Causal Inference in the Presence of Networks

Randomization and Observation

Alexander Volfovsky
Department of Statistical Science, Duke University

May 17, 2017
Graph Exploitation Symposium,

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: $(Y_i(0), Y_i(1))$ 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: $(Y_i(0), Y_i(1))$ 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.
  - ⋆ Applications: disease prevalence, social development, online advertising, business transactions.

Causal inference with networks

- How does the classical causal inference setting extend to these?

Causal inference with networks

- How does the classical causal inference setting extend to these?
- Some problems:

Causal inference with networks

- How does the classical causal inference setting extend to these?
- Some problems:
  - Homophily.

Causal inference with networks

- How does the classical causal inference setting extend to these?
- Some problems:
  - Homophily.
  - Interference.

Causal inference with networks

- How does the classical causal inference setting extend to these?
- Some problems:
  - Homophily.
  - Interference.
  - Entangled treatments.

Causal inference with networks

- 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

- 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

- 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

- 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 $(Z_1, \dots, Z_n)$ is given by $Y_i(Z_1, \dots, Z_n)$.
- Estimands of interest:
  - Total network: "maximal effect"
  - Direct effect: value of isolation

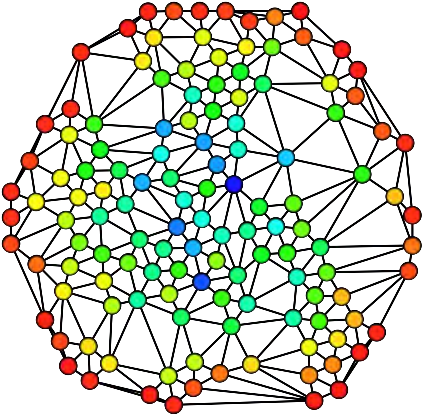
Causal inference with networks

- 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

- 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 $(Z_1, \dots, Z_n)$ is given by $Y_i(Z_1, \dots, Z_n)$.
- 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.

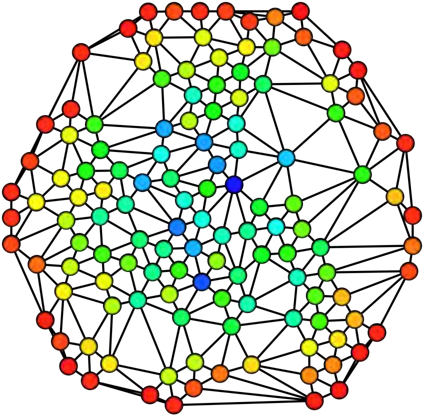
Some context: Facebook



- Facebook wants to change its' ad algorithm.

