A graph-prediction-based approach for debiasing underreported data

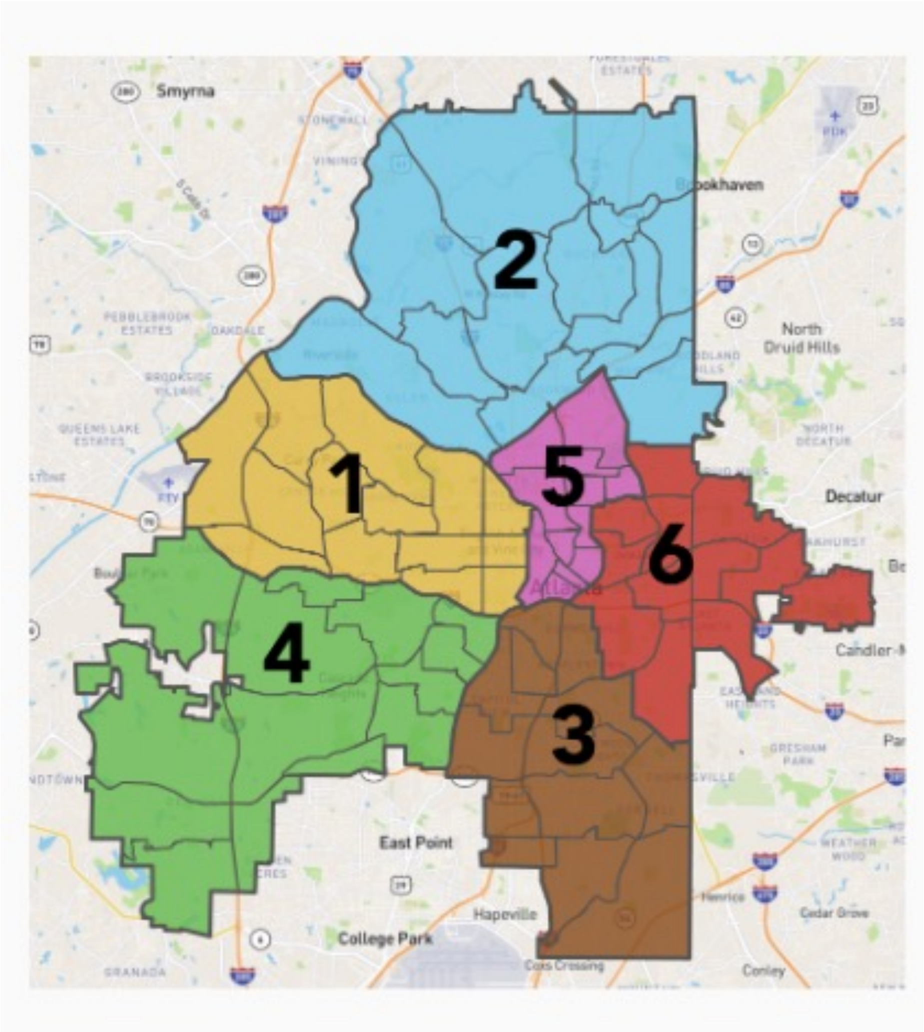

Hanyang Jiang, Yao Xie

Georgia Institute of Technology

H. Milton Stewart School of Industrial and Systems Engineering

April 16, 2024

Police operation



- Atlanta is the **6th largest** metropolitan area in the US with 6.08 million population, a booming city, and a fast-growing economy
- Atlanta Police Department (APD) is the major law enforcement agency in metro Atlanta, with **1,700 officers** and responses to over **3,000 calls** per day.
- Spatial regions are divided into **beats** such that the total "911" call **workload is balanced**
- Atlanta has 6 zones, 78 beats total

