R Package `iglm`: Regression under Interference in Connected Populations

Michael Schweinberger and Cornelius Fritz

with: Subhankar Bhadra, Jonathan R. Stewart, and David R. Hunter

The Pennsylvania State University

Structure

Interference in Connected Populations

Scalable Models

Interpretable Models

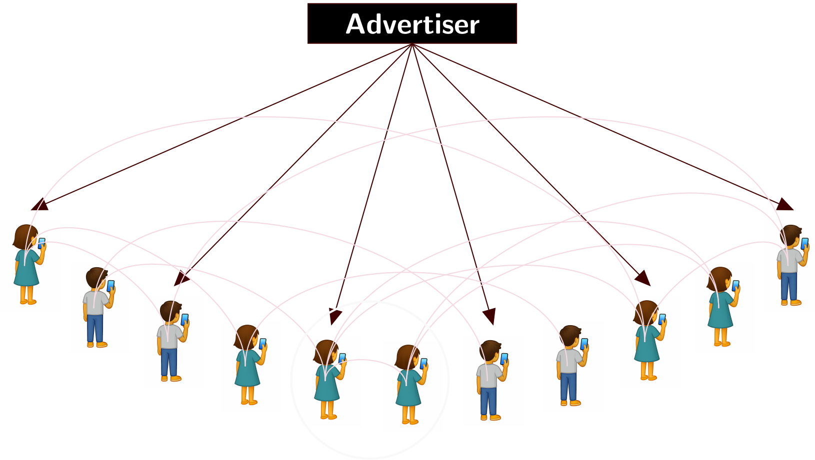
Scalable Methods

Theoretical Guarantees

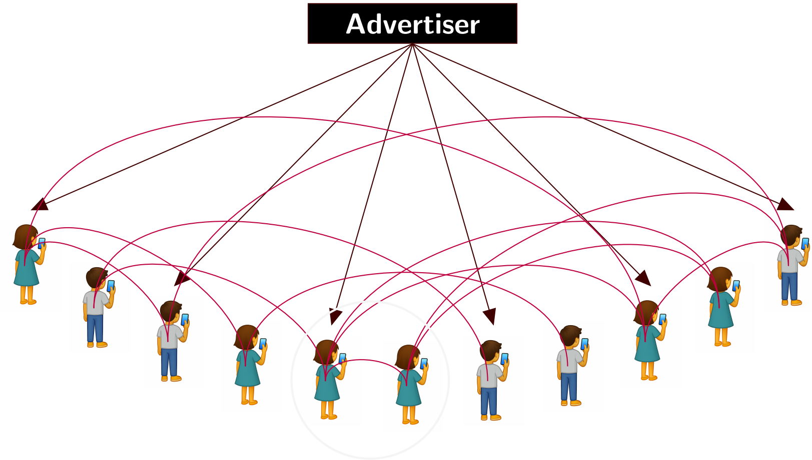
Hate Speech on X

To study relationships among attributes in connected populations with interference, a comprehensive regression framework for attributes $(\boldsymbol{X}, \boldsymbol{Y})$ and connections $\boldsymbol{Z}$ is needed.

Example

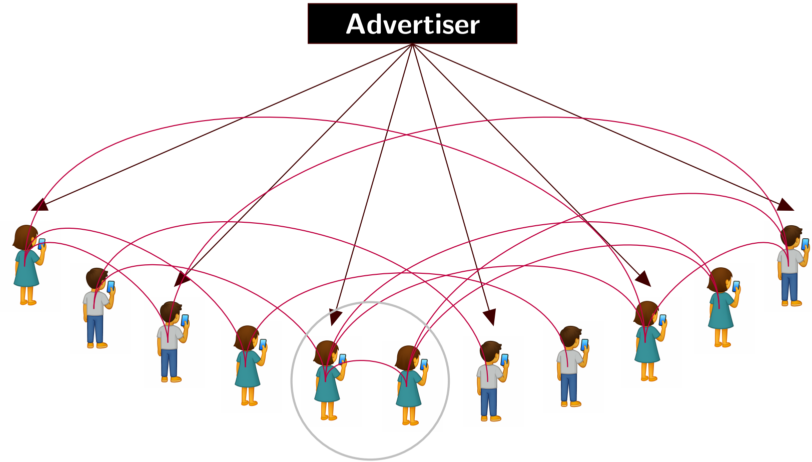


Teenagers have friends:

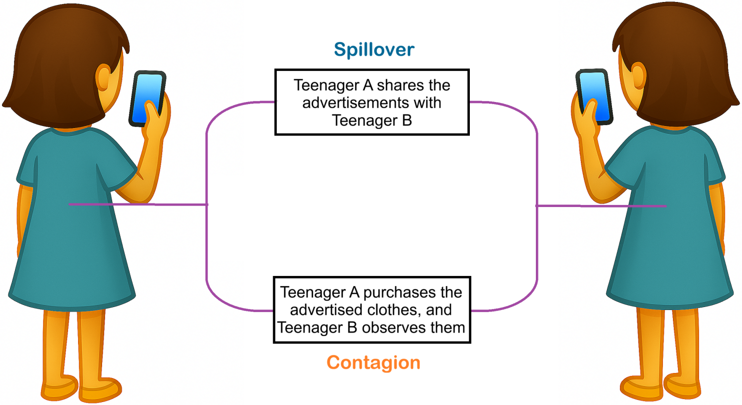


Example

Teenagers have friends:



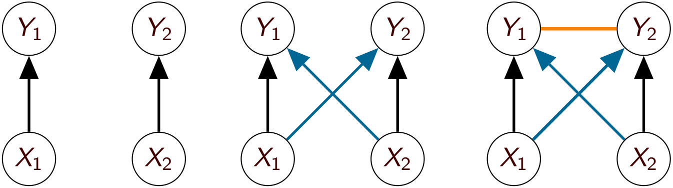
Example


Spillover
Teenager A shares the advertisements with Teenager B
Teenager A purchases the advertised clothes, and Teenager B observes them
Contagion