Source: Wikimedia

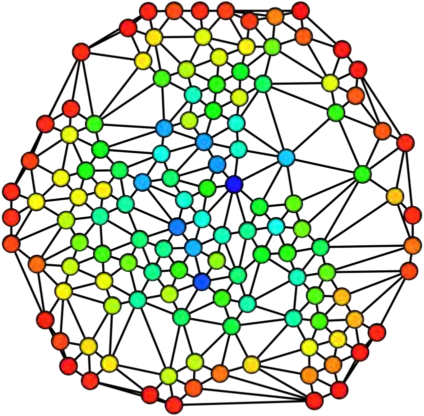
Some context: Facebook



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

Source: Wikimedia

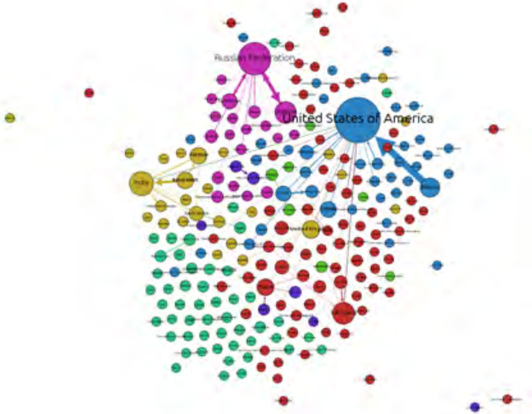
Some context: Facebook



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

Source: Wikimedia

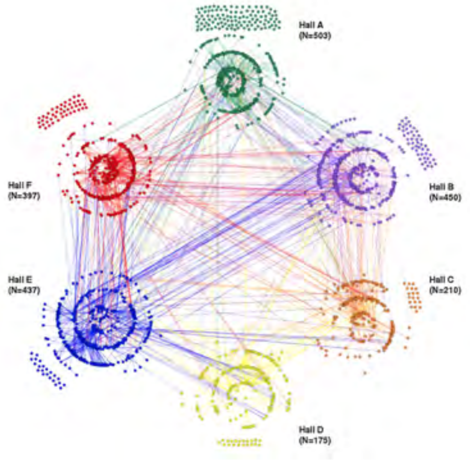
Some context: (im)migration



- 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/

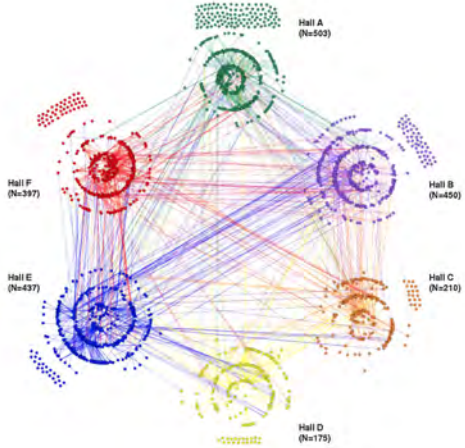
Some context: disease spread



- 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

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

Experimental design with networks

Perfect assignment:

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

Experimental design with networks

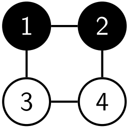
Perfect assignment:

- 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:

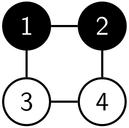
- Assign nodes to treatment to balance interference between treated and untreated nodes.
- Natural to think of as a quasi-coloring problem on graphs.



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.

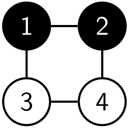


- Nodes 1,2: treated, 1 treated neighbor, 1 untreated.
- Nodes 3,4: untreated, 1 treated neighbor, 1 untreated.

Experimental design with networks

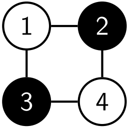
Perfect assignment:

- 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.

Perfect quasi-colorings

Do they exist?

Perfect quasi-colorings

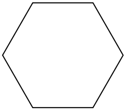
Do they exist?

- Short answer: yes.

Perfect quasi-colorings

Do they exist?

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

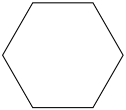


Hexagon

Perfect quasi-colorings

Do they exist?

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



Hexagon

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

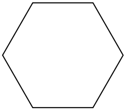
  B: $\{1, 2, 3\}$ then $d_B(2) = 2$ and $\nexists w : d_B(w) = 2$.

  B: $\{1, 2, 4\}$ then $d_B(3) = 2$ and $\nexists b : d_B(b) = 2$.

Perfect quasi-colorings

Do they exist?

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



Hexagon

- The nodes are $V(G) = \{1, \ldots, 6\}$:
  B: $\{1, 2, 3\}$ then $d_B(2) = 2$ and $\nexists w : d_B(w) = 2$.
  B: $\{1, 2, 4\}$ then $d_B(3) = 2$ and $\nexists b : d_B(b) = 2$.
- Do we need perfection?

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 Lipschitz (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 Lipschitz (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 Lipschitz (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 Lipschitz (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 Lipschitz (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 Lipschitz (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

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

What about variance?

- Simplification and complication...

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.
- 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.
- Let $\overrightarrow{d}(v) = (|T \cap \mathcal{N}(v)|, |\mathcal{N}(v) \setminus T|)$ be the bidegree of v.
- The bound on the variance is given as a function of

What about variance?

