Scalable Spatial Scan Statistics through Sampling

Michael Matheny Raghvendra Singh Liang Zhang
Kaiqiang Wang Jeff M. Phillips

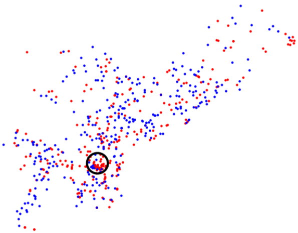
School of Computing
University of Utah

ACM SigSpatial, 2016

Spatial Scan Statistics

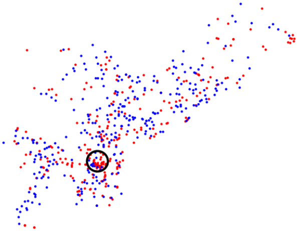
Sampled Philadelphia crime data

- Theft
- All crimes in red and blue



Spatial Scan Statistics

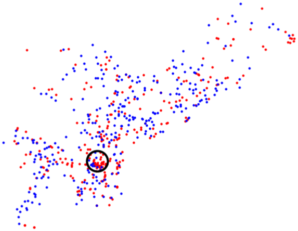
- Data set $X \subseteq \mathbb{R}^2$ and for each $x \in X$
  - $m(x)$ is a measured value. $m(x) = 1$ for theft otherwise 0.
  - $b(x)$ is a baseline value. $b(x) = 1$ for all points.
- Sets defined by regions $\mathcal{A} \subset 2^X$.
  - Disks
  - Rectangles
- Find region that maximizes function ϕ.



Spatial Scan Statistics

Want to find regions corresponding to:

- Disease outbreaks
- High regions of crime
- Environmental causes for cancer
- Wildfires, earthquakes, and other natural disasters.



Anomaly Detection Pipeline

- Formulate a model of the data and choose a corresponding measure ϕ to score the likelihood of an anomaly in a region.

Anomaly Detection Pipeline

- Formulate a model of the data and choose a corresponding measure ϕ to score the likelihood of an anomaly in a region.
- Scan the data set to find a region A which maximizes ϕ.

Anomaly Detection Pipeline

- Formulate a model of the data and choose a corresponding measure ϕ to score the likelihood of an anomaly in a region.
- Scan the data set to find a region A which maximizes ϕ.
- Assess whether the score $\phi(A)$ indicates A is significant.

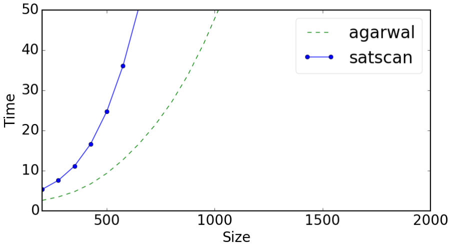
Anomaly Detection Pipeline

- Formulate a model of the data and choose a corresponding measure ϕ to score the likelihood of an anomaly in a region.
- **Scan the data set to find a region A which maximizes ϕ.**
- Assess whether the score $\phi(A)$ indicates A is significant.

Existing Approaches

For set $|X| = m$.

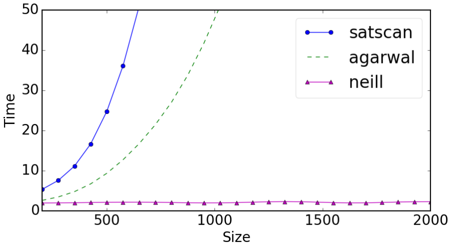
- SatScan [?] [?]
  - Commonly used.
  - Scans all disks.
  - $O(m^3 \log(m))$ runtime.
- Agarwal [?]
  - Approximation using linear functions.
  - Faster and works on rectangles.
  - $O(\frac{1}{\varepsilon} m^2 \log^2(m))$ runtime.



Existing Approaches

For set $|X| = m$.

- Neill [?]
  - Aggregates to grid.
  - Can miss anomalies if dense clusters of points exist.
  - Performance depends on data.
  - Best Case $O(g^2 \log(g))$, Worst Case $O(g^4)$.



Data is usually a Sample

Methods assume entire data set is available, but...

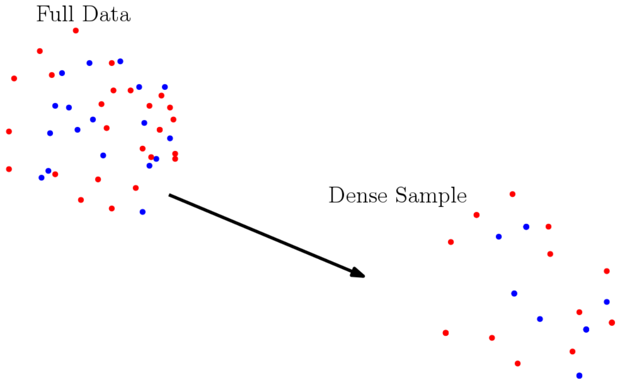
- Only reported crimes.
- Census samples population.
- 1% feed of geolocated tweets.

