Causal clustering: design of cluster experiments under network interference

Davide Viviano (Harvard University) w/ Lihua Lei, Guido Imbens, Brian Karrer, Okke Schrijvers, Liang Shi

November, 2023

Davide Viviano Suusi clusterine, destru of cluster reporting November, 2023 1/18

Example

- Ex1 Informing users on platforms for voting (Bond et al., 2012)
- Ex2 Informing farmers exposed to environmental disasters to increase insurance take-up in rural China (Cai et al., 2015)

Davide Viviano Sassa distance design of classic constants November, 2023 2 / 18

Example

- Ex1 Informing users on platforms for voting (Bond et al., 2012)
- Ex2 Informing farmers exposed to environmental disasters to increase insurance take-up in rural China (Cai et al., 2015)
 - Q: Effect of scaling up the treatment on the pop? (overall effect)
 - ⇒ Challenge: information spillovers betw/ friends



Example

- Ex1 Informing users on platforms for voting (Bond et al., 2012)
- Ex2 Informing farmers exposed to environmental disasters to increase insurance take-up in rural China (Cai et al., 2015)
 - Q: Effect of scaling up the treatment on the pop? (overall effect)
 - ⇒ Challenge: information spillovers betw/ friends
 - ⇒ Cluster designs (Karrer et al., 2022; Duflo and Banerjee, 2018)
 - (i) Researchers partition the population into subgroups (e.g., villages)
 - (ii) They randomize treatments at the cluster level
 - (iii) Estimator as difference in means betw/ treated and controls

Davide Viviano Causel clusterius, design of cluster experimer November, 2023 2 / 18

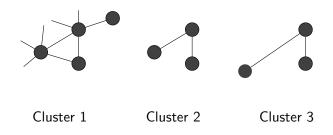


- Many independent clusters to capture overall effects
 - ⇒ Means difference betw/ treated and controls unbiased for overall effect

3 / 18

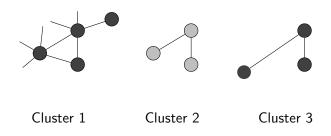
Davide Viviano Susal clustering: design of cluster experimer November, 2023

- Many independent clusters to capture overall effects
 - ⇒ Means difference betw/ treated and controls unbiased for overall effect



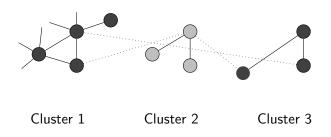
Davide Viviano November, 2023 3/18

- Many independent clusters to capture overall effects
 - ⇒ Means difference betw/ treated and controls unbiased for overall effect



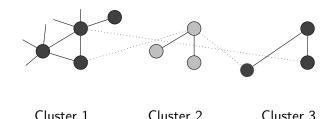
Davide Viviano November, 2023 4 / 18

- Many independent clusters to capture overall effects
 - ⇒ Means difference betw/ treated and controls unbiased for overall effect



Davide Viviano Sausal clusterine, design of cluster experiment November, 2023 4/18

- Many independent clusters to capture overall effects
 - ⇒ Means difference betw/ treated and controls unbiased for overall effect



In practice...

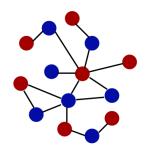
- ⇒ Not many clusters available (e.g., only one network)
- ⇒ Uncertainty over spillover (and direct) effects
- ⇒ Trade-offs in the choice and number of the clusters

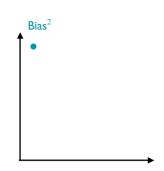
Davide Viviano

Consel clustering design of cluster according

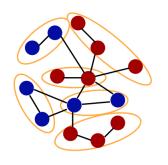
November, 2023

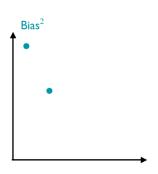
4/18



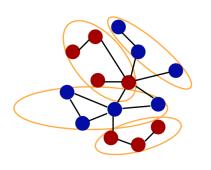


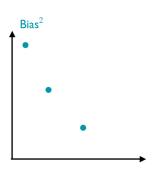
Davide Viviano Sussi clustering: dissign of cluster experimes November, 2023 5/18



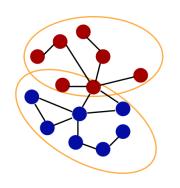


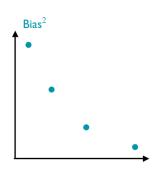
Davide Viviano Gusta dustarings dissign of cluster respectives November, 2023 5/18



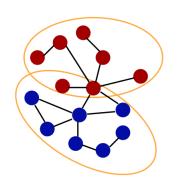


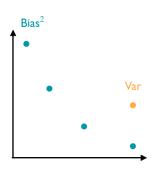
Davide Viviano Sussi clustering: design of cluster experimer November, 2023 5/18



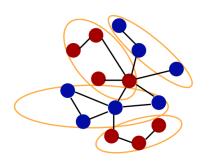


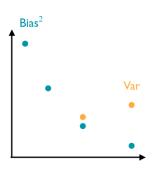
Davide Viviano Sussi clustering: dissign of cluster experimes November, 2023 5/18



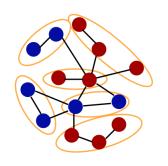


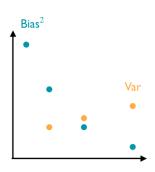
Davide Viviano Gusta dustarings dissign of cluster respectives November, 2023 5/18





Davide Viviano Gusta dustarings dissign of cluster respectives November, 2023 5/18





Davide Viviano Sussi clustering: design of cluster experimer November, 2023 5/18

This paper: clustering for global effects estimation

- When should you run a cluster experiment?
 - Novel characterization of worst-case bias and variance
 - Comparison with Bernoulli (i.i.d.) design with worst-case MSE

Davide Viviano Sausal Glasterius, design of Glaster experiment November, 2023 6/18

This paper: clustering for global effects estimation

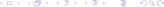
- When should you run a cluster experiment?
 - Novel characterization of worst-case bias and variance
 - Comparison with Bernoulli (i.i.d.) design with worst-case MSE
- How should you design the clusters?
 - MSE-Optimal clustering as simple optimization (sd program)
 - Application to unique data from Facebook graph

Davide Viviano Causal clustering, design of cluster experimen November, 2023 6 / 18

This paper: clustering for global effects estimation

- When should you run a cluster experiment?
 - Novel characterization of worst-case bias and variance
 - Comparison with Bernoulli (i.i.d.) design with worst-case MSE
- How should you design the clusters?
 - MSE-Optimal clustering as simple optimization (sd program)
 - Application to unique data from Facebook graph
- Conceptual contribution: clustering for treatment effect estimation
 - Provide intuitive "cluster purity" measure as bias justification
- Theoretical contribution: model spillovers through local asymptotics
 - Variance only depends on average cluster size variation

Davide Viviano Sausal clusterine, design of cluster experiment November, 2023 6/18



- Experimental design with spillovers [Baird et al., 2018; Basse and Airoldi, 2018; Johari et al. (2020); Viviano (2020); Pouget-Abadie et al. (2019, 2023) ...]
- Inference under interference [Hudgens and Halloran (2008); Aronow and Samii (2017); Leung (2020); Athey et al. (2018); Savje et al. (2021); Liu and Wager, 2022]
- Inference with clusters [Abadie et al., 2021; Ibragimov and Mueller, 2016; Canay et al., 2017]