- Simplification and complication...
- Consider $f_v(S) = f(|S|, |\mathcal{N}(v) \setminus S)$ – this is symmetric interference.
- Let $\overrightarrow{d}(v) = (|T \cap \mathcal{N}(v)|, |\mathcal{N}(v) \setminus T|)$ be the bidegree of v.
- The bound on the variance is given as a function of
  - The partition: $P = \{\{w_1, w_1'\}, \ldots, \{w_n, w_n'\}\}$

What about variance?

- Simplification and complication...
- Consider $f_v(S) = f(|S|, |\mathcal{N}(v) \setminus S)$ – this is symmetric interference.
- Let $\overrightarrow{d}(v) = (|T \cap \mathcal{N}(v)|, |\mathcal{N}(v) \setminus T|)$ be the bidegree of v.
- The bound on the variance is given as a function of
  - The partition: $P = \{\{w_1, w_1'\}, \ldots, \{w_n, w_n'\}\}$
  - Constant K_1, K_2 that describe the cost of unbalanced treatment.

What about variance?

- Simplification and complication...
- Consider $f_v(S) = f(|S|, |\mathcal{N}(v) \setminus S)$ – this is symmetric interference.
- Let $\overrightarrow{d}(v) = (|T \cap \mathcal{N}(v)|, |\mathcal{N}(v) \setminus T|)$ be the bidegree of v.
- The bound on the variance is given as a function of
  - The partition: $P = \{\{w_1, w_1'\}, \ldots, \{w_n, w_n'\}\}$
  - Constant K_1, K_2 that describe the cost of unbalanced treatment.
  - $C_P = \frac{1}{d_{\max}} \sum_{\{w,w'\} \in P} |d(w) - d(w')|$

What about variance?

- Simplification and complication...
- Consider $f_v(S) = f(|S|, |\mathcal{N}(v) \setminus S)$ – this is symmetric interference.
- Let $\overrightarrow{d}(v) = (|T \cap \mathcal{N}(v)|, |\mathcal{N}(v) \setminus T|)$ be the bidegree of v.
- The bound on the variance is given as a function of
  - The partition: $P = \{\{w_1, w_1'\}, \dots, \{w_n, w_n'\}\}$
  - Constant K_1, K_2 that describe the cost of unbalanced treatment.
  - $C_P = \frac{1}{d_{\max}} \sum_{\{w,w'\} \in P} |d(w) - d(w')|$
  - Average of $1/\sqrt{d(v)}$.

What about variance?

- Simplification and complication...
- Consider $f_v(S) = f(|S|, |\mathcal{N}(v) \setminus S)$ – this is symmetric interference.
- Let $\overrightarrow{d}(v) = (|T \cap \mathcal{N}(v)|, |\mathcal{N}(v) \setminus T|)$ be the bidegree of v.
- The bound on the variance is given as a function of
  - The partition: $P = \{\{w_1, w'_1\}, \ldots, \{w_n, w'_n\}\}$
  - Constant K_1, K_2 that describe the cost of unbalanced treatment.
  - $C_P = \frac{1}{d_{\max}} \sum_{\{w,w'\} \in P} |d(w) - d(w')|$
  - Average of $1/\sqrt{d(v)}$.
- The bound is:

$$\frac{K_1}{n} C_p + \frac{2K_2}{n} \sum_{v \in V(G)} \frac{1}{\sqrt{d(v)}}$$

What about variance?

- Simplification and complication...
- Consider $f_v(S) = f(|S|, |\mathcal{N}(v) \setminus S)$ – this is symmetric interference.
- Let $\overrightarrow{d}(v) = (|T \cap \mathcal{N}(v)|, |\mathcal{N}(v) \setminus T|)$ be the bidegree of v.
- The bound on the variance is given as a function of
  - The partition: $P = \{\{w_1, w_1'\}, \ldots, \{w_n, w_n'\}\}$
  - Constant K_1, K_2 that describe the cost of unbalanced treatment.
  - $C_P = \frac{1}{d_{\max}} \sum_{\{w,w'\} \in P} |d(w) - d(w')|$
  - Average of $1/\sqrt{d(v)}$.
- The bound is:

$$\frac{K_1}{n} C_p + \frac{2K_2}{n} \sum_{v \in V(G)} \frac{1}{\sqrt{d(v)}}$$

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

What about variance?

