Differential Privacy for Network Connectedness Indices

Amin Rahimian
Industrial Engineering
Institute for Cyber Law, Policy and Security
Intelligent Systems Program
University of Pittsburgh

Digital Fellow, MIT Initiative on Digital Economy

Privacy and Public Policy Conference
Georgetown University; February 9, 2026

with Tom Rutter (Stanford) and Yuxin Liu (Pitt)

Why connected indices?

Measures of social capital that explain intergenerational mobility

Social Capital and Intergenerational Mobility

Social scientists have long suggested that social capital – the structure of the network of social relationships – is a key input to economic outcomes such as intergenerational mobility.

- **Intergenerational mobility: the average adult income of children growing up to parents at the 25th percentile of the national income distribution.**

Chetty et al. (2022) and Harris et al. (2025) found that one measure of social capital is strongly related to intergenerational mobility: ***economic connectedness***.

Economic connectedness captures the extent to which low-SES individuals in an area have high-SES friends.

Measuring Socioeconomic Status

Construct an index of socioeconomic status (SES) by combining several proxies: ZIP code, college, phone model price, …

Baseline measure: combination that best predicts median household income in block group (available for a subset of users) using a machine learning model

Rank users in the **national** distribution based on their predicted SES ranks relative to others in their cohort.

Measuring Economic Connectedness Across Subgroups

Facebook data have sufficiently large samples to allow us to disaggregate across subgroups (ZIP codes, high schools, colleges, etc.)

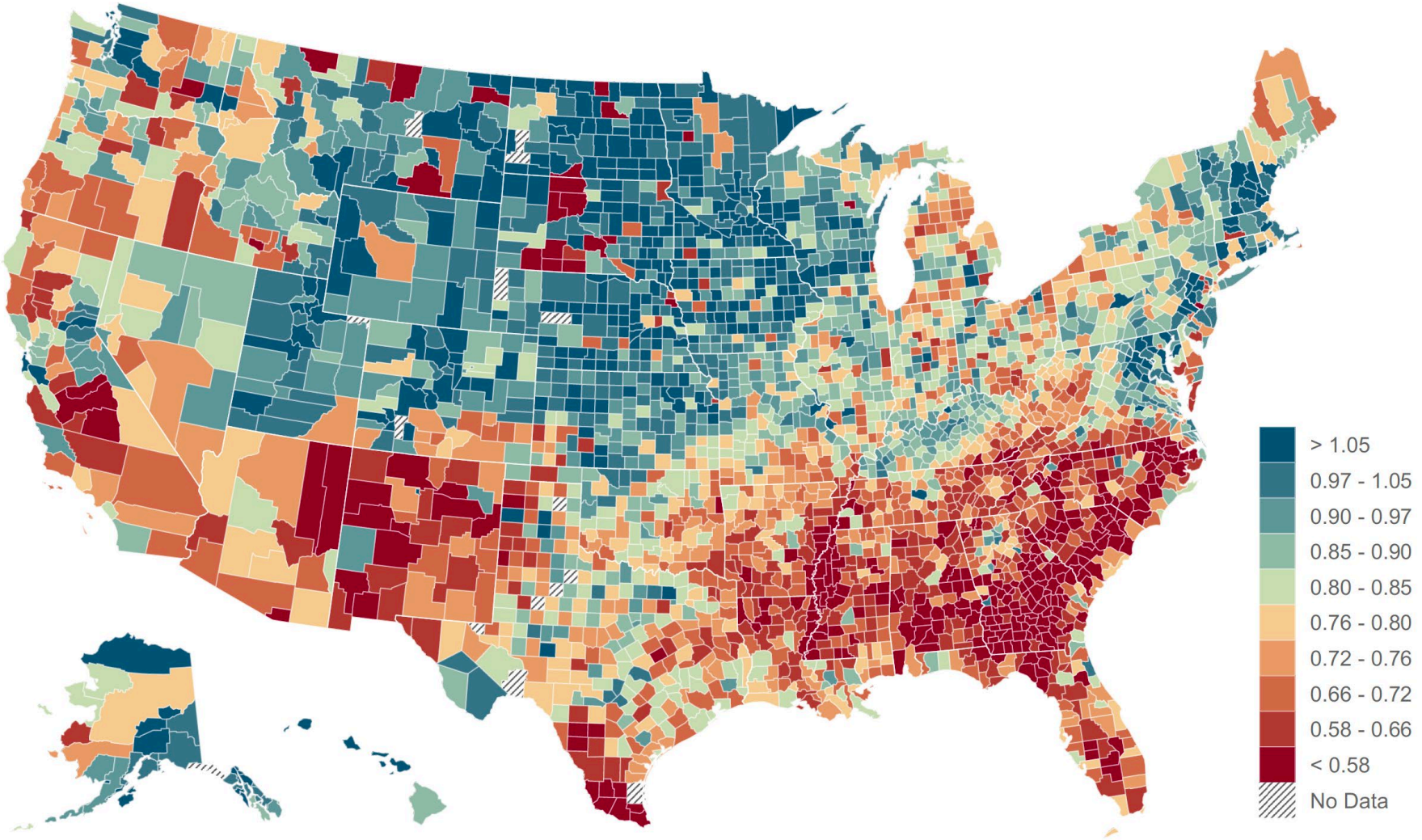
Summarize the degree to which low-SES people in a given group are connected to high-SES people using the following statistic:

$$EC = \frac{\text{Number of friends with above}-\text{median SES}}{\text{Total number of friends}}$$

Mean EC nationally = 0.39: **22% under-representation** of high-SES friends relative to random-friending benchmark

Economic Connectedness of Low-SES Individuals by County

Normalized Share of Above-Median Friends Among Below-Median People

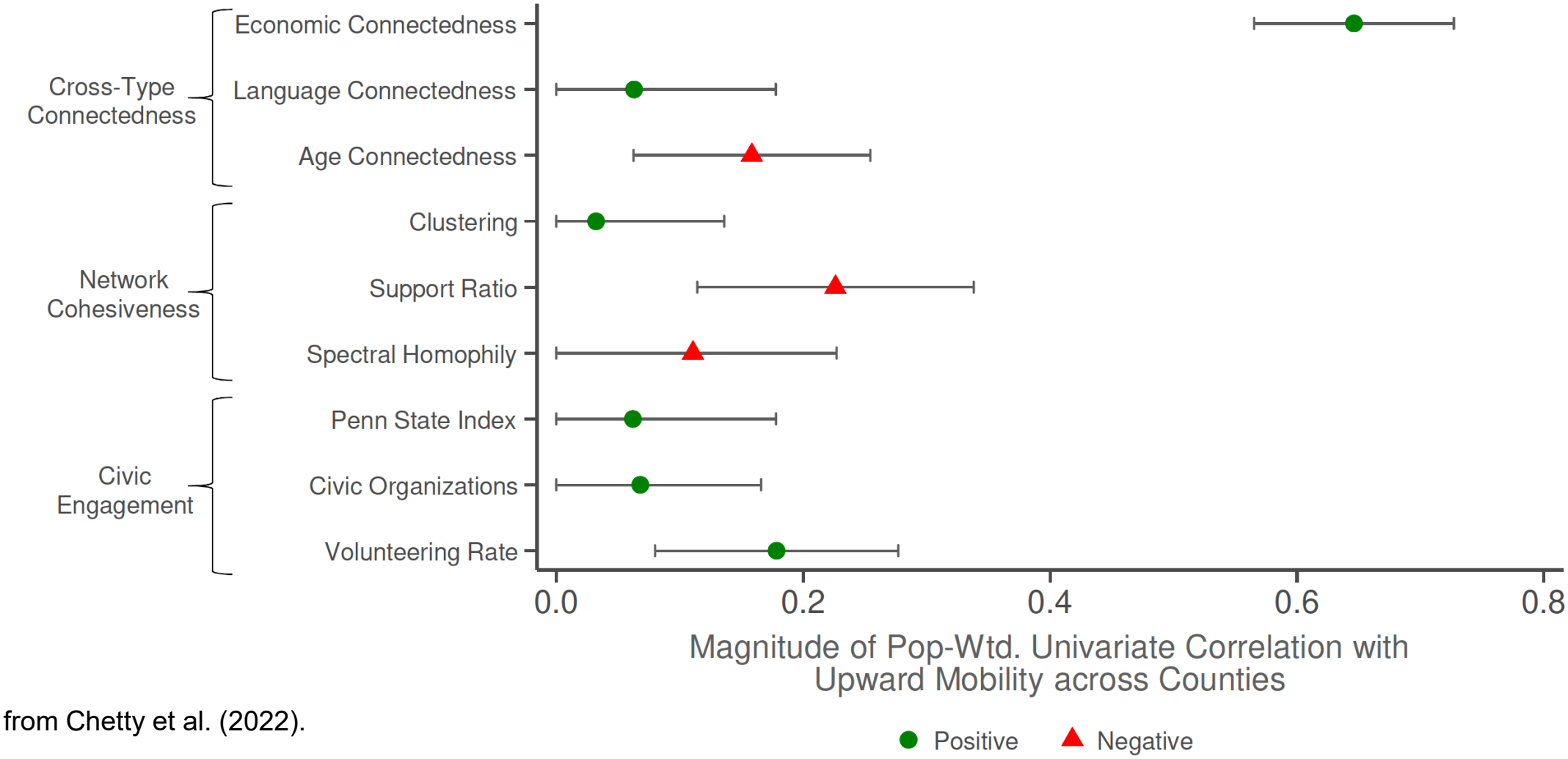


Taken from Chetty et al. (2022).