Interference

Population of $N = 2$ units $1, 2$ with connection $Z_{1,2} = 1$:

Predictors X_1, X_2 and outcomes Y_1, Y_2 of units $1, 2$:

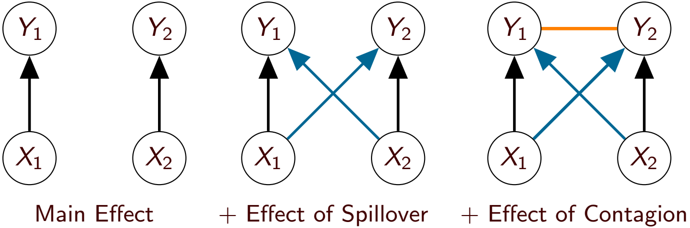


Main Effect   + Effect of Spillover   + Effect of Contagion

An empirical study by Lee & Ogburn (2021) highlights the consequences of ignoring dependence among outcomes due to contagion and other real-world processes in connected populations.

Population of $N = 2$ units $1, 2$ with connection $Z_{1,2} = 1$:

Predictors X_1, X_2 and outcomes Y_1, Y_2 of units $1, 2$:



An empirical study by Lee & Ogburn (2021) highlights the consequences of ignoring dependence among outcomes due to contagion and other real-world processes in connected populations.

Literature

- Conditional models of connections $\boldsymbol{Z} \mid \boldsymbol{X}$ (e.g., ERGMs):
  Who connects with whom?
- Conditional models of outcomes $\boldsymbol{Y} \mid (\boldsymbol{X}, \boldsymbol{Z})$ (e.g., ALAAMs):
  Who influences whom?
- Combining models of $\boldsymbol{Z} \mid \boldsymbol{X}$ and $\boldsymbol{Y} \mid (\boldsymbol{X}, \boldsymbol{Z})$ assumes that connections $\boldsymbol{Z}$ are not affected by outcomes $\boldsymbol{Y}$.
- Models of $(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z})$ (e.g., ERNMs and SAOMs):

Literature

- Conditional models of connections $\boldsymbol{Z} \mid \boldsymbol{X}$ (e.g., ERGMs):
  Who connects with whom?

- Conditional models of outcomes $\boldsymbol{Y} \mid (\boldsymbol{X}, \boldsymbol{Z})$ (e.g., ALAAMs):
  Who influences whom?

- Combining models of $\boldsymbol{Z} \mid \boldsymbol{X}$ and $\boldsymbol{Y} \mid (\boldsymbol{X}, \boldsymbol{Z})$ assumes that connections $\boldsymbol{Z}$ are not affected by outcomes $\boldsymbol{Y}$.

- Models of $(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z})$ (e.g., ERNMs and SAOMs):

Literature

- Conditional models of connections $\boldsymbol{Z} \mid \boldsymbol{X}$ (e.g., ERGMs):
  Who connects with whom?
- Conditional models of outcomes $\boldsymbol{Y} \mid (\boldsymbol{X}, \boldsymbol{Z})$ (e.g., ALAAMs):
  Who influences whom?
- Combining models of $\boldsymbol{Z} \mid \boldsymbol{X}$ and $\boldsymbol{Y} \mid (\boldsymbol{X}, \boldsymbol{Z})$ assumes that connections $\boldsymbol{Z}$ are not affected by outcomes $\boldsymbol{Y}$.
- Models of $(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z})$ (e.g., ERNMs and SAOMs):
  - neither models nor methods scale from small to large N;
  - lack theoretical guarantees: we do not know whether the error of estimation decays as N increases, and at which rate.

Literature

- Conditional models of connections $\boldsymbol{Z} \mid \boldsymbol{X}$ (e.g., ERGMs): **Who connects with whom?**
- Conditional models of outcomes $\boldsymbol{Y} \mid (\boldsymbol{X}, \boldsymbol{Z})$ (e.g., ALAAMs): **Who influences whom?**
- Combining models of $\boldsymbol{Z} \mid \boldsymbol{X}$ and $\boldsymbol{Y} \mid (\boldsymbol{X}, \boldsymbol{Z})$ assumes that connections $\boldsymbol{Z}$ are not affected by outcomes $\boldsymbol{Y}$.
- Models of $(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z})$ (e.g., ERNMs and SAOMs):
  - — neither models nor methods scale from small to large N;
  - — lack theoretical guarantees: we do not know whether the error of estimation decays as N increases, and at which rate.

Literature

- Conditional models of connections $\boldsymbol{Z} \mid \boldsymbol{X}$ (e.g., ERGMs):
  Who connects with whom?
- Conditional models of outcomes $\boldsymbol{Y} \mid (\boldsymbol{X}, \boldsymbol{Z})$ (e.g., ALAAMs):
  Who influences whom?
- Combining models of $\boldsymbol{Z} \mid \boldsymbol{X}$ and $\boldsymbol{Y} \mid (\boldsymbol{X}, \boldsymbol{Z})$ assumes that connections $\boldsymbol{Z}$ are not affected by outcomes $\boldsymbol{Y}$.
- Models of $(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z})$ (e.g., ERNMs and SAOMs):
  - neither models nor methods scale from small to large N;
  - lack theoretical guarantees: we do not know whether the error of estimation decays as N increases, and at which rate.

Literature

- Conditional models of connections $\boldsymbol{Z} \mid \boldsymbol{X}$ (e.g., ERGMs):
  Who connects with whom?

- Conditional models of outcomes $\boldsymbol{Y} \mid (\boldsymbol{X}, \boldsymbol{Z})$ (e.g., ALAAMs):
  Who influences whom?

