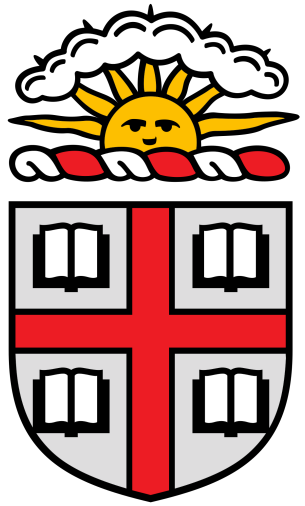
Quantifying and Reducing Bias in Maximum Likelihood Estimation of Structured Anomalies

Uthsav Chitra[1], Kimberly Ding[1], Jasper C.H. Lee[2], Benjamin J. Raphael[1]

[1] Princeton University, [2] Brown University

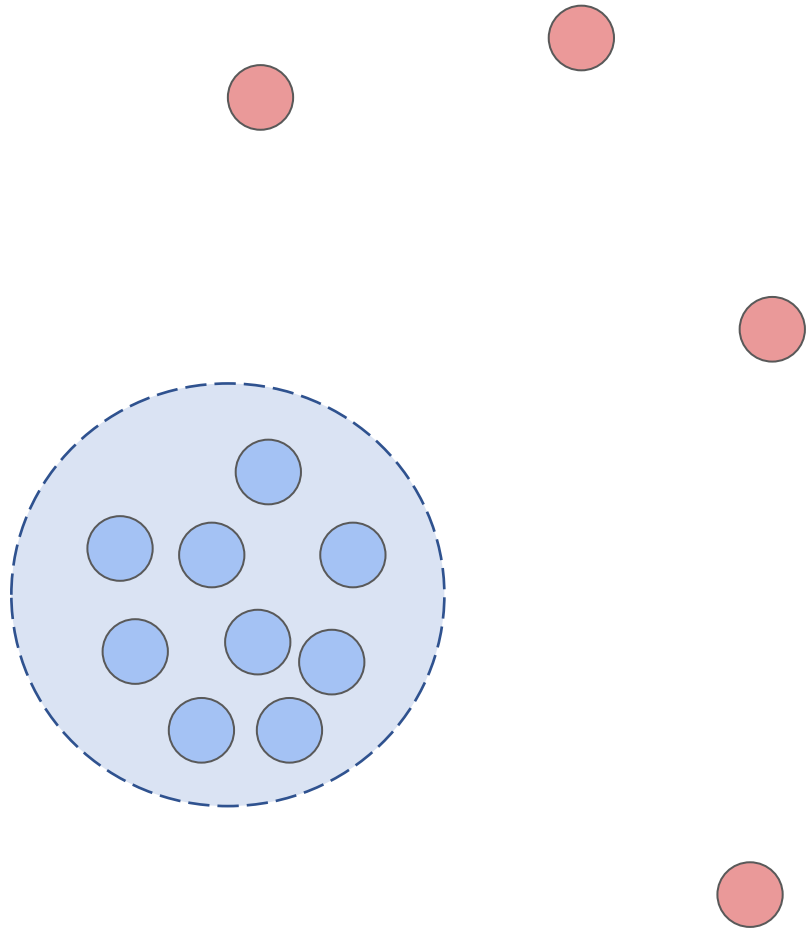


Anomaly Detection

The identification of rare, irregular, or otherwise aberrant patterns (i.e. *anomalies*) in data

Many applications in ML and statistics:

- Real-time system monitoring
  - identifying anomalies in real-time data (ML model data, sensors, …)
- Healthcare/biological data
  - identifying groups of patients with anomalous reactions to certain drugs
- Anomaly detection in graphs
  - identifying disease outbreak regions (e.g. COVID) or anomalous activity in social networks

