Causal inference on networks

modern experimental design

Alexander Volfovsky
Department of Statistical Science, Duke University

IIT Kanpur Mathematics and Statistics, October 8, 2018

A casual stroll through causal inference

- Neyman, 1923, Rubin, 1974, etc.
- n units are potentially assigned to treatments $(Z_1, \ldots, Z_n)$.
- The potential outcome of unit i is given by $Y_i(Z_1, \ldots, Z_n)$.
- Standard assumption: $Y_i(Z_1, \ldots, Z_n) = Y_i(Z_i)$ and so $(Y_i(0), Y_i(1))$ are the PO of unit i.
- Frequently interested in the average treatment effect (ATE):

$$ATE = \frac{1}{n} \sum_{i=1}^{n} Y_i(1) - Y_i(0)$$

A casual stroll through causal inference

- Neyman, 1923, Rubin, 1974, etc.
- n units are potentially assigned to treatments $(Z_1, \ldots, Z_n)$.
- The potential outcome of unit i is given by $Y_i(Z_1, \ldots, Z_n)$.
- Standard assumption: $Y_i(Z_1, \ldots, Z_n) = Y_i(Z_i)$ and so $(Y_i(0), Y_i(1))$ are the PO of unit i.
- Frequently interested in the average treatment effect (ATE):

$$ATE = \frac{1}{n} \sum_{i=1}^{n} Y_i(1) - Y_i(0)$$

- Networks make the above hard! Need:

⋆ Randomization schemes to control interference and homophily.

⋆ Matching methods for observational studies with networks.

Causal inference with networks

- Applications: disease prevalence, social development, online advertising, business transactions.

Causal inference with networks

- Applications: disease prevalence, social development, online advertising, business transactions.
- How does the classical causal inference setting extend to these?

Causal inference with networks

- Applications: disease prevalence, social development, online advertising, business transactions.
- How does the classical causal inference setting extend to these?
- Some problems:

Causal inference with networks

- Applications: disease prevalence, social development, online advertising, business transactions.
- How does the classical causal inference setting extend to these?
- Some problems:
  - Homophily.

Causal inference with networks

- Applications: disease prevalence, social development, online advertising, business transactions.
- How does the classical causal inference setting extend to these?
- Some problems:
  - Homophily.
  - Interference.

Causal inference with networks

- Applications: disease prevalence, social development, online advertising, business transactions.
- How does the classical causal inference setting extend to these?
- Some problems:
  - Homophily.
  - Interference.
  - Entangled treatments.

Causal inference with networks

- Applications: disease prevalence, social development, online advertising, business transactions.
- How does the classical causal inference setting extend to these?
- Some problems:
  - Homophily.
  - Interference.
  - Entangled treatments.
- The potential outcome of unit i under assignment vector $(Z_1, \ldots, Z_n)$ is given by $Y_i(Z_1, \ldots, Z_n)$.

Causal inference with networks

- Applications: disease prevalence, social development, online advertising, business transactions.
- How does the classical causal inference setting extend to these?
- Some problems:
  - Homophily.
  - Interference.
  - Entangled treatments.
- The potential outcome of unit i under assignment vector $(Z_1, \ldots, Z_n)$ is given by $Y_i(Z_1, \ldots, Z_n)$.
- Estimands of interest:

Causal inference with networks

- Applications: disease prevalence, social development, online advertising, business transactions.
- How does the classical causal inference setting extend to these?
- Some problems:
  - Homophily.
  - Interference.
  - Entangled treatments.
- The potential outcome of unit i under assignment vector $(Z_1, \ldots, Z_n)$ is given by $Y_i(Z_1, \ldots, Z_n)$.
- Estimands of interest:
  - Total network: "maximal effect"

Causal inference with networks

- Applications: disease prevalence, social development, online advertising, business transactions.
- How does the classical causal inference setting extend to these?
- Some problems:
  - Homophily.
  - Interference.
  - Entangled treatments.
- The potential outcome of unit i under assignment vector $(Z_1, \ldots, Z_n)$ is given by $Y_i(Z_1, \ldots, Z_n)$.
- Estimands of interest:
  - Total network: "maximal effect"
  - Direct effect: value of isolation

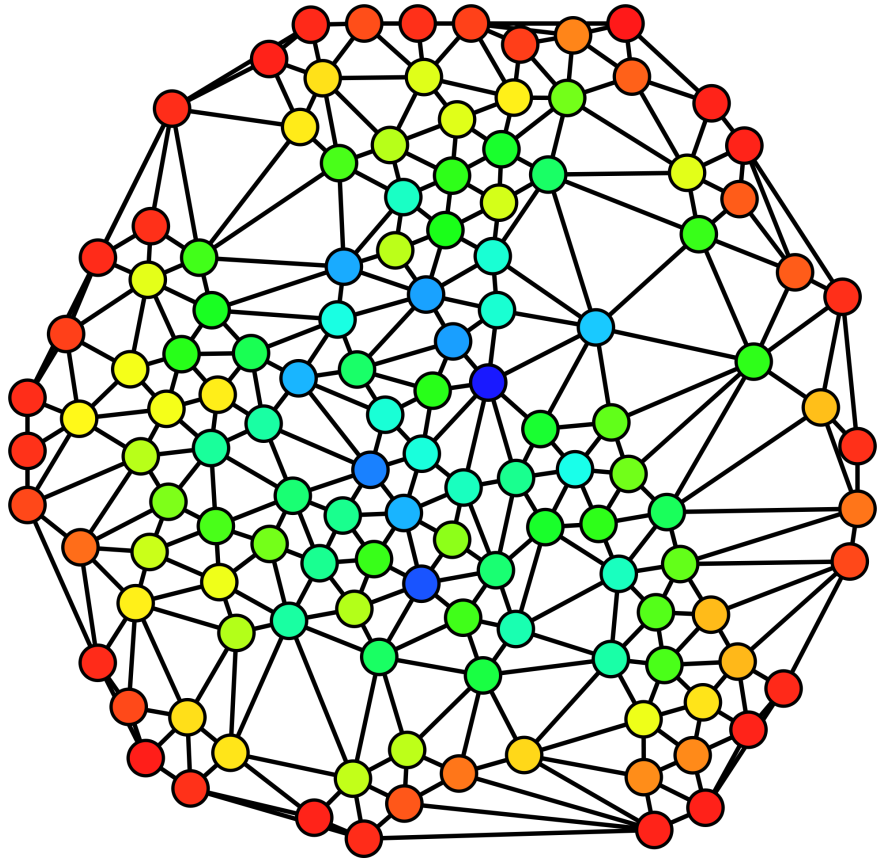
Causal inference with networks

- Applications: disease prevalence, social development, online advertising, business transactions.
- How does the classical causal inference setting extend to these?
- Some problems:
  - Homophily.
  - Interference.
  - Entangled treatments.
- The potential outcome of unit i under assignment vector $(Z_1, \ldots, Z_n)$ is given by $Y_i(Z_1, \ldots, Z_n)$.
- Estimands of interest:
  - Total network: "maximal effect"
  - Direct effect: value of isolation
  - Indirect effect: value of interactions with at least someone

Causal inference with networks

- Applications: disease prevalence, social development, online advertising, business transactions.
- How does the classical causal inference setting extend to these?
- Some problems:
  - Homophily.
  - Interference.
  - Entangled treatments.
- The potential outcome of unit i under assignment vector $(Z_1, \ldots, Z_n)$ is given by $Y_i(Z_1, \ldots, Z_n)$.
- Estimands of interest:
  - Total network: "maximal effect"
  - Direct effect: value of isolation
  - Indirect effect: value of interactions with at least someone
  - Total node: herd immunity.

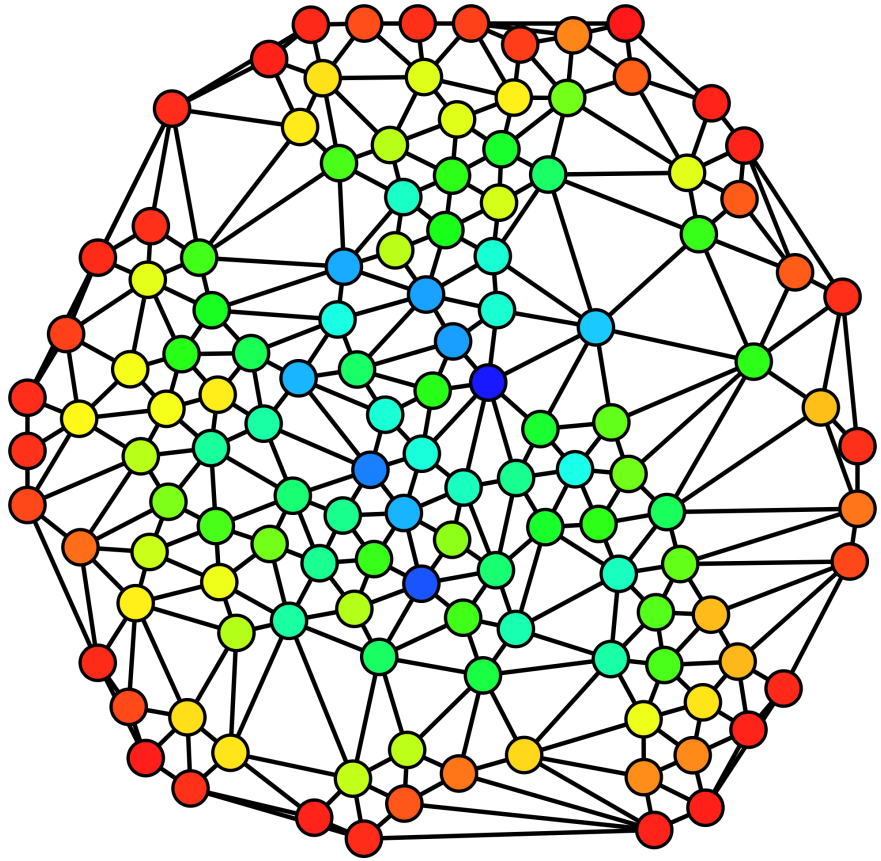
Some context: Facebook



- Facebook wants to change its' ad algorithm.

Source: Wikimedia

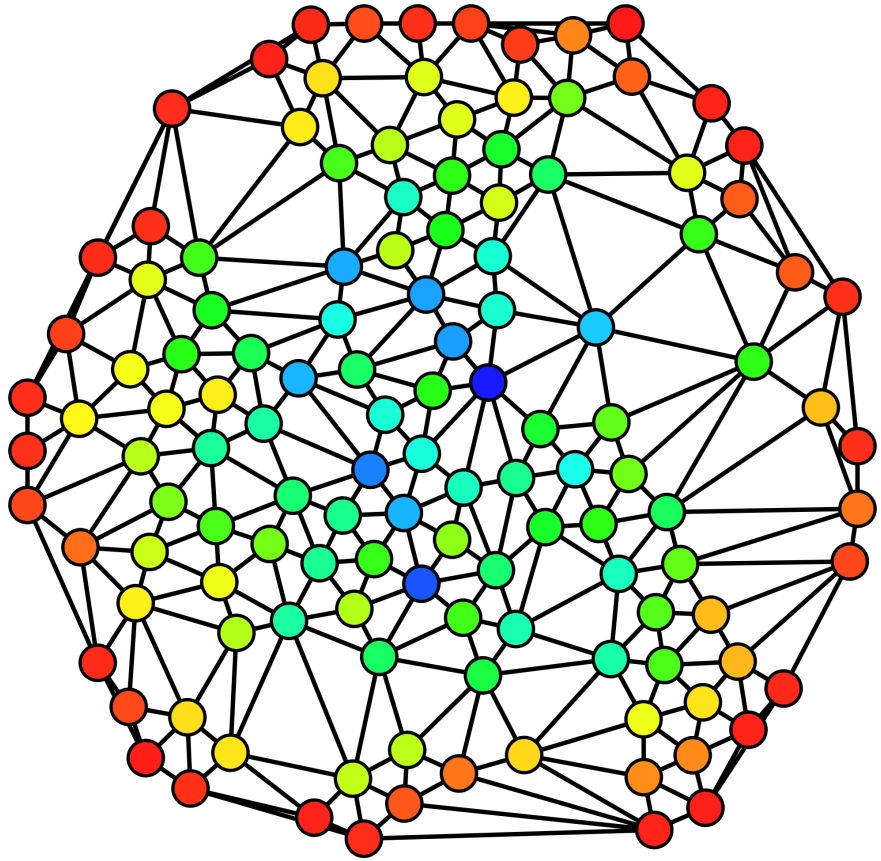
Some context: Facebook



- Facebook wants to change its' ad algorithm.
- Can't do it on the whole graph

Source: Wikimedia

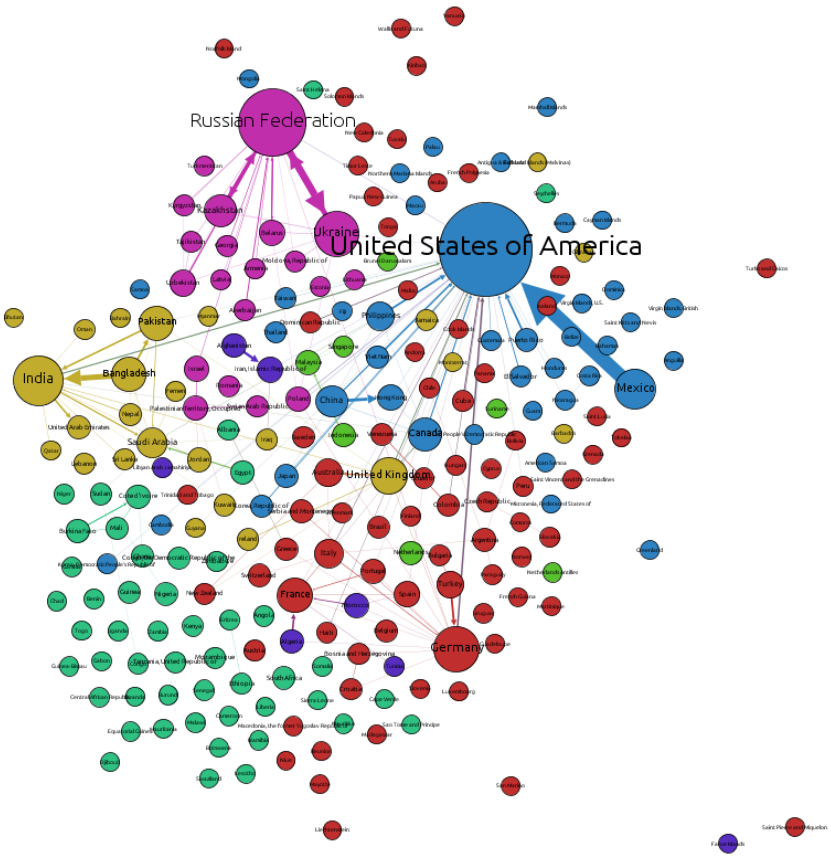
Some context: Facebook



- Facebook wants to change its' ad algorithm.
- Can't do it on the whole graph
- Need "total network effect"