- Combining models of $\boldsymbol{Z} \mid \boldsymbol{X}$ and $\boldsymbol{Y} \mid (\boldsymbol{X}, \boldsymbol{Z})$ assumes that connections $\boldsymbol{Z}$ are not affected by outcomes $\boldsymbol{Y}$.

- Models of $(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z})$ (e.g., ERNMs and SAOMs):

  — neither models nor methods scale from small to large N;

  — lack theoretical guarantees: we do not know whether the error of estimation decays as N increases, and at which rate.

Regression Under Interference

Introduce comprehensive regression framework for studying relationships among attributes $(\boldsymbol{X}, \boldsymbol{Y})$ and connections $\boldsymbol{Z}$:

- **Scalable models**, capturing heterogeneity along with dependence among $(\boldsymbol{X}, \boldsymbol{Y})$ and $\boldsymbol{Z}$ in large populations.
- **Interpretable models**, based on extensions of Generalized Linear Models (GLMs) to dependent $(\boldsymbol{X}, \boldsymbol{Y})$ and $\boldsymbol{Z}$.
- **Scalable methods**, based on convex optimization of pseudolikelihoods, implemented by a divide-and-conquer approach using minorization-maximization (MM) methods.

Regression Under Interference

Introduce comprehensive regression framework for studying relationships among attributes $(\boldsymbol{X}, \boldsymbol{Y})$ and connections $\boldsymbol{Z}$:

- **Scalable models,** capturing heterogeneity along with dependence among $(\boldsymbol{X}, \boldsymbol{Y})$ and $\boldsymbol{Z}$ in large populations.
- **Interpretable models,** based on extensions of Generalized Linear Models (GLMs) to dependent $(\boldsymbol{X}, \boldsymbol{Y})$ and $\boldsymbol{Z}$.
- **Scalable methods,** based on convex optimization of pseudolikelihoods, implemented by a divide and conquer approach using minorization-maximization (MM) methods.
- **Provable theoretical guarantees,** based on a single observation of dependent random variables $(\boldsymbol{X}, \boldsymbol{Y})$ and $\boldsymbol{Z}$.
- **Software:** R package `iglm`, which extends R functions `lm()` and `glm()` to regression under interference and captures spillover, contagion, and other network-mediated phenomena.

Regression Under Interference

Introduce comprehensive regression framework for studying relationships among attributes $(\boldsymbol{X}, \boldsymbol{Y})$ and connections $\boldsymbol{Z}$:

- **Scalable models,** capturing heterogeneity along with dependence among $(\boldsymbol{X}, \boldsymbol{Y})$ and $\boldsymbol{Z}$ in large populations.
- **Interpretable models,** based on extensions of Generalized Linear Models (GLMs) to dependent $(\boldsymbol{X}, \boldsymbol{Y})$ and $\boldsymbol{Z}$.
- **Scalable methods,** based on convex optimization of pseudolikelihoods, implemented by a divide and conquer approach using minorization-maximization (MM) methods.
- **Provable theoretical guarantees,** based on a single observation of dependent random variables $(\boldsymbol{X}, \boldsymbol{Y})$ and $\boldsymbol{Z}$.
- **Software:** R package `iglm`, which extends R functions `lm()` and `glm()` to regression under interference and captures spillover, contagion, and other network-mediated phenomena.

Regression Under Interference

Introduce comprehensive regression framework for studying relationships among attributes $(\boldsymbol{X}, \boldsymbol{Y})$ and connections $\boldsymbol{Z}$:

- **Scalable models,** capturing heterogeneity along with dependence among $(\boldsymbol{X}, \boldsymbol{Y})$ and $\boldsymbol{Z}$ in large populations.
- **Interpretable models,** based on extensions of Generalized Linear Models (GLMs) to dependent $(\boldsymbol{X}, \boldsymbol{Y})$ and $\boldsymbol{Z}$.
- **Scalable methods,** based on convex optimization of pseudolikelihoods, implemented by a divide and conquer approach using minorization-maximization (MM) methods.
- **Provable theoretical guarantees,** based on a single observation of dependent random variables $(\boldsymbol{X}, \boldsymbol{Y})$ and $\boldsymbol{Z}$.
- **Software:** R package `iglm`, which extends R functions `lm()` and `glm()` to regression under interference and captures spillover, contagion, and other network-mediated phenomena.

Regression Under Interference

Introduce comprehensive regression framework for studying relationships among attributes $(\boldsymbol{X}, \boldsymbol{Y})$ and connections $\boldsymbol{Z}$:

- **Scalable models,** capturing heterogeneity along with dependence among $(\boldsymbol{X}, \boldsymbol{Y})$ and $\boldsymbol{Z}$ in large populations.
- **Interpretable models,** based on extensions of Generalized Linear Models (GLMs) to dependent $(\boldsymbol{X}, \boldsymbol{Y})$ and $\boldsymbol{Z}$.
- **Scalable methods,** based on convex optimization of pseudolikelihoods, implemented by a divide and conquer approach using minorization-maximization (MM) methods.
- **Provable theoretical guarantees,** based on a single observation of dependent random variables $(\boldsymbol{X}, \boldsymbol{Y})$ and $\boldsymbol{Z}$.
- **Software:** R package `iglm`, which extends R functions `lm()` and `glm()` to regression under interference and captures spillover, contagion, and other network-mediated phenomena.