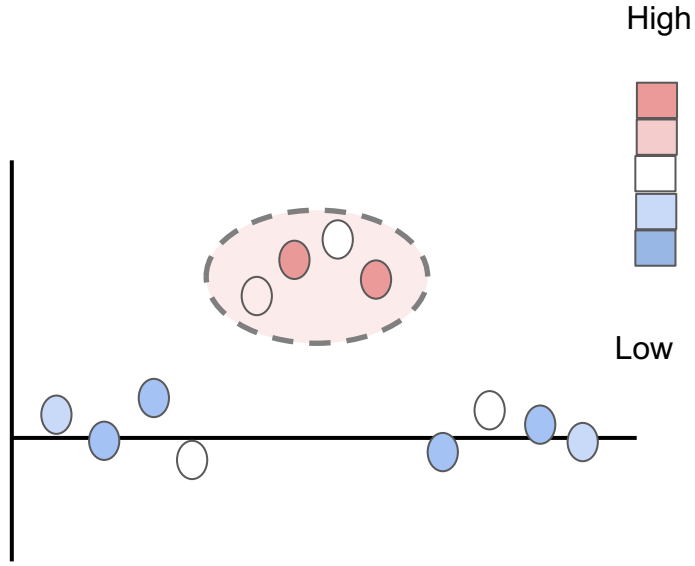


Structured Anomaly Detection

Depending on type of data, anomaly often has specific **structure**

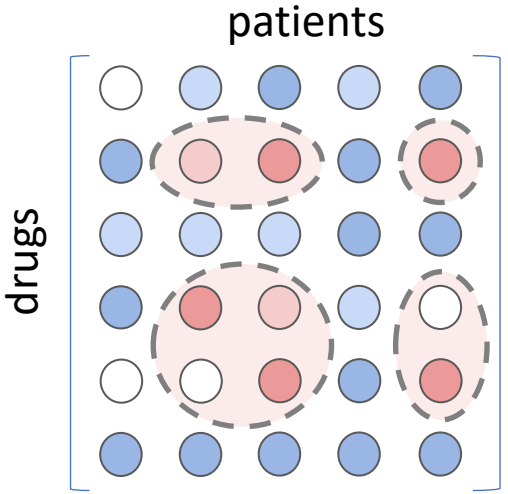
- Real-time system monitor
  - Identifying anomalies in real-time data

Anomalies are time intervals



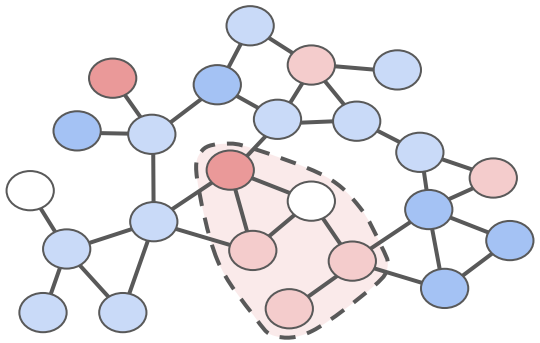
- Healthcare/biological data
  - Identifying drugs w/ anomalous reactions for specific groups of patients

Anomalies are submatrices



- Anomalies in graphs
  - Identifying disease outbreak hotspots or anomalous activity in social networks

Anomalies are connected subgraphs



Structured Anomaly Detection

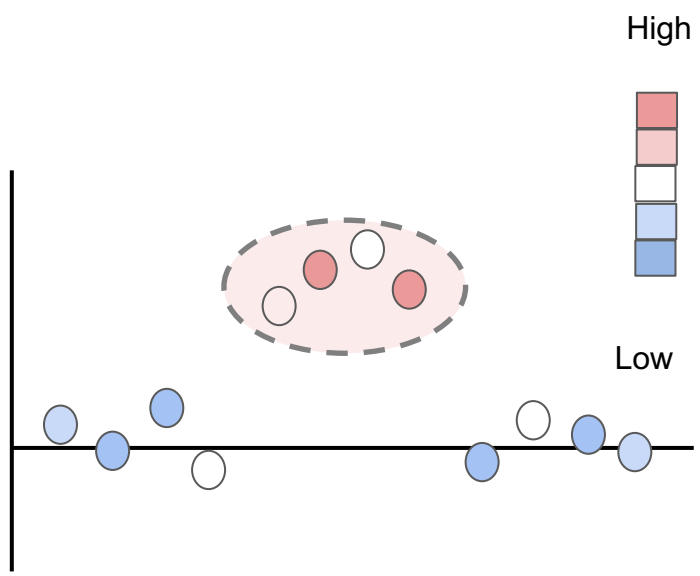
Depending on type of data, anomaly often has specific **structure** described by an **anomaly family** $\mathcal{S}$ or set of all possible anomalies