Source: Wikimedia

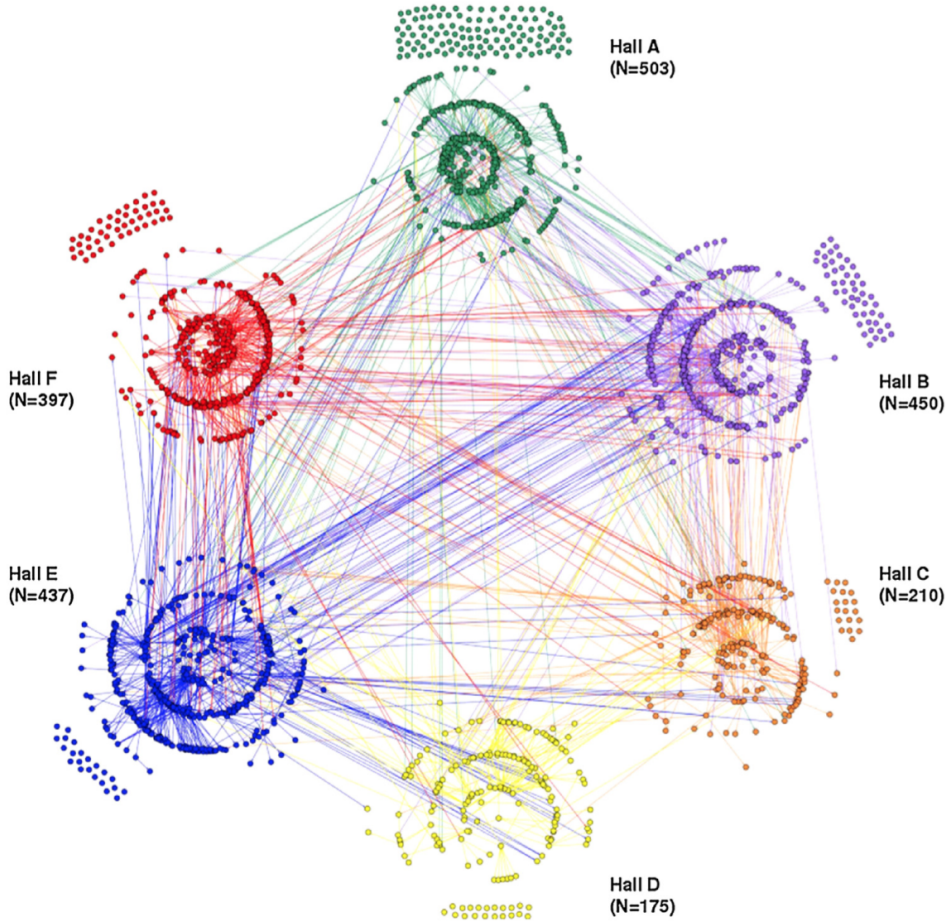
Some context: (im)migration



- Want to know how regime change affects population.
- Politicians during election years care about direct effects.

Source: http://openscience.alpine-geckos.at/courses/social-network-analyses/empirical-network-analysis/

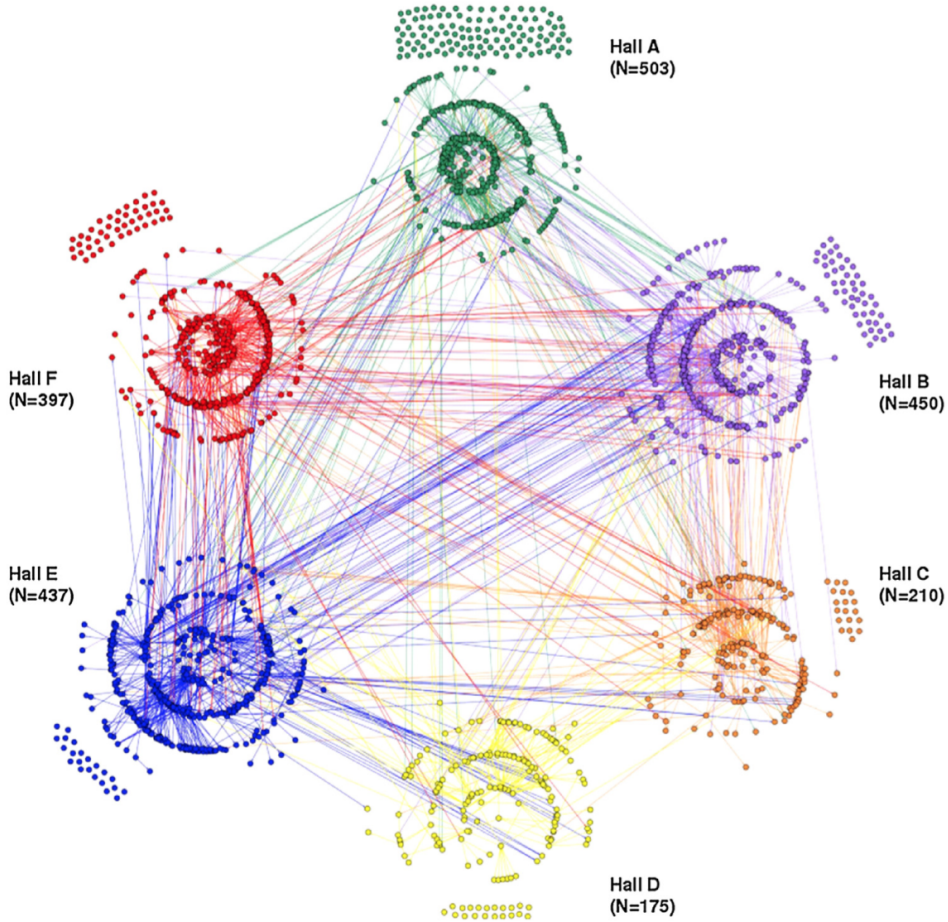
Some context: disease spread



- Want to study efficacy of isolation as treatment for influenza-like illness.

Source: Figure 9 of "Design and methods of a social network isolation study for reducing respiratory infection transmission: The eX-FLU cluster randomized trial" by Aiello et al.

Some context: disease spread



- Want to study efficacy of isolation as treatment for influenza-like illness.
- Interested in spread, duration of illness, etc.

Source: Figure 9 of "Design and methods of a social network isolation study for reducing respiratory infection transmission: The eX-FLU cluster randomized trial" by Aiello et al.

Experimental design with networks

- Interference (and homophily) lead to problems...

Experimental design with networks

- Interference (and homophily) lead to problems...
- Early work by Sobel 2006, Hudgens and Halloran 2008, Tchetgen Tchetgen and VanderWeele 2012 (and others) consider two stage randomization of groups into treatment regimes and then randomization with groups.

Experimental design with networks

- Interference (and homophily) lead to problems...
- Early work by Sobel 2006, Hudgens and Halloran 2008, Tchetgen Tchetgen and VanderWeele 2012 (and others) consider two stage randomization of groups into treatment regimes and then randomization with groups.
- Estimand of interest should guide the randomization strategy.

Experimental design with networks

- Interference (and homophily) lead to problems...
- Early work by Sobel 2006, Hudgens and Halloran 2008, Tchetgen Tchetgen and VanderWeele 2012 (and others) consider two stage randomization of groups into treatment regimes and then randomization with groups.
- Estimand of interest should guide the randomization strategy.
- Total network effect is studied by Eckles, Karrer and Ugander, 2014 – they propose graph-cluster randomization.

Experimental design with networks

- Interference (and homophily) lead to problems...
- Early work by Sobel 2006, Hudgens and Halloran 2008, Tchetgen Tchetgen and VanderWeele 2012 (and others) consider two stage randomization of groups into treatment regimes and then randomization with groups.
- Estimand of interest should guide the randomization strategy.
- Total network effect is studied by Eckles, Karrer and Ugander, 2014 – they propose graph-cluster randomization.
- We are interested in the direct effect!

Experimental design with networks

- Interference (and homophily) lead to problems...
- Early work by Sobel 2006, Hudgens and Halloran 2008, Tchetgen Tchetgen and VanderWeele 2012 (and others) consider two stage randomization of groups into treatment regimes and then randomization with groups.
- Estimand of interest should guide the randomization strategy.
- Total network effect is studied by Eckles, Karrer and Ugander, 2014 – they propose graph-cluster randomization.
- We are interested in the direct effect!
- Simplifying assumption: interference/homophily is restricted to the neighborhood of a node.

Model and estimators

joint work with Natesh Pillai and Ravi Jagadeesan at Harvard

- We have a graph G with $|V(G)| = 2n$ nodes.

Model and estimators

joint work with Natesh Pillai and Ravi Jagadeesan at Harvard

- We have a graph G with $|V(G)| = 2n$ nodes.
- $\mathcal{N}(v)$ denotes the neighbors of node v.

Model and estimators

joint work with Natesh Pillai and Ravi Jagadeesan at Harvard

- We have a graph G with $|V(G)| = 2n$ nodes.
- $\mathcal{N}(v)$ denotes the neighbors of node v.
- $d(v) = |\mathcal{N}(v)|$ is the degree of v.

Model and estimators

joint work with Natesh Pillai and Ravi Jagadeesan at Harvard

- We have a graph G with $|V(G)| = 2n$ nodes.
- $\mathcal{N}(v)$ denotes the neighbors of node v.
- $d(v) = |\mathcal{N}(v)|$ is the degree of v.
- For each vertex $v \in V(G)$ we have
  - t_v: direct treatment effect
  - $f_v : 2^{\mathcal{N}(v)} \to \mathbb{R}$ is a function such that $f_v(\emptyset) = 0$.
  - x_v: vertex covariates

Model and estimators

joint work with Natesh Pillai and Ravi Jagadeesan at Harvard

- We have a graph G with $|V(G)| = 2n$ nodes.
- $\mathcal{N}(v)$ denotes the neighbors of node v.
- $d(v) = |\mathcal{N}(v)|$ is the degree of v.
- For each vertex $v \in V(G)$ we have
  - t_v: direct treatment effect
  - $f_v : 2^{\mathcal{N}(v)} \to \mathbb{R}$ is a function such that $f_v(\emptyset) = 0$.
  - x_v: vertex covariates
- Let $T \subset V(G)$ be the set of treated units.