- Experimental design with spillovers [Baird et al., 2018; Basse and Airoldi, 2018; Johari et al. (2020); Viviano (2020); Pouget-Abadie et al. (2019, 2023) ...]
- Inference under interference [Hudgens and Halloran (2008); Aronow and Samii (2017); Leung (2020); Athey et al. (2018); Savje et al. (2021); Liu and Wager, 2022
- Inference with clusters Abadie et al., 2021; Ibragimov and Mueller, 2016; Canay et al., 2017
 - ⇒ None study the problem of optimal cluster designs



Davide Viviano November, 2023

- Experimental design with spillovers [Baird et al., 2018; Basse and Airoldi, 2018; Johari et al. (2020); Viviano (2020); Pouget-Abadie et al. (2019, 2023) ...]
- Inference under interference [Hudgens and Halloran (2008); Aronow and Samii (2017); Leung (2020); Athey et al. (2018); Savje et al. (2021); Liu and Wager, 2022]
- Inference with clusters [Abadie et al., 2021; Ibragimov and Mueller, 2016; Canay et al., 2017]
 - ⇒ None study the problem of optimal cluster designs
- Graph clustering: rate optimal cluster size for ε -clustering [Ugander et al., 2013; Eckles et al., 2017; Leung, 2022; Faridani and Niehaus, 2022]
 - \Rightarrow Here optimal clustering method + novel bias/variance characterization

Davide Viviano Suusi clusterine, destru of cluster reporting November, 2023 7/18

- Experimental design with spillovers [Baird et al., 2018; Basse and Airoldi, 2018; Johari et al. (2020); Viviano (2020); Pouget-Abadie et al. (2019, 2023) ...]
- Inference under interference [Hudgens and Halloran (2008); Aronow and Samii (2017); Leung (2020); Athey et al. (2018); Savje et al. (2021); Liu and Wager, 2022]
- Inference with clusters [Abadie et al., 2021; Ibragimov and Mueller, 2016; Canay et al., 2017]
 - ⇒ None study the problem of optimal cluster designs
- Graph clustering: rate optimal cluster size for ε -clustering [Ugander et al., 2013; Eckles et al., 2017; Leung, 2022; Faridani and Niehaus, 2022]
 - \Rightarrow Here optimal clustering method + novel bias/variance characterization
- Clustering in stat [Von Luxburg, 2007; Newman 2013; Lei et al. (2020), et al.]
 - ⇒ Here different focus on clustering for causal inference

Davide Viviano

Constitution of the state of

Content

- Setup
- (When) should you cluster?
- Optimal clustering
- 4 Conclusions



8 / 18

Davide Viviano Causal clustering, design of cluster experimer November, 2023

- Finite pop: *n* units connected under adjacency matrix *A* (observed)
 - Potential outcomes: $Y_i(\mathbf{d}), \mathbf{d} \in \{0,1\}^n$



Davide Viviano Causal clustering design of cluster experimer November, 2023 9 / 18

- Finite pop: *n* units connected under adjacency matrix *A* (observed)
 - Potential outcomes: $Y_i(\mathbf{d}), \mathbf{d} \in \{0,1\}^n$
 - Set of neighbors: $\mathcal{N}_i = \{j : A_{i,j} = 1\}$



Davide Viviano Causal clustering: design of cluster experimer November, 2023 9 / 18

- Finite pop: *n* units connected under adjacency matrix *A* (observed)
 - Potential outcomes: $Y_i(\mathbf{d}), \mathbf{d} \in \{0,1\}^n$
 - Set of neighbors: $\mathcal{N}_i = \{j : A_{i,j} = 1\}$
 - Local interference: $Y_i(\mathbf{d}) = \mu_i(\mathbf{d}_i, \mathbf{d}_{\mathcal{N}_i}), \mu_i \in \mathcal{M}$



9/18

Davide Viviano Causal clusterings design of cluster experimer November, 2023

- Finite pop: *n* units connected under adjacency matrix *A* (observed)
 - Potential outcomes: $Y_i(\mathbf{d}), \mathbf{d} \in \{0,1\}^n$
 - Set of neighbors: $\mathcal{N}_i = \{j : A_{i,j} = 1\}$
 - Local interference: $Y_i(\mathbf{d}) = \mu_i(\mathbf{d}_i, \mathbf{d}_{\mathcal{N}_i}), \mu_i \in \mathcal{M}$
- Cluster experiment: for cluster design C_n



9 / 18

Davide Viviano Causal clusterings design of cluster experimen November, 2023

- Finite pop: *n* units connected under adjacency matrix *A* (observed)
 - Potential outcomes: $Y_i(\mathbf{d}), \mathbf{d} \in \{0,1\}^n$
 - Set of neighbors: $\mathcal{N}_i = \{j : A_{i,j} = 1\}$
 - Local interference: $Y_i(\mathbf{d}) = \mu_i(\mathbf{d}_i, \mathbf{d}_{\mathcal{N}_i}), \mu_i \in \mathcal{M}$
- Cluster experiment: for cluster design C_n
 - Number of clusters: K_n



9 / 18

Davide Viviano Causal clustering, design of cluster experimen November, 2023

- Finite pop: *n* units connected under adjacency matrix *A* (observed)
 - Potential outcomes: $Y_i(\mathbf{d}), \mathbf{d} \in \{0,1\}^n$
 - Set of neighbors: $\mathcal{N}_i = \{j : A_{i,j} = 1\}$
 - Local interference: $Y_i(\mathbf{d}) = \mu_i(\mathbf{d}_i, \mathbf{d}_{\mathcal{N}_i}), \mu_i \in \mathcal{M}$
- Cluster experiment: for cluster design C_n
 - Number of clusters: K_n
 - Cluster of unit $i: c(i) \in \{1, \dots, K_n\}$



9 / 18

Davide Viviano Causal clusterings design of cluster experimen November, 2023

- Finite pop: n units connected under adjacency matrix A (observed)
 - Potential outcomes: $Y_i(\mathbf{d}), \mathbf{d} \in \{0,1\}^n$
 - Set of neighbors: $\mathcal{N}_i = \{j : A_{i,j} = 1\}$
 - Local interference: $Y_i(\mathbf{d}) = \mu_i(\mathbf{d}_i, \mathbf{d}_{\mathcal{N}_i}), \mu_i \in \mathcal{M}$
- Cluster experiment: for cluster design C_n
 - Number of clusters: K_n
 - Cluster of unit $i: c(i) \in \{1, \dots, K_n\}$
 - Cluster level treatment.