- Real-time system monitor
  - Identifying anomalies in real-time data

Anomalies are time intervals

Interval family

$\mathcal{S} = \mathcal{I}_n$

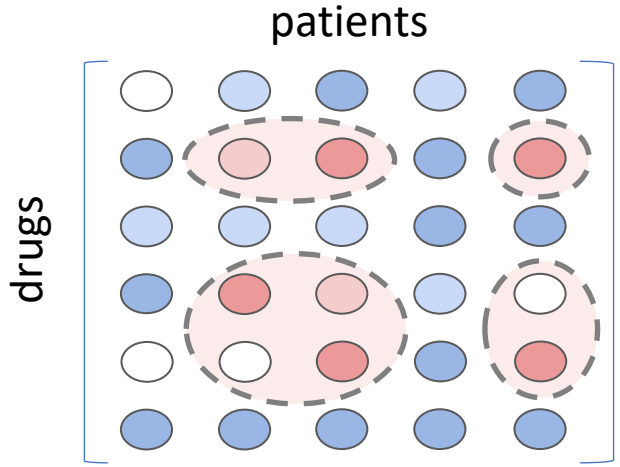


- Healthcare/biological data
  - Identifying drugs w/ anomalous reactions for specific groups of patients

Anomalies are submatrices

Submatrix family

$\mathcal{S} = \mathcal{M}_n$

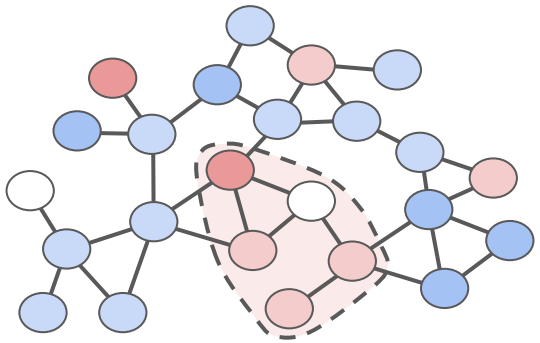


- Anomalies in graphs
  - Identifying disease outbreak hotspots or anomalous activity in social networks

Anomalies are connected subgraphs

Connected family

$\mathcal{S} = \mathcal{C}_G$

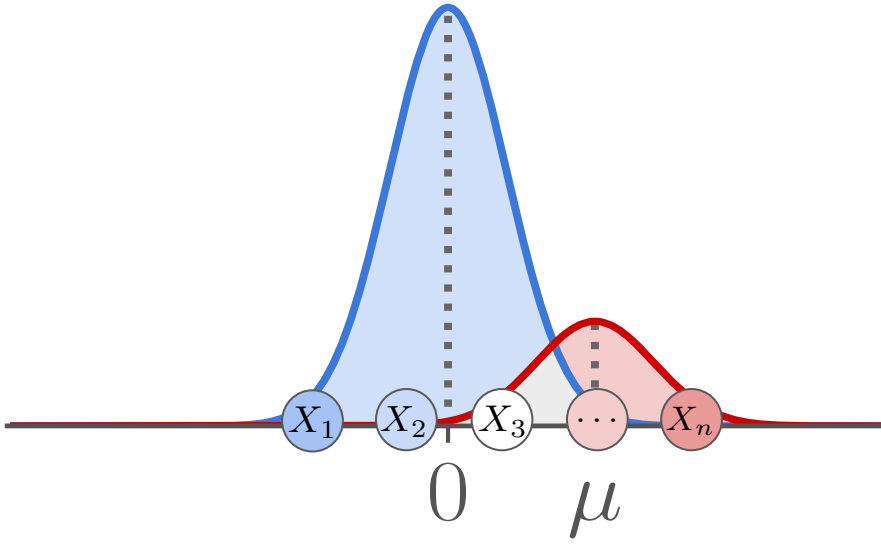


Structured Normal Means Setting

Data $X_1, \ldots, X_n$ independently distributed as

$$X_i \sim \begin{cases} N(\mu, 1) & \text{if } i \in A \\ N(0, 1) & \text{otherwise} \end{cases}$$