Model and estimators

joint work with Natesh Pillai and Ravi Jagadeesan at Harvard

- We have a graph G with $|V(G)| = 2n$ nodes.
- $\mathcal{N}(v)$ denotes the neighbors of node v.
- $d(v) = |\mathcal{N}(v)|$ is the degree of v.
- For each vertex $v \in V(G)$ we have
  - t_v: direct treatment effect
  - $f_v : 2^{\mathcal{N}(v)} \to \mathbb{R}$ is a function such that $f_v(\emptyset) = 0$.
  - x_v: vertex covariates
- Let $T \subset V(G)$ be the set of treated units.
- Consider the general linear model as motivation

$$y_v = x_v + 1_T(v)t_v + f_v(T \cap N(v))$$

Model and estimators

joint work with Natesh Pillai and Ravi Jagadeesan at Harvard

- We have a graph G with $|V(G)| = 2n$ nodes.
- $\mathcal{N}(v)$ denotes the neighbors of node v.
- $d(v) = |\mathcal{N}(v)|$ is the degree of v.
- For each vertex $v \in V(G)$ we have
  - t_v: direct treatment effect
  - $f_v : 2^{\mathcal{N}(v)} \to \mathbb{R}$ is a function such that $f_v(\emptyset) = 0$.
  - x_v: vertex covariates
- Let $T \subset V(G)$ be the set of treated units.
- Consider the general linear model as motivation

$$y_v = x_v + 1_T(v)t_v + f_v(T \cap N(v))$$

- The average treatment effect is defined as $\bar{t} = \frac{1}{2n}\sum_{v \in V(G)} t_v$

Model and estimators

joint work with Natesh Pillai and Ravi Jagadeesan at Harvard

- We have a graph G with $|V(G)| = 2n$ nodes.
- $\mathcal{N}(v)$ denotes the neighbors of node v.
- $d(v) = |\mathcal{N}(v)|$ is the degree of v.
- For each vertex $v \in V(G)$ we have
  - t_v: direct treatment effect
  - $f_v : 2^{\mathcal{N}(v)} \to \mathbb{R}$ is a function such that $f_v(\emptyset) = 0$.
  - x_v: vertex covariates
- Let $T \subset V(G)$ be the set of treated units.
- Consider the general linear model as motivation

$$y_v = x_v + 1_T(v)t_v + f_v(T \cap N(v))$$

- The average treatment effect is defined as $\bar{t} = \frac{1}{2n}\sum_{v \in V(G)} t_v$
- We study $|T| = n$ and the naive estimator

$$\hat{t} = \frac{1}{n}\sum_{v \in T} y_v - \frac{1}{n}\sum_{v \in V(G)\setminus T} y_v$$

Why does this model assist us?

$$y_v = x_v + 1_T(v)t_v + f_v(T \cap N(v))$$

- What can the linear model capture?
- Let $z_v = 1_T(v)$ and $z_{\mathcal{N}(v)}$ be the treatments of the neighbors.
- Sussman and Airoldi (2017) show that under a neighborhood interference assumption one can write:

$$\begin{aligned}
Y_v(T) =& \tilde{Y}_v(z_v, z_{\mathcal{N}(v)}) \\
=& \tilde{Y}_v(0,0) \\
&+ z_v(\tilde{Y}_v(1,0) - \tilde{Y}_v(0,0)) \\
&+ (\tilde{Y}_v(0, z_{\mathcal{N}(v)}) - \tilde{Y}_v(0,0)) \\
&+ z_v(\tilde{Y}_v(1, z_{\mathcal{N}(v)}) - \tilde{Y}_v(1,0) - (\tilde{Y}_v(0, z_{\mathcal{N}(v)}) - \tilde{Y}_v(0,0)))
\end{aligned}$$

Operational quantities

- ATE:

$$\bar{t} = \frac{1}{2n} \sum_{v \in V(G)} t_v$$

- Neymanian estimator:

$$\hat{t} = \frac{1}{n} \sum_{v \in T} y_v - \frac{1}{n} \sum_{v \in V(G) \setminus T} y_v$$

- "Ideal estimator":

$$t_{ideal} = \frac{1}{n} \sum_{v \in T} (x_v + t_v) - \frac{1}{n} \sum_{v \in V(G) \setminus T} x_v$$

- The difference

$$\xi = \hat{t} - t_{ideal} = \frac{1}{n} \sum_{v \in T} f_v(T \cap \mathcal{N}(v)) - \frac{1}{n} \sum_{v \in V(G) \setminus T} f_v(T \cap \mathcal{N}(v))$$

- Throughout we will be working on bounding expectations of ξ since $E[\hat{t}] - \bar{t} = E[\xi]$

Comparison to others...

- Neymanian estimator:

$$\hat{t} = \frac{1}{n} \sum_{v \in T} y_v - \frac{1}{n} \sum_{v \in V(G) \setminus T} y_v$$

- Horovitz-Thompson: inverse propensity score weighting ignoring interference
- Stratified estimator: weighted difference of means (post-stratifying on degree)
- Other options?

Experimental design with networks

- Partition the nodes in a graph into pairs:
  $P = \{\{w_1, w_1'\}, \ldots, \{w_n, w_n'\}\}$.

Experimental design with networks

- Partition the nodes in a graph into pairs:
  $P = \{\{w_1, w_1'\}, \ldots, \{w_n, w_n'\}\}$.
- Assign one member of a pair to treatment, one to control.

Experimental design with networks

- Partition the nodes in a graph into pairs: $P = \{\{w_1, w_1'\}, \ldots, \{w_n, w_n'\}\}$.
- Assign one member of a pair to treatment, one to control.
- Consider the function $f_v, v \in V(G)$ K_v Lipshitz (that is $|f_v(A) - f_v(B)| \leq \frac{K_v |A \Delta B|}{d(v)}$)

Experimental design with networks

- Partition the nodes in a graph into pairs: $P = \{\{w_1, w_1'\}, \ldots, \{w_n, w_n'\}\}$.
- Assign one member of a pair to treatment, one to control.
- Consider the function $f_v, v \in V(G)$ K_v Lipshitz (that is $|f_v(A) - f_v(B)| \leq \frac{K_v |A \Delta B|}{d(v)}$)
- What does that capture: K_v is an upper bound on the amount that treating a proportion of the neighbors of v can affect y_v.

Experimental design with networks

- Partition the nodes in a graph into pairs: $P = \{\{w_1, w_1'\}, \ldots, \{w_n, w_n'\}\}$.
- Assign one member of a pair to treatment, one to control.
- Consider the function $f_v, v \in V(G)$ K_v Lipshitz (that is $|f_v(A) - f_v(B)| \leq \frac{K_v |A \Delta B|}{d(v)}$)
- What does that capture: K_v is an upper bound on the amount that treating a proportion of the neighbors of v can affect y_v.
- Examples:

Experimental design with networks

- Partition the nodes in a graph into pairs: $P = \{\{w_1, w_1'\}, \ldots, \{w_n, w_n'\}\}$.
- Assign one member of a pair to treatment, one to control.
- Consider the function $f_v, v \in V(G)$ K_v Lipshitz (that is $|f_v(A) - f_v(B)| \leq \frac{K_v |A \Delta B|}{d(v)}$)
- What does that capture: K_v is an upper bound on the amount that treating a proportion of the neighbors of v can affect y_v.
- Examples:
  - $f_v(A) = \gamma |A|$ is $\gamma d(v)$ Lipschitz.

Experimental design with networks

- Partition the nodes in a graph into pairs: $P = \{\{w_1, w_1'\}, \ldots, \{w_n, w_n'\}\}$.
- Assign one member of a pair to treatment, one to control.
- Consider the function $f_v, v \in V(G)$ K_v Lipshitz (that is $|f_v(A) - f_v(B)| \leq \frac{K_v |A \Delta B|}{d(v)}$)
- What does that capture: K_v is an upper bound on the amount that treating a proportion of the neighbors of v can affect y_v.
- Examples:
  - $f_v(A) = \gamma |A|$ is $\gamma d(v)$ Lipschitz.
  - $f_v(A) = \gamma \frac{|A|}{d(v)}$ is γ Lipschitz.

Experimental design with networks

- Partition the nodes in a graph into pairs: $P = \{\{w_1, w_1'\}, \ldots, \{w_n, w_n'\}\}$.
- Assign one member of a pair to treatment, one to control.
- Consider the function $f_v, v \in V(G)$ K_v Lipshitz (that is $|f_v(A) - f_v(B)| \leq \frac{K_v |A \Delta B|}{d(v)}$)
- What does that capture: K_v is an upper bound on the amount that treating a proportion of the neighbors of v can affect y_v.
- Examples:
  - $f_v(A) = \gamma |A|$ is $\gamma d(v)$ Lipschitz.
  - $f_v(A) = \gamma \frac{|A|}{d(v)}$ is γ Lipschitz.
- The bias is bounded above by

$$\frac{1}{n} \sum_{\{w_i, w_i'\} \subseteq E(G) \cap P} \left(\frac{K_{w_i}}{d(w_i)} + \frac{K_{w_i'}}{d(w_i')} \right)$$

Parsing the result

- The bias is bounded above:

$$|E[\xi|P]| \leq \frac{1}{n} \sum_{\{w_i, w_i'\} \subseteq E(G) \cap P} \left(\frac{K_{w_i}}{d(w_i)} + \frac{K_{w_i'}}{d(w_i')} \right)$$

- If we choose a partition of the nodes such that $\{w, w'\} \in P$ when $\{w, w'\} \notin E(G)$ then $E[\hat{t}] = \bar{t}$
- How does this connect to known results?
- Define $\bar{K} = \frac{1}{2n} \sum_{v \in V(G)} K_v$
- For example, if $K_v = |\gamma| d(v)$ then $\bar{K} = |\gamma| m$ where m is the average degree.
- Sample P uniformly over all possible partitions to get

$$E_P[E[\xi|P]] \leq \frac{\bar{K}}{2n-1}$$

What about variance?

Simplification and complication

- Consider $f_v(S) = f(|S|, |\mathcal{N}(v) \setminus S|)$ – this is symmetric interference.

What about variance?

Simplification and complication

- Consider $f_v(S) = f(|S|, |\mathcal{N}(v) \setminus S|)$ – this is symmetric interference.
- Examples:

What about variance?

Simplification and complication

- Consider $f_v(S) = f(|S|, |\mathcal{N}(v) \setminus S|)$ – this is symmetric interference.
- Examples:
  - $f_v(S) = \gamma|S|$ when $f(a, b) = \gamma a$

What about variance?

Simplification and complication

- Consider $f_v(S) = f(|S|, |\mathcal{N}(v) \setminus S|)$ – this is symmetric interference.
- Examples:
  - $f_v(S) = \gamma|S|$ when $f(a, b) = \gamma a$
  - $f_v(S) = \gamma \frac{|S|}{d(v)}$ when $f(a, b) = \gamma a/(a + b)$

What about variance?

Simplification and complication

- Consider $f_v(S) = f(|S|, |\mathcal{N}(v) \setminus S|)$ – this is symmetric interference.
- Examples:
  - $f_v(S) = \gamma|S|$ when $f(a, b) = \gamma a$
  - $f_v(S) = \gamma \frac{|S|}{d(v)}$ when $f(a, b) = \gamma a/(a + b)$
  - $f_v(S) = \gamma \min\{|S|, k\}$ when $f(a, b) = \gamma \min\{a, k\}$

What about variance?

Simplification and complication

- Consider $f_v(S) = f(|S|, |\mathcal{N}(v) \setminus S|)$ – this is symmetric interference.
- Examples:
  - $f_v(S) = \gamma|S|$ when $f(a, b) = \gamma a$
  - $f_v(S) = \gamma \frac{|S|}{d(v)}$ when $f(a, b) = \gamma a/(a + b)$
  - $f_v(S) = \gamma \min\{|S|, k\}$ when $f(a, b) = \gamma \min\{a, k\}$
  - $f_v(S) = \gamma \min\{\frac{|S|}{d(v)}, p\}$ when $f(a, b) = \gamma \min\{a/(a + b), p\}$

What about variance?

Simplification and complication

- Consider $f_v(S) = f(|S|, |\mathcal{N}(v) \setminus S|)$ – this is symmetric interference.
- Examples:
  - $f_v(S) = \gamma|S|$ when $f(a,b) = \gamma a$
  - $f_v(S) = \gamma\frac{|S|}{d(v)}$ when $f(a,b) = \gamma a/(a+b)$
  - $f_v(S) = \gamma \min\{|S|, k\}$ when $f(a,b) = \gamma \min\{a, k\}$
  - $f_v(S) = \gamma \min\{\frac{|S|}{d(v)}, p\}$ when $f(a,b) = \gamma \min\{a/(a+b), p\}$
- Let $\overrightarrow{d}(v) = (|T \cap \mathcal{N}(v)|, |\mathcal{N}(v) \setminus T|)$ be the bidegree of v.

What about variance?

Simplification and complication

- Consider $f_v(S) = f(|S|, |\mathcal{N}(v) \setminus S|)$ – this is symmetric interference.
- Examples:
  - $f_v(S) = \gamma|S|$ when $f(a,b) = \gamma a$
  - $f_v(S) = \gamma\frac{|S|}{d(v)}$ when $f(a,b) = \gamma a/(a+b)$
  - $f_v(S) = \gamma \min\{|S|, k\}$ when $f(a,b) = \gamma \min\{a, k\}$
  - $f_v(S) = \gamma \min\{\frac{|S|}{d(v)}, p\}$ when $f(a,b) = \gamma \min\{a/(a+b), p\}$
- Let $\overrightarrow{d}(v) = (|T \cap \mathcal{N}(v)|, |\mathcal{N}(v) \setminus T|)$ be the bidegree of v.
- Define a finite, signed measure (of mass 0) on $\mathcal{B} = \{(a,b) : a+b \in d(V(G))\}$:

$$D_T(u) = \frac{1}{n}\sum_{v \in V}(-1)^{1_T(v)}\delta_{\overrightarrow{d}(v)}(u), \quad u \in \mathcal{B}$$

Experimental design with networks

Perfect assignment:

- Assign nodes to treatment to balance interference between treated and untreated nodes.

Experimental design with networks

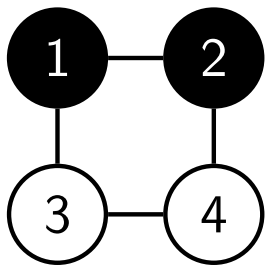
Perfect assignment:

- Assign nodes to treatment to balance interference between treated and untreated nodes.
- Natural to think of as a quasi-coloring problem on graphs.