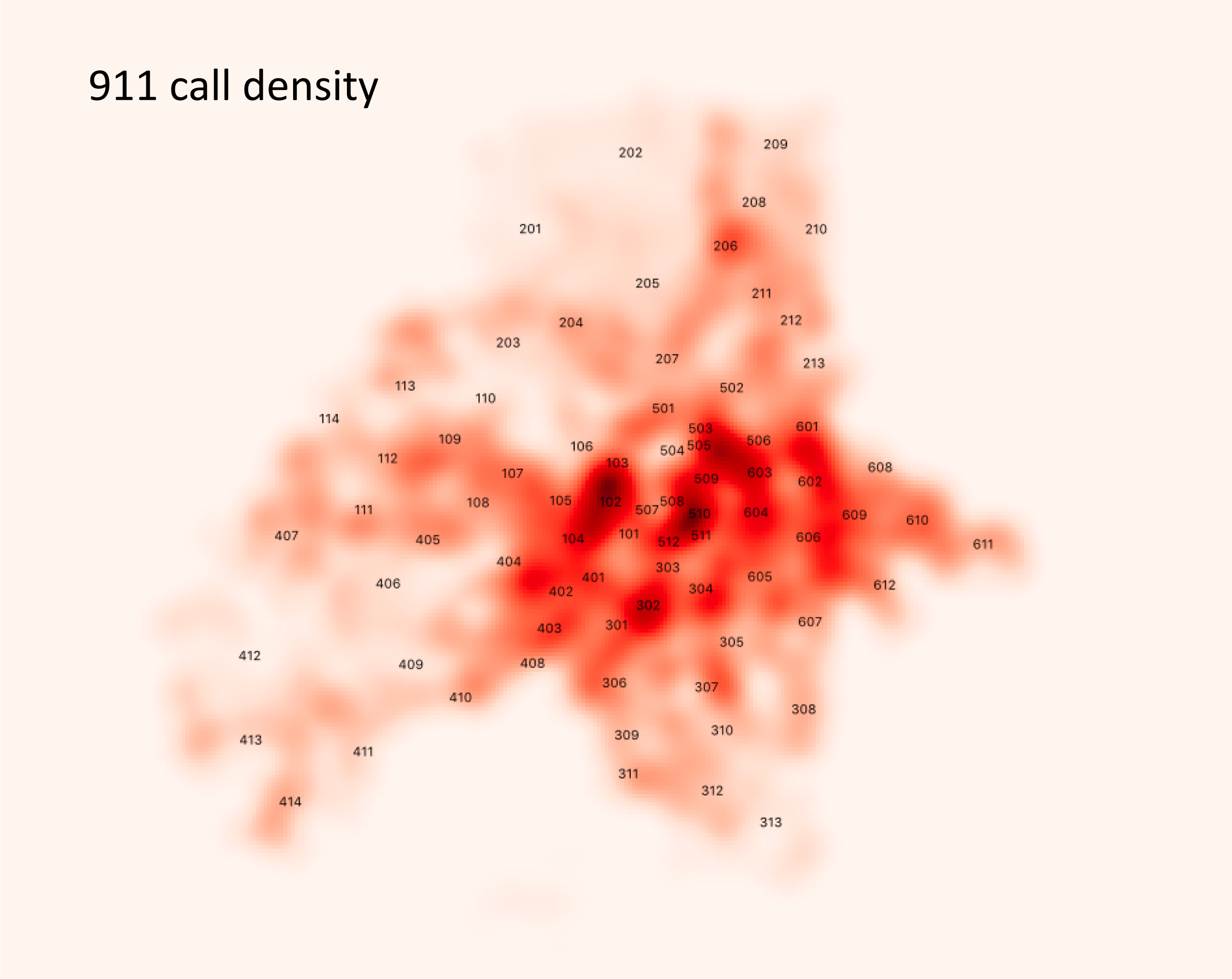
Bias in crime “counts”