Note: see the Social Capital Atlas (www.socialcapital.org) for an interactive version of this map and downloadable data

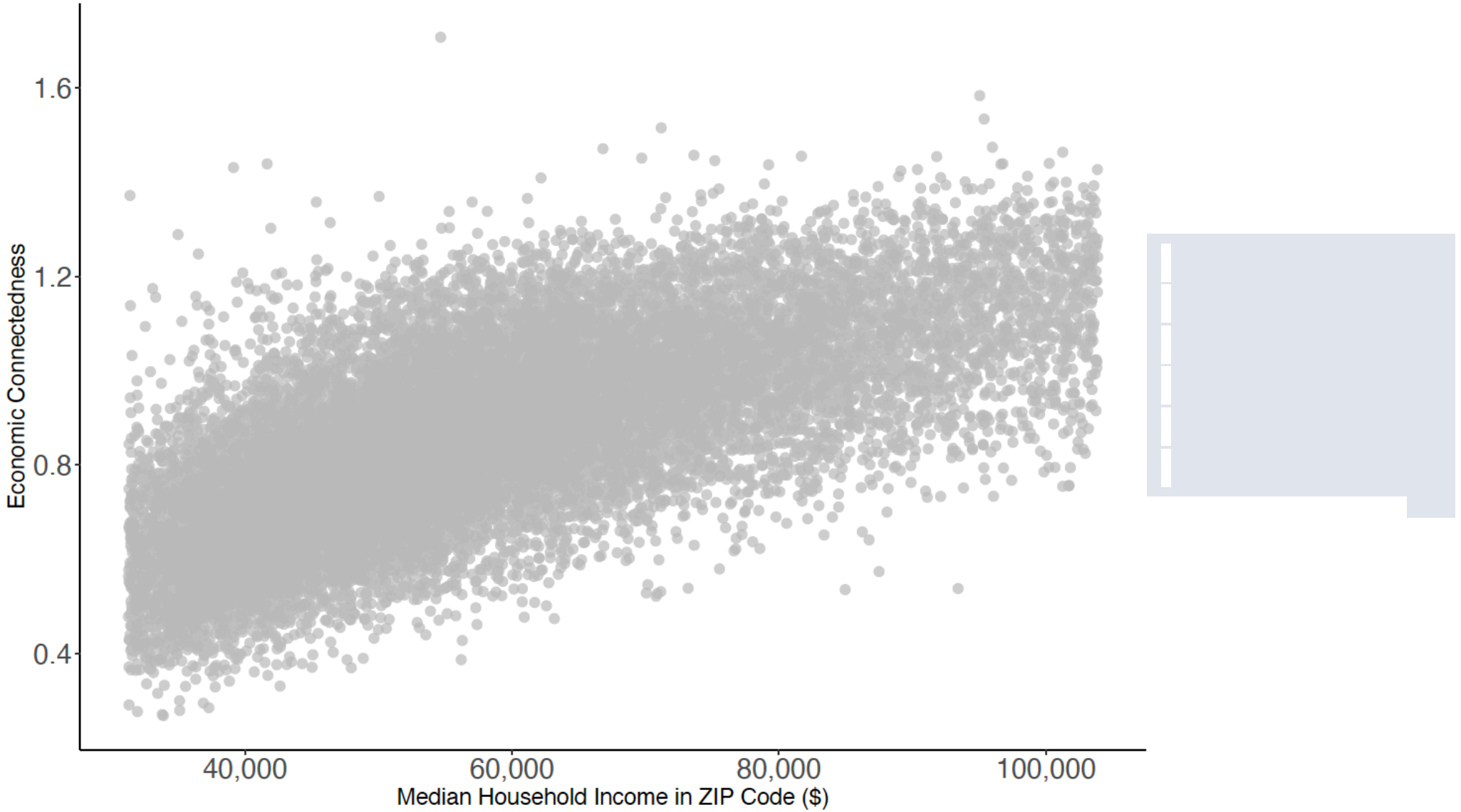
Correlations between Upward Mobility and Measures of Social Capital

County-level Univariate Correlations



Taken from Chetty et al. (2022).

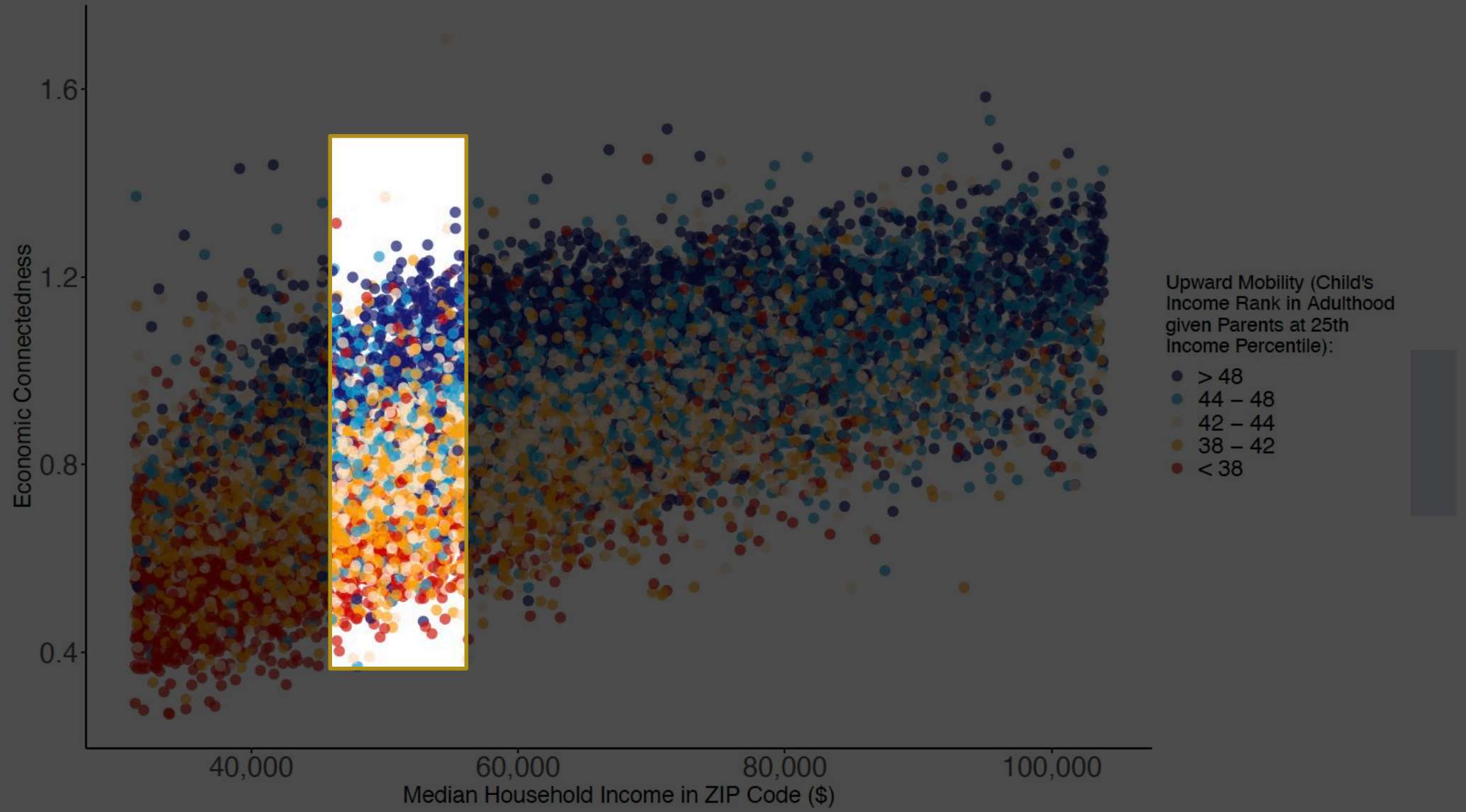
Economic Connectedness vs. Household Median Income, by ZIP Code



Taken from Chetty et al. (2022).