where anomaly $A \in \mathcal{S}$ is a member of anomaly family $\mathcal{S}$

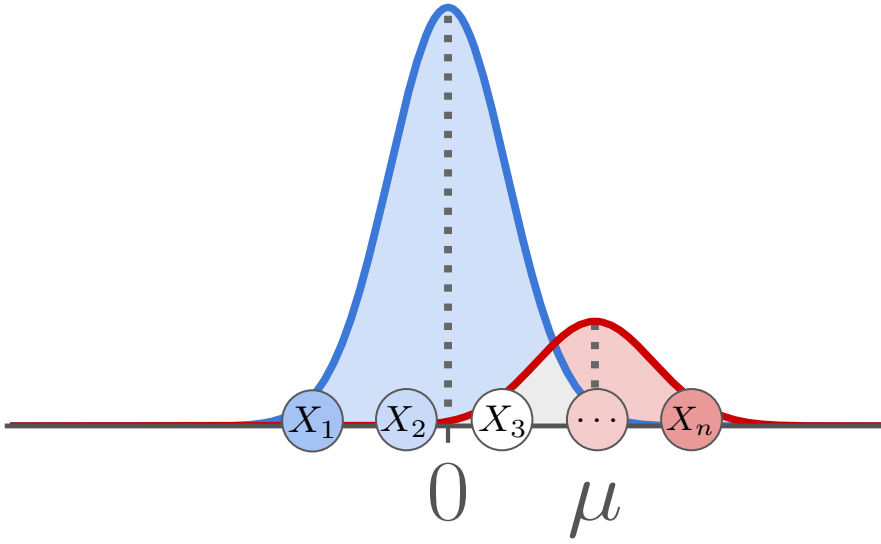


Structured Normal Means Setting

Data $X_1, \ldots, X_n$ independently distributed as

$$X_i \sim \begin{cases} N(\mu, 1) & \text{if } i \in A \\ N(0, 1) & \text{otherwise} \end{cases}$$

where anomaly $A \in \mathcal{S}$ is a member of anomaly family $\mathcal{S}$



Normal means settings have a long history in statistics, with classical methods using the normal means to model *unstructured* anomalies in p-value data

- **Localfdr/empirical Bayes** methods by Efron et al, **Higher criticism** by Donoho and Jin, …

Recent work in ML/stats study *structured* normal means settings for different anomaly families $\mathcal{S}$

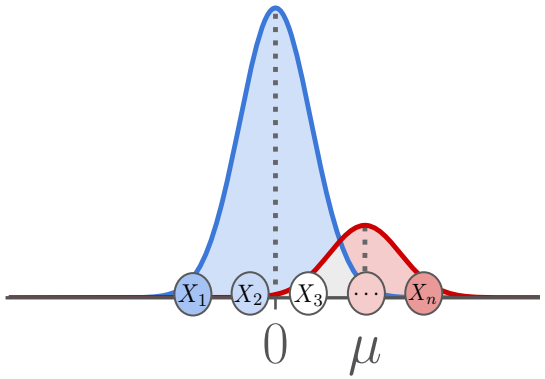
- **Intervals:** Jeng et al (JASA 2010)
- **Submatrices:** Kolar et al (NeurIPS 2011), Chen and Xu (ICML 2014), Brennan et al (COLT 2018), Liu and A-C (KDD 2019)
- **Connected subgraphs:** Qian et al (NeurIPS 2014), Aksoylar et al (ICML 2017), Cadena et al (AAAI 2018/TKDD 2019)
- **Subgraphs w/ small cut**: Sharpnack et al (NeurIPS 2013/AISTATS 2013)
- **Other:** Brennan et al (ICML 2020)

Standard approach for anomaly detection is to compute the MLE

Maximum Likelihood Estimator (MLE): $\widehat{A}_{\mathrm{MLE}} = \arg\max_{S \in \mathcal{S}} \frac{1}{\sqrt{|S|}} \sum_{i \in S} X_i$

Many papers focus on efficient algorithms for (approximately) computing the MLE.

However statistical properties of the MLE are not as well understood



Data $X_1, \ldots, X_n$ distributed as

$$X_i \sim \begin{cases} N(\mu, 1) & \text{if } i \in A \\ N(0, 1) & \text{otherwise} \end{cases}$$

where anomaly $A \in \mathcal{S}$ is a member of anomaly family $\mathcal{S}$

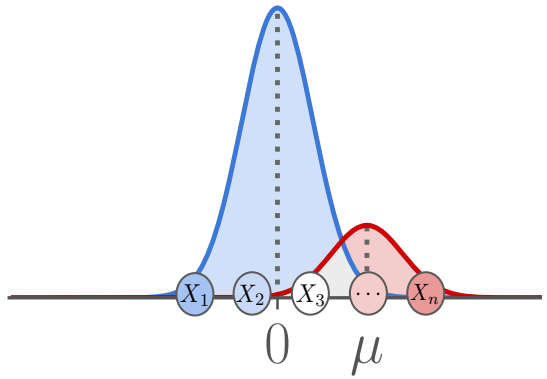
The MLE is (near-)optimal for some anomaly families…