- Under-reporting

- Over-policing

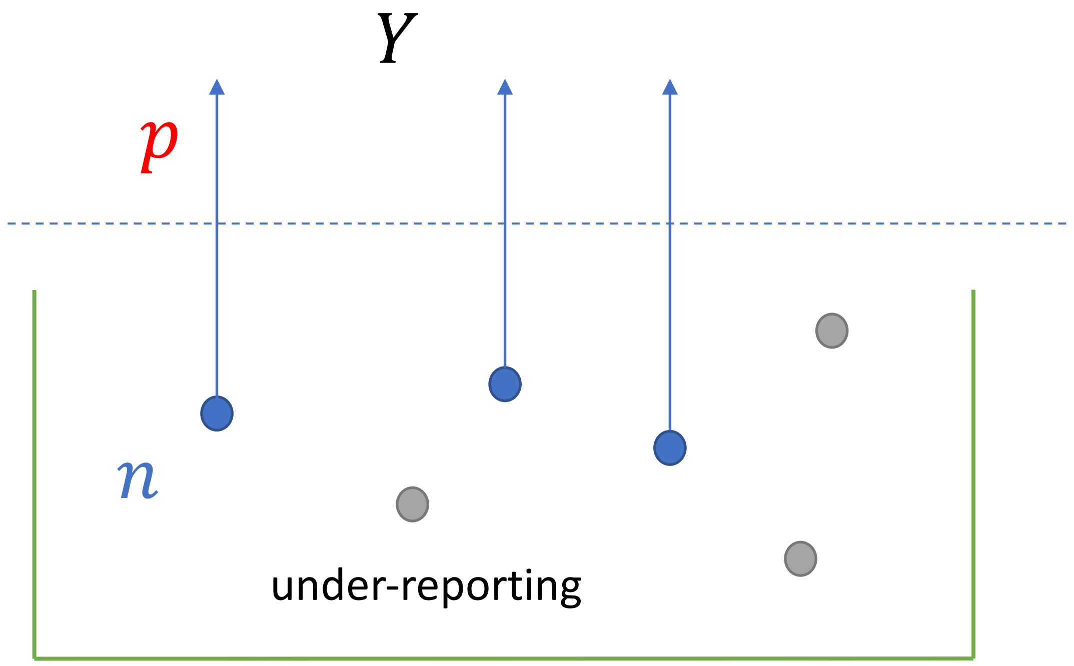
How to detect bias and correct bias?

- A common problem in **service systems**: ambulance, delivery trucks



Under-reporting: Binomial ($\boldsymbol{n},\boldsymbol{p}$) problem

- Observing $Y \sim$ binomial (n, p)
- We know on $\mathrm{E}[Y] = np$
- Identifiability issue: With one Y, we cannot estimate p and n at the same time



(Draper and Guttman 1971) (Draper, Guttman 1971) (DasGupta, Rubin 2005)

Identifiability issue with one-sample

- Observing $Y \sim$ binomial (n, p)
- We know

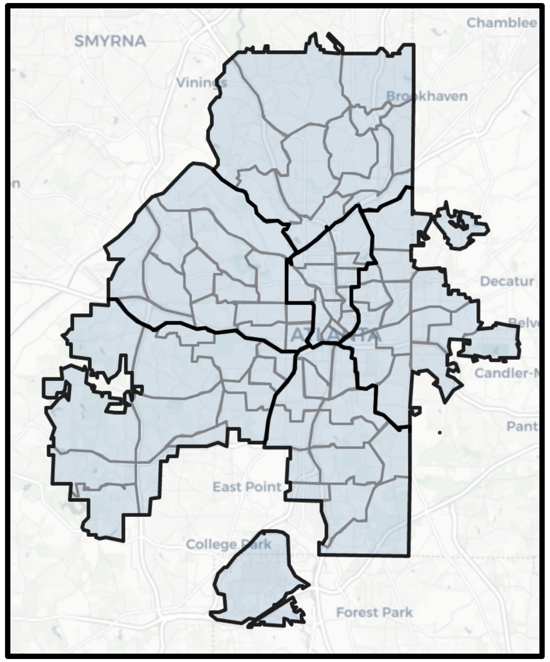
$$\mathrm{E}[Y] = np, \qquad \mathrm{VAR}(Y) = np(1 - p)$$