Economic Connectedness vs. Household Median Income, by ZIP Code

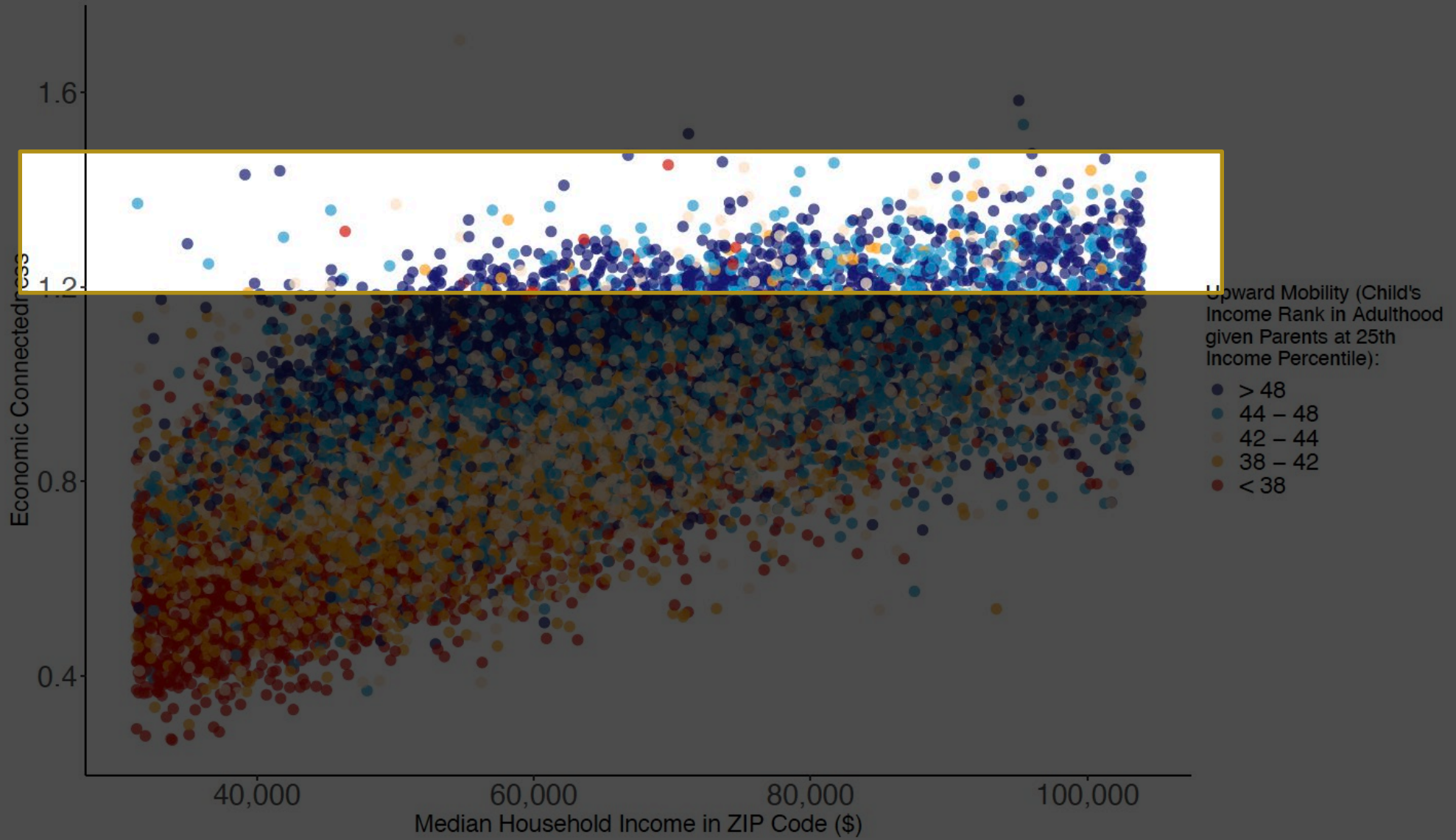
Colored by Rate of Upward Mobility



Taken from Chetty et al. (2022).

Economic Connectedness vs. Household Median Income, by ZIP Code

Colored by Rate of Upward Mobility



Taken from Chetty et al. (2022).

Upward Mobility vs. Economic Connectedness, Inequality, and Segregation

OLS Regression Estimates, Across Counties and ZIP codes

Dependent Variable:	Upward Mobility (Mean Income Rank at Age 35 for Children with Parents at 25th Percentile)	
	Across Counties	
	(1)	(2)
Income Inequality (Gini coefficient)	–0.449***	-0.103
	(-0.084)	(-0.091)
Share Black		
Economic Connectedness		0.577***
		(0.063)
Observations	2,741	2,741
R-squared	0.207	0.424

Connectedness explains the link between inequality and mobility (Great Gatsby Curve) [Corak 2013, Krueger 2016]

Taken from Chetty et al. (2022).

Upward Mobility vs. Economic Connectedness, Inequality, and Segregation

OLS Regression Estimates, Across Counties and ZIP codes

Dependent Variable:	Upward Mobility (Mean Income Rank at Age 35 for Children with Parents at 25th Percentile)		Upward Mobility for Black Individuals		Upward Mobility for White Individuals	
	Across Counties		Across ZIP Codes			
	(1)	(2)	(3)	(4)	(5)	(6)
Income Inequality (Gini coefficient)	–0.449***	-0.103				
	(-0.084)	(-0.091)				
Share Black			-0.204***	-0.014	-0.250***	0.035*
			(0.057)	(0.071)	(0.018)	(0.018)
Economic Connectedness		0.577***		0.468***		0.631***
		(0.063)		(0.083)		(0.027)
Observations	2,741	2,741	11,147	11,147	24,020	24,020
R-squared	0.207	0.424	0.042	0.224	0.063	0.380

Cutler and Glaeser (1997): “segregation is extremely harmful for blacks, but we do not have an exact understanding of why this is true.”

Lack of connectedness provides a (statistical) explanation

Taken from Chetty et al. (2022).

Cross-Type Connectedness Index

High and Low Socio-Economic Status Connectedness Indices

Partition induced by the labels (low and high SES/below and above median)

$$\mathcal{A} = \{i \in \mathcal{V}; l_i = a\}$$
$$\mathcal{B} = \{i \in \mathcal{V}; l_i = b\}$$

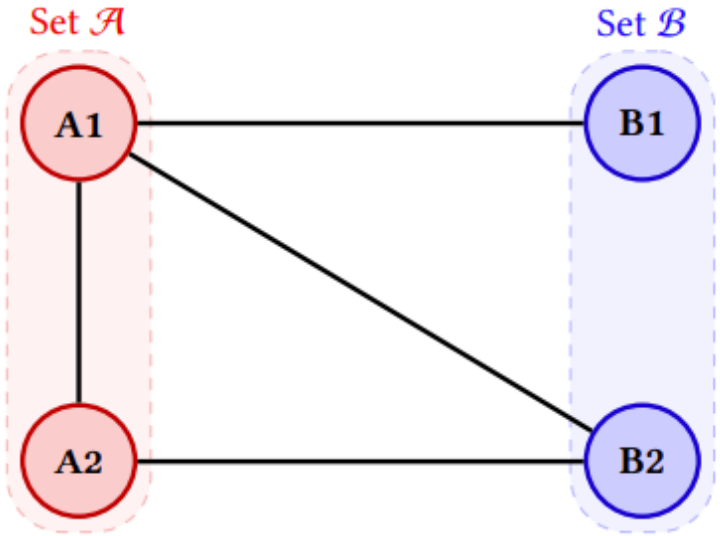
Agent 3 follows their action regardless of their private signal

$$C^{\mathcal{A}\to\mathcal{B}} = \frac{1}{|\mathcal{A}|}\sum_{i\in\mathcal{A}} \frac{\deg(i;\mathcal{B})}{\deg(i)}$$

Average over nodes in a state, county, ZIP code, school, etc.