- Jeng et al (JASA 2010) show (asymptotic) "*near-optimality*" for interval family $\mathcal{S} = \mathcal{I}_n$

- Liu and A-C (KDD 2019) show similar guarantees for submatrix family $\mathcal{S} = \mathcal{M}_N$

Maximum Likelihood Estimator (MLE):

$$\widehat{A}_{\mathrm{MLE}} = \arg\max_{S \in \mathcal{S}} \frac{1}{\sqrt{|S|}} \sum_{i \in S} X_i$$



Data $X_1, \ldots, X_n$ distributed as

$$X_i \sim \begin{cases} N(\mu, 1) & \text{if } i \in A \\ N(0, 1) & \text{otherwise} \end{cases}$$

where anomaly $A \in \mathcal{S}$ is a member of anomaly family $\mathcal{S}$

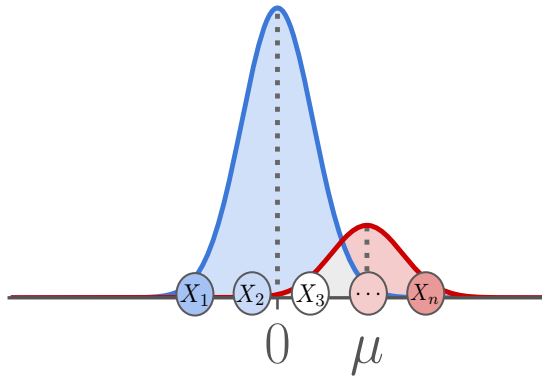
… but MLE is not optimal for other anomaly families

- Jeng et al (JASA 2010) show (asymptotic) "*near-optimality*" for interval family $\mathcal{S} = \mathcal{I}_n$
- Liu and A-C (KDD 2019) show similar guarantees for submatrix family $\mathcal{S} = \mathcal{M}_N$

In recent prior work, we (RECOMB 2020) observed that MLE is a **biased** estimator for the connected family $\mathcal{S} = \mathcal{C}_G$

Maximum Likelihood Estimator (MLE):

$$\widehat{A}_{\mathrm{MLE}} = \arg\max_{S \in \mathcal{S}} \frac{1}{\sqrt{|S|}} \sum_{i \in S} X_i$$