Experimental design with networks

Perfect assignment:

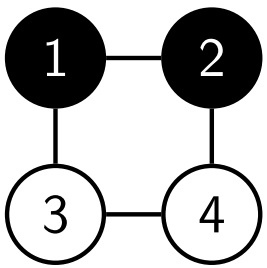
- Assign nodes to treatment to balance interference between treated and untreated nodes.
- Natural to think of as a quasi-coloring problem on graphs.



Experimental design with networks

Perfect assignment:

- Assign nodes to treatment to balance interference between treated and untreated nodes.
- Natural to think of as a quasi-coloring problem on graphs.

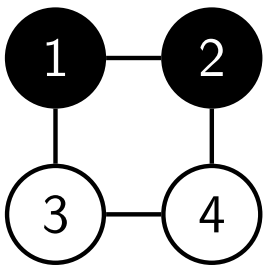


- Nodes 1,2: treated, 1 treated neighbor, 1 untreated.
- Nodes 3,4: untreated, 1 treated neighbor, 1 untreated.

Experimental design with networks

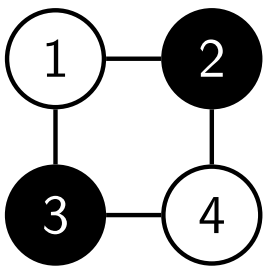
Perfect assignment:

- Assign nodes to treatment to balance interference between treated and untreated nodes.
- Natural to think of as a quasi-coloring problem on graphs.



- Nodes 1,2: treated, 1 treated neighbor, 1 untreated.
- Nodes 3,4: untreated, 1 treated neighbor, 1 untreated.

- Bad quasi-coloring:

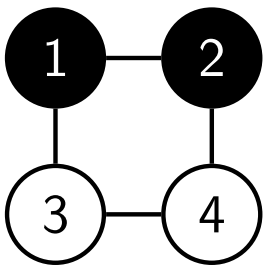


- Nodes 2,3: treated, 2 untreated neighbors.
- Nodes 1,4: untreated, 2 treated neighbor.

Experimental design with networks

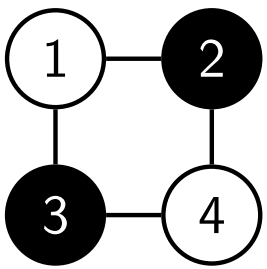
Perfect assignment:

- Assign nodes to treatment to balance interference between treated and untreated nodes.
- Natural to think of as a quasi-coloring problem on graphs.



- Nodes 1,2: treated, 1 treated neighbor, 1 untreated.
- Nodes 3,4: untreated, 1 treated neighbor, 1 untreated.

- Bad quasi-coloring:



- Nodes 2,3: treated, 2 untreated neighbors.
- Nodes 1,4: untreated, 2 treated neighbor.

- A perfect quasi-coloring is such that $D_Q = 0$

Perfect quasi-colorings

Do they exist?

Perfect quasi-colorings

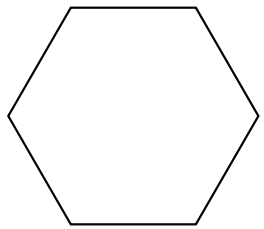
Do they exist?

- Short answer: yes.

Perfect quasi-colorings

Do they exist?

- Short answer: yes.
- Slightly longer answer: sometimes.



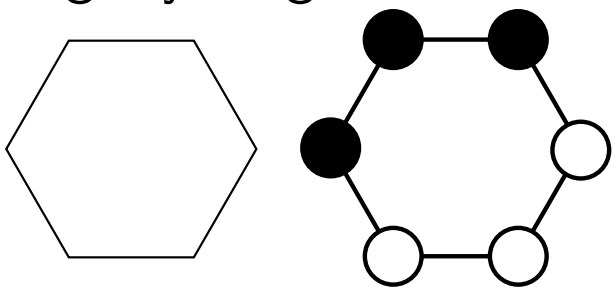
Hexagon

- The nodes are $V(G) = \{1, \ldots, 6\}$:

Perfect quasi-colorings

Do they exist?

- Short answer: yes.
- Slightly longer answer: sometimes.



Hexagon Hexagon

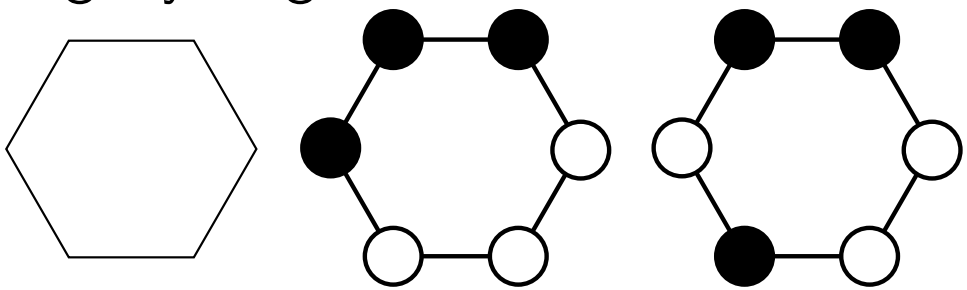
- The nodes are $V(G) = \{1, \ldots, 6\}$:

  B: $\{1, 2, 3\}$ then $|D_B(2, 0)| = 1$

Perfect quasi-colorings

Do they exist?

- Short answer: yes.
- Slightly longer answer: sometimes.



- The nodes are $V(G) = \{1, \ldots, 6\}$:

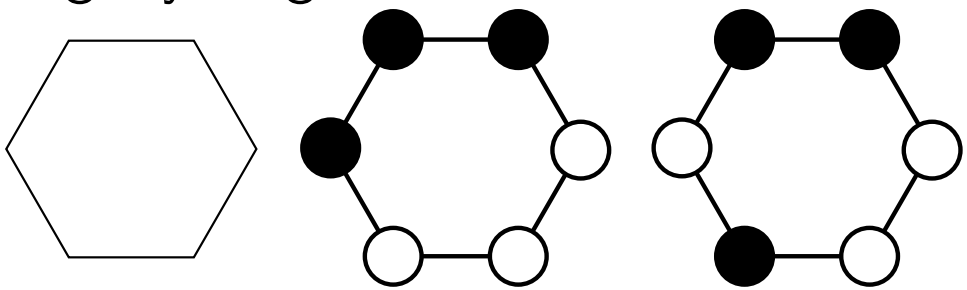
  B: $\{1, 2, 3\}$ then $|D_B(2, 0)| = 1$

  B: $\{1, 2, 4\}$ then $|D_B(0, 2)| = 1$

Perfect quasi-colorings

Do they exist?

- Short answer: yes.
- Slightly longer answer: sometimes.



- The nodes are $V(G) = \{1, \ldots, 6\}$:

  B: $\{1, 2, 3\}$ then $|D_B(2, 0)| = 1$

  B: $\{1, 2, 4\}$ then $|D_B(0, 2)| = 1$

- Do we need perfection?

Connecting quasi-colorings to estimation

- We study
$$\xi = \hat{t} - t_{ideal} = \sum_{u \in \mathcal{B}} f(u) D_T(u)$$
- If we can't find a perfect quasi-coloring, we can try controlling ξ by saying something about D_T.
- Below we consider the following metric on $\mathcal{B}$:
$$d_K((a,b),(c,d)) = K_1 \frac{|a+b-c-d|}{d_{\max}} + K_2 \left| \frac{a}{a+b} - \frac{c}{c+d} \right|$$
- K_1 captures the difference in degrees
- K_2 captures the difference in the fraction of treated neighbors
- They describe the cost of unbalanced treatment

Quantifying the notion of perfect quasi-coloring

- Let $\mathbf{d}$ be a metric on $\mathcal{B}$.
- For $f : \mathcal{B} \to \mathbb{R}$, define the Lipschitz norm

$$\|f\|_{\mathbf{d}} = \sup_{u_1,u_2\in\mathcal{B},u_1\neq u_2} \frac{|f(u_1) - f(u_2)|}{\mathbf{d}(u_1, u_2)}.$$

- For a measure $D \in \Delta^0(\mathcal{B})$, define the Wasserstein norm

$$\|D\|_{\mathbf{d}_\mathrm{w}} = \sup_{\|f\|_{\mathbf{d}}\leq 1} \left\| \int_{\mathcal{B}} f \, dD \right\|.$$

- Since the total mass is 0 for any $D \in \Delta^0(\mathcal{B})$, we have that

$$\|D\|_{\mathbf{d}_\mathrm{w}} \leq \frac{1}{2}\mathrm{diam}(\mathcal{B})\|D\|_{\mathrm{TV}},$$

  where $\| - \|_{\mathrm{TV}}$ denotes the total variation norm.
- If the interference function $f : \mathcal{B} \to \mathbb{R}$ is Lipschitz with respect to a metric $\mathbf{d}$, then

$$|\xi| \leq \|f\|_{\mathbf{d}} \, \|D_\mathrm{T}\|_{\mathbf{d}_\mathrm{w}}.$$

What about variance?

almost there...

- We can bound the L^2 norm of ξ.
- The bound is given as a function of
  - $C_P = \frac{2}{d_{\max}} \sum_{\{w,w'\} \in P} |d(w) - d(w')|$
  - Average of $1/\sqrt{d(v)}$.
  - Information about the partition P.

What about variance?

almost there...

- We can bound the L^2 norm of ξ.
- The bound is given as a function of
  - $C_P = \frac{2}{d_{\max}} \sum_{\{w,w'\} \in P} |d(w) - d(w')|$
  - Average of $1/\sqrt{d(v)}$.
  - Information about the partition P.
- The bound is:

$$\frac{K_1}{n} C_p + \frac{2K_2}{n} \sum_{v \in V(G)} \frac{1}{\sqrt{d(v)}} + \frac{K_2}{n} \sum_{(v,v') \in E(G) \cap P} \left(\frac{1}{d(v)} + \frac{1}{d(v')} \right)$$

What about variance?

almost there...

- We can bound the L^2 norm of ξ.
- The bound is given as a function of
  - $C_P = \frac{2}{d_{\max}} \sum_{\{w,w'\} \in P} |d(w) - d(w')|$
  - Average of $1/\sqrt{d(v)}$.
  - Information about the partition P.
- The bound is:

$$\frac{K_1}{n} C_p + \frac{2K_2}{n} \sum_{v \in V(G)} \frac{1}{\sqrt{d(v)}} + \frac{K_2}{n} \sum_{(v,v') \in E(G) \cap P} \left(\frac{1}{d(v)} + \frac{1}{d(v')}\right)$$

- We can use this information to build a better partition!

What about variance?

almost there...

- We can bound the L^2 norm of ξ.
- The bound is given as a function of
  - $C_P = \frac{2}{d_{\max}} \sum_{\{w,w'\} \in P} |d(w) - d(w')|$
  - Average of $1/\sqrt{d(v)}$.
  - Information about the partition P.
- The bound is:

$$\frac{K_1}{n} C_p + \frac{2K_2}{n} \sum_{v \in V(G)} \frac{1}{\sqrt{d(v)}} + \frac{K_2}{n} \sum_{(v,v') \in E(G) \cap P} \left(\frac{1}{d(v)} + \frac{1}{d(v')}\right)$$

- We can use this information to build a better partition!
- (by controlling the C_P term)

Better partitions

- Order the vertices as $V(G) = \{w_1^\star, w_1^{\star\prime}, \ldots, w_n^\star, w_n^{\star\prime}\}$ such that
$$d(w_1^\star) \geq d(w_1^{\star\prime}) \geq \cdots \geq d(w_n^\star) \geq d(w_n^{\star\prime})$$
- Define the partition as
$$P^\star = \{\{w_1^\star, w_1^{\star\prime}\}, \ldots, \{w_n^\star, w_n^{\star\prime}\}\}$$
- By definition: $C_p \leq 2$.
- The bias and L^2 norm are bounded by
$$\frac{K_1 + K_2}{d_{\min}} \text{ and } \frac{2K_1}{n} + \frac{2K_2}{\sqrt{d_{\min}}} + \frac{2K_2}{d_{\min}}$$
- In a dense graph we have $n \to \infty$ implies $d_{\min} \to \infty$.
- So MSE goes to zero!

Why is this randomization better?

Example

- Let $V(G) = \{s_1, \ldots, s_{2k}, w_1, \ldots, w_{2k}\} = S \cup W$.

Why is this randomization better?

Example

- Let $V(G) = \{s_1, \ldots, s_{2k}, w_1, \ldots, w_{2k}\} = S \cup W$.
- Let $E(G) = \{(s_i, s_j)\}$

Why is this randomization better?

Example

- Let $V(G) = \{s_1, \ldots, s_{2k}, w_1, \ldots, w_{2k}\} = S \cup W$.
- Let $E(G) = \{(s_i, s_j)\}$
- G is a disjoint union of a complete graph on $2k$ vertices and a $2k$ vertex empty graph.

Why is this randomization better?

Example

- Let $V(G) = \{s_1, \ldots, s_{2k}, w_1, \ldots, w_{2k}\} = S \cup W$.
- Let $E(G) = \{(s_i, s_j)\}$
- G is a disjoint union of a complete graph on $2k$ vertices and a $2k$ vertex empty graph.
- Consider symmetric linear interference $f(a, b) = \gamma a$.

Why is this randomization better?