How much error does sampling introduce in anomaly detection?

Algorithms

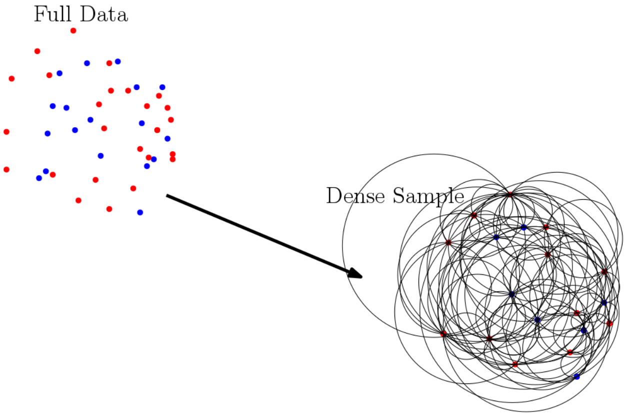
Sample Then Scan

Idea: Run SatScan on Sample.



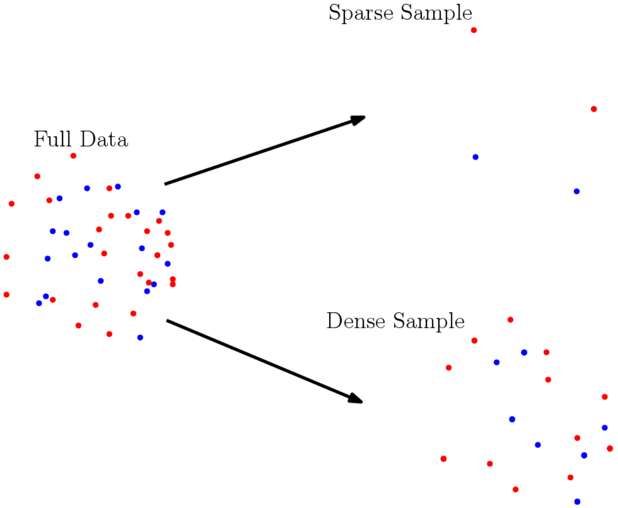
Sample Then Scan

Problem: Far too many combinatorial regions.



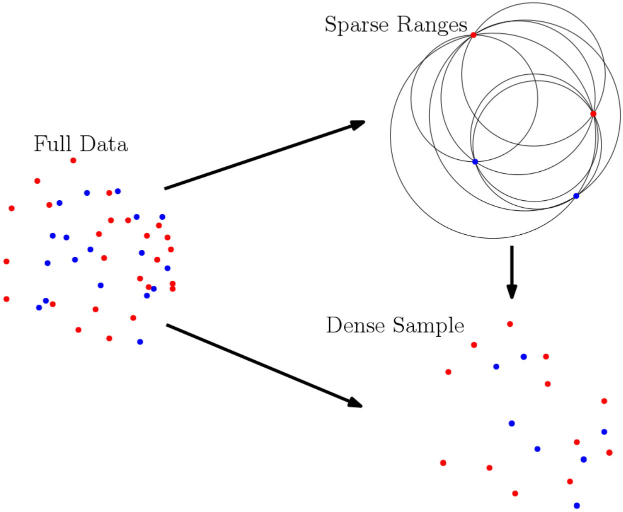
Sample Then Scan

Idea: Use smaller sample to induce regions.



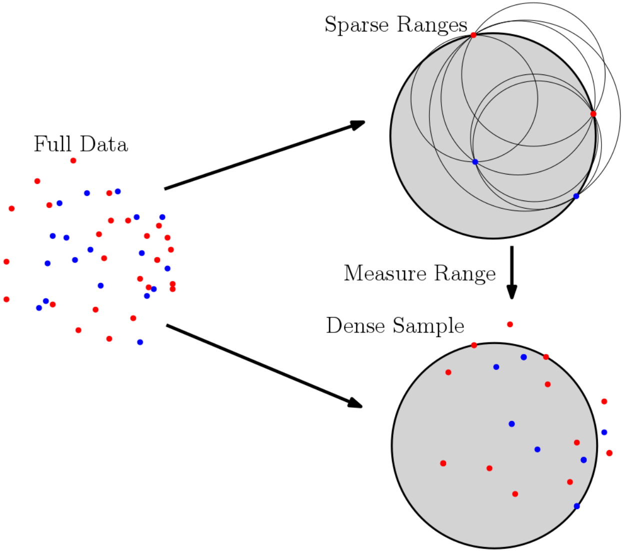
Sample Then Scan

Idea: Use smaller sample to induce regions.



