that arise because individuals in the network share common traits, environments, or experiences, rather than influencing each other.
    - ▶ Two friends might both gain weight not because they influence each other, but because they share the same lifestyle, socioeconomic status, or access to unhealthy food options.
- ▶ Importance: Incorrect methods may yield misleading conclusions about peer effects.

# Background: Causal Inference in Social Networks

- ▶ Causal inference typically assumes independent data, but network data often violates this.
- ▶ Interference: one person's treatment can affect others' outcomes.
- ▶ Uses structural equation models (SEMs) to define causal dependencies in the network.

# Types of Dependencies in Social Networks

- ▶ Transmission: Direct influence through network ties (e.g., peer effects).
- ▶ Latent Similarity: Unobserved traits shared by connected individuals (e.g., homophily).
- ▶ These types require different statistical handling for accurate causal inference.
- ▶ SEM framework accommodates both dependencies.

# Structural Equation Model (SEM) for Causal Dependence

- ▶ SEM captures causal dependencies in network data.
- ▶ Observed data: Each node  $i$  has:
  - ▶ Outcome  $Y_i$ , covariates  $C_i$ , treatment  $X_i$ .
  - ▶ Relationships represented in an adjacency matrix  $A$ .
- ▶ SEM equations:

$$C_i = f_C(\epsilon_{C_i}) \quad (1)$$

$$X_i = f_X(C_j : A_{ij} = 1, \epsilon_{X_i}) \quad (2)$$

$$Y_i = f_Y(X_j : A_{ij} = 1, C_j : A_{ij} = 1, \epsilon_{Y_i}) \quad (3)$$

- ▶  $K_i = \sum_{j=1}^n A_{ij}$  be the degree of node  $i$ , where  $f_C$ ,  $f_X$ , and  $f_Y$  are unknown & may depend on  $K_i$  and  $\epsilon_i = (\epsilon_{C_i}, \epsilon_{X_i}, \epsilon_{Y_i})$  is a vector of exogenous, unobserved errors for individual  $i$ . The errors may be correlated across units, as described below.



# Assumptions of Structural Equation Model (SEM)

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

# Structural Equation Model (SEM) for Causal Dependence

- ▶ To facilitate the definition and identification of dynamic and stochastic estimands, simplifying assumptions on the forms of  $f_X$  and  $f_Y$  are often required.
- ▶ Introduce summary functions  $s_C$  and  $s_X$ 
  - ▶  $W_i = s_{C,i}(C_j : A_{ij} = 1)$
  - ▶  $V_i = s_{X,i}(X_j : A_{ij} = 1)$
- ▶ Model relationships are written as:

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

- ▶ For example,  $s_{C,i}(C_j : A_{ij} = 1)$  represents that the exposure and outcome of node  $i$  depends on its own value and the sum of the covariate values of its alters.
- ▶  $s_{X,i}(X_j : A_{ij} = 1)$  is a summary function for  $X$ .

# Targeted Maximum Loss based Estimation (TMLE) of $\psi_n$

## ► Step 1: Compute Auxiliary Weights $H_i$

- Define the auxiliary weights as the ratio of estimated densities of  $V^*$ ,  $W$  and  $V$ ,  $W$  evaluated at the observed value  $W_i$ :

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

- $h_i(v|w) = P(V_i = v \mid W_i = w)$   
 $h_i(v, w) = P(V_i = v, W_i = w)$

## ► Step 2: Compute Initial Predicted Outcomes

- Compute initial predicted outcome values  $\hat{Y}_i \equiv m(\hat{V}_i, W_i)$  and predicted potential outcome values  $\hat{Y}_i^* \equiv m(\hat{V}_i^*, W_i)$  evaluated at the counterfactual value  $V_i^*$ .
- $m(v, w) = \mathbb{E}[Y \mid V = v, W = w] = \int y p_Y(y \mid v, w) dy$
- Under a static intervention,  $V_i^*$  is the degenerate random variable  $s_{X,i}(x^*)$ .

# Targeted Maximum Loss based Estimation (TMLE) of $\psi_n$

## ► Step 3: Construct TMLE Model Update

- Run a weighted intercept-only logistic regression model with weights  $H_i$  from Step 1, using  $Y_i$  as the outcome and including  $\hat{Y}_i$  as an offset.
- The logistic regression model is:

$$\text{logit } m(\hat{v}, w)\epsilon = \text{logit } m(\hat{v}, w) + \hat{\epsilon}$$

where  $\text{logit}(x) = \log\left(\frac{x}{1-x}\right)$ .  $\hat{\epsilon}$  as the estimate of the intercept parameter

## ► Step 4: Compute Updated Predicted Potential Outcomes

- Compute updated predicted potential outcomes  $Y_i^*$  as the fitted values from Step 3, evaluated at  $\hat{V}_i^*$  rather than  $\hat{V}_i$ :

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

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

# Targeted Maximum Loss Based Estimation (TMLE) of $\psi_n$

## ► Step 5: Compute the TMLE Estimator

- Compute the TMLE  $\hat{\psi}_n$  as:

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

## ► Doubly Robust Property

- The TMLE is doubly robust: it will be consistent for  $\psi_n$  if either the working model for  $h$  or the model for  $m$  is correctly specified.
- The resulting estimator remains consistent asymptotically normal (CAN) under assumptions (A2) and (A3) or (A4) and (A5).

# Asymptotic Properties

- ▶ Asymptotic normality of the TMLE estimator under network size  $n \rightarrow \infty$ .
- ▶ Rate of convergence depends on network degree  $K_{max,n}$ .
- ▶ Valid inference requires that  $K_{max,n}^2/n \rightarrow 0$ .
- ▶ Practical implications for large social networks (e.g., Facebook, Twitter).

# Reanalysis of FHS Study on Peer Effects for Obesity

- ▶ The original study (Christakis and Fowler, 2007) used partially reconstructed social networks of FHS participants to study peer effects for obesity.
- ▶ They fitted longitudinal logistic regression models for each individual's obesity status over multiple exams ( $k = 2, 3, 4, \dots$ ).
- ▶ The analysis was based on the obesity statuses of each individual's social contacts at the same and previous visits.
- ▶ This paper's methodology accounted for network structure and statistical dependence among subjects instead of treating each network tie independently.

# Key Findings and Methodological Insights

- ▶ The original model assumed independence across individuals, which may have led to incoherent models for the full network.
- ▶ This new method considered network dependence and estimated the expected probability of obesity under hypothetical interventions to increase obesity status among alters.
- ▶ This paper's reanalysis suggests that CF's significant results may be spurious due to unaccounted statistical dependence and/or model misspecification.
- ▶ The authors caution against interpreting the results as true causal effects due to potential unobserved confounding and the difficulty in adequately controlling for individual covariates.





# Conclusion

- ▶ This paper proposed new methods for causal and statistical inference using data from a single interconnected social network, accounting for causal and statistical dependence through network ties.
- ▶ Unlike existing methods, this approach does not require randomization of an exogenous treatment and demonstrates proven performance as the number of network ties grows with sample size.
- ▶ This analysis of peer effects for obesity in the Framingham Heart Study highlights the limitations of existing naive methods and emphasizes the importance of adopting new methods like ours for more accurate inference.

Thank You !

# References

-  Christakis, N. A., Fowler, J. H. (2007). The spread of obesity in a large social network over 32 years. *New England Journal of Medicine*, 357(4), 370-379.
-  Ogburn, E. L., Sofrygin, O., Díaz, I., van der Laan, M. J. (2020). Causal inference for social network data. *arXiv preprint arXiv:2010.16115*.