LECTURE 3: CLUSTERED RANDOMIZED EXPERIMENTS

Guido Imbens – Stanford University

Economics 272, GSB 507, Spring 2025

OUTLINE

1. To Cluster or Not To Cluster?
2. Clustering: Assignment or Sampling?
3. Cluster Randomization: Analysis
4. Cluster Randomization: Design and Power
5. Other Interesting Things in Experimental Design

TO CLUSTER OR NOT TO CLUSTER?

- Suppose you have a random sample, $N = 10,000$, from US population,
- Suppose you randomly assign M individuals to a job training program.
- Regressing the outcome on the treatment, you find

$$\hat{\tau} = \overline{Y}_T - \overline{Y}_C = 0.058 \quad \left(\text{Neyman/robust s.e. } \sqrt{\hat{\mathbb{V}}} = 0.011\right) \qquad \hat{\mathbb{V}} = \frac{s_C^2}{N-M} + \frac{s_T^2}{M}$$

- Your RA realizes you know which of the 50 states these individuals live in, and suggests clustering by state. (meaning, using Stata clustering option in regression – details in a couple of lectures).
- Question: What do you tell the RA, and why?
  - Yes, definitely cluster!
  - No, definitely do not cluster!
  - Does not matter.

VIEWS FROM THE ECONOMETRIC LITERATURE

- Hansen, 2007, p. 671

  "The clustering problem is caused by the presence of a common unobserved random shock at the group level that will lead to correlation between all observations within each group"

- Cameron and Miller, 2015, p. 333

  "The consensus is to be conservative and avoid bias and to use bigger and more aggregate clusters when possible, up to and including the point at which there is concern about having too few clusters."

EFFECT OF CLUSTERING THE STANDARD ERRORS

- The RA calculates the Liang-Zeger (STATA) clustered standard errors, comes back with:

$$\hat{\tau} = 0.058 \ \ \text{(s.e. 0.067)}$$

  (where the Neyman/robust standard error was 0.011 before).
- Are you confident that the program has a non-zero effect?
- What should you report?

TO BE PRECISE ABOUT STANDARD ERROR CALCULATION

- Set Up

$$Y_i = \beta_0 + \beta_1 W_i + \varepsilon_i = X_i^\top \beta + \varepsilon_i$$

$$\hat{\beta} = \left(\sum_{i=1}^{N} X_i X_i^\top \right)^{-1} \left(\sum_{i=1}^{N} X_i Y_i \right), \qquad \hat{\varepsilon} = Y_i - X_i^\top \hat{\beta}$$

- Neyman / Eicker-Huber-White / robust Standard errors

$$\hat{\mathbb{V}}^{\text{EHW}} = \left(\sum_{i=1}^{N} X_i X_i^\top \right)^{-1} \left(\sum_{i=1}^{N} X_i X_i^\top \hat{\varepsilon}_i^2 \right) \left(\sum_{i=1}^{N} X_i X_i^\top \right)^{-1}$$

- Liang-Zeger/ cluster-robust Standard errors

$$\hat{\mathbb{V}}^{\text{LZ}} = \left(\sum_{i=1}^{N} X_i X_i^\top \right)^{-1} \sum_{g=1}^{G} \left\{ \left(\sum_{i:G_i=g} X_i \hat{\varepsilon}_i \right) \left(\sum_{i:G_i=g} X_i \hat{\varepsilon}_i \right)^\top \right\} \left(\sum_{i=1}^{N} X_i X_i^\top \right)^{-1}$$

ROBUST AND CLUSTER-ROBUST STANDARD ERRORS FOR THE BINARY TREATMENT CASE

- With $W_i \in \{0, 1\}$ the variance estimators for the treatment effect estimator $\hat{\tau}$ simplifies to

$$\hat{\mathbb{V}}^{\text{EHW}}(\hat{\tau}) = \frac{1}{\overline{W}^2(1-\overline{W})^2} \left\{ \frac{1}{N} \sum_{i=1}^{N} \hat{\varepsilon}_i^2 (W_i - \overline{W})^2 \right\}$$

and

$$\hat{\mathbb{V}}^{\text{LZ}}(\hat{\tau}) = \frac{1}{\overline{W}^2(1-\overline{W})^2} \left\{ \frac{1}{N} \sum_{g=1}^{G} \left(\sum_{i:G_i=g} \hat{\varepsilon}_i (W_i - \overline{W}) \right)^2 \right\}$$

MODIFIED PROBLEM

- Would your answer change if the RA suggested clustering by gender or race?
- Would your answer change if I told you the residuals from the regression are uncorrelated?
- How would you explain any difference in the answers?