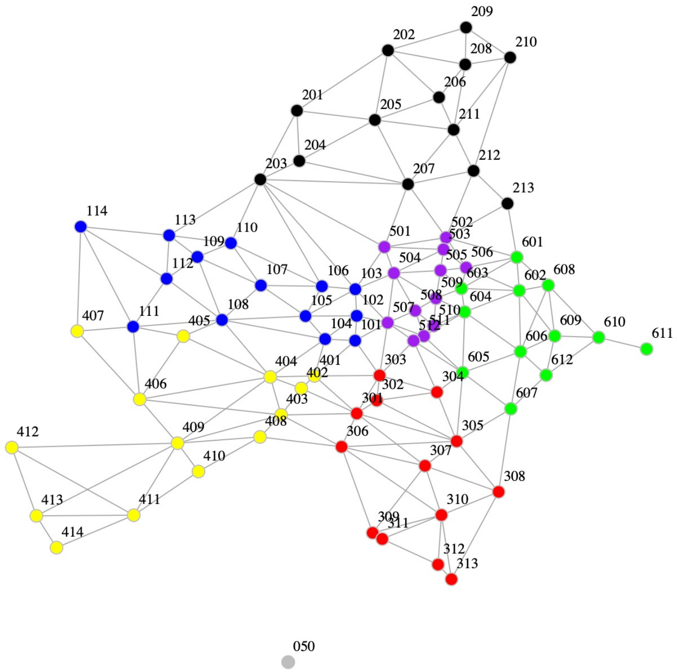
- In theory: $\mathrm{VAR}(Y)/\mathrm{E}[Y] = 1 - p$

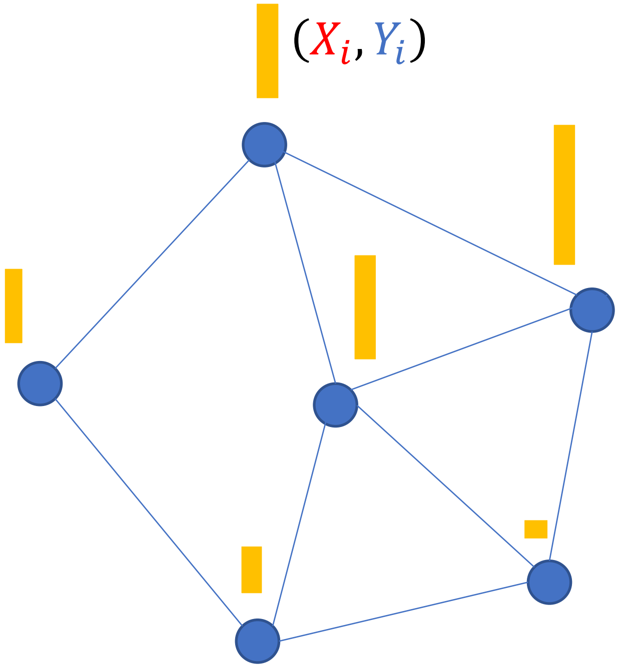
- In practice: we only observe **one sample** of X, we cannot estimate variance!

- Long history in statistics: Typical solution require prior distribution on n or p; however, can be subjective.

Graph structure







- Count Y_i for each node
- Covariate X_i for each node: Police data (911 call, GPS), census factors
- Adjacent nodes are “similar”