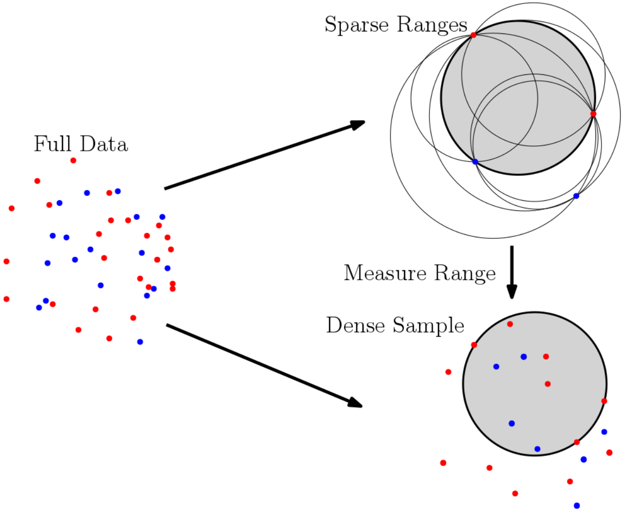
Sample Then Scan

Compute ϕ using dense sample.



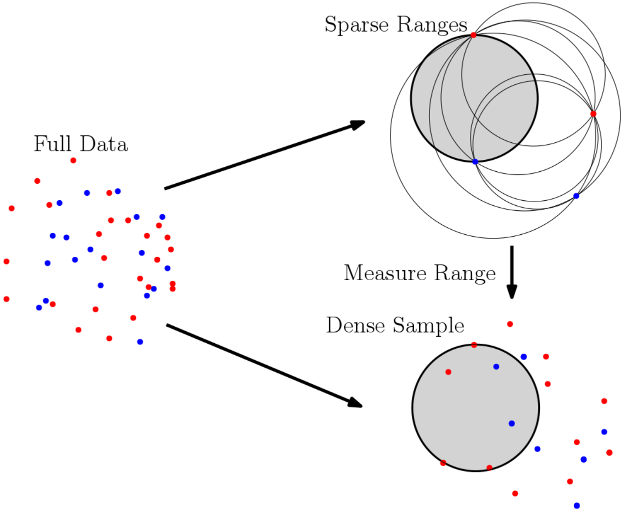
Sample Then Scan

Compute ϕ using dense sample.



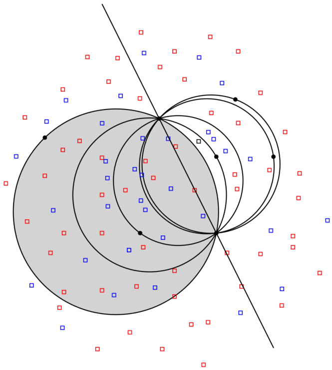
Sample Then Scan

Compute ϕ using dense sample.



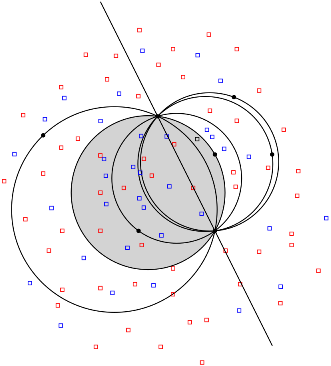
Enumerating Disks

- Split points along half space defined by points $p_1, p_2 \in N$ (sparse sample)
- Sort S (dense sample) by the order points fall into a disk passing through p_1 and p_2.



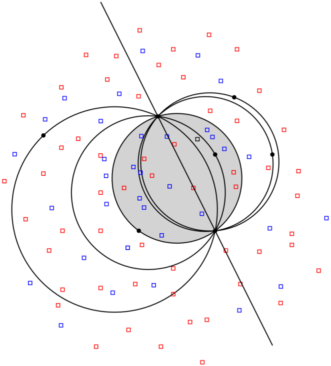
Enumerating Disks

- Split points along half space defined by points $p_1, p_2 \in N$ (sparse sample)
- Sort S (dense sample) by the order points fall into a disk passing through p_1 and p_2.