Regression Under Interference

Introduce comprehensive regression framework for studying relationships among attributes $(\boldsymbol{X}, \boldsymbol{Y})$ and connections $\boldsymbol{Z}$:

- **Scalable models,** capturing heterogeneity along with dependence among $(\boldsymbol{X}, \boldsymbol{Y})$ and $\boldsymbol{Z}$ in large populations.
- **Interpretable models,** based on extensions of Generalized Linear Models (GLMs) to dependent $(\boldsymbol{X}, \boldsymbol{Y})$ and $\boldsymbol{Z}$.
- **Scalable methods,** based on convex optimization of pseudolikelihoods, implemented by a divide and conquer approach using minorization-maximization (MM) methods.
- **Provable theoretical guarantees,** based on a single observation of dependent random variables $(\boldsymbol{X}, \boldsymbol{Y})$ and $\boldsymbol{Z}$.
- **Software:** `R` package `iglm`, which extends `R` functions `lm()` and `glm()` to regression under interference and captures spillover, contagion, and other network-mediated phenomena.

Data

Connected population of $N \geq 2$ units:

- one or more binary, count-, or real-valued predictors X_i: e.g., covariates or treatment assignments;
- binary, count-, or real-valued responses or outcomes Y_i;
- binary, count-, or real-valued connections $Z_{i,j}$, directed or not.

Structure

Scalable Models

How to extend GLMs for independent attributes (X_i, Y_i) to dependent attributes (X_i, Y_i) and connections $Z_{i,j}$?

$$\begin{aligned} f_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}) &\propto \prod_{i=1}^{N} a_{\mathcal{X}}(x_i)\, a_{\mathcal{Y}}(y_i)\, e^{\boldsymbol{\theta}_g^\top g_i(\boldsymbol{x}_i,\, y_i)} \quad \text{GLM} \\ &\times \prod_{i<j}^{N} a_{\mathcal{Z}}(z_{i,j})\, e^{\boldsymbol{\theta}_h^\top h_{i,j}(\boldsymbol{x}, y_i, y_j, \boldsymbol{z})} \quad \text{Interference} \end{aligned}$$

- Models with complex dependence can be constructed using two simple building blocks: g_i and $h_{i,j}$.
- The functions $h_{i,j}$ can induce dependence among outcomes Y_i and connections $Z_{i,j}$.
- To build scalable models that can accommodate small and large populations, each unit i has a (known) neighborhood $\mathcal{N}_i$, which can affect i's attributes (X_i, Y_i) and connections $Z_{i,j}$.

Structure

Interpretable Models

If g_i and $h_{i,j}$ are coordinatewise affine,

- conditionals of predictors X_i can be represented by GLMs;
- conditionals of outcomes Y_i can be represented by GLMs;
- conditionals of connections $Z_{i,j}$ can be represented by GLMs.

Thus, results can be interpreted along the lines of GLMs.

Background: GLMs

GLMs extend linear regression

$$Y_i \quad = \quad \alpha_{\mathcal{Y}} + \beta_{\mathcal{X},\mathcal{Y}}\, x_i + \epsilon_i$$

to binary, count-, and continuous outcomes Y_i:

- conditional distribution of outcome Y_i depends on the nature of Y_i and is an exponential family: e.g., Bernoulli (binary), Poisson (count), Gaussian (continuous);
- conditional expectation of outcome Y_i is either a linear or nonlinear function of a linear predictor: e.g., $\alpha_{\mathcal{Y}} + \beta_{\mathcal{X},\mathcal{Y}}\, x_i$.

Example: continuous outcome $Y_i \in \mathbb{R}$

Linear regression:

1. *Conditional distribution of outcome Y_i:* $N(\mu_i(\eta_i), \psi)$
2. *Conditional expectation of outcome Y_i:*

$$\mu_i(\eta_i) \;=\; \eta_i$$

where

$$\eta_i \;=\; \alpha_{\mathcal{Y}} \;+\; \beta_{\mathcal{X},\mathcal{Y}}\, x_i$$

capturing the effect of x_i on Y_i.

Linear regression assumes that the outcome Y_i of unit i is not affected by the outcome Y_j of any other unit j.

Example: count-valued outcome $Y_i \in \{0, 1, \ldots\}$

Poisson regression:

1. *Conditional distribution of outcome Y_i:* Poisson$(\mu_i(\eta_i))$
2. *Conditional expectation of outcome Y_i:*

$$\mu_i(\eta_i) \;=\; \exp(\eta_i)$$

where

$$\eta_i \;=\; \alpha_{\mathcal{Y}} \;+\; \beta_{\mathcal{X},\mathcal{Y}}\, x_i$$

capturing the effect of x_i on Y_i.

Poisson regression assumes that the outcome Y_i of unit i is not affected by the outcome Y_j of any other unit j.

Example: binary outcome $Y_i \in \{0, 1\}$

Logistic regression:

1. *Conditional distribution of outcome Y_i:* Bernoulli$(\mu_i(\eta_i))$

2. *Conditional expectation of outcome Y_i:*

$$\mu_i(\eta_i) \;=\; \frac{\exp(\eta_i)}{1 + \exp(\eta_i)}$$

   where

$$\eta_i \;=\; \alpha_{\mathcal{Y}} \;+\; \beta_{\mathcal{X},\mathcal{Y}}\, x_i$$

   capturing the effect of x_i on Y_i.

Logistic regression assumes that the outcome Y_i of unit i is not affected by the outcome Y_j of any other unit j.