$$D_i = \tilde{D}_{c(i)} \sim \mathrm{Bern}(1/2)$$



Davide Viviano November, 2023

- Finite pop: *n* units connected under adjacency matrix *A* (observed)
 - Potential outcomes: $Y_i(\mathbf{d}), \mathbf{d} \in \{0,1\}^n$
 - Set of neighbors: $\mathcal{N}_i = \{j : A_{i,j} = 1\}$
 - Local interference: $Y_i(\mathbf{d}) = \mu_i(\mathbf{d}_i, \mathbf{d}_{\mathcal{N}_i}), \mu_i \in \mathcal{M}$
- Cluster experiment: for cluster design C_n
 - Number of clusters: K_n
 - Cluster of unit $i: c(i) \in \{1, \dots, K_n\}$
 - Cluster level treatment

$$D_i = \tilde{D}_{c(i)} \sim \mathrm{Bern}(1/2)$$

• n_k : number of units in cluster k



Davide Viviano November, 2023 9 / 18

- Finite pop: *n* units connected under adjacency matrix *A* (observed)
 - Potential outcomes: $Y_i(\mathbf{d}), \mathbf{d} \in \{0,1\}^n$
 - Set of neighbors: $\mathcal{N}_i = \{j : A_{i,j} = 1\}$
 - Local interference: $Y_i(\mathbf{d}) = \mu_i(\mathbf{d}_i, \mathbf{d}_{\mathcal{N}_i}), \mu_i \in \mathcal{M}$
- Cluster experiment: for cluster design C_n
 - Number of clusters: K_n
 - Cluster of unit $i: c(i) \in \{1, \dots, K_n\}$
 - Cluster level treatment

$$D_i = \tilde{D}_{c(i)} \sim \mathrm{Bern}(1/2)$$

- n_k : number of units in cluster k
- Estimand and difference in means estimator:

$$\tau_n = \frac{1}{n} \sum_{i=1}^n \left[Y_i(\mathbf{1}) - Y_i(\mathbf{0}) \right],$$

Davide Viviano

Gruss disserting disserting disserting Parameters November, 2023 9/18



- Finite pop: *n* units connected under adjacency matrix *A* (observed)
 - Potential outcomes: $Y_i(\mathbf{d}), \mathbf{d} \in \{0,1\}^n$
 - Set of neighbors: $\mathcal{N}_i = \{j : A_{i,j} = 1\}$
 - Local interference: $Y_i(\mathbf{d}) = \mu_i(\mathbf{d}_i, \mathbf{d}_{\mathcal{N}_i}), \mu_i \in \mathcal{M}$
- Cluster experiment: for cluster design C_n
 - Number of clusters: K_n
 - Cluster of unit $i: c(i) \in \{1, \dots, K_n\}$
 - Cluster level treatment

$$D_i = \tilde{D}_{c(i)} \sim \mathrm{Bern}(1/2)$$

- n_k : number of units in cluster k
- Estimand and difference in means estimator:

$$\tau_n = \frac{1}{n} \sum_{i=1}^n \left[Y_i(\mathbf{1}) - Y_i(\mathbf{0}) \right], \quad \hat{\tau}_n = \frac{2}{n} \sum_{i=1}^n \left[Y_i D_i - Y_i(1 - D_i) \right]$$

Davide Viviano Causal clusterings design of cluster experiment November, 2023 9 / 18

Model on spillover effects

ullet Class of models: for $\mathbf{d}' \in \{\mathbf{1}, \mathbf{0}\}$

$$\sup_{\mu_i(1,\cdot)\in\mathcal{M}_1}\left|\mu_i(1,\mathbf{d})-\mu_i(1,\mathbf{d}')\right|=\frac{\bar{\phi}_n}{|\mathcal{N}_i|}\sum_{i\in\mathcal{N}_i}\left|\mathbf{d}_k'-\mathbf{d}_k\right|$$



Davide Viviano Rausa la la sagura de la sagu

Model on spillover effects

ullet Class of models: for $oldsymbol{d}' \in \{oldsymbol{1}, oldsymbol{0}\}$

$$\sup_{\mu_i(1,\cdot)\in\mathcal{M}_1}\left|\mu_i(1,\mathbf{d})-\mu_i(1,\mathbf{d}')\right|=\frac{\bar{\phi}_n}{|\mathcal{N}_i|}\sum_{i\in\mathcal{N}_i}\left|\mathbf{d}_k'-\mathbf{d}_k\right|$$

Asm Consider local asymptotics $\bar{\phi}_n = o(1)$:

- \Rightarrow signal to noise is of order smaller than $n^{1/2}$
- $\Rightarrow \bar{\phi}_n \gg n^{-1/2}$ "moderately large" spillover effects
- \Rightarrow $\bar{\phi}_{\it n} \leq \it n^{-1/2}$ "small spillover effects

◆ロト ◆個ト ◆差ト ◆差ト を めへぐ

Davide Viviano Rausa la la sagura de la sagu

Model on spillover effects

ullet Class of models: for $\mathbf{d}' \in \{\mathbf{1}, \mathbf{0}\}$

$$\sup_{\mu_i(1,\cdot)\in\mathcal{M}_1}\left|\mu_i(1,\mathbf{d})-\mu_i(1,\mathbf{d}')\right|=\frac{\bar{\phi}_n}{|\mathcal{N}_i|}\sum_{i\in\mathcal{N}_i}\left|\mathbf{d}_k'-\mathbf{d}_k\right|$$

Asm Consider local asymptotics $\bar{\phi}_n = o(1)$:

- \Rightarrow signal to noise is of order smaller than $n^{1/2}$
- $\Rightarrow \bar{\phi}_n \gg n^{-1/2}$ "moderately large" spillover effects
- $\Rightarrow \bar{\phi}_n \leq n^{-1/2}$ "small spillover effects
- E.g. Write the overall effect as $\tau_n = \Delta_n + \bar{\phi}_n$, for $\Delta_n, \bar{\phi}_n = o(1)$.
- E.g. Information experiments [Karrer et al. (2021)]

Davide Viviano

Super classarius dieler di classarius

November, 2023

10 / 18

Content

- Setup
- (When) should you cluster?
- Optimal clustering
- 4 Conclusions



11 / 18

Davide Viviano Causal clustering: design of cluster experimen November, 2023

Lem Worst-case outcome model justifies cluster purity:

$$\sup_{\mu \in \mathcal{M}} \left(\mathbb{E} \Big[\hat{\tau}_n | \mathbf{A}, \mathcal{C}_n \Big] - \tau_n \right)^2$$

Davide Viviano Susal clustering design of cluster experimen November, 2023 12 / 18

Lem Worst-case outcome model justifies cluster purity:

$$\sup_{\mu \in \mathcal{M}} \left(\mathbb{E} \Big[\hat{\tau}_n | \mathbf{A}, \mathcal{C}_n \Big] - \tau_n \right)^2 = \frac{\bar{\phi}_n}{n}$$



Davide Viviano Susal clustering design of cluster experimen November, 2023 12 / 18

Lem Worst-case outcome model justifies cluster purity:

$$\sup_{\mu \in \mathcal{M}} \left(\mathbb{E} \left[\hat{\tau}_n | \mathbf{A}, \mathcal{C}_n \right] - \tau_n \right)^2 = \frac{\bar{\phi}_n}{n} \sum_{i=1}^n \frac{1}{|\mathcal{N}_i|} \underbrace{\left| \mathcal{N}_i \bigcap \left\{ j : c(i) \neq c(j) \right\} \right|}_{\text{\#}i\text{'s friends not in }i\text{'s cluster}}$$

Davide Viviano Susal clustering design of cluster experimen November, 2023 12 / 18

Lem Worst-case outcome model justifies cluster purity:

$$\sup_{\mu \in \mathcal{M}} \left(\mathbb{E} \left[\hat{\tau}_n | \mathbf{A}, \mathcal{C}_n \right] - \tau_n \right)^2 = \frac{\bar{\phi}_n}{n} \sum_{i=1}^n \frac{1}{|\mathcal{N}_i|} \underbrace{\left| \mathcal{N}_i \bigcap \left\{ j : c(i) \neq c(j) \right\} \right|}_{\#i'\text{s friends not in } i'\text{s cluster}}$$

⇒ Clusters' purity justified as estimator's bias!