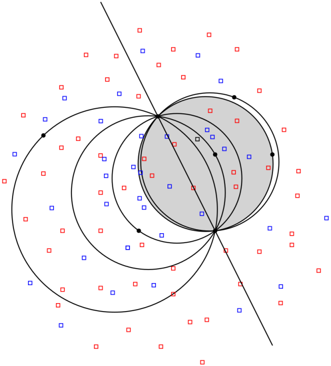
Enumerating Disks

- Split points along half space defined by points $p_1, p_2 \in N$ (sparse sample)
- Sort S (dense sample) by the order points fall into a disk passing through p_1 and p_2.



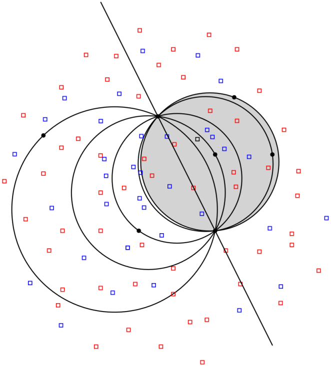
Enumerating Disks

- Split points along half space defined by points $p_1, p_2 \in N$ (sparse sample)
- Sort S (dense sample) by the order points fall into a disk passing through p_1 and p_2.



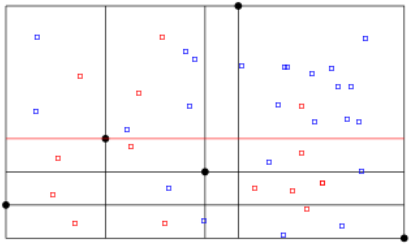
Enumerating Disks

- Split points along half space defined by points $p_1, p_2 \in N$ (sparse sample)
- Sort S (dense sample) by the order points fall into a disk passing through p_1 and p_2.
- Repeat for all p_1 and p_2.
- $|N| = n$, $|S| = s$
- $O(n^2 s \log(n))$



Enumerating Rectangles

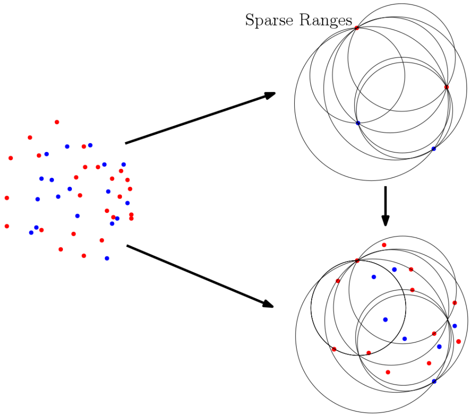
- Use N to define a grid of size n^2
- Distribute points in S into grid cells
- Enumerate over all lower and upper corners.
- $O(n^4 + s\log(n))$



Can this Work?

If this works then we help scalability and the sampling problem.

- How well does this method work in practice?
- Can we prove guarantees?

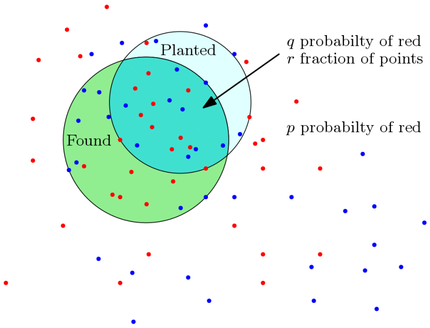


How well does this work?

Experimental Setup

- 5 million tweets.
- Algorithm ran with:
  - $|N| = n$ sparse sample.
  - $|S| = s$ dense sample.
- Planted region containing:
  - r fraction of points.
  - p measured rate outside.
  - q measured rate inside.
- Jaccard Distance

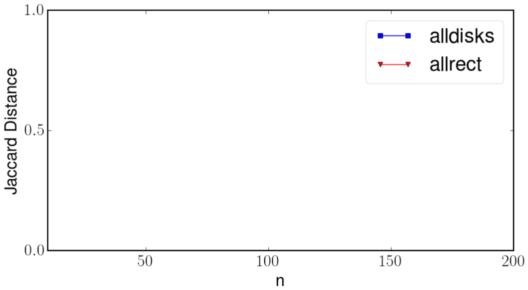
$$d(A, B) = 1 - \frac{|A \cap B|}{|A \cup B|}$$