Example of g_i and $h_{i,j}$

Models with complex dependence can be constructed using two simple building blocks:

$$g_i := \begin{pmatrix} y_i \\ x_i\, y_i \end{pmatrix} \quad \begin{pmatrix} \textit{intercept } \alpha_{\mathcal{Y}} \\ \textit{effect } \beta_{\mathcal{X},\mathcal{Y}} \textit{ of } x_i \textit{ on } Y_i \end{pmatrix}$$

$$h_{i,j} := \begin{pmatrix} \boldsymbol{e}_{i,j}\, z_{i,j} \\ -(1 - c_{i,j})\, z_{i,j} \log N \\ d_{i,j}(\boldsymbol{z})\, z_{i,j} \\ c_{i,j}\,(x_i\, y_j + x_j\, y_i)\, z_{i,j} \\ c_{i,j}\, y_i\, y_j\, z_{i,j} \end{pmatrix} \quad \begin{pmatrix} \textit{degree heterogeneity } \alpha_{\mathcal{Z},i},\ \alpha_{\mathcal{Z},j} \\ \textit{sparsity } \lambda \\ \textit{local transitivity } \gamma_{\mathcal{Z},\mathcal{Z}} \\ \textit{spillover } \gamma_{\mathcal{X},\mathcal{Y},\mathcal{Z}} \\ \textit{contagion } \gamma_{\mathcal{Y},\mathcal{Y},\mathcal{Z}} \end{pmatrix}$$

where the dimension of $\boldsymbol{\theta} := (\boldsymbol{\theta}_g,\ \boldsymbol{\theta}_h) \in \mathbb{R}^{N+6}$ increases with N.

Example: continuous outcome $Y_i \in \mathbb{R}$

Let

$$a_{\mathcal{Y}}(y_i) \quad := \quad \frac{1}{\sqrt{2\,\pi\,\psi}}\,\exp\left(-\frac{y_i^2}{2\,\psi}\right)\,\mathbb{I}(y_i \in \mathbb{R})$$

1. *Conditional distribution of outcome Y_i: $N(\mu_i(\eta_i),\, \psi)$*

2. *Conditional expectation of outcome Y_i:*

$$\mu_i(\eta_i) \quad = \quad \eta_i$$

where

$$\eta_i \quad = \quad \alpha_{\mathcal{Y}} \quad + \quad \beta_{\mathcal{X},\mathcal{Y}}\, x_i \quad + \quad \gamma_{\mathcal{X},\mathcal{Y},\mathcal{Z}} \sum_{j:\, \mathcal{N}_i \cap \mathcal{N}_j \neq \emptyset} x_j\, z_{i,j} \quad + \quad \gamma_{\mathcal{Y},\mathcal{Y},\mathcal{Z}} \sum_{j:\, \mathcal{N}_i \cap \mathcal{N}_j \neq \emptyset} y_j\, z_{i,j}$$

capturing **spillover** and **contagion.**

Example: count-valued outcome $Y_i \in \{0, 1, \ldots\}$

Let

$$a_{\mathcal{Y}}(y_i) \quad := \quad \frac{1}{y_i!}\, \mathbb{I}(y_i \in \{0, 1, \ldots\})$$

1. *Conditional distribution of outcome Y_i:* Poisson$(\mu_i(\eta_i))$
2. *Conditional expectation of outcome Y_i:*

$$\mu_i(\eta_i) \quad = \quad \exp(\eta_i)$$

where

$$\eta_i \quad = \quad \alpha_{\mathcal{Y}} \quad + \quad \beta_{\mathcal{X},\mathcal{Y}}\, x_i \quad + \quad \gamma_{\mathcal{X},\mathcal{Y},\mathcal{Z}} \sum_{j:\, \mathcal{N}_i \cap \mathcal{N}_j \neq \emptyset} x_j\, z_{i,j} \quad + \quad \gamma_{\mathcal{Y},\mathcal{Y},\mathcal{Z}} \sum_{j:\, \mathcal{N}_i \cap \mathcal{N}_j \neq \emptyset} y_j\, z_{i,j}$$

capturing **spillover** and **contagion.**

Example: binary outcome $Y_i \in \{0, 1\}$

Let

$$a_{\mathcal{Y}}(y_i) \quad := \quad \mathbb{I}(y_i \in \{0, 1\})$$

1. *Conditional distribution of outcome Y_i:* Bernoulli$(\mu_i(\eta_i))$
2. *Conditional expectation of outcome Y_i:*

$$\mu_i(\eta_i) \;=\; \frac{\exp(\eta_i)}{1 + \exp(\eta_i)}$$

where

$$\begin{aligned}\eta_i \quad = \quad \alpha_{\mathcal{Y}} \quad + \quad \beta_{\mathcal{X},\mathcal{Y}}\, x_i \quad &+ \quad \gamma_{\mathcal{X},\mathcal{Y},\mathcal{Z}} \sum_{j:\, \mathcal{N}_i \,\cap\, \mathcal{N}_j \,\neq\, \emptyset} x_j\, z_{i,j} \\ &+ \quad \gamma_{\mathcal{Y},\mathcal{Y},\mathcal{Z}} \sum_{j:\, \mathcal{N}_i \,\cap\, \mathcal{N}_j \,\neq\, \emptyset} y_j\, z_{i,j}\end{aligned}$$

capturing **spillover** and **contagion.**

Structure

Interference in Connected Populations

Scalable Models

Interpretable Models

Scalable Methods

Theoretical Guarantees

Hate Speech on X

Scalable Methods

Pseudo-loglikelihood:

$$
\begin{aligned}
\ell(\boldsymbol{\theta}; \boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}) \quad := \quad & \sum_{i=1}^{N} \log f_{\boldsymbol{\theta}}(x_i \mid \text{others}) \\
+ \quad & \sum_{i=1}^{N} \log f_{\boldsymbol{\theta}}(y_i \mid \text{others}) \\
+ \quad & \sum_{i<j}^{N} \log f_{\boldsymbol{\theta}}(z_{i,j} \mid \text{others})
\end{aligned}
$$

where $\boldsymbol{\theta}$ is the vector of all weights.

Conditionals of X_i, Y_i, and $Z_{i,j}$ are tractable (e.g., Normal, Poisson, Bernoulli).