CHALLENGE

- In previous lectures we did not make any assumptions about the potential outcomes, so there could be an arbitrary correlation structure there.
- Should we have worried about clustering in that discussion?
- Why not?

CLUSTERING: ASSIGNMENT VERSUS SAMPLING

- Suppose we have a finite population with N units, partitioned into G clusters, with $G_i \in \{1, \ldots, G\}$ denoting the cluster/group.
- Two distinct clustering problems
  - Clustered Sampling: Sample M clusters out of the set of G clusters at random and sample a fraction p of the units from the sampled clusters. Units from the remaining clusters are not observed.
  - Clustered Assignment Select G_T clusters at random from the population of G clusters and assign all units from the selected clusters to the active treatment and assign all units from the remaining $G_C = G - G_T$ clusters to the control treatment.
- Very different problems, but very similar solutions, and often conflated.

FOUR DIFFERENT COMBINATIONS

- clustered assignment and random sampling (today)
- clustered assignment and clustered sampling (later in course)
- clustered sampling and random assignment (later in course)
- random sampling and random assignment (first two lectures)

ASYMPTOTICS

- Sequence of populations, $k = 1, \ldots,$
- In population k
  - there are G_k clusters,
  - $N_{k,g}$ units in cluster g, $N_k = \sum_{g=1}^{G_k} N_{k,g}$ units total.
- Sampling:
  - Cluster g is sampled with probability q_k.
  - Units from the sampled clusters are sampled with probability p_k.
- Assignment:
  - For each cluster a probability $A_{k,g}$ is drawn randomly, with support $[0, 1]$, mean μ_k and variance σ_k^2.
  - Units in cluster g are assigned to the treatment with probability $A_{k,g}$.
- Observed is treatment assignment $W_{k,i} \in \{0, 1\}$ and outcome $Y_{k,i} = y_{k,i}(W_{k,i})$
- Interest in average effect $\tau = \sum_{i=1}^{N_k} (y_i(\mathrm{T}) - y_{k,i}(\mathrm{C}))/N_k$

SPECIAL CASES

- Asymptotics generally use $N_k \to \infty$, sometimes $G_k \to \infty$, sometimes $G_k = G, \forall k$.
- Random sampling of clusters from a large population of clusters, q_k is small (clustered sampling, not that common)
- Random sampling from a large population, $q_k = 1$, p_k is small.
- Completely random assignment, $A_{k,g} = A_k \forall g$ (first two classes)
  - robust standard errors
- Clustered random assignment, $A_{k,g} \in \{0, 1\}, \forall k, g$ (today)
  - cluster-robust standard errors
- How do we decide what case we are interested in given a sample (W_i, Y_i, G_i)?
  - This is not a statistical question, cannot be answered on the basis of the data.
- Challenge: what to do if $A_{k,g}$ has a distribution with positive variance that has support different from $\{0, 1\}$ (next class)

CLUSTERED RANDOMIZED EXPERIMENTS: ANALYSIS

- (dropping indexing by population k for ease of notation)
- Population with G groups/clusters. Cluster g has N_g units. Total number of units $N = \sum_{g=1}^{G} N_g$. $G_i \in \{1, \ldots, G\}$ indicates cluster that units i belongs to.
- We consider two objects of interest (estimands):
  - The overall average

$$\tau_{\text{pop}} = \frac{1}{N} \sum_{i=1}^{N} \Big(Y_i(\text{T}) - Y_i(\text{C}) \Big)$$

  - The average, over clusters, of the within-cluster average effects:

$$\tau_{\text{cluster}} = \frac{1}{G} \sum_{g=1}^{G} \frac{1}{N_g} \sum_{i:G_i=g} \Big(Y_i(\text{T}) - Y_i(\text{C}) \Big)$$

- $\tau_{\text{pop}} = \tau_{\text{cluster}}$ if either:
  - treatment effect is constant (implausible in practice)
  - $N_g = N/G$ for all g, equal cluster sizes, (sometimes holds).