Example

- Let $V(G) = \{s_1, \ldots, s_{2k}, w_1, \ldots, w_{2k}\} = S \cup W$.
- Let $E(G) = \{(s_i, s_j)\}$
- G is a disjoint union of a complete graph on $2k$ vertices and a $2k$ vertex empty graph.
- Consider symmetric linear interference $f(a, b) = \gamma a$.
- Fixing a treatment group T, let $\alpha = |T \cap S|$ then

$$\xi = \frac{\gamma(\alpha(\alpha - 1) - (2k - \alpha)\alpha)}{2k}$$

Why is this randomization better?

Example

- Let $V(G) = \{s_1, \ldots, s_{2k}, w_1, \ldots, w_{2k}\} = S \cup W$.
- Let $E(G) = \{(s_i, s_j)\}$
- G is a disjoint union of a complete graph on $2k$ vertices and a $2k$ vertex empty graph.
- Consider symmetric linear interference $f(a, b) = \gamma a$.
- Fixing a treatment group T, let $\alpha = |T \cap S|$ then

$$\xi = \frac{\gamma(\alpha(\alpha - 1) - (2k - \alpha)\alpha)}{2k}$$

- Now letting T be uniform on all possible partitions we have $(\alpha - k)/\sqrt{k} \to N(0, 1/2)$ and so $E\xi \to 0$ but $(E\xi^2)^{1/2} \sim \gamma\sqrt{k}$.

Why is this randomization better?

Example

- Let $V(G) = \{s_1, \ldots, s_{2k}, w_1, \ldots, w_{2k}\} = S \cup W$.
- Let $E(G) = \{(s_i, s_j)\}$
- G is a disjoint union of a complete graph on $2k$ vertices and a $2k$ vertex empty graph.
- Consider symmetric linear interference $f(a, b) = \gamma a$.
- Fixing a treatment group T, let $\alpha = |T \cap S|$ then

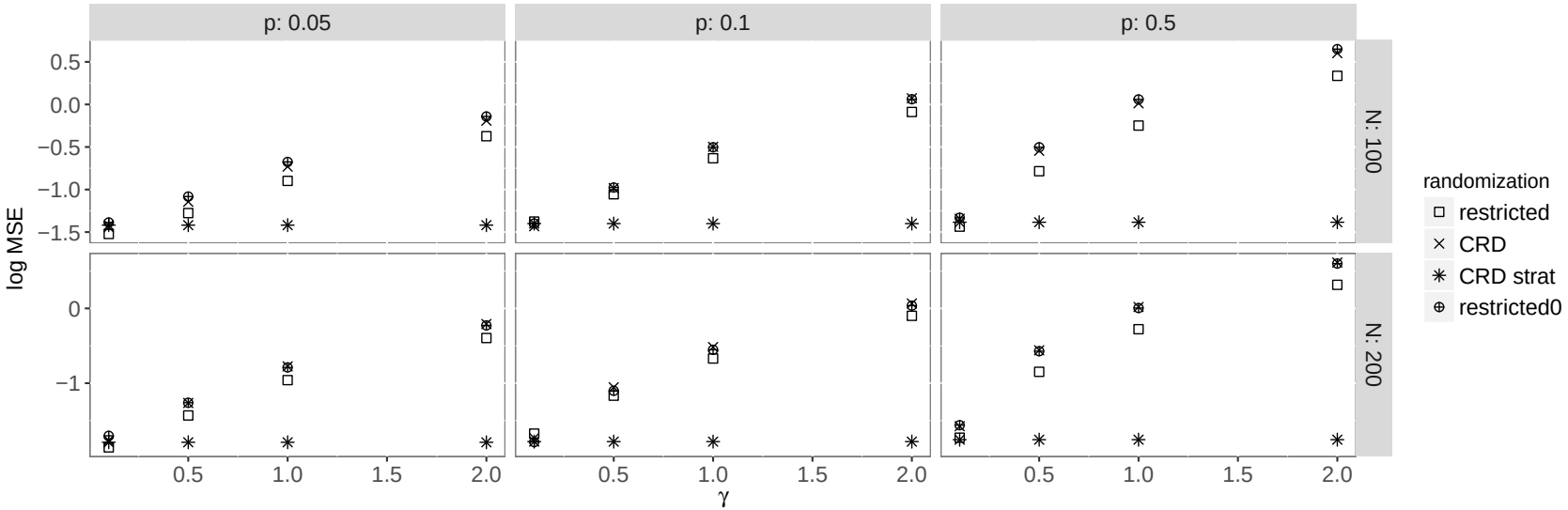
$$\xi = \frac{\gamma(\alpha(\alpha - 1) - (2k - \alpha)\alpha)}{2k}$$

- Now letting T be uniform on all possible partitions we have $(\alpha - k)/\sqrt{k} \to N(0, 1/2)$ and so $E\xi \to 0$ but $(E\xi^2)^{1/2} \sim \gamma\sqrt{k}$.
- On the other hand: using our partition scheme we have $\alpha = k$ and so $\xi = -\frac{\gamma}{2}$ is independent of the size of the graph.

Even better partitions

- Control of C_p by ordering vertices by degree.
- Control $\frac{K_2}{n} \sum_{(v,v') \in E(G) \cap P} \left(\frac{1}{d(v)} + \frac{1}{d(v')} \right)$ by making sure no pair shares an edge.
- Very simple optimization...

(Un)realistic graphs: Erdos-Renyi



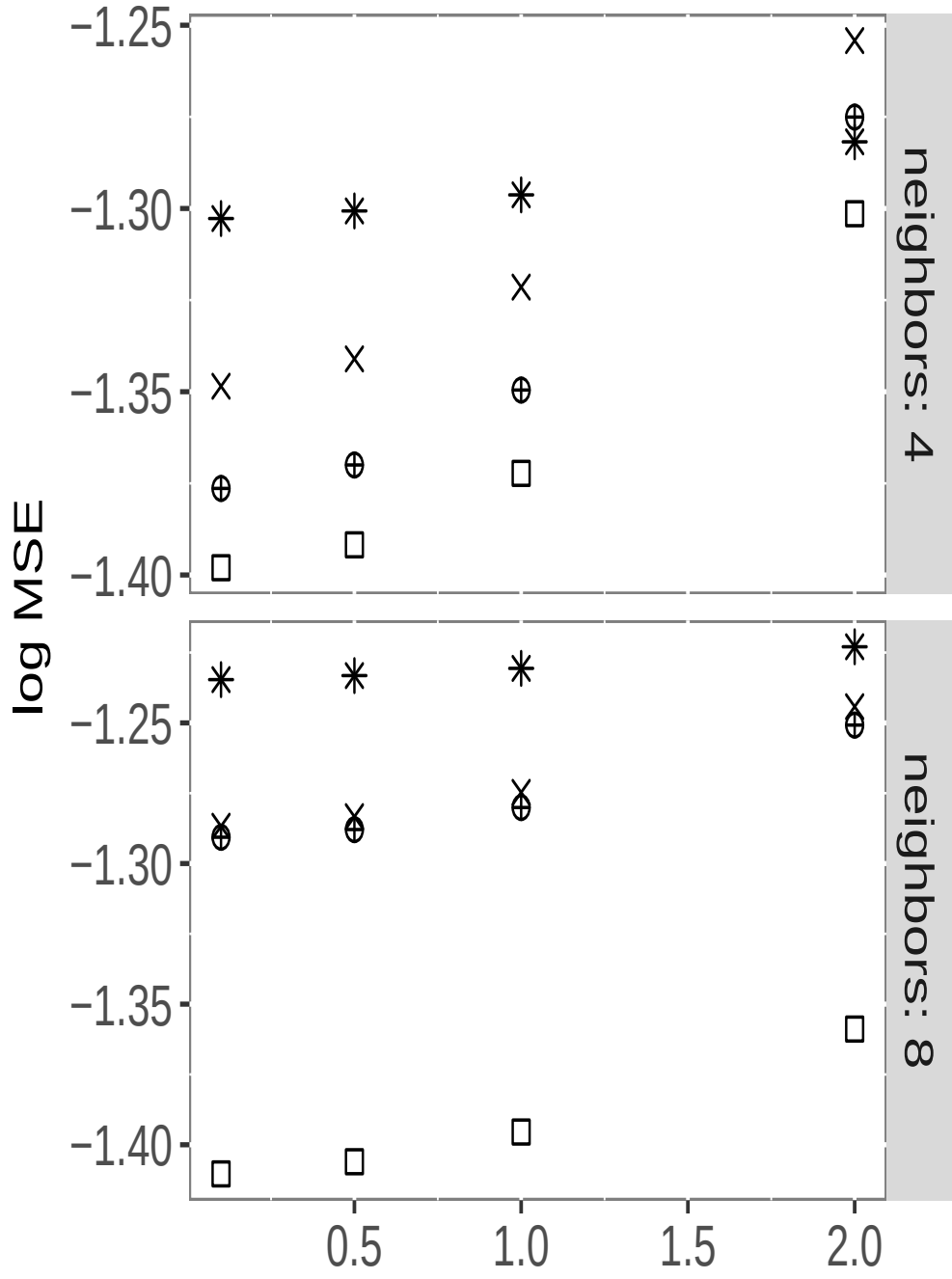
- Interference is linear $f_v(A) = \gamma|A|$
- "CRD strat" is the treated degree post-stratified estimator—it is the correct first order linear estimator!
- "restricted0" is naive partitioning
- "restricted" partitions by degree but tries to make sure there are no edges.

Realistic graphs: small world

- G is a small world graph:

  Start with n neighbors

  Rewiring probability: 0.2
- Fractional interference function



What about sparse graphs?

We need a little more math and a little more randomness to get bounds that still go to zero

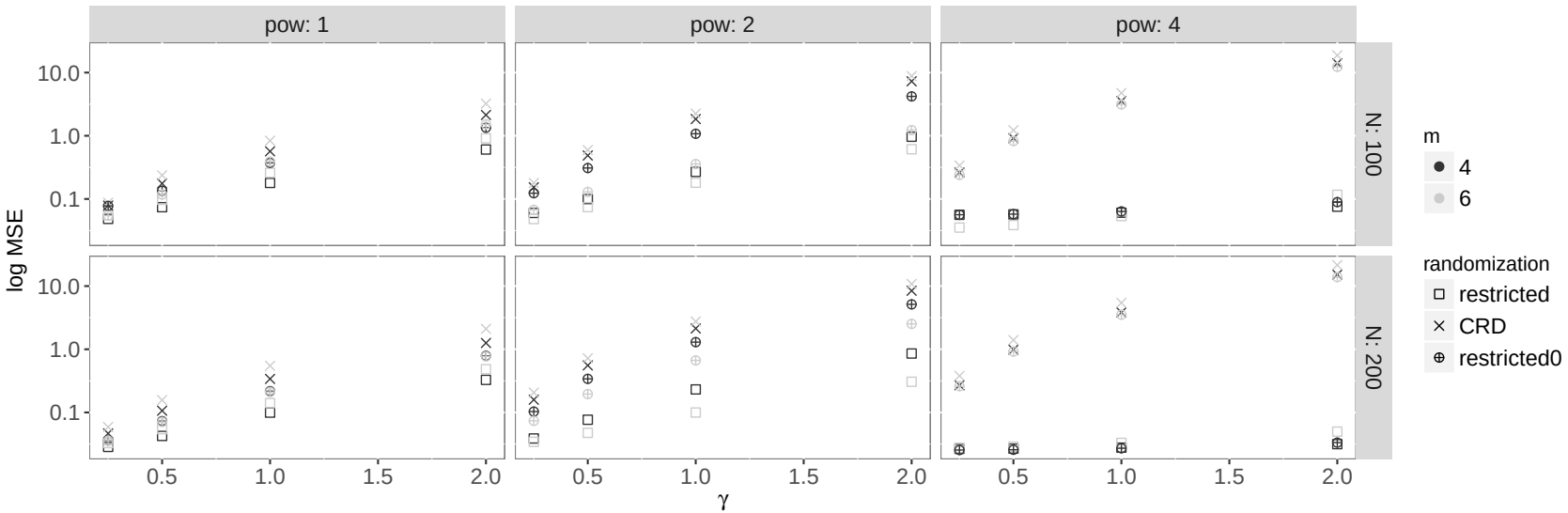
What about fancier interference?

Everything holds for $f_v = f_{\text{type}(v)}$(mutatis mutandis)

Preferential attachment graphs

$G \sim \mathrm{PA}(N, \mathrm{pow}, m)$

The new vertex forms an edge with an existing vertex v with probability proportional to $d(v)^{\mathrm{pow}}$. Each new vertex forms m new edges.



Natural extension to homophily

- Nodes are often similar in behavior because of underlying traits.

Natural extension to homophily

- Nodes are often similar in behavior because of underlying traits.
- We can sometimes typify the nodes.

Natural extension to homophily

- Nodes are often similar in behavior because of underlying traits.
- We can sometimes typify the nodes.
- Essentially working with

$$y_v = x_v + 1_T(v)t_v + f_v(T \cap N(v))$$

where x_v also includes covariate information.

Natural extension to homophily

- Nodes are often similar in behavior because of underlying traits.
- We can sometimes typify the nodes.
- Essentially working with

$$y_v = x_v + 1_T(v)t_v + f_v(T \cap N(v))$$

  where x_v also includes covariate information.
- Given such a collection of types Π and a bound on the variability of individuals inside each type σ^2 we have similar looking bounds on bias and MSE.