- Simplification and complication...
- Consider $f_v(S) = f(|S|, |\mathcal{N}(v) \setminus S)$ – this is symmetric interference.
- Let $\overrightarrow{d}(v) = (|T \cap \mathcal{N}(v)|, |\mathcal{N}(v) \setminus T|)$ be the bidegree of v.
- The bound on the variance is given as a function of
  - The partition: $P = \{\{w_1, w_1'\}, \ldots, \{w_n, w_n'\}\}$
  - Constant K_1, K_2 that describe the cost of unbalanced treatment.
  - $C_P = \frac{1}{d_{\max}} \sum_{\{w,w'\} \in P} |d(w) - d(w')|$
  - Average of $1/\sqrt{d(v)}$.
- The bound is:

$$\frac{K_1}{n} C_p + \frac{2K_2}{n} \sum_{v \in V(G)} \frac{1}{\sqrt{d(v)}}$$

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

Better partitions

- Order the vertices as $V(G) = \{w_1^\star, w_1^{\star\prime}, \dots, w_n^\star, w_n^{\star\prime}\}$ 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
$$P^\star = \{\{w_1^\star, w_1^{\star\prime}\}, \dots, \{w_n^\star, w_n^{\star\prime}\}\}$$
- By definition: $C_p \leq 1$.
- The bias and L^2 norm are bounded by
$$\frac{2K_2}{d_{\min}} \text{ and } \frac{K_1}{n} + \frac{2K_2}{\sqrt{d_{\min}}}$$
- In a dense graph we have $n \to \infty$ implies $d_{\min} \to \infty$.
- So MSE goes to zero!

What about sparse graphs?

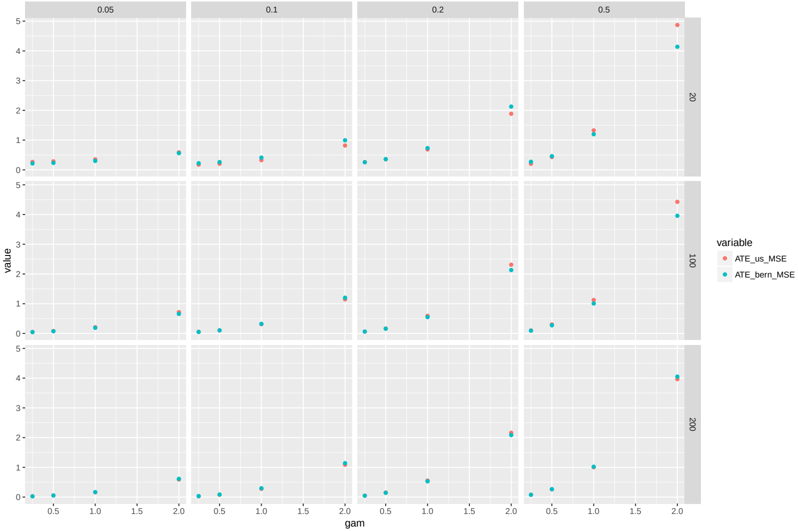
We need a little more math to get bounds that still go to zero

What about fancier interference?