Node A1
Friends in $\mathcal{B}$: 2 (B1, B2)
Total Friends: 3 (A2, B1, B2)
$\rho_{A1} = \frac{2}{3}$

Node A2
Friends in $\mathcal{B}$: 1 (B2)
Total Friends: 2 (A1, B2)
$\rho_{A2} = \frac{1}{2}$



$$C^{\mathcal{A}\to\mathcal{B}} = \frac{\frac{2}{3}+\frac{1}{2}}{2} = \frac{7}{12}$$

Privacy & connected indices

Releasing network connectedness indices with label and edge differential privacy

Differential Privacy

Differential privacy (DP): A randomized mechanism $\mathcal{M}$ is $\epsilon - DP$ if for all adjacent datasets $D \sim D'$ and all possible sets of outputs S, we have:

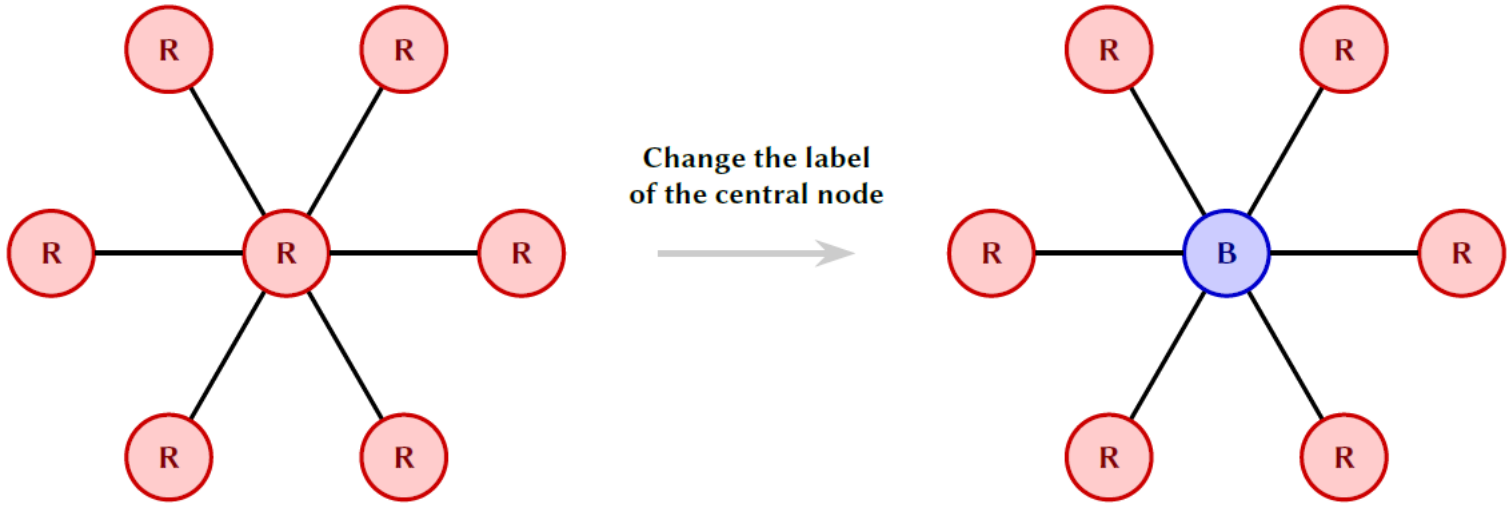
$$Pr(\mathcal{M}(D) \in S) < e^{\epsilon} Pr(\mathcal{M}(D') = S).$$

Balancing accurate decision making and individual privacy is important for ethical and effective network data release.

Edge-Adjacent Labeled Networks
Blocki et al. (2013)

n nodes in the vertex set $\mathcal{V}$, edge set $\mathcal{E}$, and with labels $L = (l_i)_{i \in \mathcal{V}}; l_i \in \{a, b\}$.

Two labeled networks are edge-adjacent if they **differ in at most one edge and at most one node's label**



A composition theorem for edge and label DP

Let $(\mathcal{V}, \mathcal{E}, L)$ and $(\mathcal{V}, \mathcal{E}', L')$ be *edge-adjacent labeled networks*. Let $\mathcal{M}_1: (\mathcal{V}, \mathcal{E}, L) \to (\mathcal{V}, \mathcal{E}, \hat{L})$ be $(\varepsilon_\ell, \delta_\ell)$-DP with respect to changing a single node attribute, and for every fixed $\hat{L}$ let $\mathcal{M}_2: (\cdot, \hat{L}): (\mathcal{V}, \mathcal{E}) \to \mathcal{R}$ be $(\varepsilon_e, \delta_e)$-DP with respect to changing a single edge. Define the composed mechanism \script $\mathcal{M} := \mathcal{M}_2 \circ \mathcal{M}_1$. Then is $(\varepsilon_\ell + \varepsilon_e, \delta_\ell + \delta_e)$ *edge-adjacent differentially private*.

- Labels are protected using a randomized response (or Laplace if continuous) mechanism
  - $\mathcal{M}_1: (\mathcal{V}, \mathcal{E}, L) \to (\mathcal{V}, \mathcal{E}, \hat{L})$; $(\boldsymbol{\varepsilon_\ell}, \boldsymbol{\delta_\ell})$**-label DP**
- Connectedness indices are constructed from noisy labels and debiased
- A final step of Laplace noise protects edges
  - $\mathcal{M}_2: (\cdot, \hat{L}): (\mathcal{V}, \mathcal{E}) \to \mathcal{R}$; $(\boldsymbol{\varepsilon_e}, \boldsymbol{\delta_e})$**-edge DP**
- $\mathcal{M}_2 \circ \mathcal{M}_1$ is $(\boldsymbol{\varepsilon_\ell + \varepsilon_e}, \boldsymbol{\delta_\ell + \delta_e})$ ***edge-adjacent differentially private*****.**

Local Label Randomization (Node-Level Privacy)

Apply randomized response to node labels with flip probability

$$p = \frac{1}{1 + e^{\varepsilon_\ell}}$$

Ensures ε_ℓ**-node differential privacy**

Introduces controlled bias that will be analytically debiased later

Algorithm 1 Differentially Private Cross-Type Connectedness Index

Require: Network $(\mathcal{V}, \mathcal{E})$ (with weights $e_{ij} \geq 0$), node labels $l_i \in \{a, b\}$, privacy parameter $\varepsilon_l, \varepsilon_e$,

Ensure: Differentially private connectedness index $\widehat{C}_{\mathrm{DP}}^{\mathcal{A}\to\mathcal{B}}$

$p \leftarrow \frac{1}{1 + e^{\varepsilon_l}}$

for each node $i \in \mathcal{V}$ **do** ▷ Randomized response on labels

$$\hat{l}_i \leftarrow \begin{cases} \text{the opposite value in } \{a, b\}, & \text{with probability } p, \\ l_i, & \text{with probability } 1 - p \end{cases}$$

end for

MVUE for individual connectedness

For each node i, define the true individual connectedness

$$\rho_i := \sum_{j \in \mathcal{V}} \alpha_{ij}\, \mathbf{1}\{l_j = b\}, \qquad \alpha_{ij} := \frac{e_{ij}}{d_i}, \quad d_i = \sum_{j \in \mathcal{V}} e_{ij},$$

and consider the observed proxy after randomized response

$$\hat{\rho}_i := \sum_{j \in \mathcal{V}} \alpha_{ij}\, \mathbf{1}\{\hat{l}_j = b\}.$$

Then the unique minimum-variance unbiased estimator (MVUE) of ρ_i based only on the privatized labels $\{\hat{l}_j\}$ is

$$\tilde{\rho}_i := \frac{\hat{\rho}_i - p}{1 - 2p}.$$

Hàjek-type weighted ratio estimator

Define the debiased weight for membership in A:

$$w_i := \frac{\mathbf{1}\{\widehat{l}_i = a\} - p}{1 - 2p}.$$

$$\mathbb{E}[w_i] = \mathbf{1}\{l_i = a\}$$

Let

$$S_0 := \sum_{i \in V} w_i, \qquad S_1 := \sum_{i \in V} w_i \, \tilde{\rho}_i,$$

where $\tilde{\rho}_i$ is the individual connectedness MVUE.
The Hájek-type estimator is

$$\widetilde{C}^{A \to B} := \frac{S_1}{S_0},$$

whenever $S_0 \neq 0$ (we will show S_0 is bounded away from 0 with probability tending to 1 under mild conditions).

Edge-Sensitivity of the Hàjek-type ratio estimator

Assumption 1 (Non-vanishing fraction of type-A nodes). There exists $\pi_A \in (0, 1]$ such that $|A|/n \to \pi_A$ as $n \to \infty$.