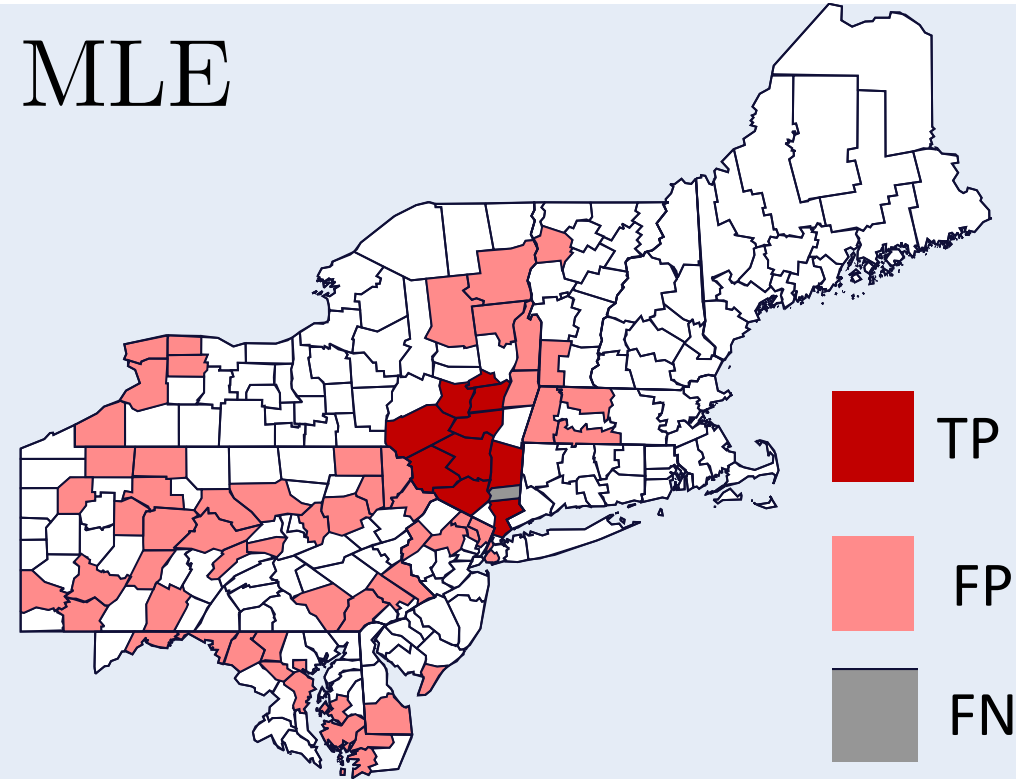


Data $X_1, \ldots, X_n$ distributed as

$$X_i \sim \begin{cases} N(\mu, 1) & \text{if } i \in A \\ N(0, 1) & \text{otherwise} \end{cases}$$

where anomaly $A \in \mathcal{S}$ is a member of anomaly family $\mathcal{S}$

MLE is **biased** for connected subgraphs

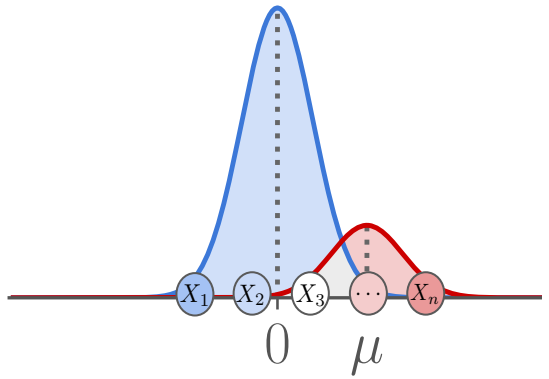


Connected anomaly A of size |A|=11 is implanted in graph of NEast USA

(Standard benchmark for spatial scan statistics)

For connected family $\mathcal{S} = \mathcal{C}_G$
MLE is **biased** estimator of size |A| of anomaly, i.e. on average

$$|\widehat{A}_{\mathrm{MLE}}| >> |A|$$



Data $X_1, \ldots, X_n$ distributed as

$$X_i \sim \begin{cases} N(\mu, 1) & \text{if } i \in A \\ N(0, 1) & \text{otherwise} \end{cases}$$

where anomaly $A \in \mathcal{S}$ is a member of anomaly family $\mathcal{S}$

Questions

1. For which anomaly families $\mathcal{S}$ is the MLE $\widehat{A}_{\mathrm{MLE}}$ biased?

2. For anomaly families $\mathcal{S}$ where MLE $\widehat{A}_{\mathrm{MLE}}$ is biased, is there a better estimator?

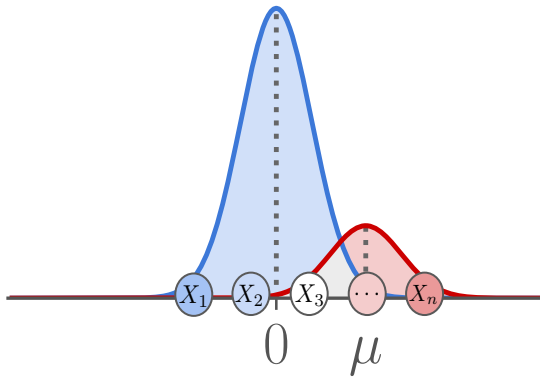
Our Contributions

1. For which anomaly families $\mathcal{S}$ is the MLE $\widehat{A}_{\mathrm{MLE}}$ biased?

Our conjecture: MLE is biased $\leftrightarrow$ number of sets in anomaly family $\mathcal{S}$ that contain the anomaly A is *exponential*

($\rightarrow$) We prove. Generalizes previous results on interval/submatrix family, which have sub-exponential size