Natural extension to homophily

- Nodes are often similar in behavior because of underlying traits.
- We can sometimes typify the nodes.
- Essentially working with

$$y_v = x_v + 1_T(v)t_v + f_v(T \cap N(v))$$

  where x_v also includes covariate information.
- Given such a collection of types Π and a bound on the variability of individuals inside each type σ^2 we have similar looking bounds on bias and MSE.
- Need information on $(x_v - \sum_{v \in \pi} x_v))^2$.

Natural extension to homophily

- Nodes are often similar in behavior because of underlying traits.
- We can sometimes typify the nodes.
- Essentially working with

$$y_v = x_v + 1_T(v)t_v + f_v(T \cap N(v))$$

  where x_v also includes covariate information.
- Given such a collection of types Π and a bound on the variability of individuals inside each type σ^2 we have similar looking bounds on bias and MSE.
- Need information on $(x_v - \sum_{v \in \pi} x_v))^2$.
- Interesting conclusion: if we can identify these "types" well then a new cluster-randomized-design is reasonable for estimating the direct effect: treat half of every "type".

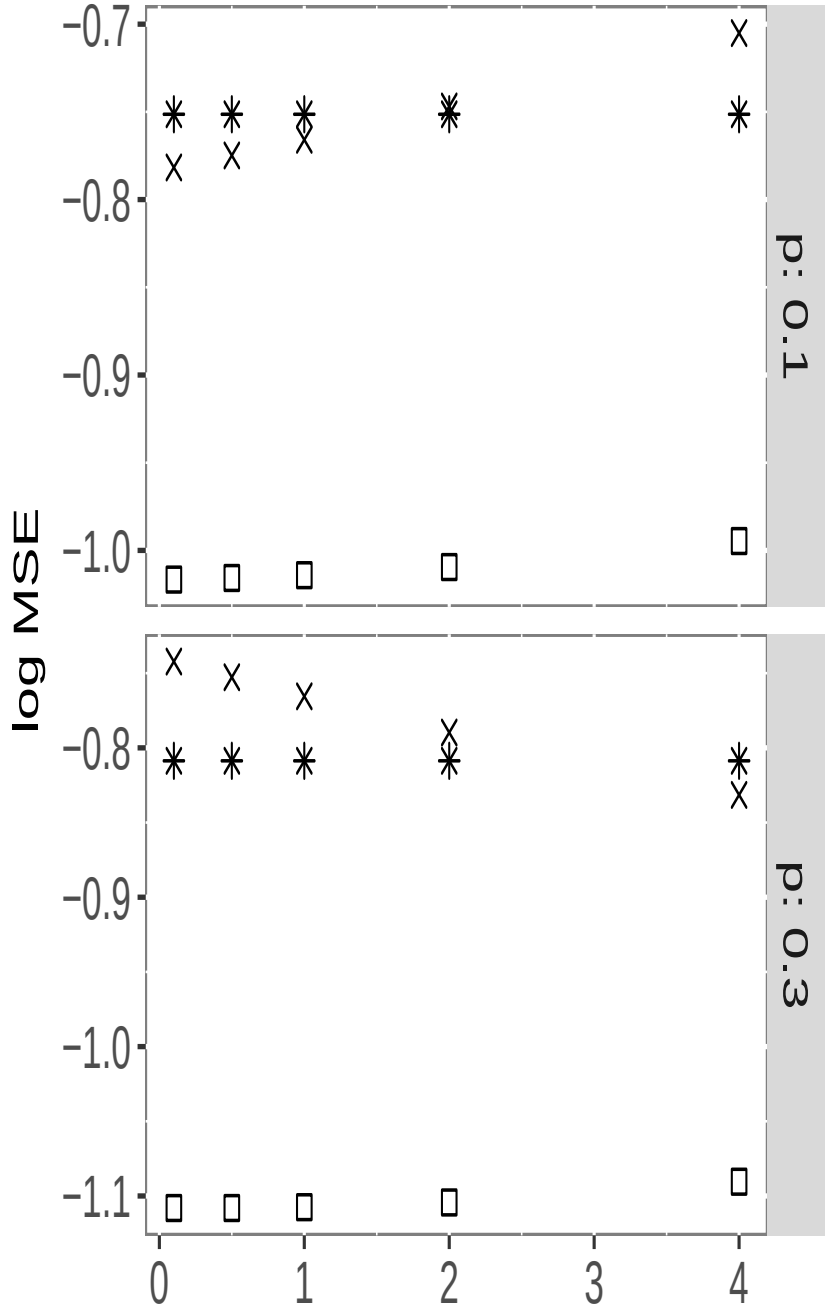
Simulation: stochastic blockmodels and homophily

- G is a stochastic blockmodel with 2 groups.
- Size of group 1: 25. Size of group 2: 75
- $p(a_{ij} = 1 | g_i, g_j)$:

  $p(a_{vv'} = 1 | 1, 1) = 0.5 + p$

  $p(a_{vv'} = 1 | 2, 2) = 0.5 - p$

  $p(a_{vv'} = 1 | 1, 2) = 0.1$
- $x_v | g_v = 1 \sim N(-2, 1)$ and $x_v | g_v = 2 \sim N(2, 1)$
- $f_v(A) = \gamma |A| / d_{\max}$

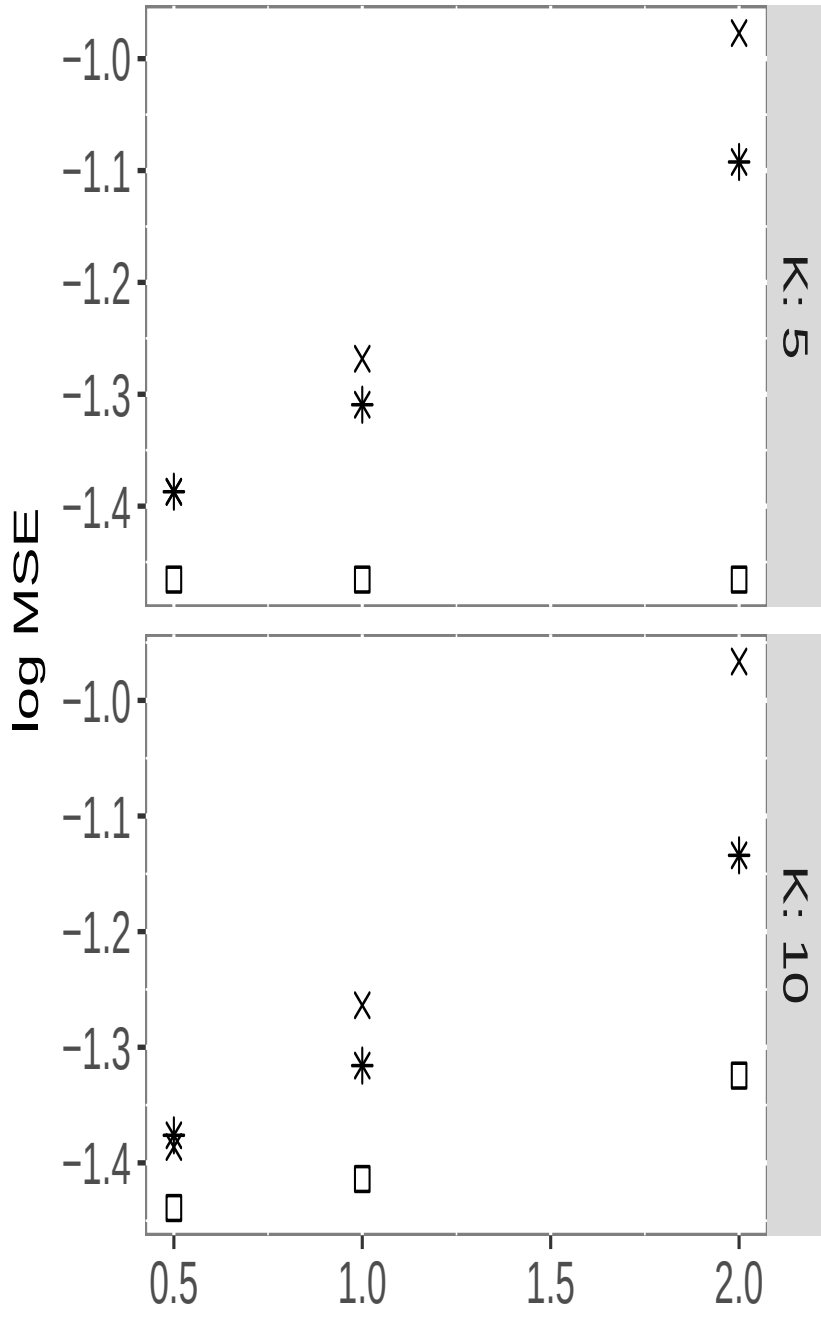


What happens under misspecification?

- G is composed of independent K stars.
- "influencer" interference function:

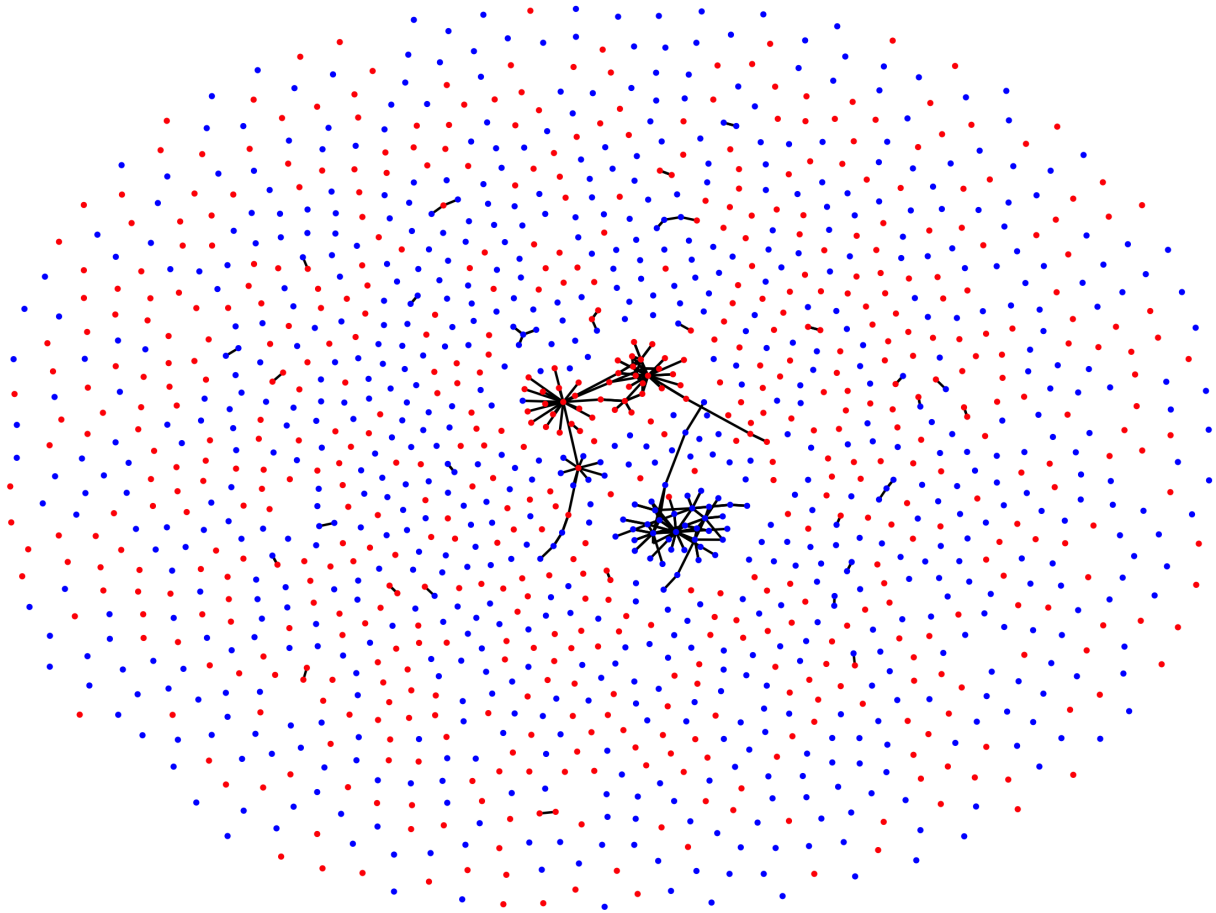
  central node interferes BUT

  cannot be interfered with
- Each influencer interferes at a random rate



Why should we care? Twitter echo chamber network

Work with Chris Bail (Duke Sociology), Mary Beth Fallin Hunzaker (NYU), Taylor Brown, Marcus Mann, Friedolin Merhout, Haohan Chen, John Bumpus, Lisa Argyle (Princeton) and Jaemin Lee.



Exposure to Opposing Views can Increase Political Polarization: Evidence from a Large-Scale Field Experiment on Social Media. Proceedings of the National Academy of Sciences (2018)

What have we done and where to do we go?

- We have a new design for experiments on networks.
  - Gives estimates of the direct effect.
  - Controls bias and MSE!

  Jagadeesan, Pillai and Volfovsky. Designs for estimating the treatment effect in networks with interference. arXiv:1705.08524
- Current/planned theory and methods projects:
  1. Graph matching methodology for matching units in observational studies
  2. Randomization schemes for peer effects
  3. Sampling perfect quasi-colorings (Markov chain)
- Current/planned applied projects:
  1. Information propagation through online networks
  2. Peer influence as a force of political polarization
  3. Efficacy of vaccine and non-pharmaceutical interventions

Thank you!