Stability

Defaults

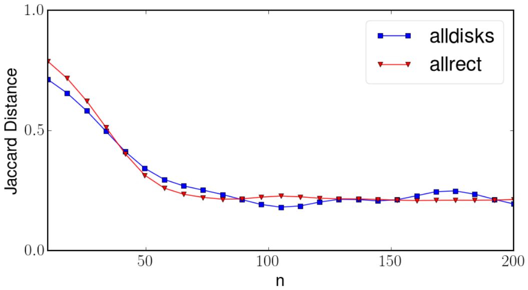
- Outside rate
  $p = .04$
- Inside rate
  $q = .08$
- Region size
  $r = .05$
- Sparse sample
  $n = 100$
- Large sample
  $s = 4000$



Stability

Defaults

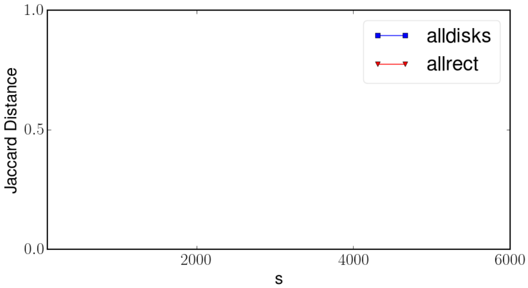
- Outside rate
  $p = .04$
- Inside rate
  $q = .08$
- Region size
  $r = .05$
- Sparse sample
  $n = 100$
- Large sample
  $s = 4000$



Stability

Defaults

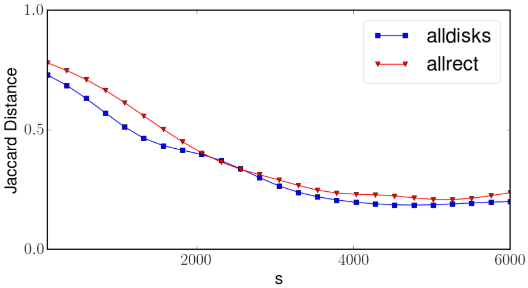
- Outside rate
  $p = .04$
- Inside rate
  $q = .08$
- Region size
  $r = .05$
- Sparse sample
  $n = 100$
- Large sample
  $s = 4000$



Stability

Defaults

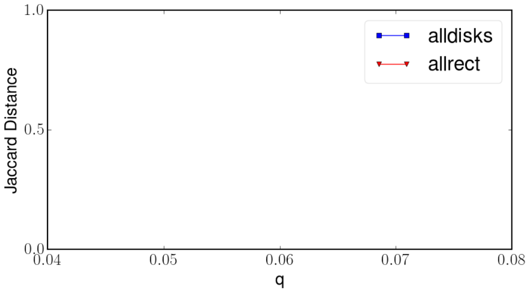
- Outside rate
  $p = .04$
- Inside rate
  $q = .08$
- Region size
  $r = .05$
- Sparse sample
  $n = 100$
- Large sample
  $s = 4000$



Stability

Defaults

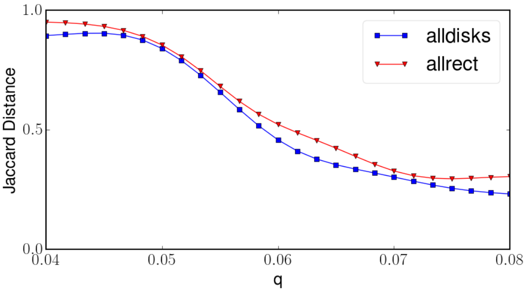
- Outside rate
  $p = .04$
- Inside rate
  $q = .08$
- Region size
  $r = .05$
- Sparse sample
  $n = 100$
- Large sample
  $s = 4000$



Stability

Defaults

- Outside rate
  $p = .04$
- Inside rate
  $q = .08$
- Region size
  $r = .05$
- Sparse sample
  $n = 100$
- Large sample
  $s = 4000$