ISSUES

- Clustering the assignment generally leads to a loss of precision/power relative to a completely randomized experiment.
- Clustering is often used to deal with spillover problems.
- (as opposed to stratification which leads to a gain in precision/power.)
- Choices regarding statistics and estimands (τ_{pop} *versus* τ_{cluster}) are important if there is much variation in N_g.
- Extreme cases:
  - few large clusters and many small clusters (*e.g.,* states, cities, countries),
  - approximately equal sized clusters (*e.g.,* classrooms).

SETTING FOR CLUSTER-RANDOMIZED EXPERIMENT

- G_T clusters are selected at random out of G cluster total. All units in those clusters receive the active treatment. The remaining $G_C = G - G_T$ units receive the control treatment. Probability of being treated is $p = G_T/G$ both for all cluster and for units.
- Different from unit-level randomization where N_T units are selected at random to receive active treatment because assignment now has complex dependence structure.
- Inference is based on the randomization distribution induced by clustered assignment. (As before, we keep the potential outcomes fixed.)
- Exact results are difficult/impossible, often asymptotic results where a sequence of populations is considered with the number of clusters getting large, with number of units per clusters fixed. (But focus remains on results for finite population estimands.)

FISHER EXACT P-VALUES

- This is conceptually straightforward extension of p-value calculation in completely randomized experiment.
  - Choose a sharp null, *e.g.,* no effect of treatment.
  - Choose a statistic.
  - Given that we know the randomization distribution, we can infer the exact p-value.
- Difference with completely randomized experiment is that the p-values are based on a different randomization distribution.

THE CHOICE OF STATISTIC

- Two natural choices (and variants thereof):
  - First

$$T_{\text{ave}} = \frac{\sum_{i=1}^{N} W_i Y_i}{\sum_{i=1}^{N} W_i} - \frac{\sum_{i=1}^{N} (1 - W_i) Y_i}{\sum_{i=1}^{N} (1 - W_i)}$$

  - and second

$$T_{\text{cluster-ave}} = \frac{1}{G_1} \sum_{g=1}^{G} \frac{\sum_{i:G_i=g}^{N} W_i Y_i}{N_g} - \frac{1}{G_0} \sum_{g=1}^{G} \frac{\sum_{i:G_i=g}^{N} (1 - W_i) Y_i}{N_g}$$

- Note: $T_{\text{ave}} = T_{\text{cluster-ave}}$ if equal cluster sizes, $N_g = N/G$ for all g.
- There could be a big difference in power for the two statistics if there is substantial variation in N_g, even under the same alternative of a constant additive treatment effect.

DESIGN AND POWER: A SMALL SIMULATION STUDY

- Simulations: 500 clusters, 10,000 units. 4 big clusters with 2,004 units, 496 small clusters with 4 units.
- Data generating process, with constant treatment effect $\tau = 0.1$

$$Y_i(\text{C}) = \eta_{G_i} + \varepsilon_i, \qquad Y_i(\text{T}) = Y_i(\text{C}) + 0.1 \times W_i$$

$$\eta_g \sim \mathcal{N}(0,1), \quad \varepsilon_i \sim \mathcal{N}(0,1)$$

- Test size 0.05.
- Results
  - Power based on T_{ave}: 0.064 (very little power because mainly 4 big clusters matter)
  - Power based on $T_{\text{cluster-ave}}$: 0.201 (substantial power, because all 500 clusters matter equally)

ESTIMATION OF AVERAGE EFFECTS

- Two estimators:
  - to estimate τ_{pop},

$$\hat{\tau}_{\text{pop}} = \overline{Y}_{\text{T}} - \overline{Y}_{\text{C}}$$

  - and to estimate τ_{cluster}:

$$\hat{\tau}_{\text{cluster}} = \frac{1}{G_{\text{T}}} \sum_{g=1}^{G} \frac{1}{N_g} \sum_{i:G_i=g} W_i Y_i - \frac{1}{G_{\text{C}}} \sum_{g=1}^{G} \frac{1}{N_g} \sum_{i:G_i=g} (1 - W_i) Y_i$$