< ロト < 個 ト < 重 ト < 重 ト 三 重 ・ の Q ()

12 / 18

Davide Viviano Causal clustering: design of cluster experimen November, 2023

Lem Worst-case outcome model justifies cluster purity:

$$\sup_{\mu \in \mathcal{M}} \left(\mathbb{E} \Big[\hat{\tau}_n \big| \mathbf{A}, \mathcal{C}_n \Big] - \tau_n \right)^2 = \frac{\bar{\phi}_n}{n} \sum_{i=1}^n \frac{1}{|\mathcal{N}_i|} \underbrace{\left| \mathcal{N}_i \bigcap \left\{ j : c(i) \neq c(j) \right\} \right|}_{\text{\#$i's friends not in i's cluster}}$$

Clusters' purity justified as estimator's bias!

Lem Let
$$\bar{\phi}_n \mathcal{N}_{n,\max}^2 = o(1)$$
, $\bar{\psi} = \sup_i \max\{\mu_i(\mathbf{1})^2, \mu_i(\mathbf{0})^2\}$, then

12 / 18

Davide Viviano November, 2023

Lem Worst-case outcome model justifies cluster purity:

$$\sup_{\mu \in \mathcal{M}} \left(\mathbb{E} \Big[\hat{\tau}_n \big| \mathbf{A}, \mathcal{C}_n \Big] - \tau_n \right)^2 = \frac{\bar{\phi}_n}{n} \sum_{i=1}^n \frac{1}{|\mathcal{N}_i|} \underbrace{\left| \mathcal{N}_i \bigcap \left\{ j : c(i) \neq c(j) \right\} \right|}_{\text{\#$i's friends not in i's cluster}}$$

⇒ Clusters' purity justified as estimator's bias!

Lem Let
$$\bar{\phi}_n \mathcal{N}_{n,\max}^2 = o(1)$$
, $\bar{\psi} = \sup_i \max\{\mu_i(\mathbf{1})^2, \mu_i(\mathbf{0})^2\}$, then

$$\sup_{\mu \in \mathcal{M}} \mathrm{Var} \Big(\hat{\tau} | \boldsymbol{A}, \mathcal{C}_n \Big) =$$

Davide Viviano November, 2023 12 / 18

Lem Worst-case outcome model justifies cluster purity:

$$\sup_{\mu \in \mathcal{M}} \left(\mathbb{E} \Big[\hat{\tau}_n \big| \mathbf{A}, \mathcal{C}_n \Big] - \tau_n \right)^2 = \frac{\bar{\phi}_n}{n} \sum_{i=1}^n \frac{1}{|\mathcal{N}_i|} \underbrace{\left| \mathcal{N}_i \bigcap \left\{ j : c(i) \neq c(j) \right\} \right|}_{\text{\#$i's friends not in i's cluster}}$$

Clusters' purity justified as estimator's bias!

Lem Let
$$\bar{\phi}_n \mathcal{N}_{n,\max}^2 = o(1)$$
, $\bar{\psi} = \sup_i \max\{\mu_i(\mathbf{1})^2, \mu_i(\mathbf{0})^2\}$, then

$$\sup_{\mu \in \mathcal{M}} \operatorname{Var} \left(\hat{\tau} | \mathbf{A}, \mathcal{C}_n \right) = \underbrace{\bar{\psi} \sum_{k=1}^{K_n} \frac{n_k^2}{n^2}}_{\text{Variation in clusters' size}}$$

Davide Viviano November, 2023 12 / 18

Lem Worst-case outcome model justifies cluster purity:

$$\sup_{\mu \in \mathcal{M}} \left(\mathbb{E} \Big[\hat{\tau}_n \big| \mathbf{A}, \mathcal{C}_n \Big] - \tau_n \right)^2 = \frac{\bar{\phi}_n}{n} \sum_{i=1}^n \frac{1}{|\mathcal{N}_i|} \underbrace{\left| \mathcal{N}_i \bigcap \left\{ j : c(i) \neq c(j) \right\} \right|}_{\text{\#i's friends not in i's cluster}}$$

⇒ Clusters' purity justified as estimator's bias!

Lem Let
$$\bar{\phi}_n \mathcal{N}_{n,\max}^2 = o(1)$$
, $\bar{\psi} = \sup_i \max\{\mu_i(\mathbf{1})^2, \mu_i(\mathbf{0})^2\}$, then

$$\sup_{\mu \in \mathcal{M}} \operatorname{Var} \left(\hat{\tau} | \mathbf{A}, C_n \right) = \underbrace{\bar{\psi} \sum_{k=1}^{K_n} \frac{n_k^2}{n^2}}_{} + o(1/K_n)$$

Variation in clusters' size

Davide Viviano Sausal clusterine, design of cluster experiment November, 2023 12 / 18

Lem Worst-case outcome model justifies cluster purity:

$$\sup_{\mu \in \mathcal{M}} \left(\mathbb{E} \Big[\hat{\tau}_n \big| \mathbf{A}, \mathcal{C}_n \Big] - \tau_n \right)^2 = \frac{\bar{\phi}_n}{n} \sum_{i=1}^n \frac{1}{|\mathcal{N}_i|} \underbrace{\left| \mathcal{N}_i \bigcap \left\{ j : c(i) \neq c(j) \right\} \right|}_{\text{\#$i's friends not in i's cluster}}$$

⇒ Clusters' purity justified as estimator's bias!

Lem Let
$$\bar{\phi}_n \mathcal{N}_{n,\max}^2 = o(1)$$
, $\bar{\psi} = \sup_i \max\{\mu_i(\mathbf{1})^2, \mu_i(\mathbf{0})^2\}$, then

$$\sup_{\mu \in \mathcal{M}} \operatorname{Var} \Big(\hat{\tau} | \boldsymbol{A}, \mathcal{C}_n \Big) = \underbrace{\bar{\psi} \sum_{k=1}^{K_n} \frac{n_k^2}{n^2}}_{} + o(1/K_n)$$

Variation in clusters' size

More clusters' size variation imply smaller variance

Davide Viviano 12 / 18

Lem Worst-case outcome model justifies cluster purity:

$$\sup_{\mu \in \mathcal{M}} \left(\mathbb{E} \Big[\hat{\tau}_n | \mathbf{A}, \mathcal{C}_n \Big] - \tau_n \right)^2 = \frac{\bar{\phi}_n}{n} \sum_{i=1}^n \frac{1}{|\mathcal{N}_i|} \underbrace{\left| \mathcal{N}_i \bigcap \left\{ j : c(i) \neq c(j) \right\} \right|}_{\#i' \text{s friends not in } i' \text{s cluster}}$$

⇒ Clusters' purity justified as estimator's bias!