Assumption 2 (Bounded influence / bounded column sums of normalized weights). There exists $\Gamma < \infty$ such that for all n and all $k \in V$,

$$\sum_{i \in V} \alpha_{ik} = \sum_{i \in V} \frac{e_{ik}}{d_i} \leq \Gamma.$$

Lemma (Edge-sensitivity of S_1). Fix the privatized labels $(\hat{l}_i)_{i \in V}$, so that $(w_i)_{i \in V}$ is fixed. Let E and E' differ by the addition or removal of a single edge (u, v). Assume $\hat{\rho}_i(E)$ is computed row-wise as $\hat{\rho}_i(E) = \sum_j \alpha_{ij}(E)\, \mathbf{1}\{\hat{l}_j = b\}$ with $\alpha_{ij}(E) = e_{ij}/d_i$ (so only the normalized weight rows for u and v change when adding or removing (u, v)). Then,

$$\left|S_1(E) - S_1(E')\right| \leq \frac{2(1-p)}{(1-2p)^2}.$$

Consistency of the privatized Hàjek estimator

Assumption 1 (Non-vanishing fraction of type-A nodes). There exists $\pi_A \in (0, 1]$ such that $|A|/n \to \pi_A$ as $n \to \infty$.

Assumption 2 (Bounded influence / bounded column sums of normalized weights). There exists $\Gamma < \infty$ such that for all n and all $k \in V$,

$$\sum_{i \in V} \alpha_{ik} = \sum_{i \in V} \frac{e_{ik}}{d_i} \leq \Gamma.$$

Theorem (Consistency of the debiased private estimator). Under Assumptions 1 and 2 with $p \in (0, \frac{1}{2})$, define the edge-DP release

$$\widehat{C}_{\mathrm{DP}}^{A \to B} := S_1/S_0 + Z_n, \qquad Z_n \sim \mathrm{Lap}\left(0, \frac{2(1-p)}{(1-2p)^2\, \varepsilon_e\, S_0}\right),$$

where Z_n is independent of the randomized-response perturbations. Then, for any fixed $\varepsilon_e > 0$,

$$\widehat{C}_{\mathrm{DP}}^{A \to B} \xrightarrow{p} C^{A \to B} \qquad \text{as } n \to \infty.$$

Debiased Weighted Connectedness Estimation

Compute individual cross-type exposure using noisy labels

Apply analytic debiasing of label indicators and neighbor-type proportions

Construct a **Hàjek-type weighted ratio estimator**

for each node $i \in \mathcal{V}$ **do** ▷ Compute debiased weights and individual connectedness

$d_i \leftarrow \sum_{j \in \mathcal{V}} e_{ij}$

$$\hat{\rho}_i \leftarrow \begin{cases} \frac{1}{d_i} \sum_{j \in \mathcal{V}} e_{ij} \, \mathbf{1}\{\hat{l}_j = b\}, & d_i > 0, \\ 0, & d_i = 0 \end{cases}$$

$$w_i \leftarrow \frac{\mathbf{1}\{\hat{l}_i = a\} - p}{1 - 2p}$$

$$\tilde{\rho}_i \leftarrow \frac{\hat{\rho}_i - p}{1 - 2p}$$

end for

$S_0 \leftarrow \sum_{i \in \mathcal{V}} w_i$

$S_1 \leftarrow \sum_{i \in \mathcal{V}} w_i \, \tilde{\rho}_i$

$\widehat{C}_{\mathrm{DP}}^{\mathcal{A} \to \mathcal{B}} \leftarrow \frac{S_1}{S_0}$ Hájek estimator

Edge-Level Privacy via Output Perturbation

Add calibrated Laplace noise to the final ratio estimator

Noise scale depends on:

- ε_e**-edge privacy budget**
- sensitivity of the estimator

Achieves overall $(\varepsilon_\ell + \varepsilon_e)$**-edge-adjacent differential privacy**

for each node $i \in \mathcal{V}$ **do** ▷ Compute debiased weights and individual connectedness

$d_i \leftarrow \sum_{j \in \mathcal{V}} e_{ij}$

$$\hat{\rho}_i \leftarrow \begin{cases} \frac{1}{d_i} \sum_{j \in \mathcal{V}} e_{ij}\, \mathbf{1}\{\hat{l}_j = b\}, & d_i > 0, \\ 0, & d_i = 0 \end{cases}$$

$$w_i \leftarrow \frac{\mathbf{1}\{\hat{l}_i = a\} - p}{1 - 2p}$$

$$\tilde{\rho}_i \leftarrow \frac{\hat{\rho}_i - p}{1 - 2p}$$

end for

$S_0 \leftarrow \sum_{i \in \mathcal{V}} w_i$

$S_1 \leftarrow \sum_{i \in \mathcal{V}} w_i\, \tilde{\rho}_i$

$$\widehat{C}_{\mathrm{DP}}^{\mathcal{A}\to\mathcal{B}} \leftarrow \frac{S_1}{S_0} + Z_n, \quad Z_n \sim \mathrm{Lap}\left(0, \frac{2(1-p)}{(1-2p)^2\, \varepsilon_e\, S_0}\right)$$ ▷ Privatized Hájek estimator

Algorithm 1 Differentially Private Cross-Type Connectedness Index

Require: Network $(\mathcal{V}, \mathcal{E})$ (with weights $e_{ij} \geq 0$), node labels $l_i \in \{a, b\}$, privacy parameter $\varepsilon_l, \varepsilon_e$,

Ensure: Differentially private connectedness index $\widehat{C}_{\mathrm{DP}}^{\mathcal{A}\to\mathcal{B}}$

$p \leftarrow \dfrac{1}{1+e^{\varepsilon_l}}$

for each node $i \in \mathcal{V}$ **do** ▷ Randomized response on labels

$\hat{l}_i \leftarrow \begin{cases} \text{the opposite value in } \{a, b\}, & \text{with probability } p, \\ l_i, & \text{with probability } 1-p \end{cases}$

end for

for each node $i \in \mathcal{V}$ **do** ▷ Compute debiased weights and individual connectedness

$d_i \leftarrow \sum_{j\in\mathcal{V}} e_{ij}$

$\hat{\rho}_i \leftarrow \begin{cases} \dfrac{1}{d_i} \sum_{j\in\mathcal{V}} e_{ij}\, \mathbf{1}\{\hat{l}_j = b\}, & d_i > 0, \\ 0, & d_i = 0 \end{cases}$

$w_i \leftarrow \dfrac{\mathbf{1}\{\hat{l}_i = a\} - p}{1-2p}$

$\tilde{\rho}_i \leftarrow \dfrac{\hat{\rho}_i - p}{1-2p}$

end for

$S_0 \leftarrow \sum_{i\in\mathcal{V}} w_i$

$S_1 \leftarrow \sum_{i\in\mathcal{V}} w_i\, \tilde{\rho}_i$

$\widehat{C}_{\mathrm{DP}}^{\mathcal{A}\to\mathcal{B}} \leftarrow \dfrac{S_1}{S_0} + Z_n, \quad Z_n \sim \mathrm{Lap}\left(0, \dfrac{2(1-p)}{(1-2p)^2\, \varepsilon_e\, S_0}\right)$ ▷ Privatized Hájek estimator

