# Randomized Graph Cluster Randomization With R Package: RGCR

Changhao Shi, Nan Qiao, Xiaolin Xu

Renmin University of China, Statistics

December 18, 2023

- Ugander, J. and Yin, H., 2023. Randomized graph cluster randomization. *Journal of Causal Inference*, 11(1), p.20220014.
- You can download 'RGCR' from https://github.com/RUC-ChangHao/RGCR.

# Classic causal inference and randomization

- Binary treatment $Z_i \in \{0, 1\}$ ($Z \sim P(z)$, CR, Bernoulli design, etc)
- Potential outcomes $Y_i(1)$ and $Y_i(0)$ (SUTVA assumption)
- Causal effects of interest: average causal effect (ACE)

$$\text{ACE} \stackrel{\text{def}}{=} \sum_{i=1}^{n} \{Y_i(1) - Y_i(0)\}$$

- Estimators: ht (ipw), hajek, aipw, etc

$$\hat{\tau}_{\text{ht}} = \frac{1}{n} \sum_{i=1}^{n} \frac{Y_i(1) Z_i}{P(Z_i=1)} - \frac{1}{n} \sum_{i=1}^{n} \frac{Y_i(0)(1-Z_i)}{P(Z_i=0)}$$

- Technically, a missing data problem

| | 1 | 2 | 3 | $\cdots$ | $n$ |
|---|---|---|---|---|---|
| $Y_i(1)$ | 36.5 | ? | ? | $\cdots$ | 38.0 |
| $Y_i(0)$ | ? | 38.0 | 37.0 | $\cdots$ | ? |

- Question: To get better estimation of ATE, how do we assign treatment $Z$? (i.e., seek the 'optimal' experiment design $Z \sim P(z)$)

# Causal inference under interference

- Violation of SUTVA
- Common in advertising, epidemiology and educational studies
- Potential outcomes $Y_i(\mathbf{z})$, where $\mathbf{z} \in \{0,1\}^n$
- Causal effects of interest: total treatment effect (TTE)

$$\mathrm{TTE} \stackrel{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n}\{Y_i(\mathbf{1}) - Y_i(\mathbf{0})\}$$

- Classic randomization does not work here (why?)

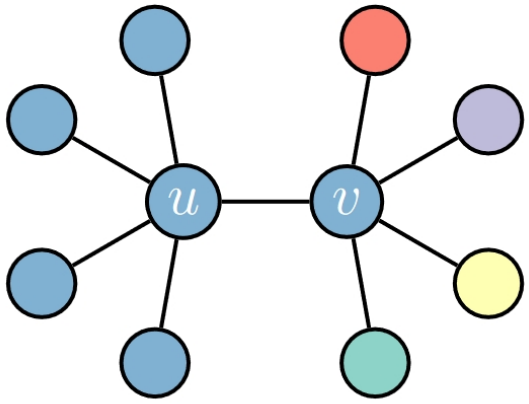
Table: Bias (SD) of HT Estimators

| sample size | $16^2$ | $24^2$ | $32^2$ | $48^2$ | $64^2$ |
|---|---|---|---|---|---|
| SUTVA | 0.00 (0.22) | -0.04 (0.16) | 0.02 (0.13) | 0.00 (0.08) | 0.01 (0.05) |
| interference | 0.09 (7.91) | -1.00 (3.74) | -0.47 (8.47) | -0.39 (2.33) | 0.22 (4.61) |

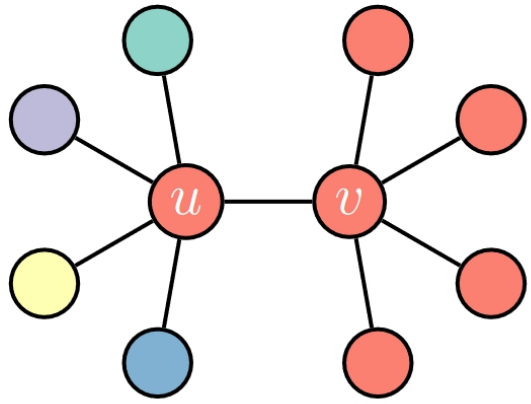
# General framework for interference

- A social network $G = (V, E)$
  - through which individuals interfere each other
  - observable and correctly measured
- An exposure mapping
  - determines the extent and intensity of the interference
  - technically reduces the number of potential outcomes
  - canonical examples (minor notation abuse)
    - (no interference) $Y_i(z) = Y_i(z_i)$
    - (neighborhood interference) $Y_i(z) = Y_i(z_{\mathcal{N}_i})$
    - (arbitrary interference) $Y_i(z) = Y_i(z)$
    - ("individualized" interference) $Y_i(z) = Y_i(?)$
- Estimators: ht, hajek, difference-in-means, etc
- Experimental designs $Z \sim P(z)$: complete randomization, bernoulli randomization, cluster randomization, etc