Lem Let
$$\bar{\phi}_n \mathcal{N}_{n,\max}^2 = o(1)$$
, $\bar{\psi} = \sup_i \max\{\mu_i(\mathbf{1})^2, \mu_i(\mathbf{0})^2\}$, then

$$\sup_{\mu \in \mathcal{M}} \operatorname{Var} \left(\hat{\tau} | \mathbf{A}, \mathcal{C}_n \right) = \underbrace{\bar{\psi} \sum_{k=1}^{K_n} \frac{n_k^2}{n^2}}_{} + o(1/K_n)$$

Variation in clusters' size

More clusters' size variation imply smaller variance

Key Bound (i,j) covariances with $\bar{\phi}_n$ for (i,j) not in same cluster

12 / 18

Davide Viviano November, 2023

Thm For $C_{0,n}$ denoting a Bernoulli design, under regularities

Davide Viviano Causal clustering dissign of cluster experiment November, 2023 13 / 18

Thm For $C_{0,n}$ denoting a Bernoulli design, under regularities

$$\lim_{n\to\infty}\sup_{\mu}\mathrm{MSE}_{\mu}(\mathcal{C}_n)-\sup_{\mu}\mathrm{MSE}_{\mu}(\mathcal{C}_0)\geq 0 \text{ if } \sqrt{K_n}\bar{\phi}_n\to 0$$



Davide Viviano Causal clustering dissign of cluster experiment November, 2023 13 / 18

Thm For $C_{0,n}$ denoting a Bernoulli design, under regularities

$$\begin{split} & \lim_{n \to \infty} \sup_{\mu} \mathrm{MSE}_{\mu}(\mathcal{C}_n) - \sup_{\mu} \mathrm{MSE}_{\mu}(\mathcal{C}_0) \geq 0 \text{ if } \sqrt{K_n} \bar{\phi}_n \to 0 \\ & \lim_{n \to \infty} \sup_{\mu} \mathrm{MSE}_{\mu}(\mathcal{C}_n) - \sup_{\mu} \mathrm{MSE}_{\mu}(\mathcal{C}_0) {\leq} 0 \end{split}$$



Davide Viviano Course clustering design of cluster experimen November, 2023 13 / 18

Thm For $C_{0,n}$ denoting a Bernoulli design, under regularities

$$\lim_{n\to\infty}\sup_{\mu}\mathrm{MSE}_{\mu}(\mathcal{C}_n)-\sup_{\mu}\mathrm{MSE}_{\mu}(\mathcal{C}_0)\geq 0 \text{ if } \sqrt{K_n}\bar{\phi}_n\to 0$$

$$\lim_{n\to\infty}\sup_{\mu}\mathrm{MSE}_{\mu}(\mathcal{C}_n)-\sup_{\mu}\mathrm{MSE}_{\mu}(\mathcal{C}_0){\leq}0 \text{ if } \sqrt{K_n}\bar{\phi}_n\to\infty$$



Davide Viviano Causal clustering, design of cluster experimer November, 2023 13 / 18

Thm For $C_{0,n}$ denoting a Bernoulli design, under regularities

$$\begin{split} &\lim_{n\to\infty}\sup_{\mu}\mathrm{MSE}_{\mu}(\mathcal{C}_n)-\sup_{\mu}\mathrm{MSE}_{\mu}(\mathcal{C}_0)\geq 0 \text{ if } \sqrt{K_n}\bar{\phi}_n\to 0 \\ &\lim_{n\to\infty}\sup_{\mu}\mathrm{MSE}_{\mu}(\mathcal{C}_n)-\sup_{\mu}\mathrm{MSE}_{\mu}(\mathcal{C}_0){\leq}0 \text{ if } \sqrt{K_n}\bar{\phi}_n\to\infty \end{split}$$

Exact threshold in the paper (rule of thumb $\sqrt{K_n}\bar{\phi}_n > 2.3$)

13 / 18

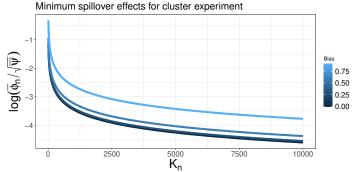
Davide Viviano November, 2023

Thm For $C_{0,n}$ denoting a Bernoulli design, under regularities

$$\lim_{n\to\infty}\sup_{\mu}\mathrm{MSE}_{\mu}(\mathcal{C}_n)-\sup_{\mu}\mathrm{MSE}_{\mu}(\mathcal{C}_0)\geq 0 \text{ if } \sqrt{K_n}\bar{\phi}_n\to 0$$

$$\lim_{n\to\infty}\sup_{\mu}\mathrm{MSE}_{\mu}(\mathcal{C}_n)-\sup_{\mu}\mathrm{MSE}_{\mu}(\mathcal{C}_0)\leq 0 \text{ if } \sqrt{K_n}\bar{\phi}_n\to \infty$$

 \Rightarrow Exact threshold in the paper (rule of thumb $\sqrt{K_n}ar{\phi}_n>2.3)$



Davide Viviano Causal siusenine, design of cluster experiment November, 2023

13 / 18

Thm For $C_{0,n}$ denoting a Bernoulli design, under regularities

$$\begin{split} &\lim_{n\to\infty}\sup_{\mu}\mathrm{MSE}_{\mu}(\mathcal{C}_n)-\sup_{\mu}\mathrm{MSE}_{\mu}(\mathcal{C}_0)\geq 0 \text{ if } \sqrt{K_n}\bar{\phi}_n\to 0\\ &\lim_{n\to\infty}\sup_{\mu}\mathrm{MSE}_{\mu}(\mathcal{C}_n)-\sup_{\mu}\mathrm{MSE}_{\mu}(\mathcal{C}_0)\leq 0 \text{ if } \sqrt{K_n}\bar{\phi}_n\to \infty \end{split}$$

Description	Mathematical Formulation	Implication
Spillovers are small and # of clusters is small	$\sqrt{K_n}ar{\phi}_n=o(1)$	Run Bernoulli design

Davide Viviano Gustel clusterings design of cluster examinar November, 2023 13/18

Thm For $C_{0,n}$ denoting a Bernoulli design, under regularities

$$\begin{split} &\lim_{n\to\infty}\sup_{\mu}\mathrm{MSE}_{\mu}(\mathcal{C}_n)-\sup_{\mu}\mathrm{MSE}_{\mu}(\mathcal{C}_0)\geq 0 \text{ if } \sqrt{K_n}\bar{\phi}_n\to 0\\ &\lim_{n\to\infty}\sup_{\mu}\mathrm{MSE}_{\mu}(\mathcal{C}_n)-\sup_{\mu}\mathrm{MSE}_{\mu}(\mathcal{C}_0)\leq 0 \text{ if } \sqrt{K_n}\bar{\phi}_n\to \infty \end{split}$$

Description	Mathematical Formulation	Implication
Spillovers are non-negligible and # of clusters is large	$\sqrt{K_n} \overline{\phi}_n o \infty$	Run Cluster design

Davide Viviano November, 2023 13 / 18

Thm For $C_{0,n}$ denoting a Bernoulli design, under regularities

$$\begin{split} &\lim_{n\to\infty}\sup_{\mu}\mathrm{MSE}_{\mu}(\mathcal{C}_n)-\sup_{\mu}\mathrm{MSE}_{\mu}(\mathcal{C}_0)\geq 0 \text{ if } \sqrt{K_n}\bar{\phi}_n\to 0 \\ &\lim_{n\to\infty}\sup_{\mu}\mathrm{MSE}_{\mu}(\mathcal{C}_n)-\sup_{\mu}\mathrm{MSE}_{\mu}(\mathcal{C}_0)\leq 0 \text{ if } \sqrt{K_n}\bar{\phi}_n\to \infty \end{split}$$