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

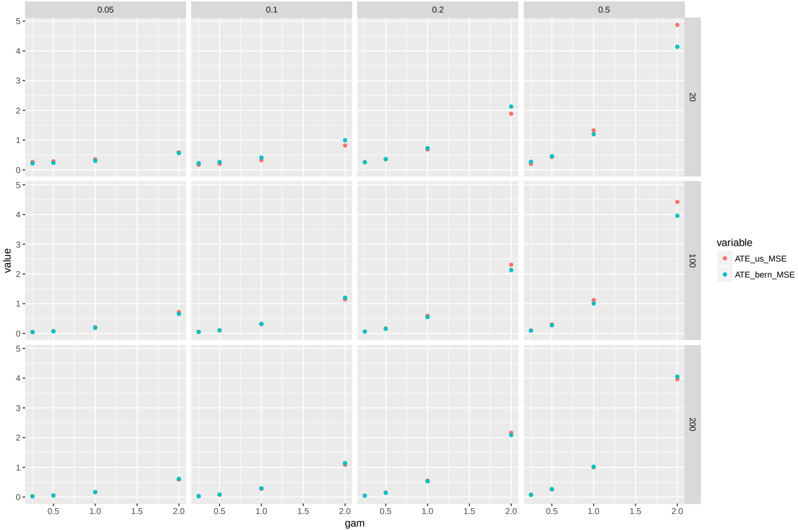
Small simulation

Linear interference, Erdos-Renyi graph



Small simulation

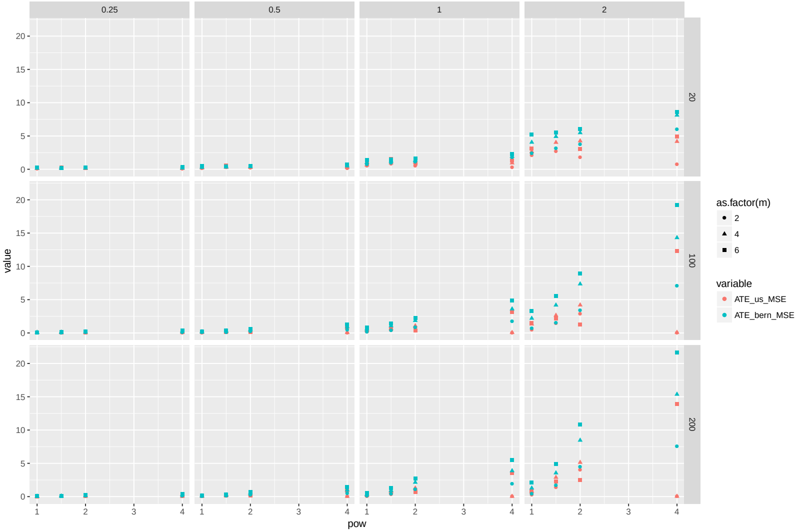
Linear interference, Erdos-Renyi graph



ER graphs have lots of nodes with the same degree...