# A toy example




clustering $\boldsymbol{c}_1$




clustering $\boldsymbol{c}_2$

- Under full neighborhood interference

$$\mathrm{TTE} \stackrel{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n}\{Y_i(\mathbf{1}) - Y_i(\mathbf{0})\} = \frac{1}{n}\sum_{i=1}^{n}\{Y_i(\mathbf{1}_{\mathcal{N}_i}) - Y_i(\mathbf{0}_{\mathcal{N}_i})\}$$

  - what really matters is valid sample size
  - what's a good design? intuitively (under proper social networks)
    - gives valid sample size a lower bound (increasing with sample size)
    - gives every node a positive exposure probability (due to ht estimators)

# Randomized graph cluster randomization

- Randomized graph clustering
  - a weight indicating which nodes are more important
    - related to the variance of $\hat{\tau}_{\text{ht}}$, roughly speaking, related to the inverse exposure probabilities $1/\pi_i^1$ and $1/\pi_i^0$
    - types: uniform, degree, spectral weighting
  - a randomized clustering design deduced by the weight
    - types: 3-net and 1-hop-max (heuristically)
  - a clustering results generated from the clustering design
- Randomization at the level of clusters
  - types: complete design and bernoulli design
- Remark: no literature really solves this problem, all algorithms are heuristic in some sense

$$d^*, \hat{\tau}^* = \arg \min_{d \in \mathcal{D}, \hat{\tau} \in \mathcal{M}} \max_{\{Y_i(z)\} \subseteq \mathcal{Y}} E_{\{Y_i(z)\}, d, \hat{\tau}} \left[ \text{dist} \left( \hat{\tau}_d(Y, Z), \tau(\{Y_i(z)\}) \right) \right]$$

# Weighted 3-net clustering

**Algorithm 4:** Weighted 3-net clustering.

**Input:** Graph $G = (V, E)$, node weights $\boldsymbol{w} \in \mathbb{R}_+^n$.
**Output:** Graph clustering $\boldsymbol{c} \in \mathbb{R}^n$

**for** $i \in V$ **do**
  $X_i \leftarrow \beta(w_i, 1)$
$\pi \leftarrow \arg\operatorname{sort}([X_i]_{i \in V}, descend)$
$S \leftarrow \emptyset$, unmark all nodes
**for** $i \in \pi$ **do**
  **if** $i$ *is unmarked* **then**
    $S \leftarrow S \cup \{i\}$
    **for** $j \in B_2(i)$ **do**
      mark node $j$ if it is unmarked yet
**for** $i \in V$ **do**
  $c_i \leftarrow \arg\min\{j \in S, j \to \mathbf{dist}(i, j)\}$, i.e., the id of the node in $S$ with shortest graph distance to $i$ (arbitrary tie breaking)
**return** $\boldsymbol{c}$