Of interest

- SAMSI programs on Machine Learning and Causal Inference happening in Fall 2019 and Spring 2020.
- NIPS, UAI, ICML all have causal workshops.
- We are looking to bring students to Duke for summer projects.

Machine Learning and Statistics Projects at Duke

Many interested faculty!

- Alexander Volfovsky—networks, causal inference
- David Dunson—scalable inference
- Galen Reeves—information theory, networks
- Jason Xu—constrained estimation, optimization
- Sayan Mukherjee—geometric data analysis

Come work with us!

Sample projects

Jason Xu

- statistical methodology for estimation with constraints.
- broad framework forencoding common constraints (sparsity, low-rank, shape, and others) as projections onto sets
- algorithms that generalize EM work well even on non-convex objectives
- current work:characterizing statistical rates, improving scalability,and developing theory toward developing a testing framework

Example project: Non-parametric regression/posterior inference over the space of dissimilarity/distance matrices. This generalizes morecommon approaches to metric learning, and has immediate applicationsin latent network models (related toHoff's work) andlarge-scale matching for causal inference (related toVolfovsky's work), among others

Applications

- We will send out information about applications shortly.
- Due November or December.
- We will have some projects listed but we are not limited to those.
- Students with and without funding should apply.
- Directed towards students between years 1 and 2 of masters but applications are open to everyone interested in pursuing a PhD in statistics and machine learning.

Proposition: total variation bound

The following proposition bounds the L^2 norm of $\|D_{\mathrm{T}}\|_{\mathbf{d}_{\mathrm{w}}}$.

Proposition

Fix $\mathcal{P} \in \binom{V(G)}{r,\ldots,r}$ *and let* $\mathrm{T} = \mathrm{T}_{\vec{B},\mathcal{P}}$. *We have*

$$\sqrt{\mathbb{E}_{\vec{B}} \|D_{\mathrm{T}}\|_{\mathbf{d}_{\mathrm{w}}}^2} \leq \frac{K_1}{\sqrt{pq}n} C_{\mathcal{P}} + \frac{1}{rn} \sum_{v \in V(G)} \frac{4K_2}{\sqrt{d(v)}} + \frac{1}{pqn} \sum_{v \in V(G)} \frac{|\mathcal{P}_v \cap \mathcal{N}(v)|}{d(v)}$$

The idea behind the proof of the Proposition is to bound the contributions of each vertex to the left-hand-side, and use the fact that $\mathrm{T} \cap S_i$ and $\mathrm{T} \cap S_j$ are independent for $i \neq j$, where $\mathcal{P} = (S_1, \ldots, S_n)$.

More (different) science

Interference/homophily makes network experiments hard...

More (different) science

Interference/homophily makes network experiments hard... and some randomization schemes are better than others.

More (different) science

Interference/homophily makes network experiments hard... and some randomization schemes are better than others.

Observational studies with network data are hard even without any formal interference or homophily...

Observational studies and entangled treatments

joint work with Panos Toulis at Chicago Booth and Edoardo Airoldi at Harvard

- Most work concentrates on questions of interference of outcomes.
- Lets take a step back from that — what if the treatments are entangled?

Observational studies and entangled treatments

joint work with Panos Toulis at Chicago Booth and Edoardo Airoldi at Harvard

- Most work concentrates on questions of interference of outcomes.
- Lets take a step back from that — what if the treatments are entangled?

- Treatment: number of new friends in an online game.
- Treatment: popularity measure of a website due to new links.
- Treatment: number of new professional connections.
- Treatment: number of new people in a working group.

pre-treatment network

There are two individuals and the pre-treatment period network G^- is disconnected:

Toy example

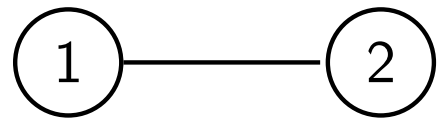
pre-treatment network

There are two individuals and the pre-treatment period network G^- is disconnected:

post-treatment network

No one is treated, that is $Y_1(0,0)$ and $Y_2(0,0)$ are observed.

OR



Both are treated, that is $Y_1(1,1)$ and $Y_2(1,1)$ are observed.

Toy example

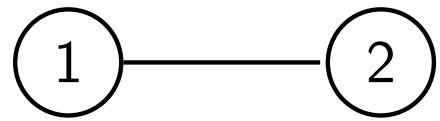
pre-treatment network

There are two individuals and the pre-treatment period network G^- is disconnected:

post-treatment network

No one is treated, that is $Y_1(0,0)$ and $Y_2(0,0)$ are observed.

OR



Both are treated, that is $Y_1(1,1)$ and $Y_2(1,1)$ are observed.

The treatment is "number of new friends" which is an edge count — and we can't observe one person with an edge and one without.

Causal inference with networks

- We will use the potential outcomes framework.
- There are n units that are connected in some network G^-.

Causal inference with networks

- We will use the potential outcomes framework.
- There are n units that are connected in some network G^-.
- Treatment is a function of a change of the network G^- to a network G^+.

  For example $Z_i = f_i(G^-, G^+) = d_i(G^+) - d_i(G^-)$

Causal inference with networks

- We will use the potential outcomes framework.
- There are n units that are connected in some network G^-.
- Treatment is a function of a change of the network G^- to a network G^+.

  For example $Z_i = f_i(G^-, G^+) = d_i(G^+) - d_i(G^-)$
- What's special about this world?

Causal inference with networks

- We will use the potential outcomes framework.
- There are n units that are connected in some network G^-.
- Treatment is a function of a change of the network G^- to a network G^+.

  For example $Z_i = f_i(G^-, G^+) = d_i(G^+) - d_i(G^-)$
- What's special about this world?
- No interference but notation still requires us to write $Y_i(Z_1, \ldots, Z_n)$ as the potential outcome of individual i under treatment vector $Z = (Z_1, \ldots, Z_n)$.

Causal inference with networks

- We will use the potential outcomes framework.
- There are n units that are connected in some network G^-.
- Treatment is a function of a change of the network G^- to a network G^+.

  For example $Z_i = f_i(G^-, G^+) = d_i(G^+) - d_i(G^-)$
- What's special about this world?
- No interference but notation still requires us to write $Y_i(Z_1, \ldots, Z_n)$ as the potential outcome of individual i under treatment vector $Z = (Z_1, \ldots, Z_n)$.
- Still in an observational framework so need to understand how to perform matching/weighting.

Causal inference with networks

- We will use the potential outcomes framework.
- There are n units that are connected in some network G^-.
- Treatment is a function of a change of the network G^- to a network G^+.

  For example $Z_i = f_i(G^-, G^+) = d_i(G^+) - d_i(G^-)$
- What's special about this world?
- No interference but notation still requires us to write $Y_i(Z_1, \ldots, Z_n)$ as the potential outcome of individual i under treatment vector $Z = (Z_1, \ldots, Z_n)$.
- Still in an observational framework so need to understand how to perform matching/weighting.
- Many estimands of interest:

$$\tau_m = E(Y_i(m+1)) - E(Y_i(m)).$$

So what goes wrong?

- Classical methods assume that no interference means we can write $Y_i(Z_i)$ and will in turn model the following propensity:

$$e(k, X_i) = P(Z_i = k | X_i, G^-)$$

So what goes wrong?

- Classical methods assume that no interference means we can write $Y_i(Z_i)$ and will in turn model the following propensity:

$$e(k, X_i) = P(Z_i = k | X_i, G^-)$$

- What's the problem here? These $e(k, X_i)$ are actually estimated conditional on the post treatment network G^+!

So what goes wrong?

- Classical methods assume that no interference means we can write $Y_i(Z_i)$ and will in turn model the following propensity:

$$e(k, X_i) = P(Z_i = k | X_i, G^-)$$

- What's the problem here? These $e(k, X_i)$ are actually estimated conditional on the post treatment network G^+!
- We need to marginalize over the post treatment network:

$$P(Z_i = k | X, G^-) = \int_{f_i(G^-, G^+)=k} p(G^+ | G^-, X) d\mu(G^+)$$

So what goes wrong?

- Classical methods assume that no interference means we can write $Y_i(Z_i)$ and will in turn model the following propensity:

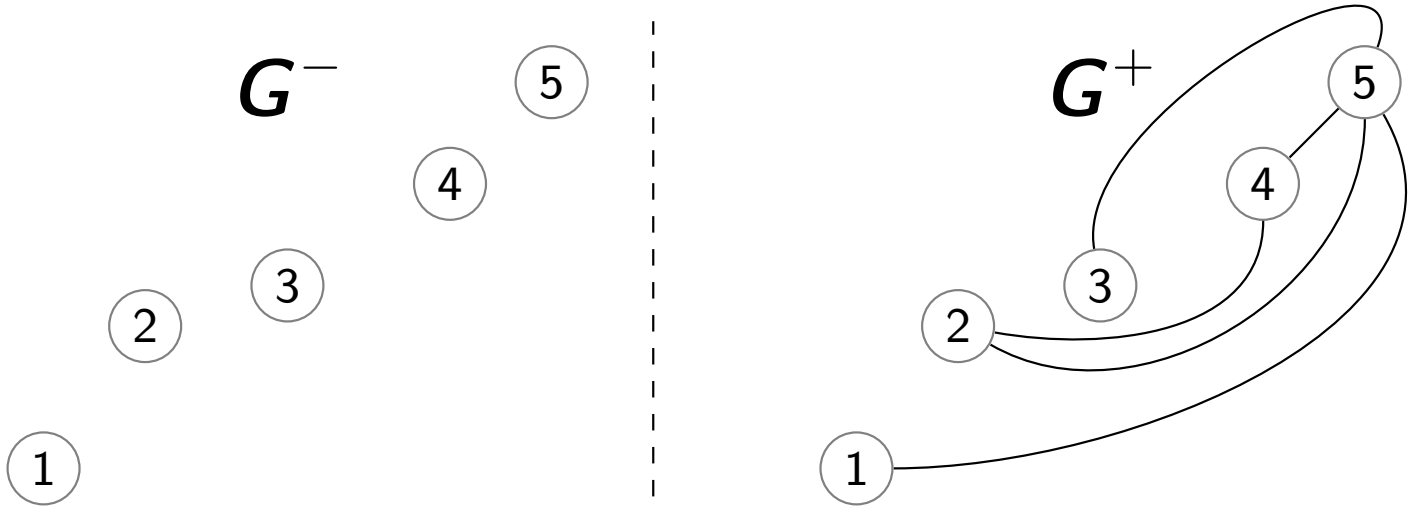
$$e(k, X_i) = P(Z_i = k | X_i, G^-)$$

- What's the problem here? These $e(k, X_i)$ are actually estimated conditional on the post treatment network G^+!
- We need to marginalize over the post treatment network:

$$P(Z_i = k | X, G^-) = \int_{f_i(G^-, G^+)=k} p(G^+ | G^-, X) d\mu(G^+)$$

- This accounts for the uncertainty in the treatment due to the network evolving from G^- to G^+.

Numerical example

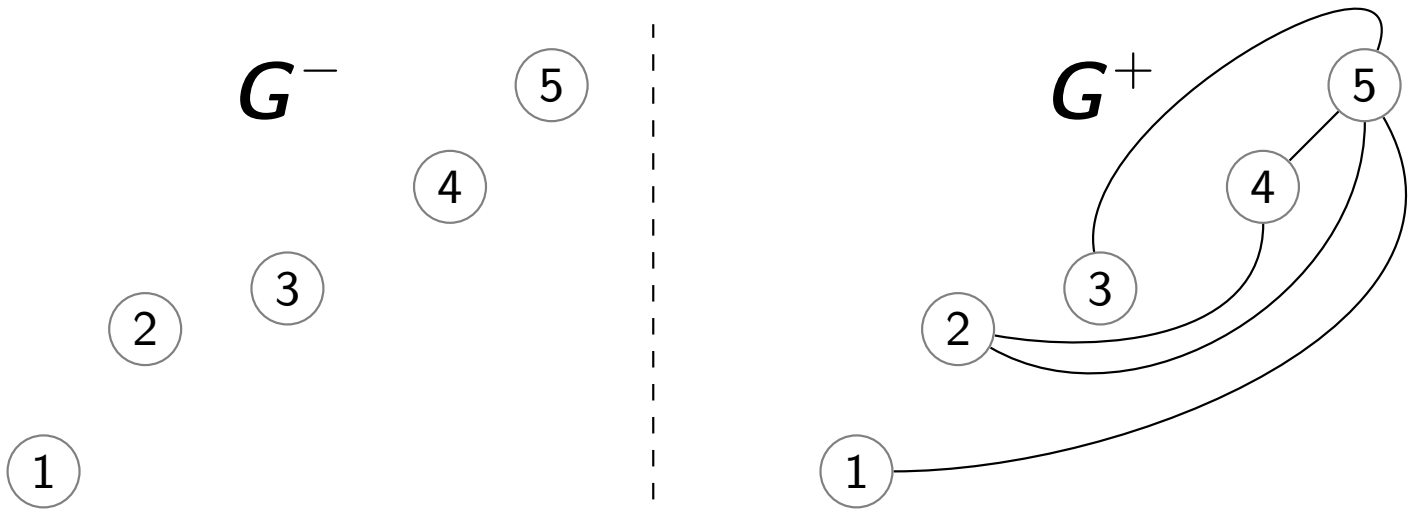


The network G^- is empty and G^+ has independent edges, each of which has probability

$$P(g_{ij}^+ = 1 | G^-, X) \propto \exp(X_i X_j + 1).$$

unit	X_i	Z_i	Y_i^{obs}
1	-5	1	0
2	-1	2	0
3	0	1	1
4	3	2	1
5	10	4	0

Numerical example



The network G^- is empty and G^+ has independent edges, each of which has probability

$$P(g_{ij}^+ = 1 | G^-, X) \propto \exp(X_i X_j + 1).$$

unit	X_i	Z_i	Y_i^{obs}
1	-5	1	0
2	-1	2	0
3	0	1	1
4	3	2	1
5	10	4	0

Numerical example – results

Ignoring information about the network we fit

$$P(Z_i = k|X_i) \propto \mathrm{Pois}(\lambda_i),\ \log \lambda_i = \alpha_\beta X_i$$

Numerical example – results

Ignoring information about the network we fit

$$P(Z_i = k|X_i) \propto \mathrm{Pois}(\lambda_i),\ \log \lambda_i = \alpha_\beta X_i$$

	propensity score for $Z_i = \ldots$						
unit (i)	0	1	2	3	4	5	...
1	0.37	0.37	0.18	0.06	0.02	0.00	...
2	0.24	0.34	0.25	0.12	0.04	0.01	...
3	0.21	0.33	0.26	0.13	0.05	0.02	...
4	0.13	0.26	0.27	0.19	0.10	0.04	...
5	0.02	0.08	0.15	0.20	0.20	0.15	...