Description	Mathematical Formulation	Implication
Spillovers are small	$K_n \propto n, \bar{\phi}_n \propto n^{-1/3}$	Run Cluster design
and # of clusters is very large		

Davide Viviano November, 2023 13 / 18

Thm For $C_{0,n}$ denoting a Bernoulli design, under regularities

$$\begin{split} &\lim_{n\to\infty}\sup_{\mu}\mathrm{MSE}_{\mu}(\mathcal{C}_n)-\sup_{\mu}\mathrm{MSE}_{\mu}(\mathcal{C}_0)\geq 0 \text{ if } \sqrt{K_n}\bar{\phi}_n\to 0\\ &\lim_{n\to\infty}\sup_{\mu}\mathrm{MSE}_{\mu}(\mathcal{C}_n)-\sup_{\mu}\mathrm{MSE}_{\mu}(\mathcal{C}_0)\leq 0 \text{ if } \sqrt{K_n}\bar{\phi}_n\to \infty \end{split}$$

Description	Mathematical Formulation	Implication
Spillovers are not that small and # of clusters is very small	$K_n = \mathcal{O}(1), \bar{\phi}_n = o(1)$	Run Bernoulli design

Davide Viviano November, 2023 13 / 18

Content

- Setup
- (When) should you cluster?
- Optimal clustering
- 4 Conclusions



Davide Viviano Canal Junguage Gapa of Glass Canal November, 2023 14 / 18

Thm For any clustering C_n , for $\bar{\phi}_n = o(1), \xi_n = \frac{\bar{\psi}}{\bar{\phi}_n^2}$,

Davide Viviano Caucal clusterings design of cluster constituer November, 2023 15 / 18

Thm For any clustering
$$C_n$$
, for $\bar{\phi}_n = o(1), \xi_n = \frac{\bar{\psi}}{\bar{\phi}_n^2}$,

$$\sup_{\mu} \mathrm{MSE}_n(\mathcal{C}) \propto$$



Davide Viviano Sussi clustering design of cluster experimen November, 2023 15 / 18

Thm For any clustering C_n , for $\bar{\phi}_n = o(1), \xi_n = \frac{\bar{\psi}}{\bar{\phi}_n^2}$,

$$\sup_{\mu} \mathrm{MSE}_n(\mathcal{C}) \propto \left(\frac{1}{n} \sum_{i=1}^n \frac{1}{|\mathcal{N}_i|} \Big| j \in \mathcal{N}_i : c(i) \neq c(j) \Big| \right)^2 +$$



Davide Viviano Causal clustering: design of cluster experimer November, 2023 15 / 18

Thm For any clustering C_n , for $\bar{\phi}_n = o(1), \xi_n = \frac{\bar{\psi}}{\bar{\phi}_n^2}$,

$$\sup_{\mu} \mathrm{MSE}_n(\mathcal{C}) \propto \left(\frac{1}{n} \sum_{i=1}^n \frac{1}{|\mathcal{N}_i|} \Big| j \in \mathcal{N}_i : c(i) \neq c(j) \Big| \right)^2 + \xi_n \sum_{k=1}^{K_n} \frac{n_k^2}{n^2}$$

4□ > 4□ > 4 = > 4 = > = 90

Davide Viviano Causal clustering: design of cluster experimer November, 2023 15 / 18

Thm For any clustering C_n , for $\bar{\phi}_n = o(1), \xi_n = \frac{\bar{\psi}}{\bar{\phi}_n^2}$,

$$\sup_{\mu} \mathrm{MSE}_n(\mathcal{C}) \propto \left(\frac{1}{n} \sum_{i=1}^n \frac{1}{|\mathcal{N}_i|} \middle| j \in \mathcal{N}_i : c(i) \neq c(j) \middle| \right)^2 + \xi_n \sum_{k=1}^{K_n} \frac{n_k^2}{n^2} + o(1/K_n)$$

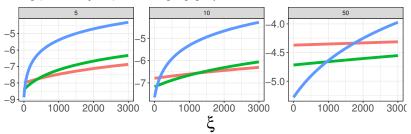


Davide Viviano Causal clustering: design of cluster experimer November, 2023 15 / 18

Thm For any clustering C_n , for $\bar{\phi}_n = o(1), \xi_n = \frac{\bar{\psi}}{\bar{\phi}_n^2}$,

$$\sup_{\mu} \mathrm{MSE}_n(\mathcal{C}) \propto \left(\frac{1}{n} \sum_{i=1}^n \frac{1}{|\mathcal{N}_i|} \middle| j \in \mathcal{N}_i : c(i) \neq c(j) \middle| \right)^2 + \xi_n \sum_{k=1}^{K_n} \frac{n_k^2}{n^2} + o(1/K_n)$$

$log(Bias^2 + \xi Var)$: Messaging graph



Clustering — Louvain Type 1 — Louvain Type 2 — Louvain Type 3

Davide Viviano Gusal clusterines design of cluster recognition November, 2023 15 / 18

Thm For any clustering
$$C_n$$
, for $\bar{\phi}_n = o(1), \xi_n = \frac{\bar{\psi}}{\bar{\phi}_n^2}$,

$$\sup_{\mu} \mathrm{MSE}_n(\mathcal{C}) \propto \mathrm{Bias}_n(\mathcal{C}_n)^2 + \xi_n \sum_{k=1}^{K_n} \frac{n_k^2}{n^2} + o(1/K_n)$$



Davide Viviano Coust clusterings design of cluster experiment November, 2023 15 / 18

Thm For any clustering
$$C_n$$
, for $\bar{\phi}_n = o(1), \xi_n = \frac{\bar{\psi}}{\bar{\phi}_n^2}$,

$$\sup_{\mu} \mathrm{MSE}_n(\mathcal{C}) \propto \mathrm{Bias}_n(\mathcal{C}_n)^2 + \xi_n \sum_{k=1}^{K_n} \frac{n_k^2}{n^2} + o(1/K_n)$$

Idea Focus on (surrogate)
$$|\operatorname{Bias}_n(\mathcal{C}_n)| + \xi_n' \sum_{k=1}^{K_n} \frac{n_k^2}{n^2}$$

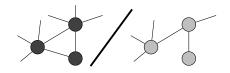
15 / 18

Davide Viviano Causal clustering design of cluster experimen November, 2023

Thm For any clustering C_n , for $\bar{\phi}_n = o(1), \xi_n = \frac{\psi}{\bar{\phi}^2}$,

$$\sup_{\mu} \mathrm{MSE}_n(\mathcal{C}) \propto \mathrm{Bias}_n(\mathcal{C}_n)^2 + \xi_n \sum_{k=1}^{K_n} \frac{n_k^2}{n^2} + o(1/K_n)$$

Idea Focus on (surrogate) $|\operatorname{Bias}_n(\mathcal{C}_n)| + \xi_n' \sum_{k=1}^{K_n} \frac{n_k^2}{n^2}$

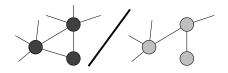


Davide Viviano November, 2023 15 / 18

Thm For any clustering C_n , for $\bar{\phi}_n = o(1), \xi_n = \frac{\bar{\psi}}{\bar{\phi}_n^2}$,

$$\sup_{\mu} \mathrm{MSE}_n(\mathcal{C}) \propto \mathrm{Bias}_n(\mathcal{C}_n)^2 + \xi_n \sum_{k=1}^{K_n} \frac{n_k^2}{n^2} + o(1/K_n)$$