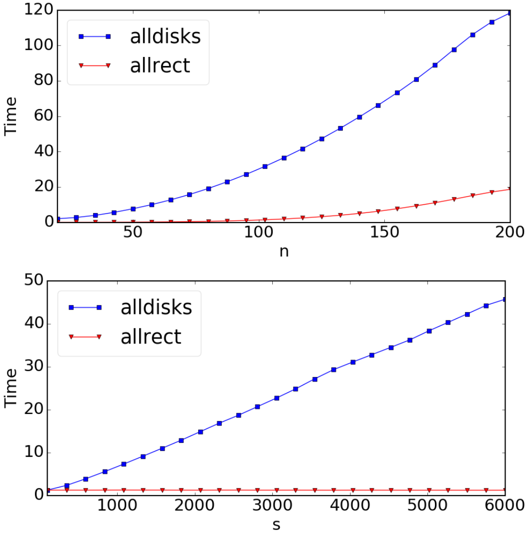


Running Time

Defaults

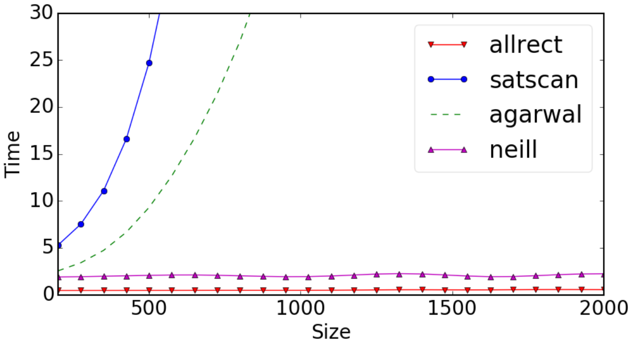
- Outside rate
  $p = .04$
- Inside rate
  $q = .08$
- Region size
  $r = .05$
- Sparse sample
  $n = 100$
- Large sample
  $s = 4000$

alldisks: $O(n^2 s \log(n))$
allrect: $O(n^4 + s \log(n))$



Running Time

Method compares favorably with existing algorithms when using similar error.



Unlike griding our methods have guarantees since sample N adapts to data.

Experiment Summary

- Reasonable sample sizes.
- Finds region with high overlap.
- Stable results till threshold.
- Very fast.

Why does this work?

Lipschitz Bounds

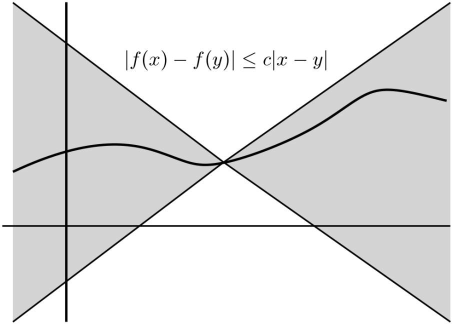
Need approximation on the Kulldorff Scan Statistic

$$\phi_X(A) = m_A \ln \frac{m_A}{b_A} + (1 - m_A) \ln \frac{1 - m_A}{1 - b_A}.$$

If:

- $\frac{\varepsilon\rho}{2} \geq |m_A - \hat{m}_A|$
- $\frac{\varepsilon\rho}{2} \geq |b_A - \hat{b}_A|$
- ρ-boundary conditions.

Then
$|\phi(m_A, b_A) - \phi(\hat{m}_A, \hat{b}_A)| \leq \varepsilon.$



Lipschitz Bounds

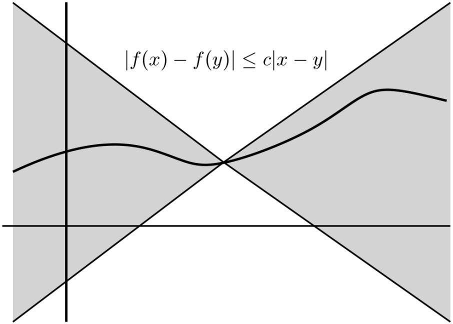
Need approximation on the Kulldorff Scan Statistic

$$\phi_X(A) = m_A \ln \frac{m_A}{b_A} + (1 - m_A) \ln \frac{1 - m_A}{1 - b_A}.$$

If:

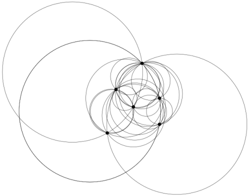
- $\frac{\varepsilon\rho}{2} \geq |\boldsymbol{m_A} - \boldsymbol{\hat{m}_A}|$
- $\frac{\varepsilon\rho}{2} \geq |\boldsymbol{b_A} - \boldsymbol{\hat{b}_A}|$
- ρ-boundary conditions.

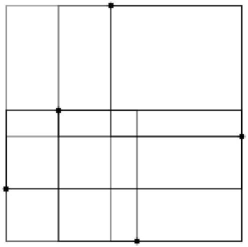
Then
$|\phi(m_A, b_A) - \phi(\hat{m}_A, \hat{b}_A)| \leq \varepsilon$.



Range Spaces

- Data set $X \subseteq \mathbb{R}^2$.
- Set of ranges $\mathcal{A} \subset 2^X$.
- Range space $R = (X, \mathcal{A})$.
  - $|\mathcal{A}| = O(|X|^3)$ for disks.
  - $|\mathcal{A}| = O(|X|^4)$ for rectangles.