Graph Binomial $(\boldsymbol{n}, \boldsymbol{p})$ problem

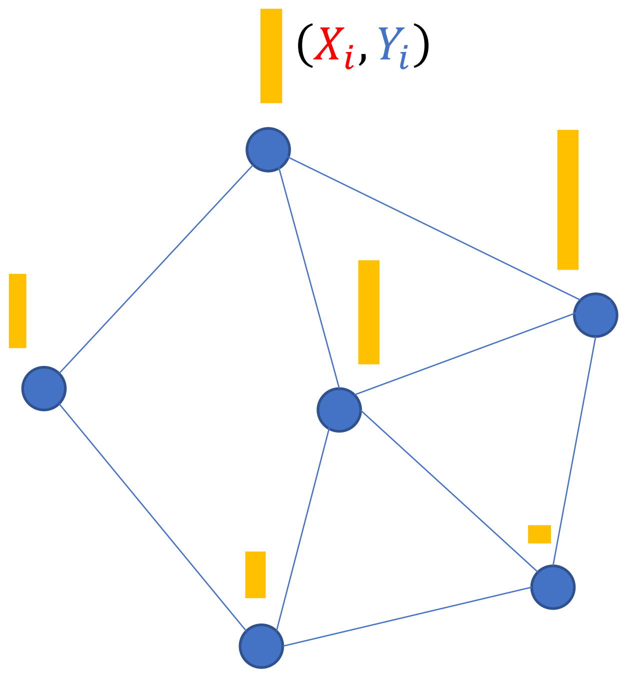
- Node response Y_i: count in each node
- Binomial model

$$Y_i \sim Binomial(n_i, p_i)$$

p_i: probability of discovery for node i

n_i: true count for node i

- Leverage **graph smoothness** on p_i over graph
- "Instrumental variable": True count n_i related to node feature x_i
- **Without using Bayesian priors**

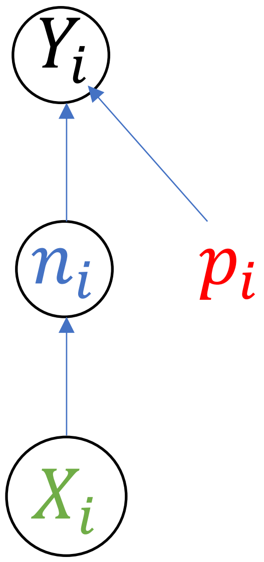


Graph prediction problem

- Data: $(X_i, Y_i), i \in V$
- Goal: estimate $(n_i, p_i), i \in V$

$$\tilde{Y}_i := \log Y_i \approx \underbrace{\log n_i}_{\tilde{n}_i} + \underbrace{\log p_i}_{\tilde{p}_i}$$

- Known: Graph with adjacency matrix A



Assumption 1: Graph smoothness on $\tilde{p}_i$
Assumption 2: Covariates: n_i is related to X_i

$$\tilde{\boldsymbol{n}} \approx \boldsymbol{X\beta}$$

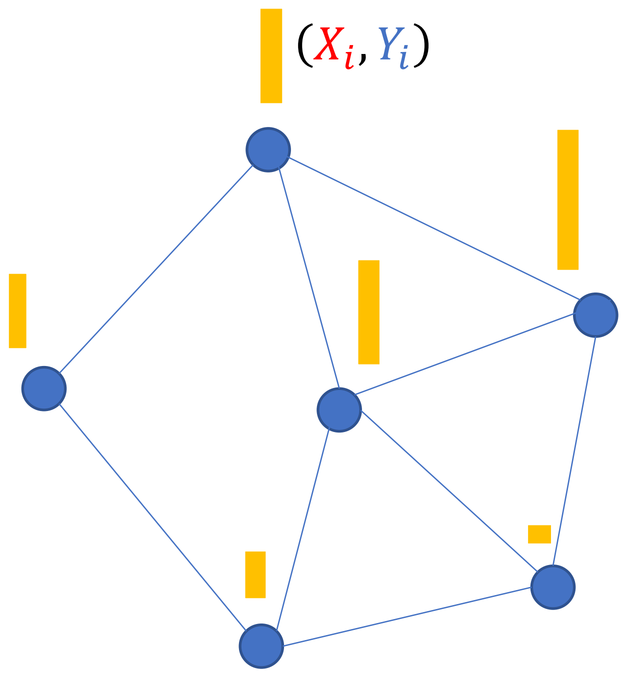
Graph smoothness

- Graph Laplacian contains topology information

$$L = D - A$$

- Graph smoothness:

$$z^T L z = \frac{1}{2} \sum_{(i,j)\in E} \left(z_i - z_j\right)^2$$



Convex reformulation

- Solve the following optimization problem

$$\min_{\widetilde{\boldsymbol{n}}, \widetilde{\boldsymbol{p}}, \boldsymbol{\beta}} ||\widetilde{\boldsymbol{Y}} - \widetilde{\boldsymbol{n}} - \widetilde{\boldsymbol{p}}||^2 + \lambda_1 \widetilde{\boldsymbol{p}}^T L \widetilde{\boldsymbol{p}} + \lambda_2 ||\widetilde{\boldsymbol{n}} - \boldsymbol{X}\boldsymbol{\beta}||_2^2$$

- Regression:

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \widetilde{\boldsymbol{n}}, \boldsymbol{H} = \underbrace{\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T}_{\text{covariates information}}$$

- Reduce to convex problem **(GRAUD)**

$$\min_{\widetilde{\boldsymbol{n}}, \widetilde{\boldsymbol{p}}} ||\widetilde{\boldsymbol{Y}} - \widetilde{\boldsymbol{n}} - \widetilde{\boldsymbol{p}}||^2 + \lambda_1 \widetilde{\boldsymbol{p}}^T \boldsymbol{L} \widetilde{\boldsymbol{p}} + \lambda_2 \widetilde{\boldsymbol{n}}^T \boldsymbol{H} \widetilde{\boldsymbol{n}}$$

- Can be solved efficiently using first-order method to global solution

Recovery guarantee

- Assumption 1: Ground truth signals satisfy

$$|\tilde{\boldsymbol{p}}_0^T \boldsymbol{L} \tilde{\boldsymbol{p}}_0| \leq \varepsilon_p, \left|\tilde{\boldsymbol{n}}_0^T \boldsymbol{H} \tilde{\boldsymbol{n}}_0\right| \leq \varepsilon_n \text{ are small}$$

- Assumption 2: $\mathrm{Null}(\boldsymbol{L}) \cap \mathrm{Null}(\boldsymbol{H}) = 0$
- Assumption 3: bounded observation "noise" $||\tilde{\boldsymbol{Y}} - \tilde{\boldsymbol{n}}_0 - \tilde{\boldsymbol{p}}_0||^2 \leq \varepsilon$

Theorem: Under Assumptions 1-3, solution $(\tilde{\boldsymbol{p}}^*, \tilde{\boldsymbol{n}}^*)$ to **GRAUD** satisfies

$$||\tilde{\boldsymbol{p}}^* - \tilde{\boldsymbol{p}}_0|| \leq c_1\varepsilon + \varepsilon_p, ||\tilde{\boldsymbol{n}}^* - \tilde{\boldsymbol{n}}_0|| \leq c_2\varepsilon + \varepsilon_n$$