Idea Focus on (surrogate) $|\operatorname{Bias}_n(\mathcal{C}_n)| + \xi_n' \sum_{k=1}^{K_n} \frac{n_k^2}{n^2}$



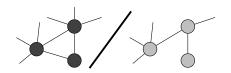
Thm Let $\mathbf{M}_{c,K}^{(i,k)} = 1\{c(i) = k\}.$

Davide Viviano Causal clusterius, distan al cluster suscilius November, 2023 15/18

Thm For any clustering C_n , for $\bar{\phi}_n = o(1), \xi_n = \frac{\bar{\psi}}{\bar{\phi}_n^2}$,

$$\sup_{\mu} \mathrm{MSE}_n(\mathcal{C}) \propto \mathrm{Bias}_n(\mathcal{C}_n)^2 + \xi_n \sum_{k=1}^{K_n} \frac{n_k^2}{n^2} + o(1/K_n)$$

Idea Focus on (surrogate) $|\operatorname{Bias}_n(\mathcal{C}_n)| + \xi_n' \sum_{k=1}^{K_n} \frac{n_k^2}{n^2}$



Thm Let $\mathbf{M}_{c,K}^{(i,k)} = 1\{c(i) = k\}$. For $K_n = K$, optimization is equivalent to

$$\max_c \mathrm{tr}\Big((\mathbf{L} - \frac{\xi_n'}{n} \mathbf{1}_n \mathbf{1}_n) \mathbf{M}_{c,K} \mathbf{M}_{c,K}^\top \Big), \quad \mathbf{L} = \mathrm{diag}(\mathbf{A} \mathbf{1}_n)^{-1} \mathbf{A}$$

Davide Viviano Guesi clustering design of cluster expaning November, 2023 15 / 18

Implementation

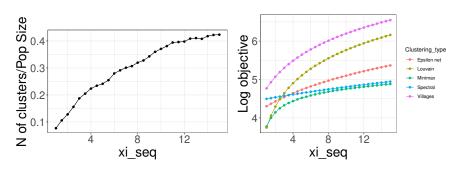
- Choose $n\xi'_n$ (e.g., 3.29)
- Solution for sequence of K_n
- ullet Trace-optimization solved via SDP $oldsymbol{\mathsf{M}}_{c,\mathcal{K}}oldsymbol{\mathsf{M}}_{c,\mathcal{K}}^{ op}$ + k-means
- \bullet Compute best K_n and clustering by comparing objectives

Davide Viviano

Guesi clusterary distanci (sluster securities November, 2023 16/18

Implementation

- Choose $n\xi'_n$ (e.g., 3.29)
- Solution for sequence of K_n
- ullet Trace-optimization solved via SDP $oldsymbol{\mathsf{M}}_{c,\mathcal{K}} oldsymbol{\mathsf{M}}_{c,\mathcal{K}}^{ op} + \mathsf{k}$ -means
- \bullet Compute best K_n and clustering by comparing objectives



Davide Viviano Causal dustarings design of cluster experimes November, 2023 16 / 18

Content

- Setup
- (When) should you cluster?
- Optimal clustering
- 4 Conclusions



17 / 18

Davide Viviano Causal clustering: design of cluster experimen November, 2023

Conclusions

- Novel causal clustering algorithm motivated by ATE estimation
- Bias/Variance motivate measures of cluster purity/variation
- Comparisons Cluster vs Bernoulli design
- Algorithm to compute optimal clustering
- Application using data from Meta (and field experiment)

4□ Þ 4₫ Þ 4 ₺ Þ ₺ ₺ ₹ У)Q(Y

18 / 18

Davide Viviano Causal obstacione design of obstacles experiment November, 2023

Conclusions

- Novel causal clustering algorithm motivated by ATE estimation
- Bias/Variance motivate measures of cluster purity/variation
- Comparisons Cluster vs Bernoulli design
- Algorithm to compute optimal clustering
- Application using data from Meta (and field experiment)

Thanks very much, questions? :)



18 / 18

Davide Viviano Causal clustering design of cluster experimen November, 2023

Worst case variance depends on

$$\operatorname{Var}\left(\hat{\tau}|\mathbf{A},\mathcal{C}_{n}\right) = \frac{4}{n^{2}} \sum_{i,j} \underbrace{\operatorname{Cov}\left(\mu_{i}(\mathbf{D})(2D_{i}-1), \mu_{j}(\mathbf{D})(2D_{j}-1)\right)}_{=\sigma_{i,j}}$$

hmm... and now?



Davide Viviano Causal susseines design of cluster execution November, 2023 18 / 18

Worst case variance depends on

$$\operatorname{Var}\left(\hat{\tau}|\mathbf{A},\mathcal{C}_{n}\right) = \frac{4}{n^{2}} \sum_{i,j} \underbrace{\operatorname{Cov}\left(\mu_{i}(\mathbf{D})(2D_{i}-1), \mu_{j}(\mathbf{D})(2D_{j}-1)\right)}_{=\sigma_{i,j}}$$

hmm... and now?

Lem
$$|\sigma_{i,j}| = \mathcal{O}\Big(\mathrm{Bias}_n(\mathcal{C}_n)\Big)$$
 if



Davide Viviano Causal clusterings design of cluster experimen November, 2023 18 / 18

Worst case variance depends on

$$\operatorname{Var}\left(\hat{\tau}|\mathbf{A},\mathcal{C}_{n}\right) = \frac{4}{n^{2}} \sum_{i,j} \underbrace{\operatorname{Cov}\left(\mu_{i}(\mathbf{D})(2D_{i}-1), \mu_{j}(\mathbf{D})(2D_{j}-1)\right)}_{=\sigma_{i,j}}$$

hmm... and now?

Lem $|\sigma_{i,j}| = \mathcal{O}\big(\mathrm{Bias}_n(\mathcal{C}_n)\big)$ if (i,j) are (a) not friends or friends of friends,



Davide Viviano Rosentos designos designos designos November, 2023 18 / 18

Worst case variance depends on

$$\operatorname{Var}\left(\hat{\tau}|\mathbf{A},\mathcal{C}_{n}\right) = \frac{4}{n^{2}} \sum_{i,j} \underbrace{\operatorname{Cov}\left(\mu_{i}(\mathbf{D})(2D_{i}-1), \mu_{j}(\mathbf{D})(2D_{j}-1)\right)}_{=\sigma_{i,j}}$$

hmm... and now?

Lem $|\sigma_{i,j}| = \mathcal{O}\big(\mathrm{Bias}_n(\mathcal{C}_n)\big)$ if (i,j) are (a) not friends or friends of friends, (b) in different clusters, and

18 / 18

Worst case variance depends on

$$\operatorname{Var}\left(\hat{\tau}|\mathbf{A},\mathcal{C}_{n}\right) = \frac{4}{n^{2}} \sum_{i,j} \underbrace{\operatorname{Cov}\left(\mu_{i}(\mathbf{D})(2D_{i}-1), \mu_{j}(\mathbf{D})(2D_{j}-1)\right)}_{=\sigma_{i,j}}$$

hmm... and now?