# Weighted 1-hop-max clustering

**Algorithm 3:** Weighted 1-hop-max clustering.

**Input:** Graph $G = (V, E)$, node weights $\boldsymbol{w} \in \mathbb{R}_+^n$.

**Output:** Graph clustering $\boldsymbol{c} \in \mathbb{R}^n$

**for** $i \in V$ **do**

  $X_i \leftarrow \beta(w_i, 1)$

**for** $i \in V$ **do**

  $c_i \leftarrow \max([X_j \text{ for } j \in B_1(i)])$

**return** $\boldsymbol{c}$

# R package 'RGCR'

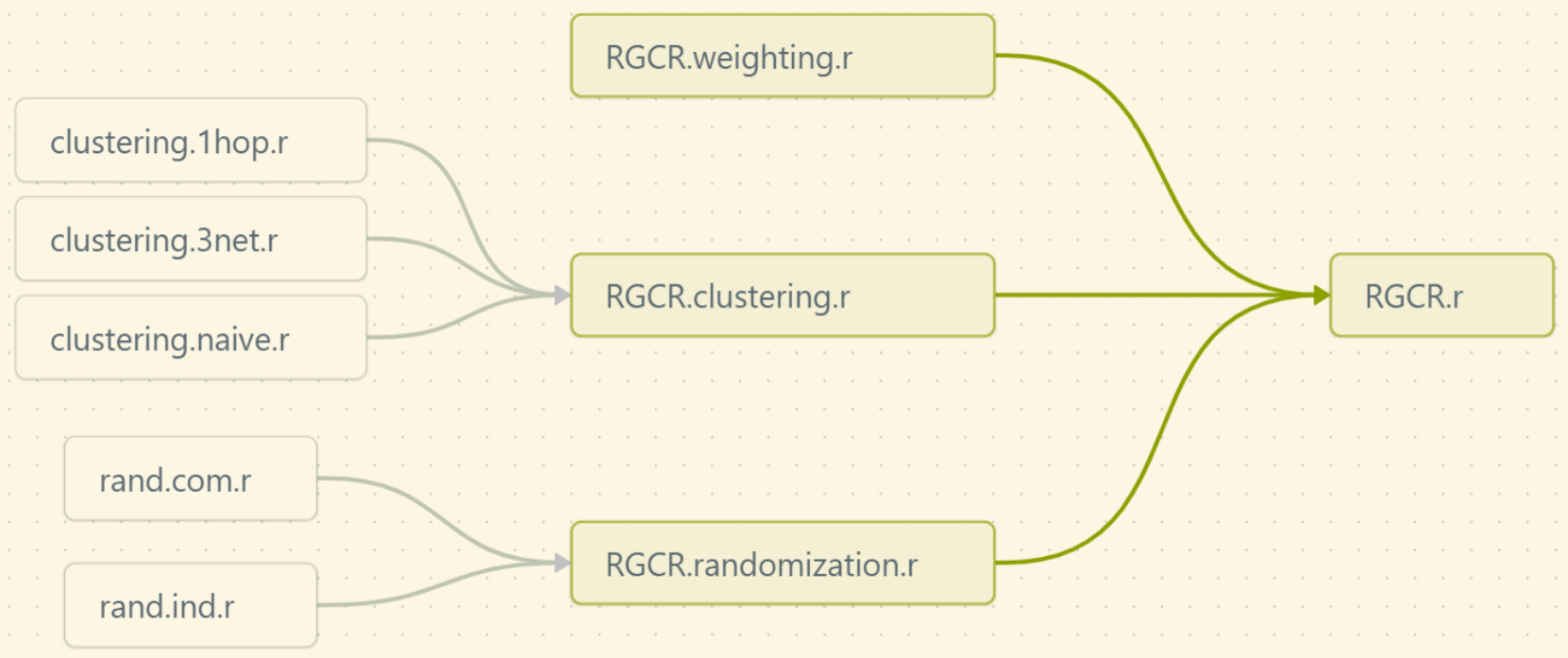
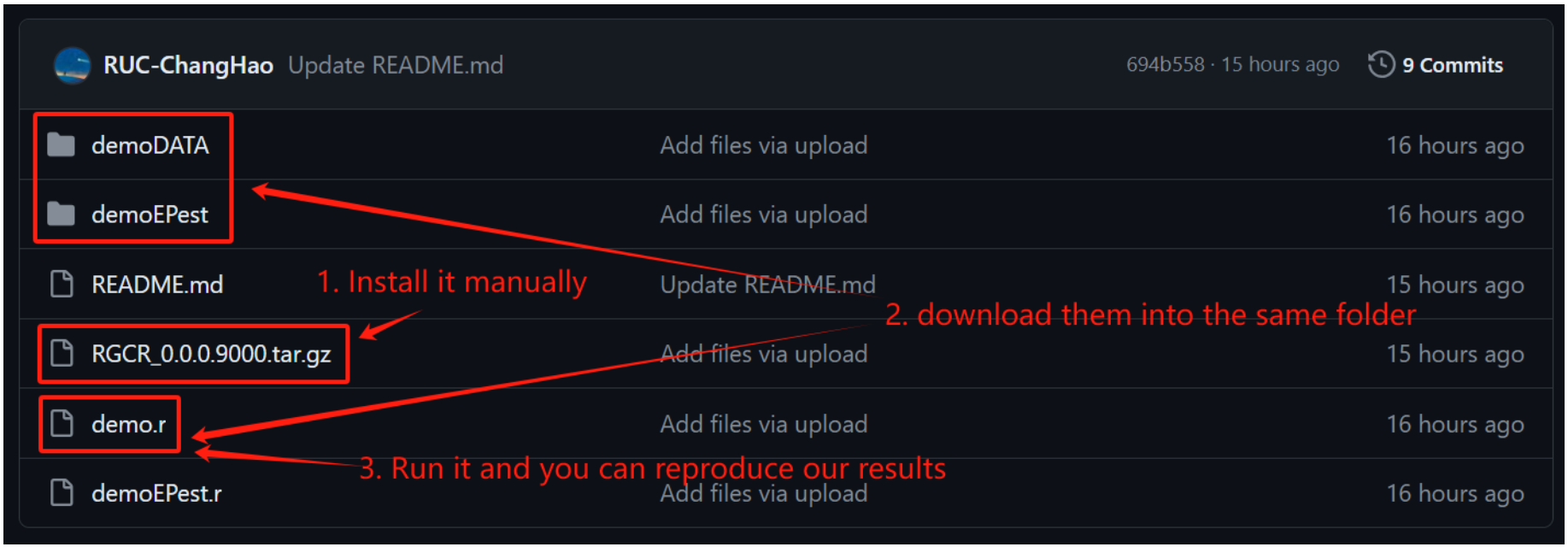