- Recall: if equal cluster sizes, then:

$$\hat{\tau}_{\text{pop}} = \hat{\tau}_{\text{cluster}}$$

ESTIMATING THE CLUSTER AVERAGE τ_{cluster}

- Conceptually straightforward
- Define cluster average:

$$\overline{Y}_g = \frac{1}{N_g} \sum_{i:G_i=g} Y_i, \qquad \overline{Y}_g(w) = \frac{1}{N_g} \sum_{i:G_i=g} Y_i(w)$$

$$\overline{W}_g = \frac{1}{N_g} \sum_{i:G_i=g} W_i$$

- Note: $\overline{W}_g \in \{0, 1\}$ because we assign at the cluster level.
- Then:

$$\hat{\tau}_{\text{cluster}} = \frac{1}{G_T} \sum_{g=1}^{G} \overline{W}_g \overline{Y}_g - \frac{1}{G_T} \sum_{g=1}^{G} (1 - \overline{W}_g) \overline{Y}_g$$

ESTIMATING THE VARIANCE OF THE CLUSTER AVERAGE ESTIMATOR $\hat{\tau}_{\text{cluster}}$

- This is just like a completely randomized experiment, but with clusters as units of analysis, and analysis is straightforward.
- The exact variance, based on Neyman variance calculations from completely randomized experiment, is

$$\mathbb{V}(\hat{\tau}_{\text{cluster}}) = \frac{S_C^2}{G_C} + \frac{S_T^2}{G_T} - \frac{S_{CT}^2}{G} \qquad S_w^2 = \sum_{g=1}^{G} (\overline{Y}_g(w) - \overline{\overline{Y}_g(w)})^2/(G-1)$$

- Estimated (conservative) variance is

$$\hat{\mathbb{V}}(\hat{\tau}_{\text{cluster}}) = \frac{s_C^2}{G_C} + \frac{s_T^2}{G_T} \qquad s_w^2 = \frac{1}{(G_w - 1)} \sum_{g:\overline{W}_g=w} (\overline{Y}_g - \overline{\overline{Y}_{g_w}})^2$$

COMPARISON TO STANDARD ROBUST VARIANCE

- Note that this variance estimator $\hat{\mathbb{V}}$ is very close to the standard robust variance estimator for regression models, based on using the clusters as units.
- Consider a linear regression model with cluster-level outcomes and regressors:

$$Y_g = X_g^\top \beta + \varepsilon_g,$$

with least squares estimator

$$\hat{\beta} = \left(\frac{1}{G} \sum_{g=1}^{G} X_g X_g^\top \right)^{-1} \left(\frac{1}{G} \sum_{g=1}^{G} X_g Y_g \right).$$

COMPARISON TO STANDARD ROBUST VARIANCE (CTD)

- Under standard conditions we have

$$\sqrt{G}(\hat{\beta} - \beta) \xrightarrow{d} \mathcal{N}(0, V),$$

- $\mathbb{V}$ can be estimated as

$$\hat{\mathbb{V}} = \left(\frac{1}{G}\sum_{g=1}^{G} X_g X_g^\top\right)^{-1} \left(\frac{1}{G}\sum_{g=1}^{G} (Y_g - X_g^\top \hat{\beta})^2 X_g X_g^\top\right) \left(\frac{1}{G}\sum_{g=1}^{G} X_g X_g^\top\right)^{-1}$$

- Simplify this to the case where $X_i \in \{0, 1\}$, so that the slope is $\hat{\beta}_1 = \overline{Y}_1 - \overline{Y}_0$,
- the variance estimator for $\hat{\beta}_1$ is

$$\hat{\mathbb{V}} = \frac{\tilde{s}_0^2}{G_0} + \frac{\tilde{s}_1^2}{G_1} \qquad s_w^2 = \frac{1}{G_w} \sum_{g:\overline{W}_g = w} (\overline{Y}_g - \overline{\overline{Y}_{g_w}})^2$$

- (no degrees of freedom adjustment)

ESTIMATING τ_{pop}

- The simple difference in means estimator

$$\hat{\tau}_{pop} = \overline{Y}_T - \overline{Y}_C = \frac{\sum_{i=1}^N W_i Y_i}{\sum_{i=1}^N W_i} - \frac{\sum_{i=1}^N (1 - W_i) Y_i}{\sum_{i=1}^N (1 - W_i)}$$

  is unbiased for τ_{pop} because all units have probability of being treated equal to G_T/G.