ε-Samples

Given a range space $(X, \mathcal{A})$ with constant VC dimension then for $\forall A \in \mathcal{A}$ a random sample $S \subseteq X$ with constant probability will be an:

- ε-Sample
  - if $|S| = O(\frac{1}{\varepsilon^2})$
  - then $\left| \frac{|X \cap A|}{|X|} - \frac{|S \cap A|}{|S|} \right| \leq \varepsilon$

Just Sample Approach

Idea: Sample full data X and run SatScan on sample. For function ϕ with constant probability need:

- $|S| = O\left(\frac{1}{(\rho\varepsilon)^2}\right)$ for additive error bound on ϕ.
- Disks enumerated in $O\left(\left(\frac{1}{\varepsilon\rho}\right)^6 \log \frac{1}{\varepsilon\rho}\right)$
- Rectangles enumerated in $O\left(\left(\frac{1}{\varepsilon\rho}\right)^8\right)$

Not good.

ε-Samples and ε-Nets

Given a range space $(X, \mathcal{A})$ with constant VC dimension then for $\forall A \in \mathcal{A}$ a random sample $S \subseteq X$ with constant probability will be an:

- ε-Sample
  - if $|S| = O(\frac{1}{\varepsilon^2})$
  - then $\left| \frac{|X \cap A|}{|X|} - \frac{|S \cap A|}{|S|} \right| \leq \varepsilon$

ε-Samples and ε-Nets

Given a range space $(X, \mathcal{A})$ with constant VC dimension then for $\forall A \in \mathcal{A}$ a random sample $S \subseteq X$ with constant probability will be an:

- ε-Sample
  - if $|S| = O(\frac{1}{\varepsilon^2})$
  - then $\left|\frac{|X \cap A|}{|X|} - \frac{|S \cap A|}{|S|}\right| \leq \varepsilon$
- ε-Net
  - if $|S| = O(\frac{1}{\varepsilon} \log(\frac{1}{\varepsilon}))$
  - and if $\frac{|X \cap A|}{|X|} \geq \varepsilon$ then $|S \cap A| \geq 1$

Using Nets

Consider range space $(X, \mathcal{A})$ then random samples of X:

- N of size $n = O(\frac{1}{\varepsilon} \log \frac{1}{\varepsilon})$ and
- S of size $s = O(\frac{1}{\varepsilon^2} \log \frac{1}{\delta})$.

Then with constant probability for $\forall A \in \mathcal{A}$ then $\exists A' \in \{A \cap N | A \in \mathcal{A}\}$ such that

$$\left| \frac{|A \cap X|}{|X|} - \frac{|\psi(A') \cap S|}{|S|} \right| \leq \varepsilon$$

Note: Some restrictions beyond VC dimension required that rectangles and disks satisfy. See paper for details on ψ.

Theory Summary

Combine sample bound with Lipschitz bound.

- $|N| = O\left(\frac{1}{\varepsilon\rho}\log\frac{1}{\varepsilon\rho}\right)$
- $|S| = O\left(\frac{1}{(\varepsilon\rho)^2}\right)$.

Runtime with constant probability:

- Disks: $O\left(|X| + \frac{1}{(\varepsilon\rho)^4}\log^3\left(\frac{1}{\varepsilon\rho}\right)\right)$
- Rectangles: $O\left(|X| + \left(\frac{1}{\varepsilon\rho}\log\frac{1}{\varepsilon\rho}\right)^4\right)$

Attain error bound $|\phi - \phi_{N,S}| \leq \varepsilon$.

Summary

Theory

Sample sizes:

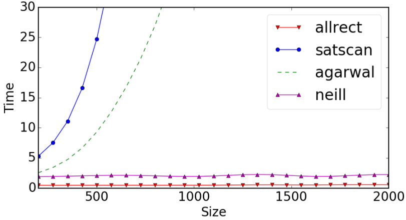
- $|N| = O\left(\frac{1}{\varepsilon\rho}\log\frac{1}{\varepsilon\rho}\right)$
- $|S| = O\left(\frac{1}{(\varepsilon\rho)^2}\right)$.

Runtime with constant probability:

- Disks: $O\left(|X| + \frac{1}{(\varepsilon\rho)^4}\log^3\left(\frac{1}{\varepsilon\rho}\right)\right)$
- Rectangles: $O\left(|X| + \left(\frac{1}{\varepsilon\rho}\log\frac{1}{\varepsilon\rho}\right)^4\right)$

Error bound $|\phi - \phi_{N,S}| \leq \varepsilon$.

Experimental



Can be even faster

- Orthogonal to [?] approach.
- Can be combined with [?].