Numerical example – results

Ignoring information about the network we fit

$$P(Z_i = k|X_i) \propto \mathrm{Pois}(\lambda_i),\ \log \lambda_i = \alpha_\beta X_i$$

	propensity score for $Z_i = \ldots$						
unit (i)	0	1	2	3	4	5	...
1	0.37	0.37	0.18	0.06	0.02	0.00	...
2	0.24	0.34	0.25	0.12	0.04	0.01	...
3	0.21	0.33	0.26	0.13	0.05	0.02	...
4	0.13	0.26	0.27	0.19	0.10	0.04	...
5	0.02	0.08	0.15	0.20	0.20	0.15	...

Numerical example – results

Ignoring information about the network we fit

$$P(Z_i = k|X_i) \propto \text{Pois}(\lambda_i), \ \log \lambda_i = \alpha_\beta X_i$$

	propensity score for $Z_i = \ldots$						
unit (i)	0	1	2	3	4	5	...
1	0.37	0.37	0.18	0.06	0.02	0.00	...
2	0.24	0.34	0.25	0.12	0.04	0.01	...
3	0.21	0.33	0.26	0.13	0.05	0.02	...
4	0.13	0.26	0.27	0.19	0.10	0.04	...
5	0.02	0.08	0.15	0.20	0.20	0.15	...

Numerical example – results

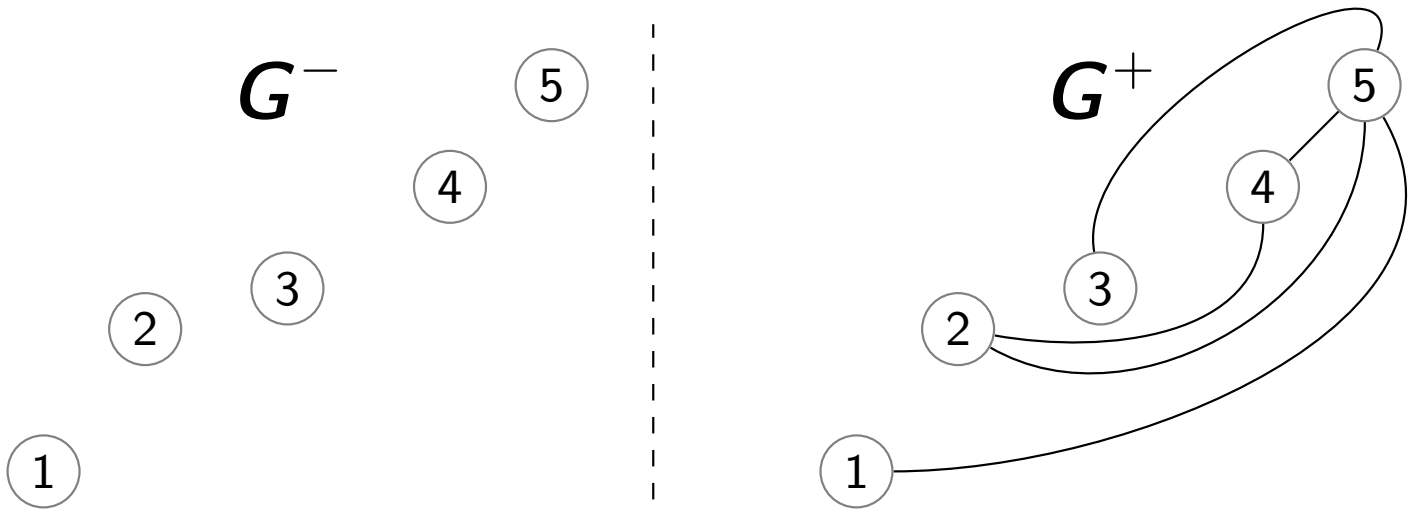
Ignoring information about the network we fit

$$P(Z_i = k|X_i) \propto \text{Pois}(\lambda_i), \ \log \lambda_i = \alpha_\beta X_i$$

	propensity score for $Z_i = \ldots$						
unit (i)	0	1	2	3	4	5	...
1	0.37	0.37	0.18	0.06	0.02	0.00	...
2	0.24	0.34	0.25	0.12	0.04	0.01	...
3	0.21	0.33	0.26	0.13	0.05	0.02	...
4	0.13	0.26	0.27	0.19	0.10	0.04	...
5	0.02	0.08	0.15	0.20	0.20	0.15	...

Set of units that have similar propensities to make one connection or two connections: $\mathcal{S} = \{1, 2, 3, 4\}$

Numerical example



The network G^- is empty and G^+ has independent edges, each of which has probability

$$P(g_{ij}^+ = 1 | G^-, X) \propto \exp(X_i X_j + 1).$$

unit	X_i	Z_i	Y_i^{obs}
1	-5	1	0
2	-1	2	0
3	0	1	1
4	3	2	1
5	10	4	0

Numerical example – results

Using the information about the network:

	propensity for $Z_i = \ldots$				
unit (i)	0	1	2	3	4
1	0.00	0.27	0.73	0.00	0.00
2	0.00	0.24	0.67	0.09	0.00
3	0.01	0.06	0.23	0.42	0.28
4	0.00	0.24	0.68	0.09	0.00
5	0.00	0.27	0.73	0.00	0.00

Set of units that have similar propensities to make one connection or two connections: $\mathcal{S} = \{1, 2, 4, 5\}$

Numerical example – results

Using the information about the network:

	propensity for $Z_i = \ldots$				
unit (i)	0	1	2	3	4
1	0.00	0.27	0.73	0.00	0.00
2	0.00	0.24	0.67	0.09	0.00
3	0.01	0.06	0.23	0.42	0.28
4	0.00	0.24	0.68	0.09	0.00
5	0.00	0.27	0.73	0.00	0.00

Set of units that have similar propensities to make one connection or two connections: $\mathcal{S} = \{1, 2, 4, 5\}$

Practical guide

- Integral is usually analytically intractable.
- Fit favorite model for $G^+|G^-, X$.
- Sample J networks from the fitted model.
- Use the samples $\{G^+_{(j)}\}$, $j = 1, \ldots, J$ to compute estimates $\hat{e}(k, X)$ of the propensity score $e(k, X)$:

$$\hat{e}(k, X) = \frac{1}{J} \sum_{j=1}^{J} \mathbb{I}\{f_i(G^-, G^+_{(j)}) = k\}$$

- Group according to estimated propensity scores.
- Compute estimates within groups and combine information across groups.

All nice in practice, but how does it work in theory?

- Define (approximate) similarity between propensity score models as

$$J(e,m) = |E[\frac{\nabla e(X)^t \nabla m(X)}{\|\nabla e(X)\|\|\nabla m(X)\|}]|$$

- $0 \leq J(e,m) \leq 1$ and $J(e,m) = 1 \implies e = \pm m$
- Let the probability of edge between two people be $\text{expit}(a + bX_i^t X_j)$. Let $X_i \sim N(0, \tau^2 I)$ and define treatment as

$$\begin{aligned} E[Z_i|X_i] &= \sum_{j \neq i} E[\text{expit}(a + bX_i^t X_j)|X_i] \\ &= (n-1)E[\text{expit}(a + b||X_i||\tau U)] = r(a, \|X_i\|) \end{aligned}$$

where $U \sim N(0,1)$ and r is monotone with respect to $\|X_i\|$.

All nice in practice, but how does it work in theory?

- We then have $\nabla e(X_i) = (n-1)\frac{\partial r(a,\|X_i\|)}{\partial \|X_i\|}\frac{1}{\|X_i\|}X_i$ and so

$$\begin{aligned}\nabla_e &\equiv E[\nabla e(X_i)/\|\nabla e(X_i)\|] \\ &= \text{sign}(\frac{\partial(a,\|X_i\|)}{\partial\|X_i\|})E[X_i/\|X_i\|] \\ &\propto E[X_i/\|X_i\|] = 0\end{aligned}$$

  by symmetry.

- Consider an alternative misspecified model $m_\beta(X) = h(\beta^t X)$ where h is monotone, then

$$J[e, m_\beta] = |\nabla_e^t \nabla_{m_\beta}| = |\frac{\beta^t \nabla_e}{\|\beta\|}$$

- In this case $J[e, m_\beta] = 0$.

What is going on?

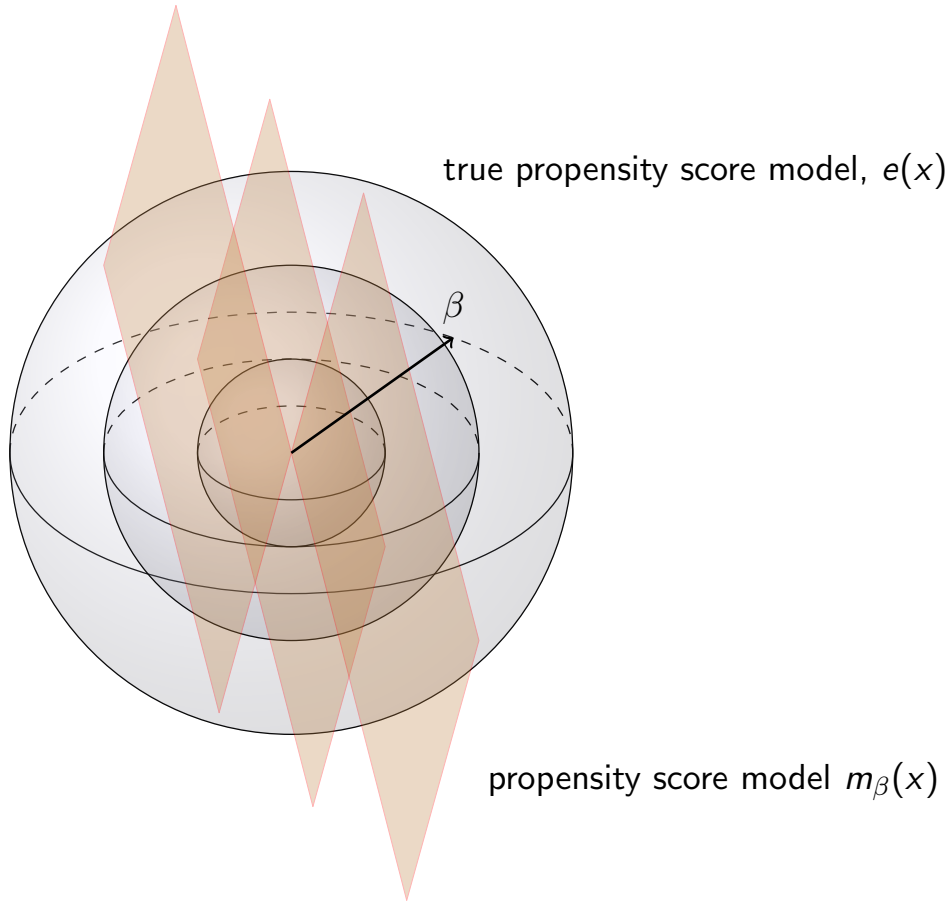


Figure: The contour surfaces for the true propensity score model, $e(x)$, are spherical because they only depend on the norm $||x||$. The contour surfaces of linear model $m_\beta(x)$ are hyperplanes oriented by vector β.

Some thoughts about moving forward

- We have a new design for experiments on networks.
  - Gives estimates of the direct effect.
  - Controls bias and MSE!
- How do we port this to observational studies?

- We develop entangled treatments in observational studies.
  - Theory for balancing of covariates.
  - Random graph connects with network analysis.
- How do we port this to randomization schemes?

- How do we do any of this fast?
- How do we communicate these ideas to practitioners?

Thank you!