- But, the assignments are not independent and that means the standard robust variance estimator is not appropriate.

- Unlike for $\tau_{cluster}$ exact variance calculations are not feasible because the number of treated and control units N_T and N_C (the denominators in $\overline{Y}_T$ and $\overline{Y}_C$), are stochastic if there is variation in cluster sizes:

$$N_T - \frac{G_T}{G} N = \sum_{g=1}^{G} \left(W_g - \frac{G_T}{G} \right) N_g$$

VARIANCE CALCULATION BASED ON ABADIE ET AL (2023)

- Consider a sequence of populations $k = 1, 2, \ldots$. In population k we sample all units, and assign cluster g to the treatment group with probability μ and to the control group with probability $1 - \mu$, independently across clusters.
- We estimate the average effect τ_{pop} as $\hat{\tau}_{\text{pop}} = \overline{Y}_{\text{T}} - \overline{Y}_{\text{C}}$.
- In the sequence of populations $k = 1, 2, \ldots$ the number of clusters G_k increases, and the total number of units N_k increases (but the number of units per cluster $N_{k,g}$ can stay fixed or increase).
- Define the residuals

$$\varepsilon_{ki}(\text{C}) = y_{ki}(\text{C}) - \frac{1}{n_k}\sum_{j=1}^{n_k} y_{kj}(\text{C}) \qquad \varepsilon_{ki}(\text{T}) = y_{ki}(\text{T}) - \frac{1}{n_k}\sum_{j=1}^{n_k} y_{kj}(\text{T})$$

$$\varepsilon_{ki} = \varepsilon_{ki}(\text{T})W_{ki} + \varepsilon_{ki}(\text{C})(1 - W_{ki})$$

- $G_{k,i}$ indicates which cluster a unit is from.

VARIANCE CALCULATION BASED ON ABADIE ET AL (2023)

- Then

$$\sqrt{N_k}\left(\frac{\hat{\tau}_{\text{pop}} - \tau_{\text{pop}}}{\sqrt{v_k}}\right) \xrightarrow{d} \mathcal{N}(0,1)$$

- where

$$v_k = \frac{1}{N_k}\sum_{i=1}^{N_k}\left(\frac{\varepsilon_{k,i}^2(\text{T})}{\mu} + \frac{\varepsilon_{k,i}^2(\text{C})}{1-\mu}\right) - \frac{1}{N_k}\sum_{i=1}^{N_k}\left(\varepsilon_{k,i}(\text{T}) - \varepsilon_{k,i}(\text{C})\right)^2$$

$$-\mu(1-\mu)\frac{1}{N_k}\sum_{i=1}^{N_k}\left(\frac{\varepsilon_{k,i}(\text{T})}{\mu} + \frac{\varepsilon_{k,i}(\text{C})}{1-\mu}\right)^2$$

$$+\mu(1-\mu)\frac{1}{N_k}\sum_{g=1}^{G}\left(\sum_{i|G_{k,i}=g}\left(\frac{\varepsilon_{k,i}(\text{T})}{\mu} + \frac{\varepsilon_{k,i}(\text{C})}{1-\mu}\right)\right)^2.$$

VARIANCE ESTIMATION

- A conservative estimator can be based on the standard cluster variance for the regression

$$Y_{k,i} = \alpha + W_{k,i}\tau + \varepsilon_{k,i} \qquad \hat{\varepsilon}_{ki} = Y_{k,i} - \hat{\alpha} - W_{k,i}\hat{\beta}$$