Divide and Conquer

Partition the parameter vector $\boldsymbol{\theta} \coloneqq (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \in \mathbb{R}^{N+6}$ into

- $\boldsymbol{\theta}_1 \coloneqq (\alpha_{\mathcal{Z},1}, \ldots, \alpha_{\mathcal{Z},N}) \in \mathbb{R}^N$
- $\boldsymbol{\theta}_2 \coloneqq (\alpha_{\mathcal{Y}}, \beta_{\mathcal{X},\mathcal{Y}}, \gamma_{\mathcal{Z},\mathcal{Z}}, \gamma_{\mathcal{X},\mathcal{Y},\mathcal{Z}}, \gamma_{\mathcal{Y},\mathcal{Y},\mathcal{Z}}, \lambda) \in \mathbb{R}^6$

and partition the negative Hessian of ℓ in accordance:

$$-\nabla^2_{\boldsymbol{\theta}}\, \ell(\boldsymbol{\theta}; \boldsymbol{y}, \boldsymbol{z}) \quad \coloneqq \quad \begin{pmatrix} \boldsymbol{A}(\boldsymbol{\theta}) & \boldsymbol{B}(\boldsymbol{\theta}) \\ \boldsymbol{B}(\boldsymbol{\theta})^\top & \boldsymbol{C}(\boldsymbol{\theta}) \end{pmatrix}$$

The block $\boldsymbol{A}(\boldsymbol{\theta}) \in \mathbb{R}^{N \times N}$ contains N^2 elements.

Divide and Conquer with MM

Iterate:

Step 1: Given $\boldsymbol{\theta}_2^{(t)}$, find $\boldsymbol{\theta}_1^{(t+1)}$ such that

$$\ell(\boldsymbol{\theta}_1^{(t+1)}, \boldsymbol{\theta}_2^{(t)}; \boldsymbol{y}, \boldsymbol{z}) \quad \geq \quad \ell(\boldsymbol{\theta}_1^{(t)}, \boldsymbol{\theta}_2^{(t)}; \boldsymbol{y}, \boldsymbol{z})$$

Step 2: Given $\boldsymbol{\theta}_1^{(t+1)}$, find $\boldsymbol{\theta}_2^{(t+1)}$ such that

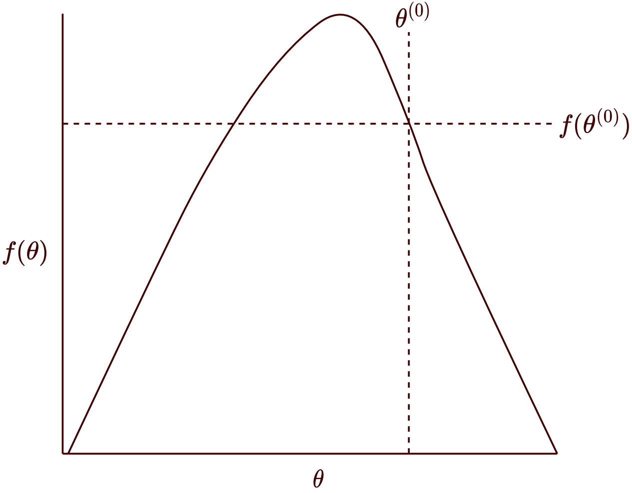
$$\ell(\boldsymbol{\theta}_1^{(t+1)}, \boldsymbol{\theta}_2^{(t+1)}; \boldsymbol{y}, \boldsymbol{z}) \quad \geq \quad \ell(\boldsymbol{\theta}_1^{(t+1)}, \boldsymbol{\theta}_2^{(t)}; \boldsymbol{y}, \boldsymbol{z})$$

Step 1 is implemented by a MM algorithm, replacing the inverse Hessian of ℓ (cost: $O(N^3)$) by a constant matrix (cost: $O(N)$).

Step 2 is implemented by a NR algorithm.

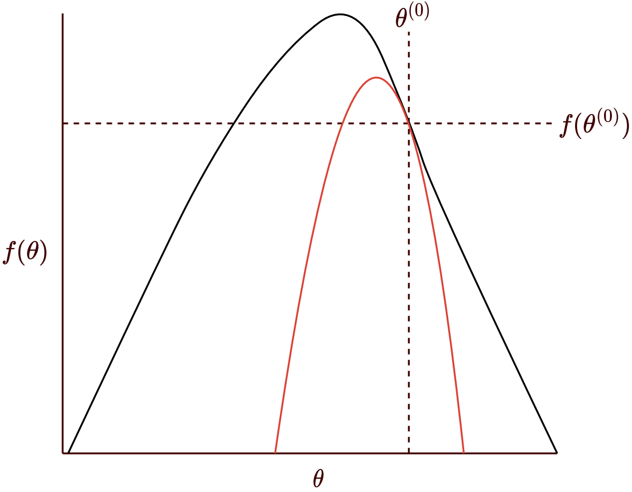
MM: Minorize-Maximize

Consider a generic objective function $f : \mathbb{R} \mapsto \mathbb{R}$:



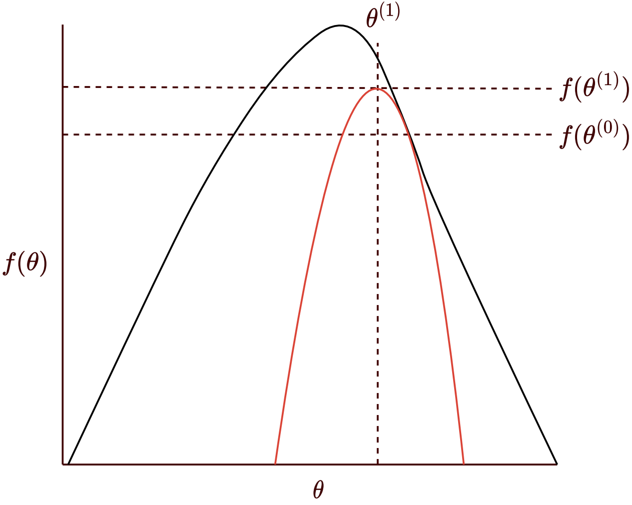
MM: Minorize-Maximize

Consider a generic objective function $f : \mathbb{R} \mapsto \mathbb{R}$:



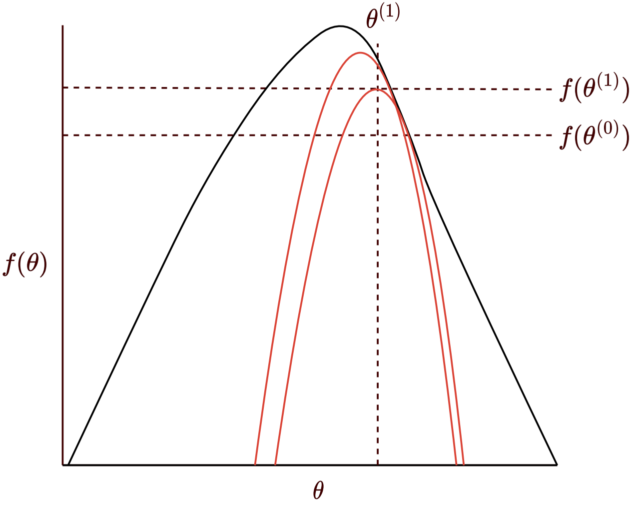
MM: Minorize-Maximize

Consider a generic objective function $f : \mathbb{R} \mapsto \mathbb{R}$:



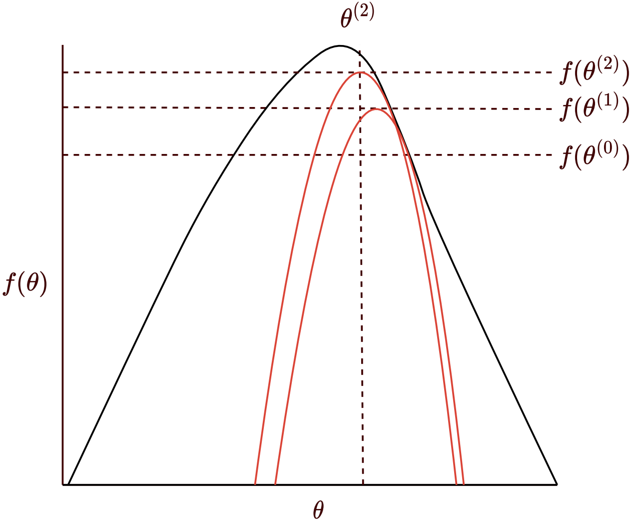
MM: Minorize-Maximize

Consider a generic objective function $f : \mathbb{R} \mapsto \mathbb{R}$:



MM: Minorize-Maximize

Consider a generic objective function $f : \mathbb{R} \mapsto \mathbb{R}$:



Structure

Interference in Connected Populations

Scalable Models

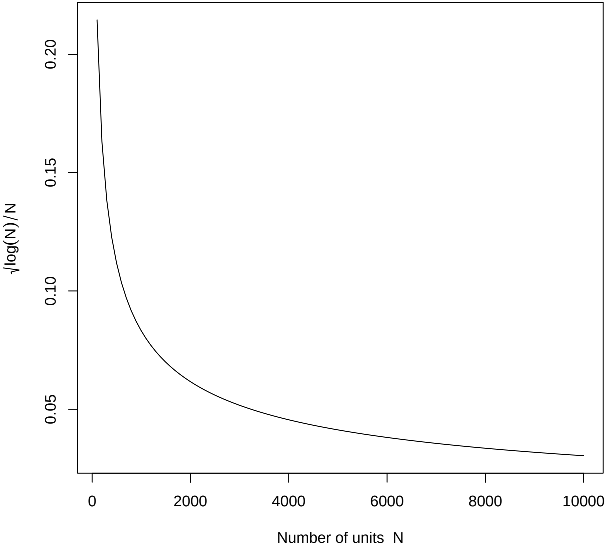
Interpretable Models

Scalable Methods

Theoretical Guarantees

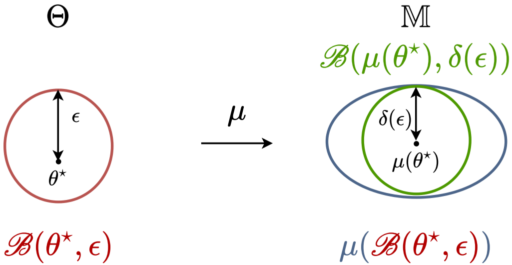
Hate Speech on X

A careful mathematical argument reveals that the error of estimation decays at rate $\sqrt{\log(N)/N}$ in the example model:



Lemma 1 of S & S (Annals of Statistics 2026)

$$\frac{\epsilon}{\sup_{\boldsymbol{\theta} \in \mathcal{B}(\boldsymbol{\theta}^\star, \epsilon)} \, \|\mathcal{I}(\boldsymbol{\theta})^{-1}\|} \quad \leq \quad \delta(\epsilon)$$



Lemma 1 relates

- the radius ϵ of balls in $\boldsymbol{\Theta}$ (unobservables: weights)
- the radius $\delta(\epsilon)$ of balls in $\mathbb{M}$ (observables: sums of g_i, $h_{i,j}$),

helping establish the rate at which the error of estimation decays.

