Penalized Fast Subset Scanning

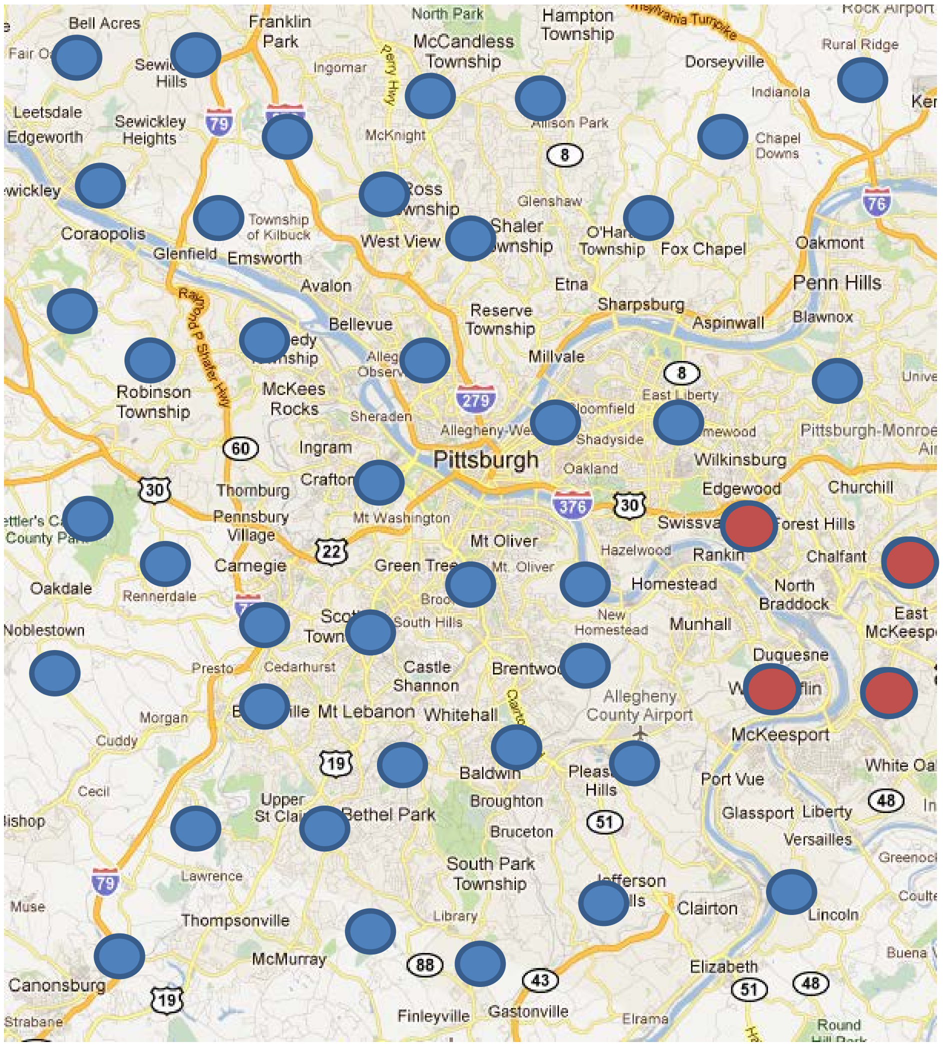
Skyler Speakman, Sriram Somanchi,
Edward McFowland III, and Daniel B. Neill*

Event and Pattern Detection Laboratory
Carnegie Mellon University

*Contact author: neill@cs.cmu.edu

Partially supported by NSF grants IIS-0916345, IIS-0911032, and IIS-0953330.