Questions

C++ implementation with Python wrapper is available at:
https://github.com/michaelmathen/SampleScan

Theory

Sample sizes:

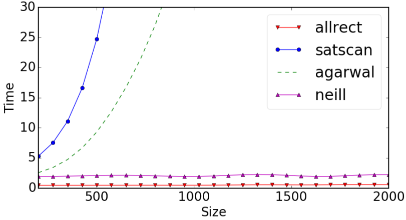
- $|N| = O\left(\frac{1}{\varepsilon\rho}\log\frac{1}{\varepsilon\rho}\right)$
- $|S| = O\left(\frac{1}{(\varepsilon\rho)^2}\right)$.

Runtime with constant probability:

- Disks: $O\left(|X| + \frac{1}{(\varepsilon\rho)^4}\log^3\left(\frac{1}{\varepsilon\rho}\right)\right)$
- Rectangles: $O\left(|X| + \left(\frac{1}{\varepsilon\rho}\log\frac{1}{\varepsilon\rho}\right)^4\right)$

Error bound $|\phi - \phi_{N,S}| \leq \varepsilon$.

Experimental



Can be even faster

- Orthogonal to [?] approach.
- Can be combined with [?].

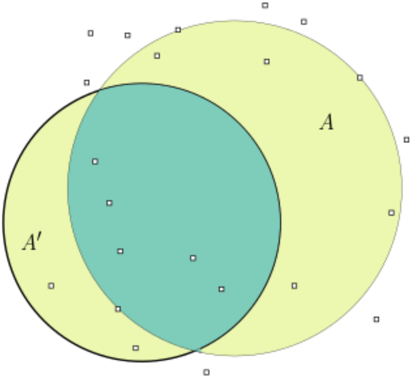
For Further Reading I

Linearization

Sample Range Approach

Symmetric Difference Range Space

- Consider a range space $(X, S_{\mathcal{A}})$ where $S_{\mathcal{A}} = \{A \triangle A' | A, A' \in \mathcal{A}\}$.
- Has VC dimension bounded by $\nu \log(\nu)$.

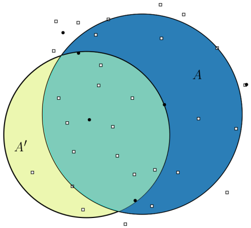


ε-Net over Symmetric Difference

- Define a conforming geometric mapping $\psi(A \cap N) \subset \mathbb{R}^2$ such that
  - $\forall A \in \mathcal{A}$ then $\psi(A \cap N) \cap N = A \cap N$
  - $\psi(A) \cap X \in \mathcal{A}$

Lemma

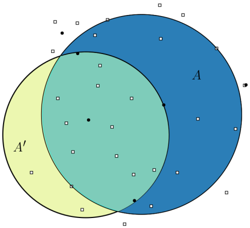
Given an ε-net N over $(X, \mathcal{S}_{\mathcal{A}})$, a geometric mapping ψ conforming to $\mathcal{A}$, then for any range $A \in (X, \mathcal{A})$, there exists a range $\psi(A') \cap X$ for $A' \in \{N \cap A| \in \mathcal{A}\}$ such that $|A \triangle (\psi(A') \cap X)| \leq \varepsilon |X|$.



ε-Net over Symmetric Difference

Use mapping to find approximate count in S.

$$2\varepsilon \geq \left| \frac{|A \cap X|}{|X|} - \frac{|\psi(A') \cap X|}{|X|} \right| + \left| \frac{|\psi(A') \cap X|}{|X|} - \frac{|\psi(A') \cap S|}{|S|} \right| \geq \left| \frac{|A \cap X|}{|X|} - \frac{|\psi(A') \cap S|}{|S|} \right|$$



Scan Statistic

- Data set $X \subseteq \underline{R}^2$ and for each $x \in X$
  - $m(x)$ is a measured value.
  - $b(x)$ is a baseline value.
- For each region $A \in \mathcal{A}$ define
  $m_X(A) = \frac{\sum_{x \in A} m(x)}{\sum_{x \in X} m(x)}, \quad b_X(A) = \frac{\sum_{x \in A} b(x)}{\sum_{x \in X} b(x)}$
- Kulldorff Scan Statistic:

$$\phi_X(A) = m_X(A) \ln \frac{m_X(A)}{b_X(A)} + (1 - m_X(A)) \ln \frac{1 - m_X(A)}{1 - b_X(A)}.$$

- Gaussian, Bernoulli, Gamma, etc versions also exist.

Matched Error Experiments