Structure

Interference in Connected Populations

Scalable Models

Interpretable Models

Scalable Methods

Theoretical Guarantees

Hate Speech on X

Hate Speech on X

Data shared by B. Desmarais: 109,974 posts by $N = 2{,}191$ U.S. state legislators between June 6, 2020 and January 6, 2021:

- $x_{i,1} = 1$ if legislator i is Republican and is 0 otherwise,
- $Y_i = 1$ if legislator i made at least one post with hate speech and is 0 otherwise based on Large Language Models (LLMs),
- $Z_{i,j} = 1$ if legislator i mentioned or retweeted legislator j in a tweet and is 0 otherwise,
- $\mathcal{N}_i$ is the set of legislators followed by legislator i, i.e., the set of legislators who can influence i.

Other predictors: gender ($x_{i,2}$), race ($x_{i,3}$), and state ($x_{i,4}$).

Hate Speech on X

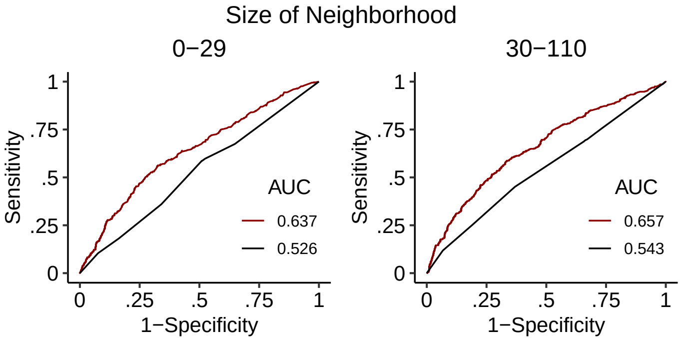
Hate Speech on X

Weight	Estimate	S.E.	Weight	Estimate	S.E.
$\alpha_{\mathcal{Y}}$	$-.893$	.143	$\gamma_{\mathcal{Z},\mathcal{Z},1}$	.604	.037
$\beta_{\mathcal{X},\mathcal{Y},1}$	$-.257$	.159	$\gamma_{\mathcal{Z},\mathcal{Z},2}$	2.57	.031
$\beta_{\mathcal{X},\mathcal{Y},2}$	$-.034$	.117	$\gamma_{\mathcal{X},\mathcal{Z},1}$	.035	.005
$\beta_{\mathcal{X},\mathcal{Y},3}$	.069	.095	$\gamma_{\mathcal{X},\mathcal{Z},2}$	.236	.015
$\gamma_{\mathcal{Y},\mathcal{Z}}$	$-.007$	.055	$\gamma_{\mathcal{X},\mathcal{Z},3}$	.756	.028
$\gamma_{\mathcal{X},\mathcal{Y},\mathcal{Z}}$	.038	.014	$\gamma_{\mathcal{X},\mathcal{Z},4}$	4.729	.041
			λ	.184	.005

The positive sign of $\widehat{\gamma}_{\mathcal{X},\mathcal{Y},\mathcal{Z}} = .038$ *suggests that the more Republicans interact with legislator i, the higher is the conditional probability of the event that legislator i uses offensive text in a post, holding everything else constant.*

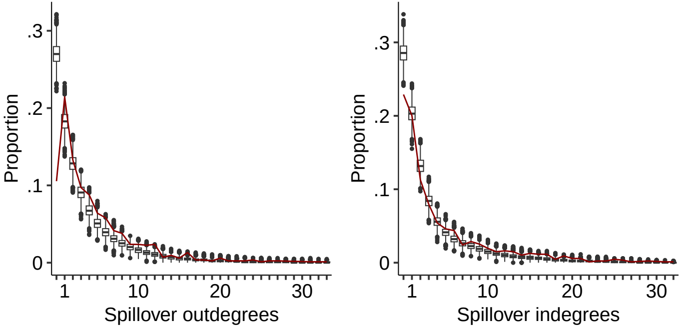
Compare model-based predictions:

- $Y_i \mid \boldsymbol{X}_i = \boldsymbol{x}_i$: logistic regression
- $Y_i \mid (\boldsymbol{X}, \boldsymbol{Y}_{-i}, \boldsymbol{Z}) = (\boldsymbol{x}, \boldsymbol{y}_{-i}, \boldsymbol{z})$



Hate Speech on X

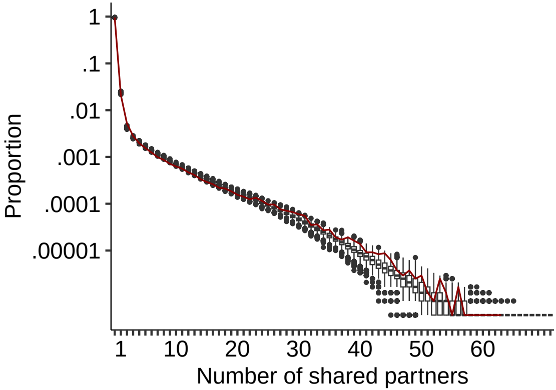
Model-based predictions of spillover in- and outdegrees of U.S. legislators in the subnetwork with i being Republican, j using offensive language, and the neighborhoods of i and j overlapping:



By construction, the connections in the subnetwork act as potential channels of spillover.

Hate Speech on X

Model-based predictions of the shared partners distribution of the network of repost and mention interactions of U.S. legislators:



Capturing closure in the network of U.S. legislators.

References

Fritz, Schweinberger, Bhadra & Hunter (2026). *A regression framework for studying relationships among attributes under network interference.* Journal of the American Statistical Association, Theory & Methods.

Stewart & Schweinberger (2026). *Pseudo-likelihood-based M-estimation of random graphs with dependent edges and parameter vectors of increasing dimension.* The Annals of Statistics, 54, 74–92.

Fritz & Schweinberger (2026). `R` *package* `iglm`*: Regression under interference in connected populations.* `arXiv:2604.22791`

Support: NSF award DMS-2515763 (with Jonathan Stewart) and ARO award W911NF-21-1-0237 (with Johan Koskinen and Pavel Krivitsky).

Thank you!