return $\widehat{C}_{\mathrm{DP}}^{\mathcal{A}\to\mathcal{B}}$

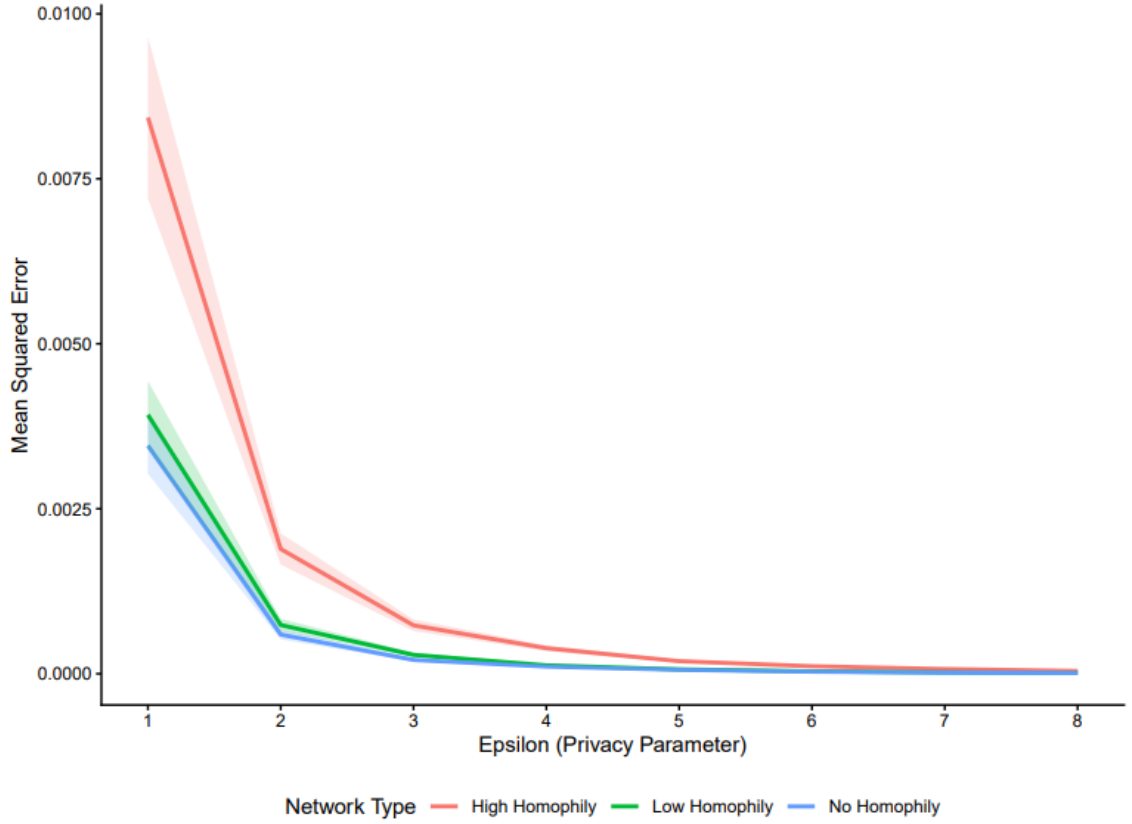
Application to Stochastic Block Models - epsilon

A network of 2,000 nodes.

In all cases, the unconditional edge probability is 0.04.

In the low homophily case, the probability of an edge is 0.06 if the two nodes are the same type, 0.02 otherwise.

In the high homophily case, the probability of an edge is 0.08 if the two nodes are of the same type, and 0 otherwise.



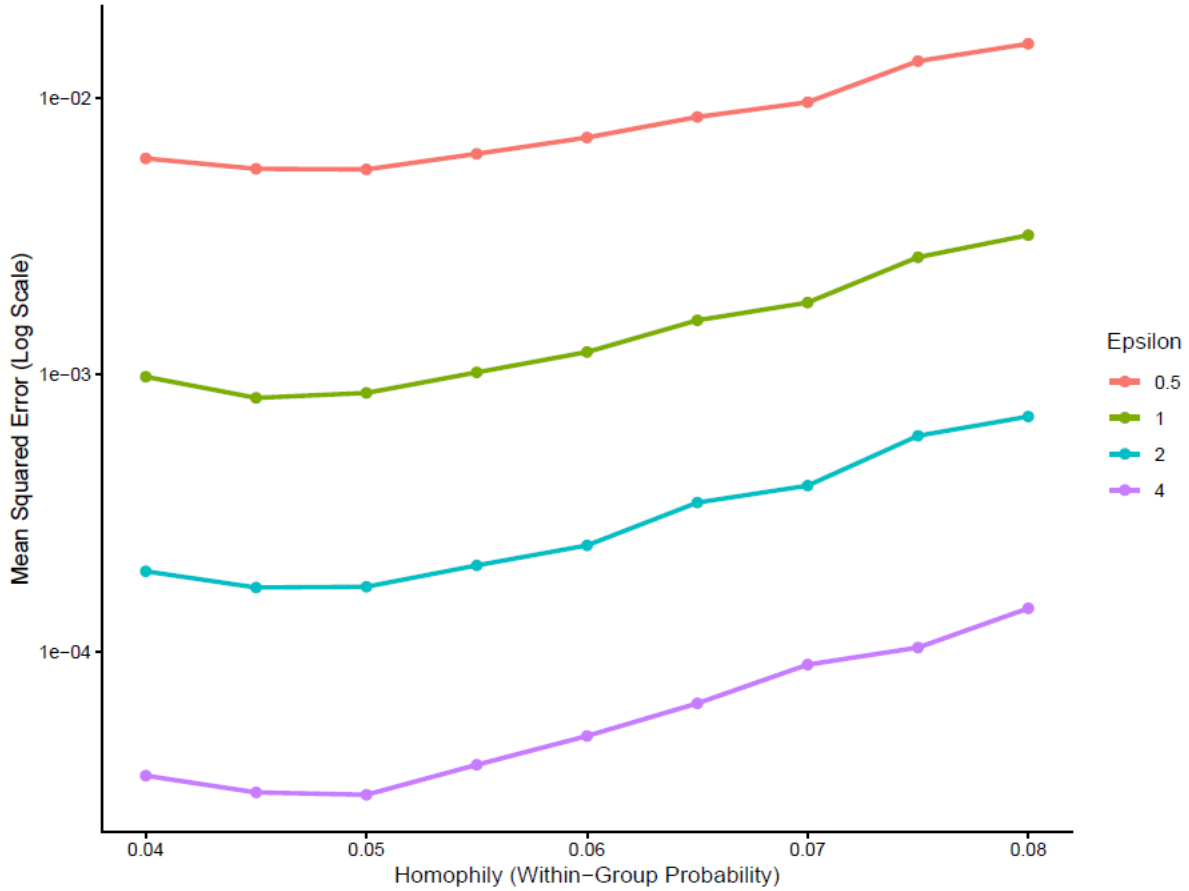
Application to Stochastic Block Models - homophily

A network of 2,000 nodes.

In all cases, the unconditional edge probability is 0.04.

In the low homophily case, the probability of an edge is 0.06 if the two nodes are the same type, 0.02 otherwise.

In the high homophily case, the probability of an edge is 0.08 if the two nodes are of the same type, and 0 otherwise.



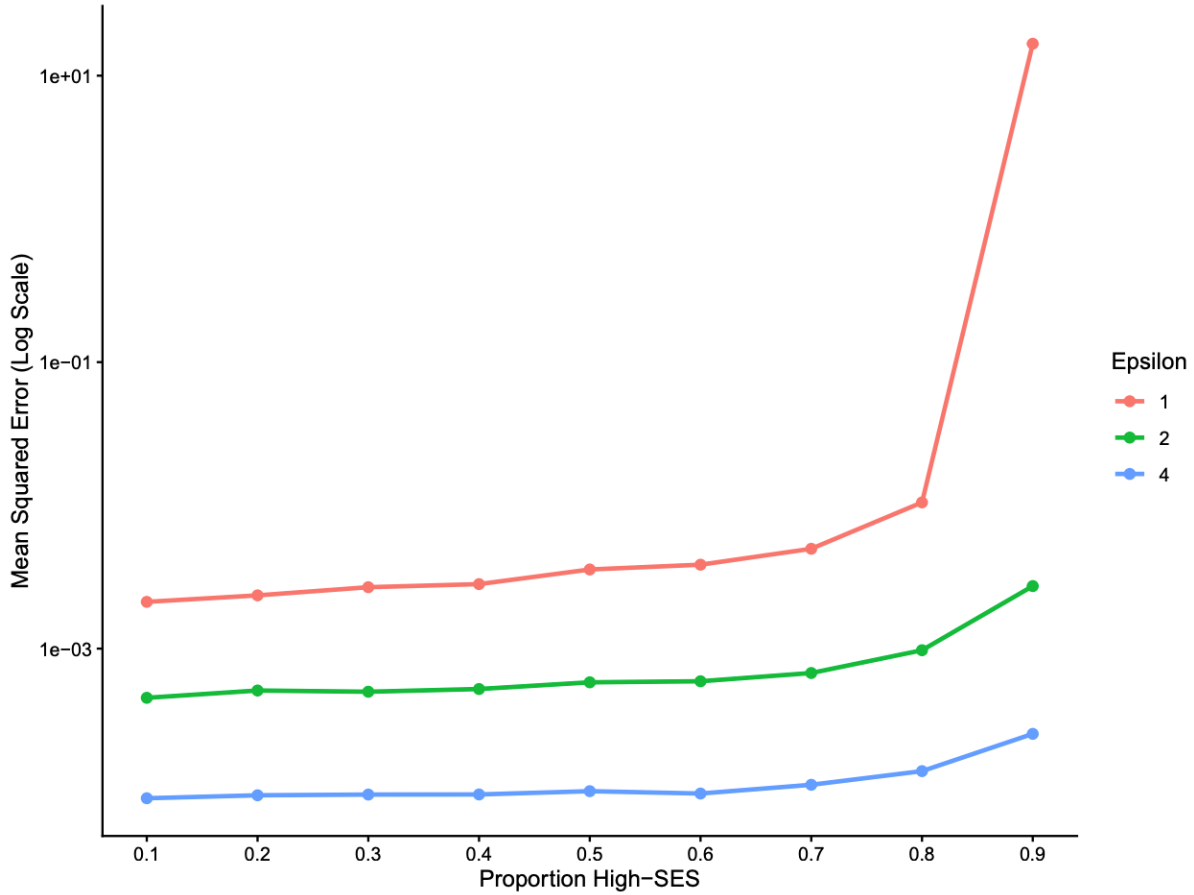
Application to Stochastic Block Models - imbalance

Graph of 2,000 nodes.