Figure: The Structure of functions

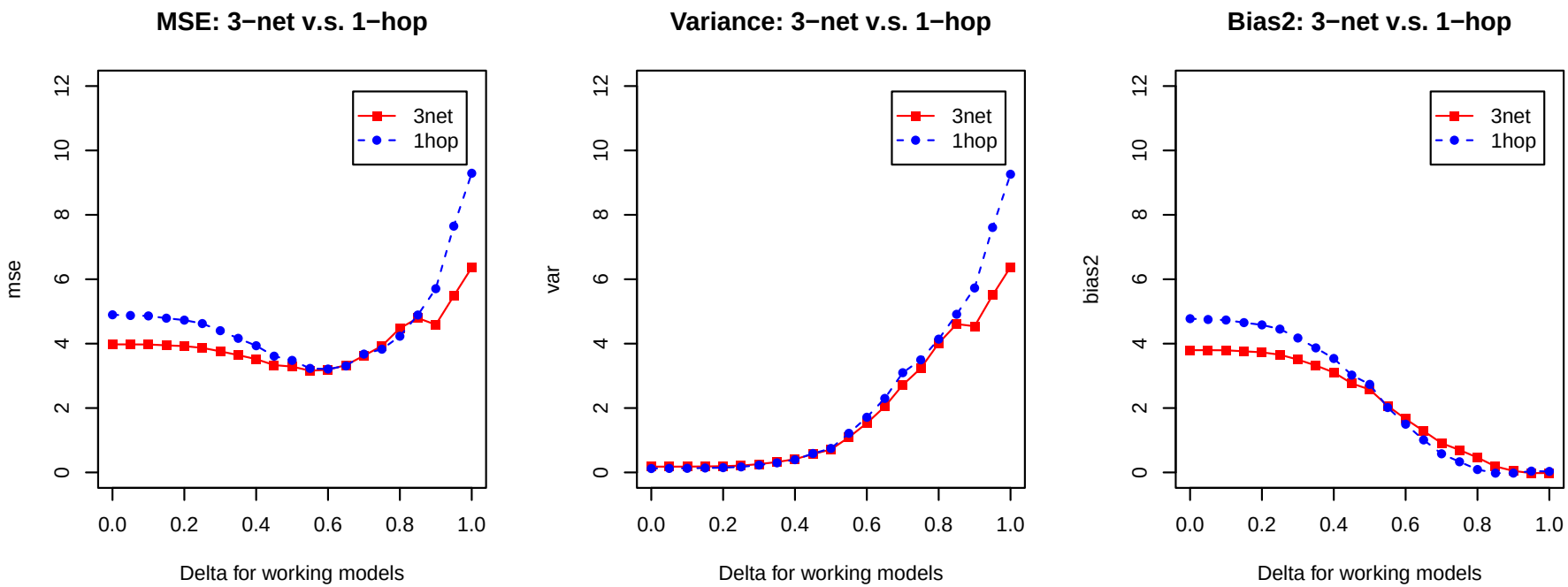
- Source: https://github.com/RUC-ChangHao/RGCR

# A short guidance to use 'RGCR'



# Application of 'RGCR'

- The real data analysis of experimental design is actually a simulation




- Conclusion: the design 'spectral weighting + 3-net clustering + complete randomization' is recommended (after numerous simulations...)

Thank you!