Lem $|\sigma_{i,j}| = \mathcal{O}\big(\mathrm{Bias}_n(\mathcal{C}_n)\big)$ if (i,j) are (a) not friends or friends of friends, (b) in different clusters, and (c) j is not in a cluster with a friend of i

18 / 18

Worst case variance depends on

$$\operatorname{Var}\left(\hat{\tau}|\mathbf{A},\mathcal{C}_{n}\right) = \frac{4}{n^{2}} \sum_{i,j} \underbrace{\operatorname{Cov}\left(\mu_{i}(\mathbf{D})(2D_{i}-1), \mu_{j}(\mathbf{D})(2D_{j}-1)\right)}_{=\sigma_{i,j}}$$

hmm... and now?

Lem $|\sigma_{i,j}| = \mathcal{O}\left(\operatorname{Bias}_n(\mathcal{C}_n)\right)$ if (i,j) are (a) not friends or friends of friends, (b) in different clusters, and (c) j is not in a cluster with a friend of i

$$\Rightarrow \operatorname{Var}\!\left(\hat{\tau}|\boldsymbol{A},\mathcal{C}_{n}\right)$$

Davide Viviano Sausal clusterine, design of cluster experiment November, 2023 18/18

Worst case variance depends on

$$\operatorname{Var}\left(\hat{\tau}|\mathbf{A},\mathcal{C}_{n}\right) = \frac{4}{n^{2}} \sum_{i,j} \underbrace{\operatorname{Cov}\left(\mu_{i}(\mathbf{D})(2D_{i}-1), \mu_{j}(\mathbf{D})(2D_{j}-1)\right)}_{=\sigma_{i,j}}$$

hmm... and now?

Lem $|\sigma_{i,j}| = \mathcal{O}\big(\mathrm{Bias}_n(\mathcal{C}_n)\big)$ if (i,j) are (a) not friends or friends of friends, (b) in different clusters, and (c) j is not in a cluster with a friend of i

$$\Rightarrow \operatorname{Var}\Big(\hat{\tau}|\mathbf{A},\mathcal{C}_n\Big) \leq \underbrace{\bar{\phi}_n/n^2}_{\operatorname{Bias}} \times$$

18 / 18

Worst case variance depends on

$$\operatorname{Var}\left(\hat{\tau}|\mathbf{A},\mathcal{C}_{n}\right) = \frac{4}{n^{2}} \sum_{i,j} \underbrace{\operatorname{Cov}\left(\mu_{i}(\mathbf{D})(2D_{i}-1), \mu_{j}(\mathbf{D})(2D_{j}-1)\right)}_{=\sigma_{i,j}}$$

hmm... and now?

Lem $|\sigma_{i,j}| = \mathcal{O}\big(\mathrm{Bias}_n(\mathcal{C}_n)\big)$ if (i,j) are (a) not friends or friends of friends, (b) in different clusters, and (c) j is not in a cluster with a friend of i

$$\Rightarrow \mathrm{Var}\Big(\hat{\tau}|\mathbf{A},\mathcal{C}_n\Big) \leq \underbrace{\bar{\phi}_n/n^2}_{\mathsf{Bias}} \times \mathcal{O}\Big(\underbrace{n \times \mathcal{N}_{n,\mathsf{max}}^2}_{\# \ 2\mathsf{nd} \ \mathsf{degree}} \times$$

18 / 18

Worst case variance depends on

$$\operatorname{Var}\left(\hat{\tau}|\mathbf{A},\mathcal{C}_{n}\right) = \frac{4}{n^{2}} \sum_{i,j} \underbrace{\operatorname{Cov}\left(\mu_{i}(\mathbf{D})(2D_{i}-1), \mu_{j}(\mathbf{D})(2D_{j}-1)\right)}_{=\sigma_{i,j}}$$

hmm... and now?

Lem $|\sigma_{i,j}| = \mathcal{O}\big(\mathrm{Bias}_n(\mathcal{C}_n)\big)$ if (i,j) are (a) not friends or friends of friends, (b) in different clusters, and (c) j is not in a cluster with a friend of i

$$\Rightarrow \operatorname{Var}\Big(\hat{\tau}|\boldsymbol{A},\mathcal{C}_n\Big) \leq \underbrace{\bar{\phi}_n/n^2}_{\text{Bias}} \times \mathcal{O}\Big(\underbrace{n \times \mathcal{N}_{n,\text{max}}^2}_{\text{\# 2nd degree}} \times \underbrace{\frac{\bar{\gamma}\,n}{K_n}}_{\text{largest cluster size}}\Big) +$$

Davide Viviano November, 2023 18 / 18

Worst case variance depends on

$$\operatorname{Var}\left(\hat{\tau}|\mathbf{A},\mathcal{C}_{n}\right) = \frac{4}{n^{2}} \sum_{i,j} \underbrace{\operatorname{Cov}\left(\mu_{i}(\mathbf{D})(2D_{i}-1), \mu_{j}(\mathbf{D})(2D_{j}-1)\right)}_{=\sigma_{i,j}}$$

hmm... and now?

Lem $|\sigma_{i,j}| = \mathcal{O}\big(\mathrm{Bias}_n(\mathcal{C}_n)\big)$ if (i,j) are (a) not friends or friends of friends, (b) in different clusters, and (c) j is not in a cluster with a friend of i

$$\Rightarrow \operatorname{Var}\left(\hat{\tau}|\mathbf{A},\mathcal{C}_{n}\right) \leq \underbrace{\bar{\phi}_{n}/n^{2}}_{\text{Bias}} \times \mathcal{O}\left(\underbrace{n \times \mathcal{N}_{n,\max}^{2}}_{\text{\# 2nd degree}} \times \underbrace{\frac{\bar{\gamma}n}{K_{n}}}_{\text{largest cluster size}}\right) + \underbrace{\frac{\bar{\psi}}{n^{2}} \sum_{i=1}^{n} |c(i)|}_{\text{ppl in same cluster}}$$

Davide Viviano November, 2023 18 / 18

Worst case variance depends on

$$\operatorname{Var}\left(\hat{\tau}|\mathbf{A},\mathcal{C}_{n}\right) = \frac{4}{n^{2}} \sum_{i,j} \underbrace{\operatorname{Cov}\left(\mu_{i}(\mathbf{D})(2D_{i}-1), \mu_{j}(\mathbf{D})(2D_{j}-1)\right)}_{=\sigma_{i,j}}$$

hmm... and now?

Lem $|\sigma_{i,j}| = \mathcal{O}\big(\mathrm{Bias}_n(\mathcal{C}_n)\big)$ if (i,j) are (a) not friends or friends of friends, (b) in different clusters, and (c) j is not in a cluster with a friend of i

$$\Rightarrow \mathrm{Var}\Big(\hat{\tau}|\mathbf{A},\mathcal{C}_n\Big) \leq \underbrace{\bar{\phi}_n/n^2}_{\mathsf{Bias}} \times \mathcal{O}\Big(\underbrace{n \times \mathcal{N}_{n,\mathsf{max}}^2}_{\text{$\#$ 2nd degree}} \times \underbrace{\frac{\bar{\gamma}\,n}{K_n}}_{\mathsf{largest cluster size}}\Big) + \underbrace{\frac{\bar{\psi}}{n^2} \sum_{i=1}^n |c(i)|}_{\mathsf{ppl in same cluster}}$$

$$=o(1/K_n)+\bar{\psi}\sum_{k=1}^{K_n}\frac{n_k^2}{n^2}$$

Davide Viviano 18 / 18