Link probability of 0.04 between every pair of nodes.

On the x-axis, we vary the proportion of nodes in the high-SES set.

Higher error with fewer nodes in the low-SES set since the noise is effectively averaged over fewer nodes.



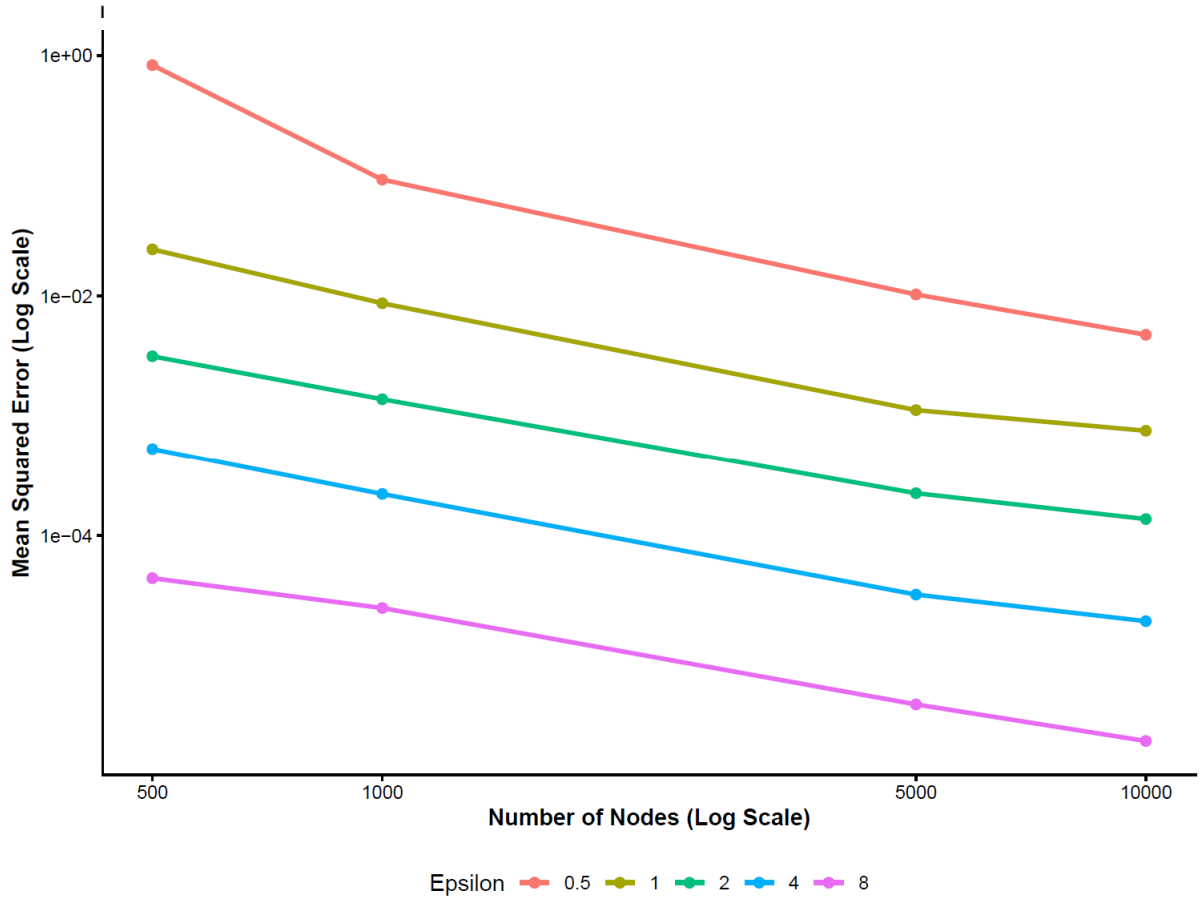
Application to Stochastic Block Models - size

Two-equally sized groups

Keeping the average degree fixed at 20.

On the x-axis, we vary the network size.

Error decreases as network size grows, consistent with the diminishing edge sensitivity and the averaging of label noise.

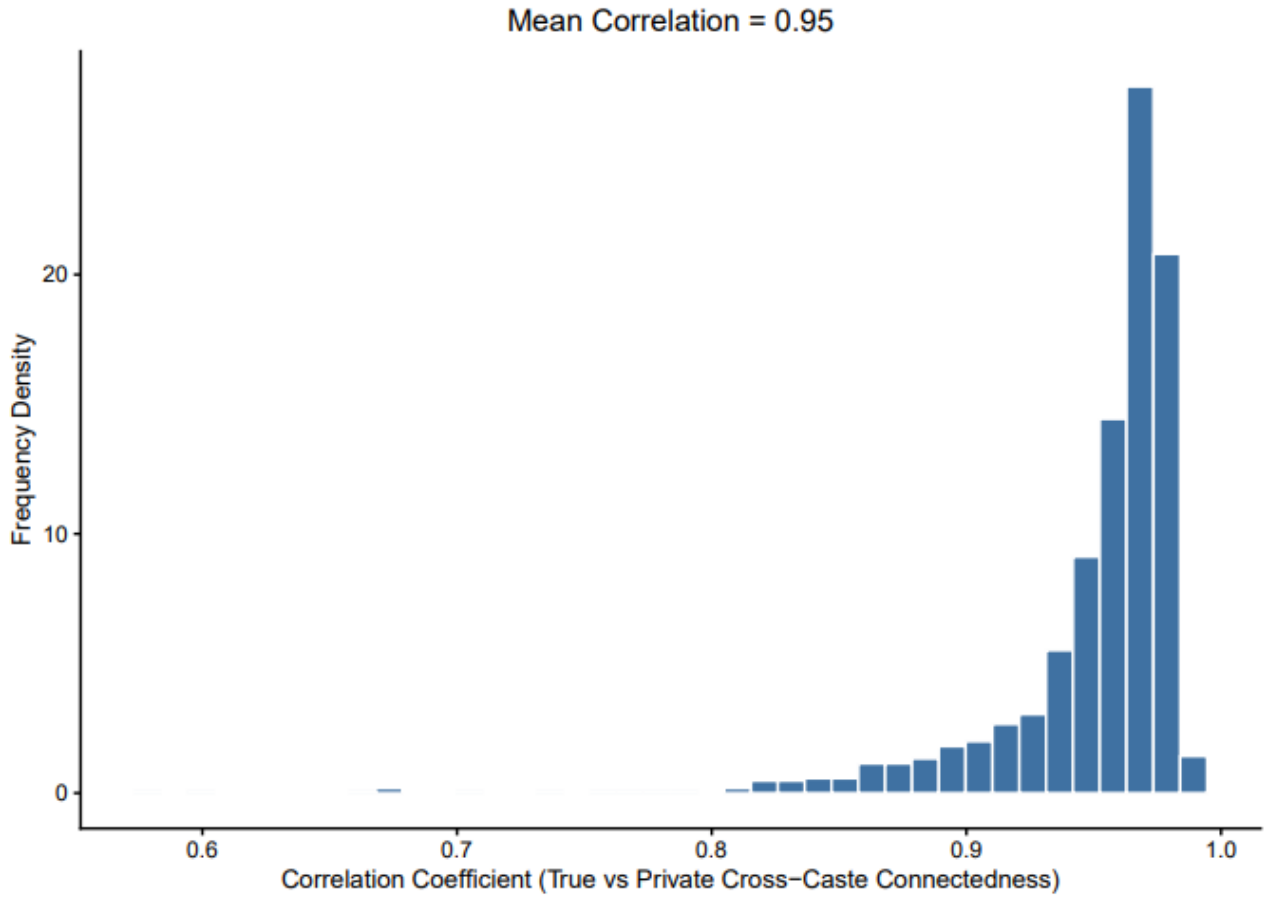


Application to Banerjee et al. (2013) (1/3)

Diffusion Networks in Rural India

Each village consists of around 200 nodes (households), with an average degree of 10.