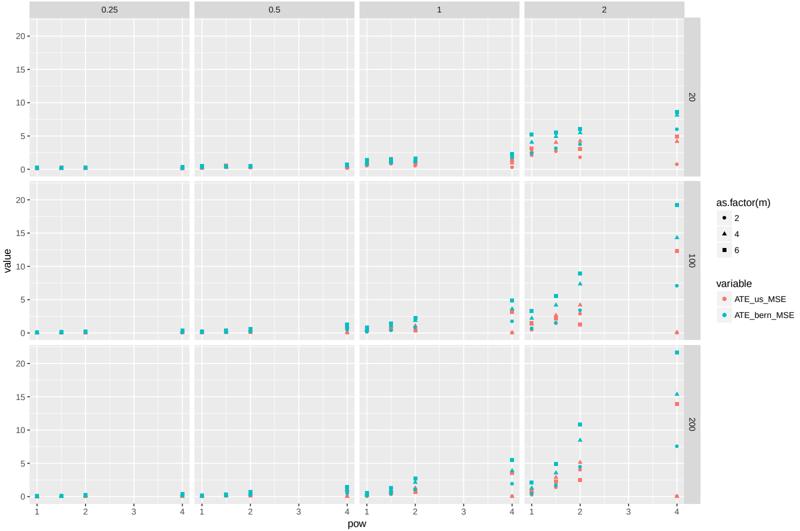
Small simulation

Linear interference, Preferential-Attachment graph



Small simulation

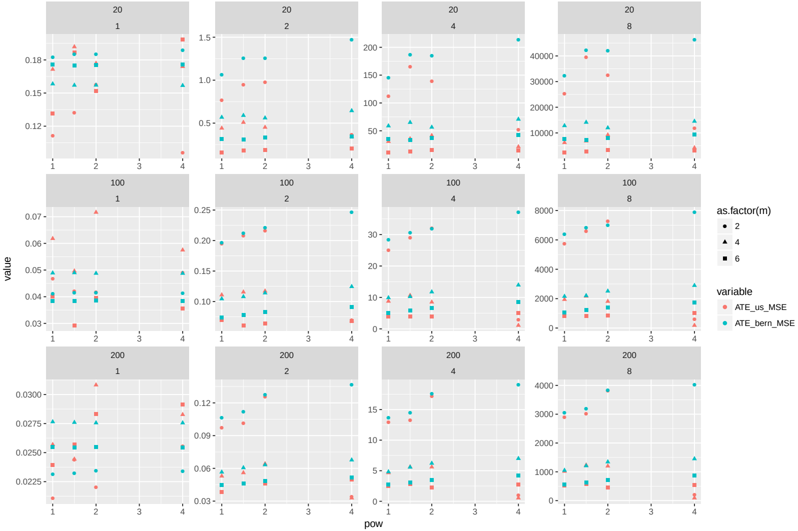
Linear interference, Preferential-Attachment graph



Lots of degree heterogeneity...

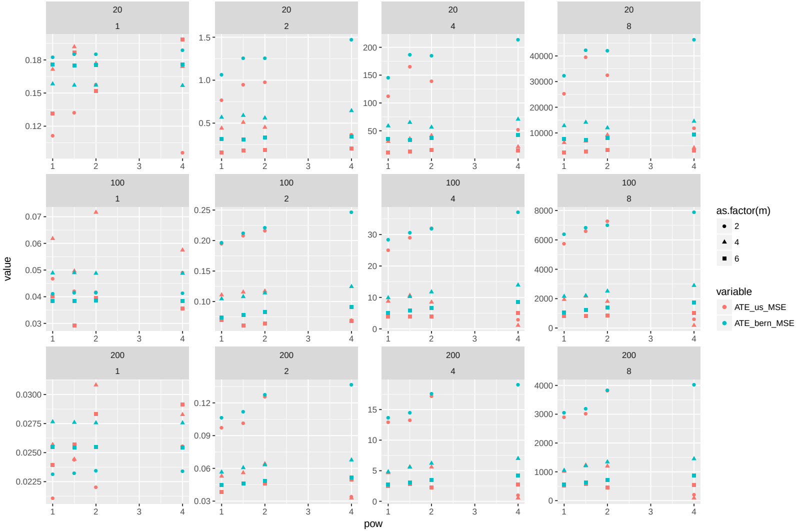
Small simulation

Generic interference, Preferential-Attachment graph



Small simulation

Generic interference, Preferential-Attachment graph



Lots of degree heterogeneity... and possibly huge interference

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

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

Natural extension to homophily

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

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

- 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

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

- 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

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

- 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".

Interference/homophily makes network experiments hard...

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

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 Edo 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 Edo 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.

Toy example

pre-treatment network

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

Toy example

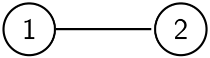
pre-treatment network

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

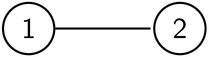
pre-treatment network

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:

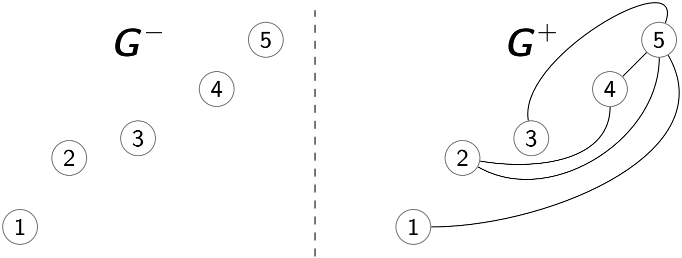
$$e(k, X_i) = P(Z_i = k | X_i, G^-)$$

- 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

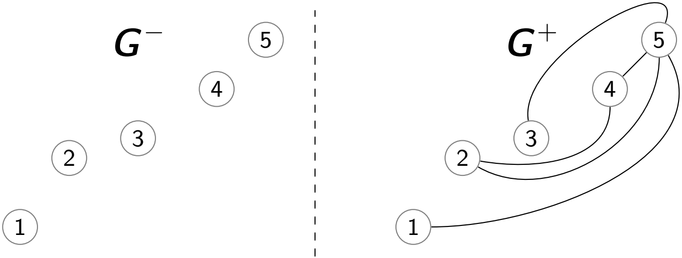


The network G^- is empty and G^+ has independent edges, each of which has probability