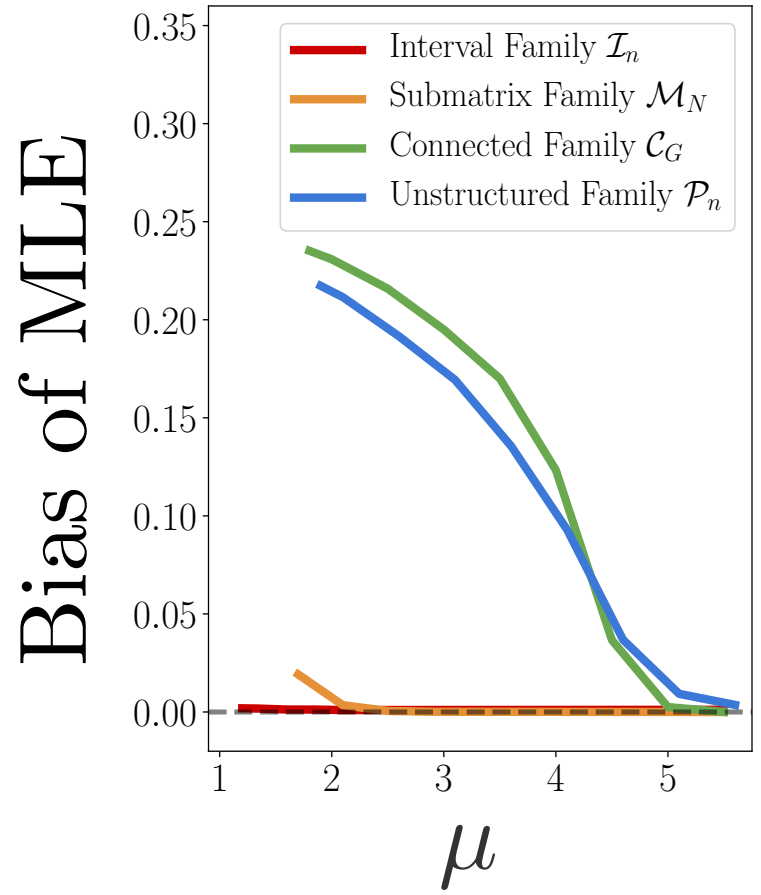
($\leftarrow$) Give partial proof/empirical evidence



Data $X_1, \ldots, X_n$ distributed as

$$X_i \sim \begin{cases} N(\mu, 1) & \text{if } i \in A \\ N(0, 1) & \text{otherwise} \end{cases}$$