- Cluster variance estimator

$$\hat{\mathbb{V}}_{\text{LZ}} =$$

$$\left(\frac{1}{N_k}\sum_{i=1}^{N_k}(W_{ki}-\overline{W}_k)^2\right)^{-1}\frac{1}{N_k}\sum_{g=1}^{G}\left(\sum_{i|G_{k,i}=g}\hat{\varepsilon}_{ki}(W_{ki}-\overline{W}_k)\right)^2\left(\frac{1}{N_k}\sum_{i=1}^{N_k}(W_{ki}-\overline{W}_k)^2\right)^{-1}$$

INTUITION

- Suppose

$$Y_i(\mathrm{C}) = \alpha + \varepsilon_i, \qquad Y_i(\mathrm{T}) = Y_i(\mathrm{C}) + \tau$$

- Suppose cluster sizes are large, $N_{k,g} \to \infty$, in the sequence of populations.
- Then the first three terms in v_k converge to constants.
- The fourth term is proportional to

$$\frac{1}{N_k} \times G_k \times \left(\frac{N_k}{G_k}\right)^2 = \frac{N_k}{G_k}$$

  the average cluster size, which diverges, so it is the largest term.

INTUITION

- Under the constant treatment effect assumption $\varepsilon_{k,i} = \varepsilon_{k,i}(\text{C}) = \varepsilon_{k,i}(\text{T})$ the last term in the true variance simplifies to

$$+\mu(1-\mu)\frac{1}{N_k}\sum_{g=1}^{G}\left(\sum_{i|G_{k,i}=g}\left(\frac{\varepsilon_{k,i}}{\mu}+\frac{u_{k,i}}{1-\mu}\right)\right)^2$$

$$=\mu(1-\mu)\frac{1}{N_k}\sum_{g=1}^{G}\left(\sum_{i|G_{k,i}=g}\frac{\varepsilon_{k,i}}{\mu(1-\mu)}\right)^2$$

$$=\frac{1}{\mu(1-\mu)}\frac{1}{N_k}\sum_{g=1}^{G}\left(\sum_{i|G_{k,i}=g}\varepsilon_{k,i}\right)^2$$

INTUITION

- The estimated variance is

$$\left(\frac{1}{N_k}\sum_{i=1}^{N_k}(W_{k,i}-\overline{W}_k)^2\right)^{-1}\frac{1}{N_k}\sum_{g=1}^{G}\left(\sum_{i|G_{k,i}=g}\hat{\varepsilon}_{ki}(W_{k,i}-\overline{W}_k)\right)^2\left(\frac{1}{N_k}\sum_{i=1}^{N_k}(W_{k,i}-\overline{W}_k)^2\right)$$

$$\approx\left(\frac{1}{N_k}\sum_{i=1}^{N_k}(W_{k,i}-\overline{W}_k)^2\right)^{-1}\frac{1}{N_k}\sum_{g=1}^{G}\left(\sum_{i|G_{k,i}=g}\varepsilon_{ki}(W_{k,i}-\overline{W}_k)\right)^2\left(\frac{1}{N_k}\sum_{i=1}^{N_k}(W_{k,i}-\overline{W}_k)^2\right)$$

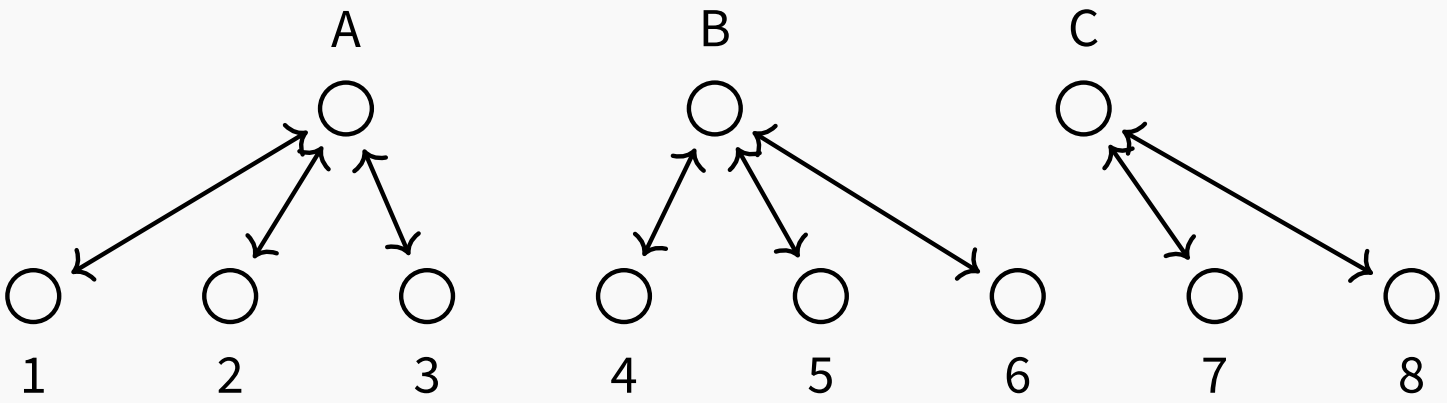
Note that $(1/N_k)\sum_{i=1}^{N_k}(W_{k,i}-\overline{W}_k)^2\approx\mu(1-\mu)$. In addition, $W_{k,i}$ is independent of $\varepsilon_{k,i}$ and $G_{k,i}$, so the middle factor is approximately