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

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

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

	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 $Z_i = \dots$						
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

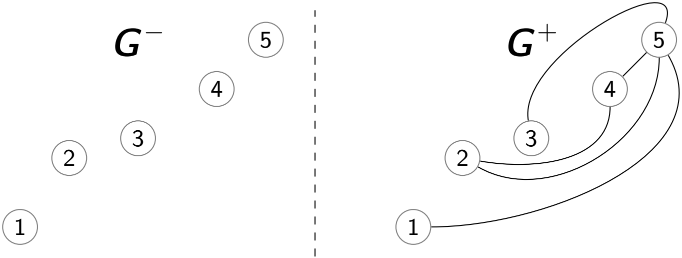
Ignoring information about the network we fit

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

	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	...

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 $Z_i = \dots$				
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:

unit (i)	propensity for $Z_i = \ldots$ 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?

- Matching on the correct propensity score means the assignment is ignorable.
- In that case the conditional distributions of covariates for treated and control are the same.
- What happens if you condition on the wrong propensity score?
- Let $S_c(X), S_w(X)$ be the correct and wrong propensity score given covariates X. For $\mu(s) = E(S_c|S_w = s)$ we have

$$
\begin{aligned}
E(X|Z = 1, S_w = s) - E(X|Z = 0, S_w = s) \\
= \frac{cov(S_c(X), X|S_w = s)}{\mu(s)(1 - \mu(s))}
\end{aligned}
$$

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!

Why is this randomization better?

Example

- Let $V(G) = \{v_1, \ldots, v_{2k}, w_1, \ldots, w_{2k}\} = V \cup W$.

Why is this randomization better?

Example

- Let $V(G) = \{v_1, \ldots, v_{2k}, w_1, \ldots, w_{2k}\} = V \cup W$.
- Let $E(G) = \{(v_i, v_j)\}$

Why is this randomization better?

Example

- Let $V(G) = \{v_1, \ldots, v_{2k}, w_1, \ldots, w_{2k}\} = V \cup W$.
- Let $E(G) = \{(v_i, v_j)\}$
- 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 $V(G) = \{v_1, \ldots, v_{2k}, w_1, \ldots, w_{2k}\} = V \cup W$.
- Let $E(G) = \{(v_i, v_j)\}$
- 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 $V(G) = \{v_1, \ldots, v_{2k}, w_1, \ldots, w_{2k}\} = V \cup W$.
- Let $E(G) = \{(v_i, v_j)\}$
- 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 $\alpha = |T \cap V|$ then

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

Why is this randomization better?

Example

- Let $V(G) = \{v_1, \ldots, v_{2k}, w_1, \ldots, w_{2k}\} = V \cup W$.
- Let $E(G) = \{(v_i, v_j)\}$
- 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 $\alpha = |T \cap V|$ then

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

- When we use our partition scheme we have $\alpha = k$ and so $\xi = -\frac{\gamma}{2}$ is independent of the size of the graph.

Why is this randomization better?

Example

- Let $V(G) = \{v_1, \ldots, v_{2k}, w_1, \ldots, w_{2k}\} = V \cup W$.
- Let $E(G) = \{(v_i, v_j)\}$
- 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 $\alpha = |T \cap V|$ then

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

- When we use our partition scheme we have $\alpha = k$ and so $\xi = -\frac{\gamma}{2}$ is independent of the size of the graph.
- Now letting T be uniform on all possible partitions it is clear that while $E\xi \to 0$ we have a growing variance!