Simulation

10 nodes
3 features for each node

$$p_i = 0.6 + 0.1N(0,1)$$

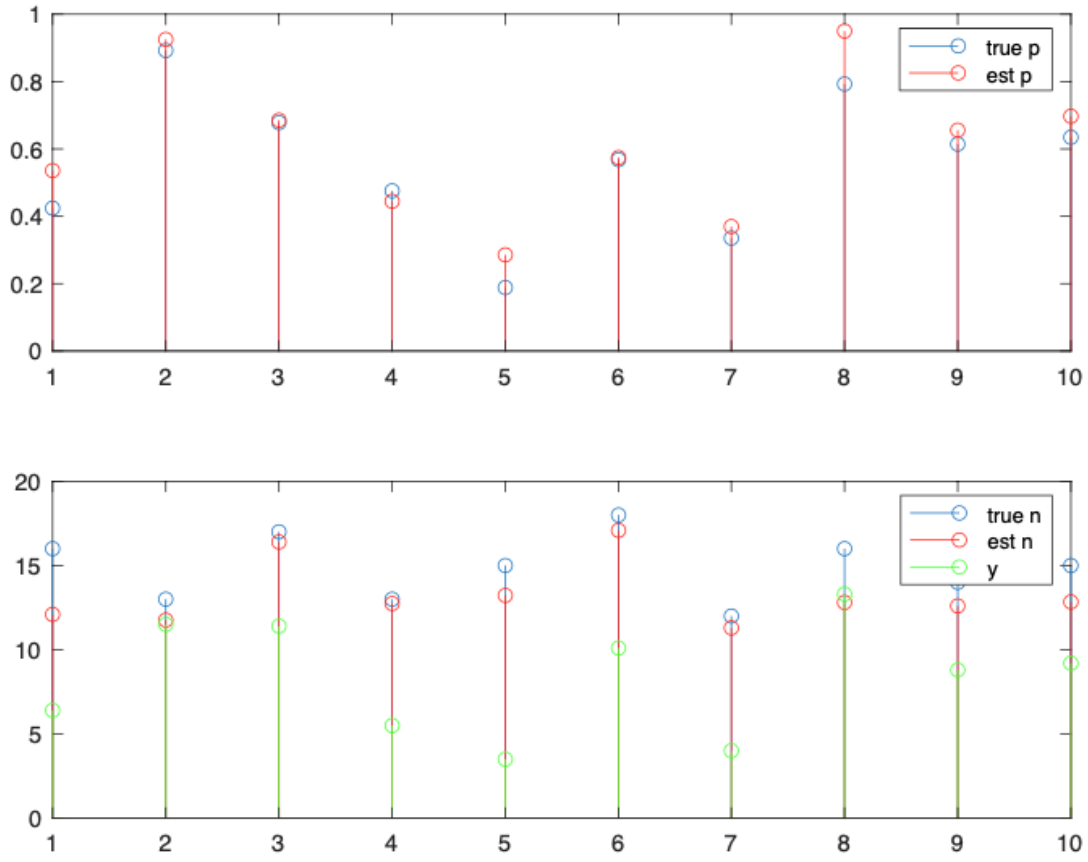


Figure 2: True p_i and n_i, and recovered $\hat{p}_i$, and $\hat{n}_i$ (from observation y).