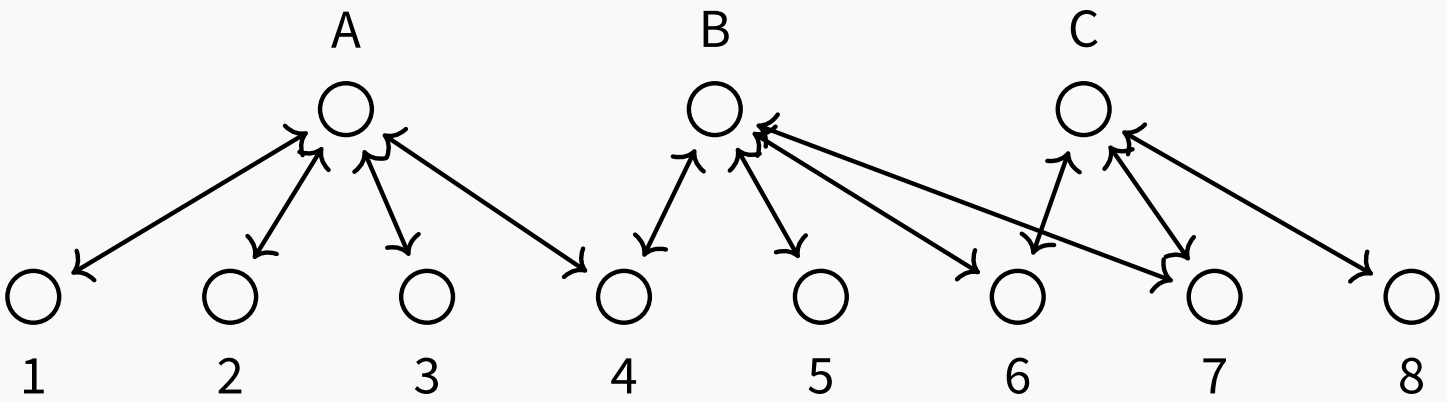
$$\frac{1}{N_k}\sum_{g=1}^{G}\left(\sum_{i|G_{k,i}=g}\varepsilon_{k,i}(W_{k,i}-\overline{W}_k)\right)^2\approx\mu(1-\mu)\frac{1}{N_k}\sum_{g=1}^{G}\left(\sum_{i|G_{k,i}=g}\varepsilon_{k,i}\right)^2$$

Putting this together gives the equality.