where anomaly $A \in \mathcal{S}$ is a member of anomaly family $\mathcal{S}$

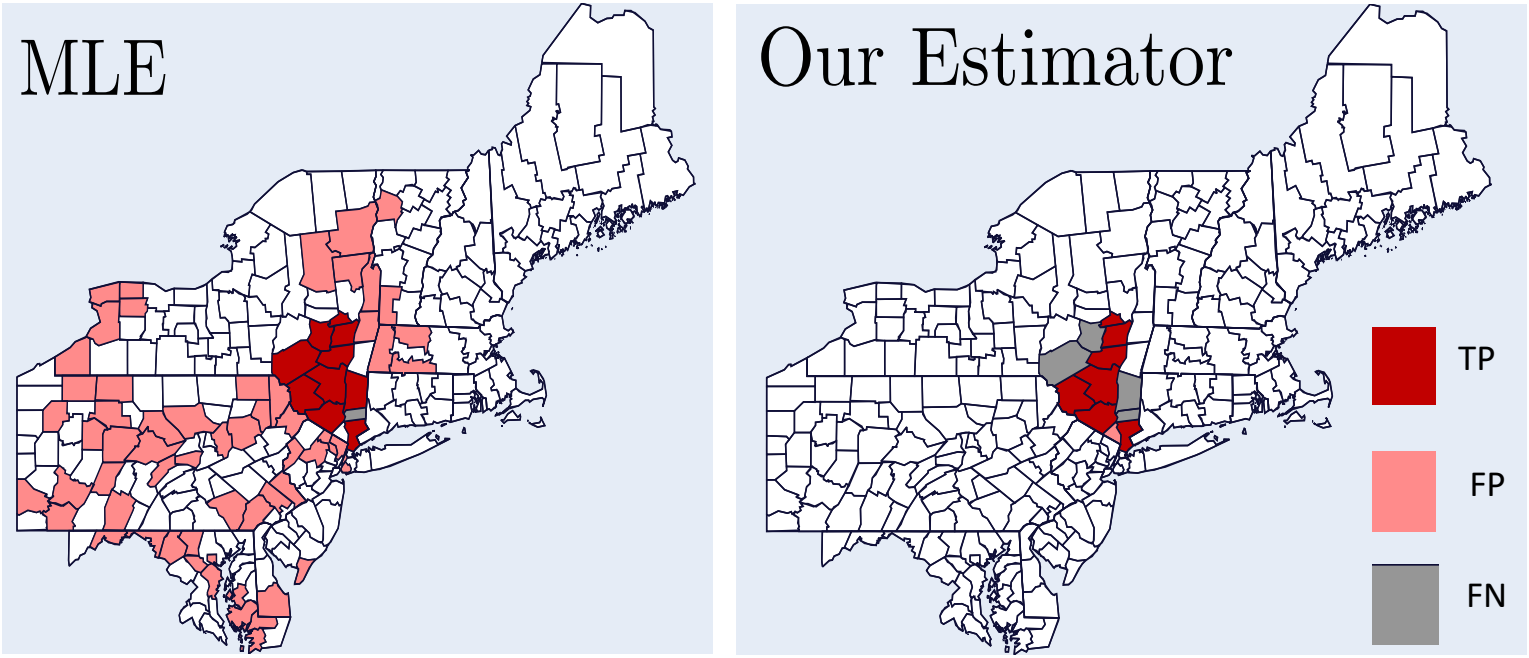


Our Contributions

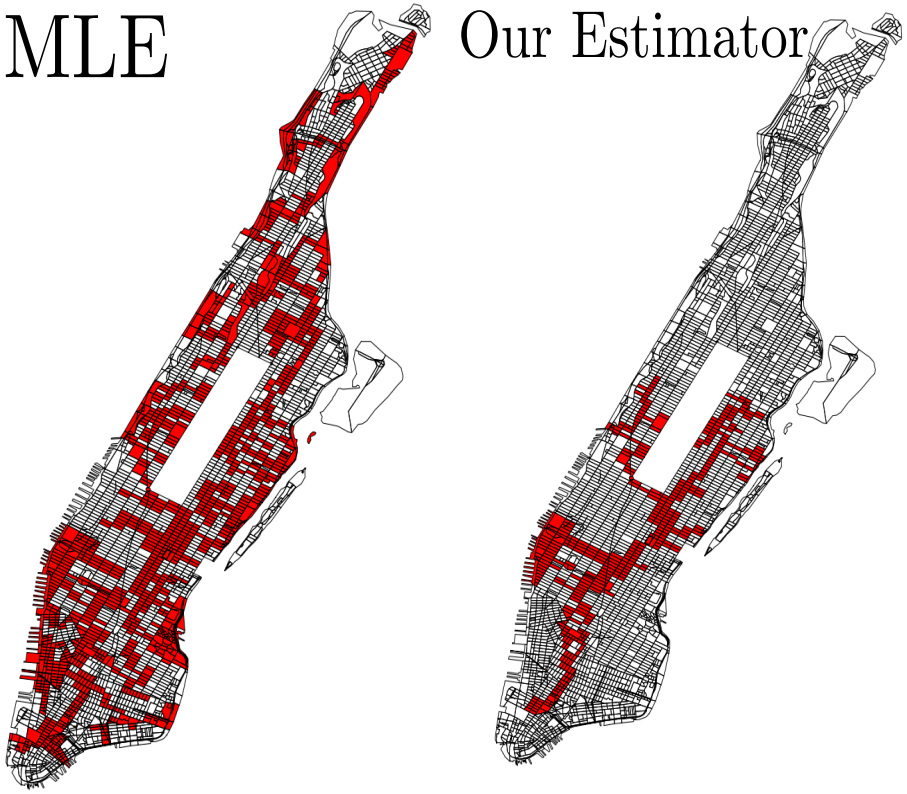
2. For anomaly families $\mathcal{S}$ where MLE $\widehat{A}_{\mathrm{MLE}}$ is biased, is there a better estimator?

Our work: asymptotically unbiased estimator for **all** anomaly families $\mathcal{S}$

Key idea: Estimate anomaly size |A| by fitting data to mixture model



Simulated Data



Real Data (Breast Cancer in NYC)

Thank you for listening!

If you have any questions or comments, stop by the poster session ☺

Scan the QR code for the arXiv

Paper link: https://arxiv.org/abs/2007.07878