Real-data

- Atlanta, data in 2019
- Lower staffing area has lower discovery rate

p_i

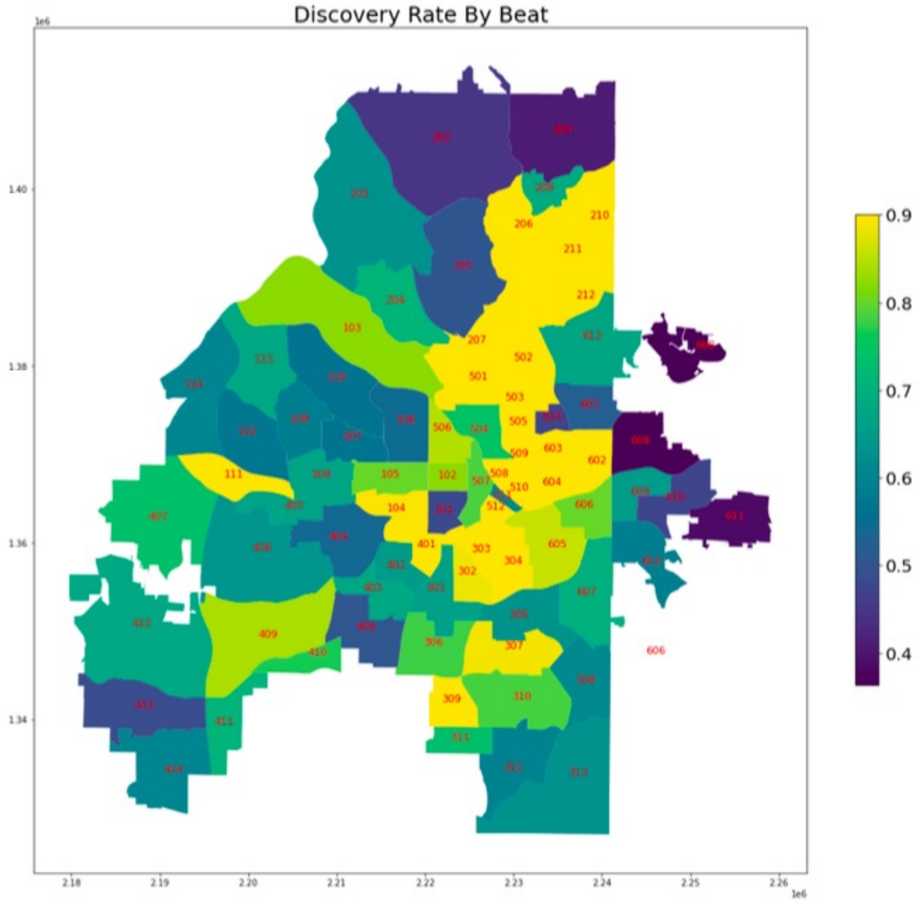
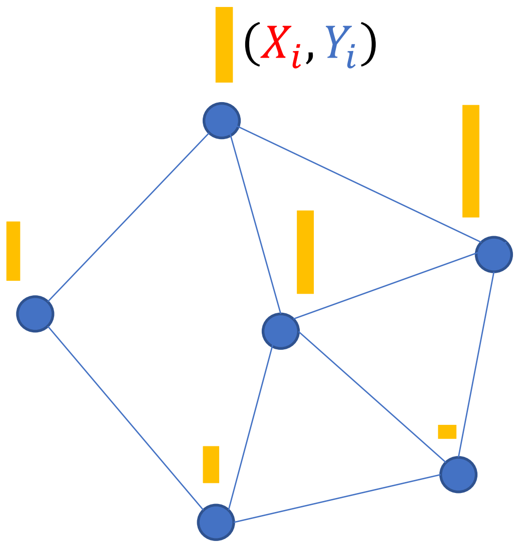


Fig. 2. The estimated p_i in each beat when initializing with the discovery probability of all 0.8.

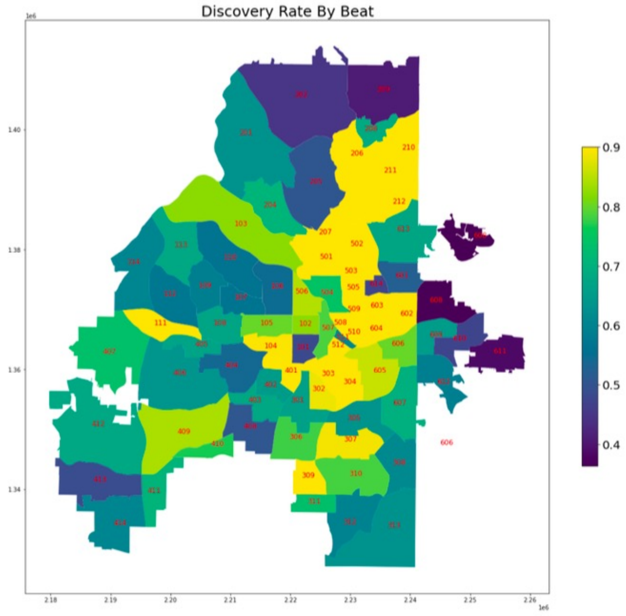
Summary



- A new graph prediction formulation for solving spatial binomial (n, p) problem to correct **undercount** bias in data

- Convex reformulation leads to efficient algorithm and recovery guarantee

- On-going: time-series observations



A graph-prediction-based approach for debiasing underreported data. Hanyang Jiang, Yao Xie. ICASSP 2024.