CHALLENGES WITH UNKNOWN CLUSTERING, PART I

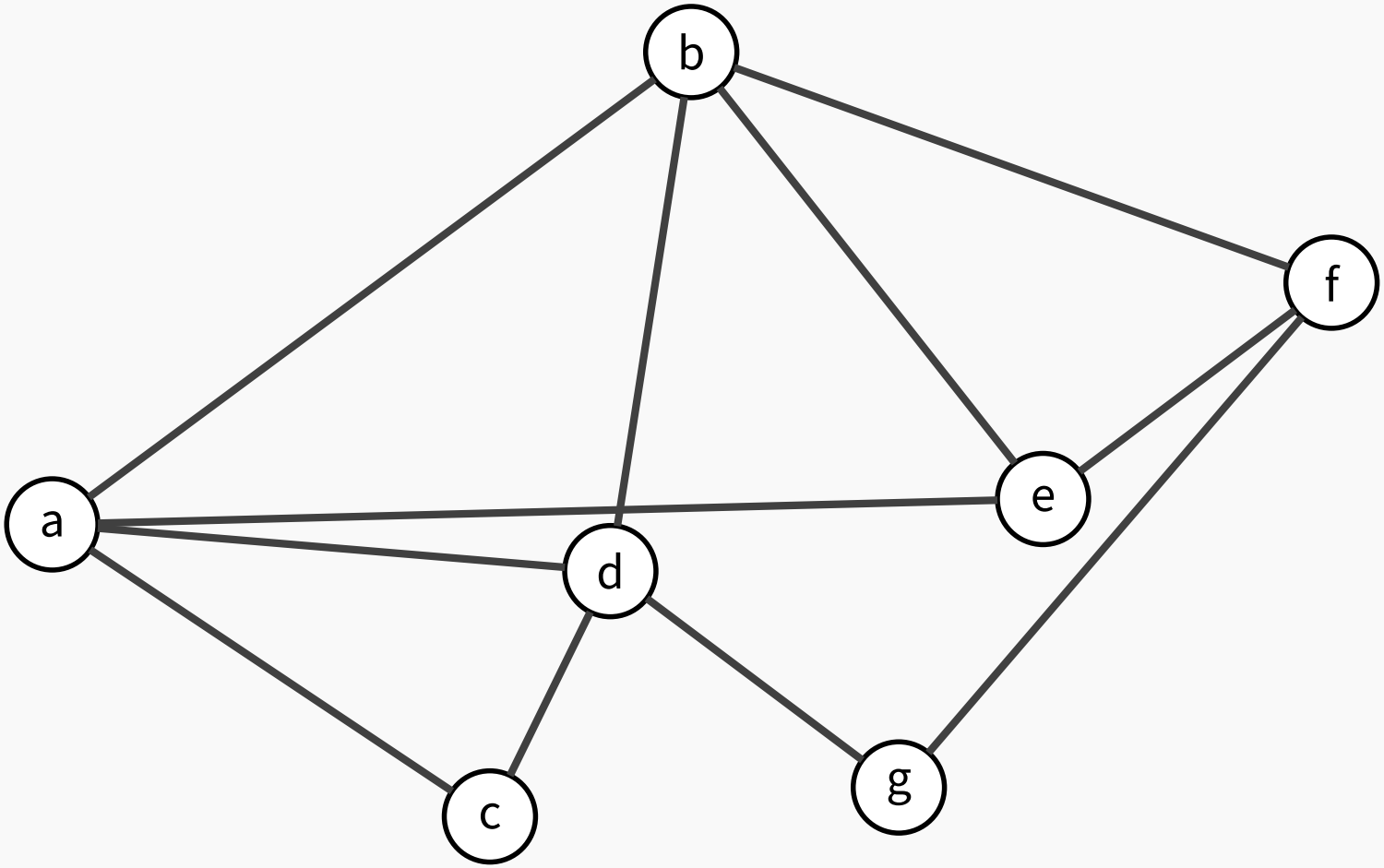
- Challenges arise when clusters are not known.
- Sometimes mediated through bipartite graph (*e.g.*, advertising auction setting, numbered units are search queries, alphabetical units are products)
- How do you create clusters with minimal broken links?





CHALLENGES WITH UNKNOWN CLUSTERING, PART II

- Here a network setting (*e.g.*, Facebook)
- How to cut the graph to create clusters while minimizing cutting links



REFERENCES

- Abadie, Alberto, Susan Athey, Guido W. Imbens, and Jeffrey M. Wooldridge. "When should you adjust standard errors for clustering?." *The Quarterly Journal of Economics* 138, no. 1 (2023): 1-35.
- Cameron, A. Colin, and Douglas L. Miller. "A practitioner's guide to cluster-robust inference." *Journal of human resources* 50, no. 2 (2015): 317-372.
- Hansen, Christian B. "Generalized least squares inference in panel and multilevel models with serial correlation and fixed effects." Journal of econometrics 140, no. 2 (2007): 670-694.