- We bucket households into historically disadvantaged vs historically non-disadvantaged castes and compute the connectedness index from historically disadvantaged to historically non-disadvantaged castes.
- $\varepsilon_\ell = 4, \varepsilon_e = 4$

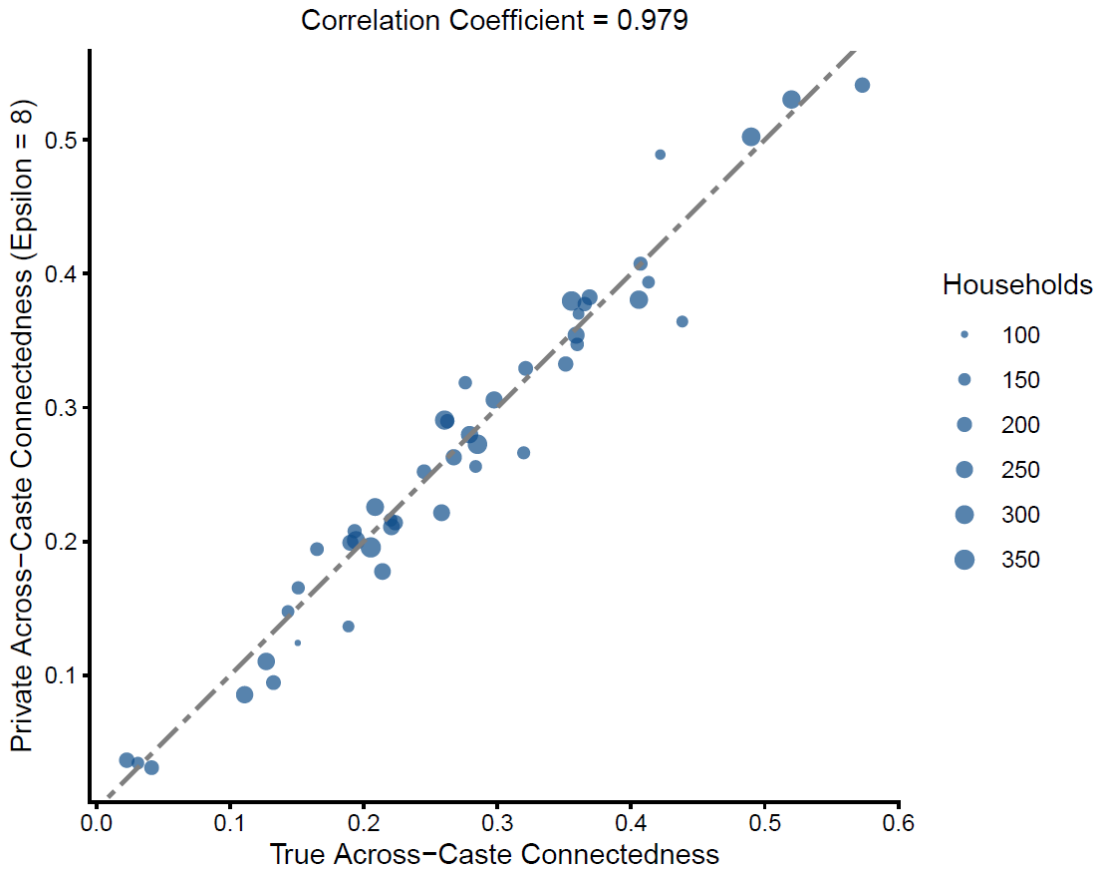


Application to Banerjee et al. (2013) (2/3)

Diffusion Networks in Rural India

Each village consists of around 200 nodes (households), with an average degree of 10.

- We bucket households into historically disadvantaged vs historically non-disadvantaged castes and compute the connectedness index from historically disadvantaged to historically non-disadvantaged castes.
- $\varepsilon_\ell = 4, \varepsilon_e = 4; \varepsilon = 8$

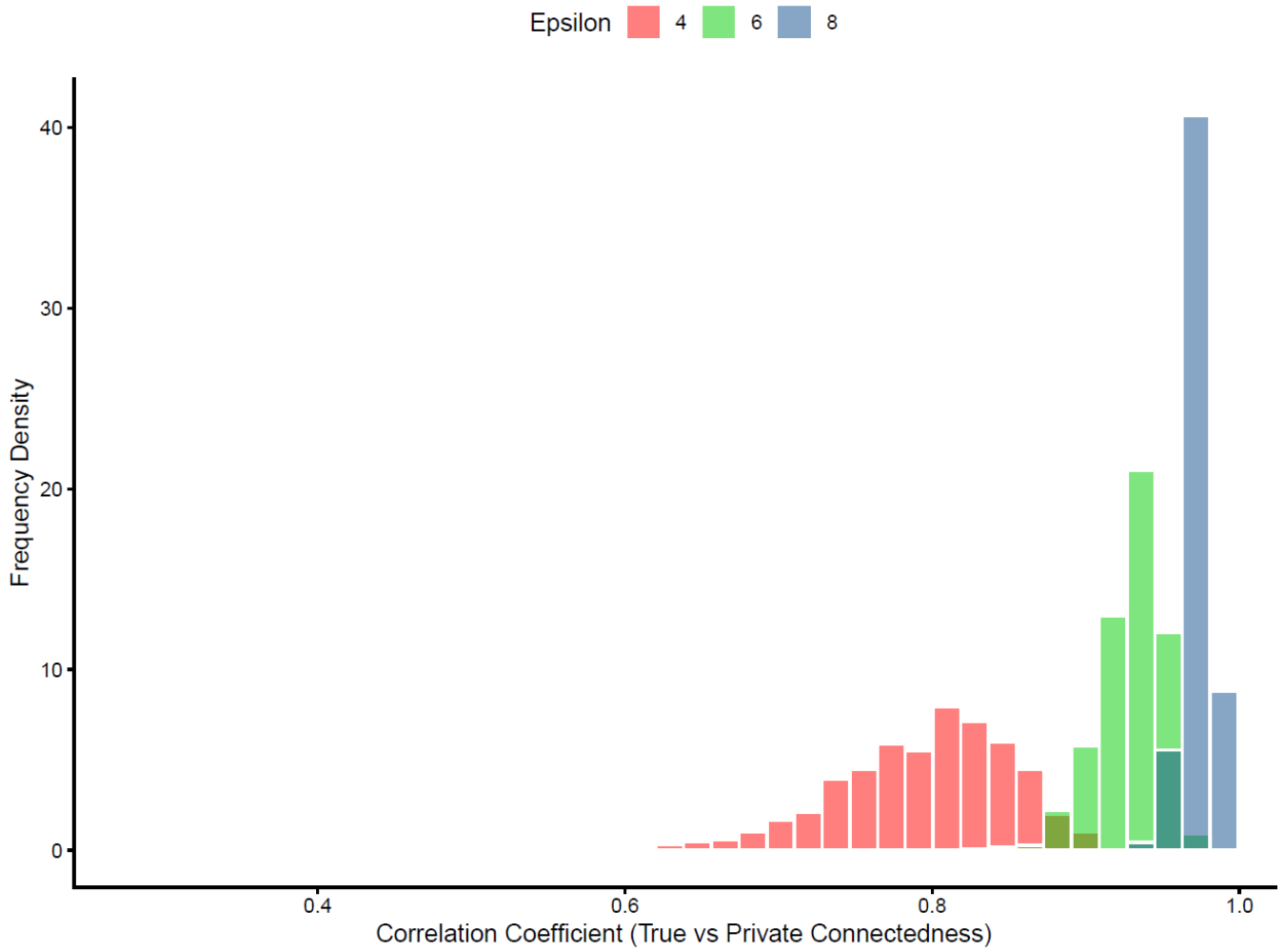


Application to Banerjee et al. (2013) (3/3)

Diffusion Networks in Rural India

Each village consists of around 200 nodes (households), with an average degree of 10.

- We bucket households into historically disadvantaged vs historically non-disadvantaged castes and compute the connectedness index from historically disadvantaged to historically non-disadvantaged castes.
- Varying $\varepsilon = 4, 6, 8$



Network connectedness indices provide valuable statistics for investigating economic outcomes, such as intergenerational mobility.

We can release these statics in a differentially private manner that protects individual labels and their network connections.

A decomposition theorem allows us to calibrate noise to label and edge privacy budgets separately and design unbiased and consistent estimators that satisfy edge-adjacent differential privacy for labeled networks.