Detecting Disease Clusters

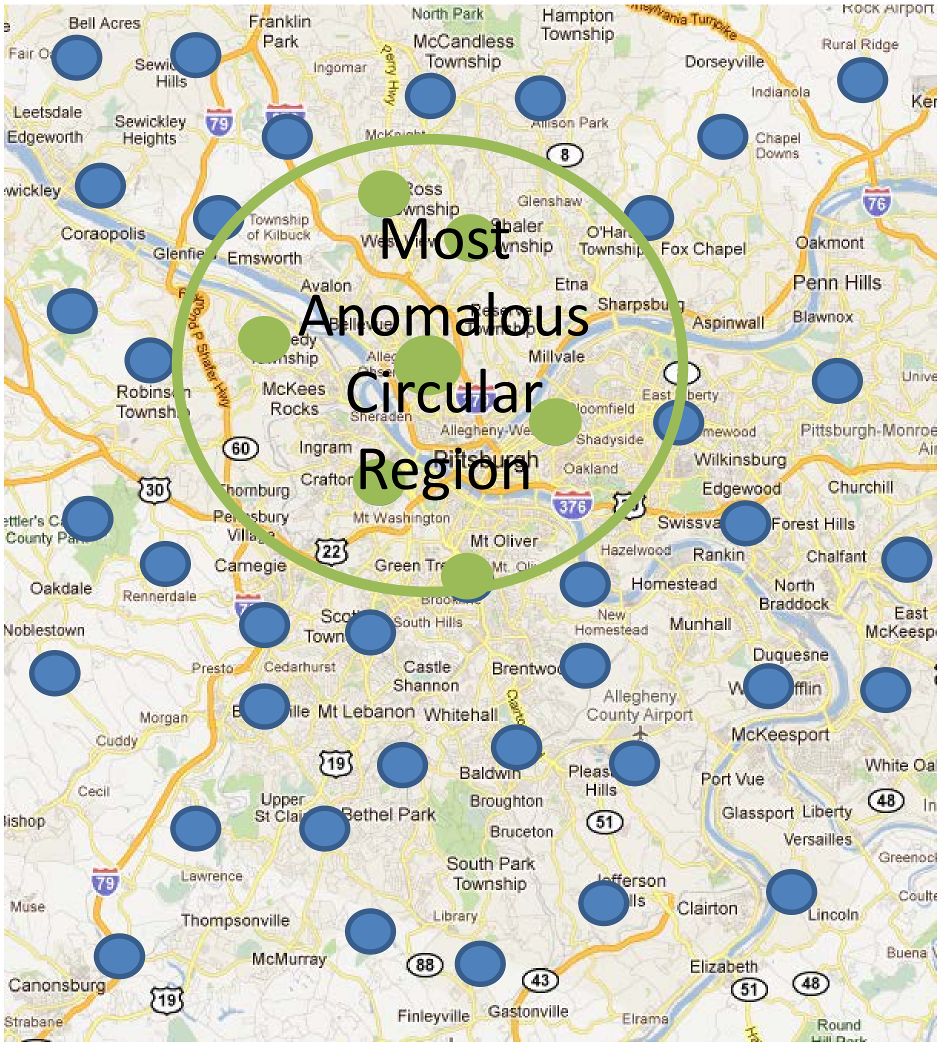


- Location of an informative data stream
  - \# of ER visits per Zip Code
  - \# of OTC Drug sales per retailer
  - Other novel data sources ...

In the presence of an outbreak, we expect counts of the affected locations to increase.

Effective methods should have high *detection power.*

Detecting Disease Clusters



(Kulldorff, 1997)

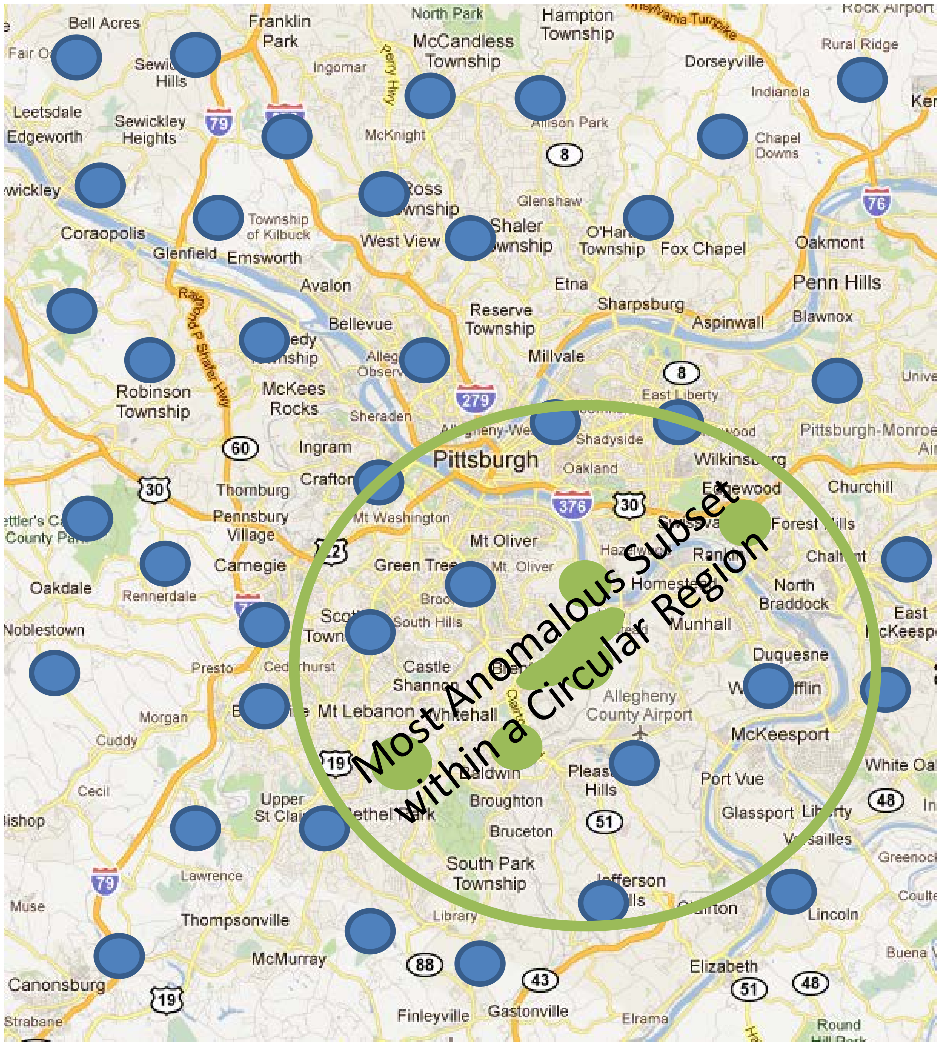
Spatial Scan Statistic (Circles)

Clusters locations by regions constrained by shape

High power to detect disease clusters of the corresponding shape

But what about irregular shaped clusters?

Detecting *Irregular* Disease Clusters



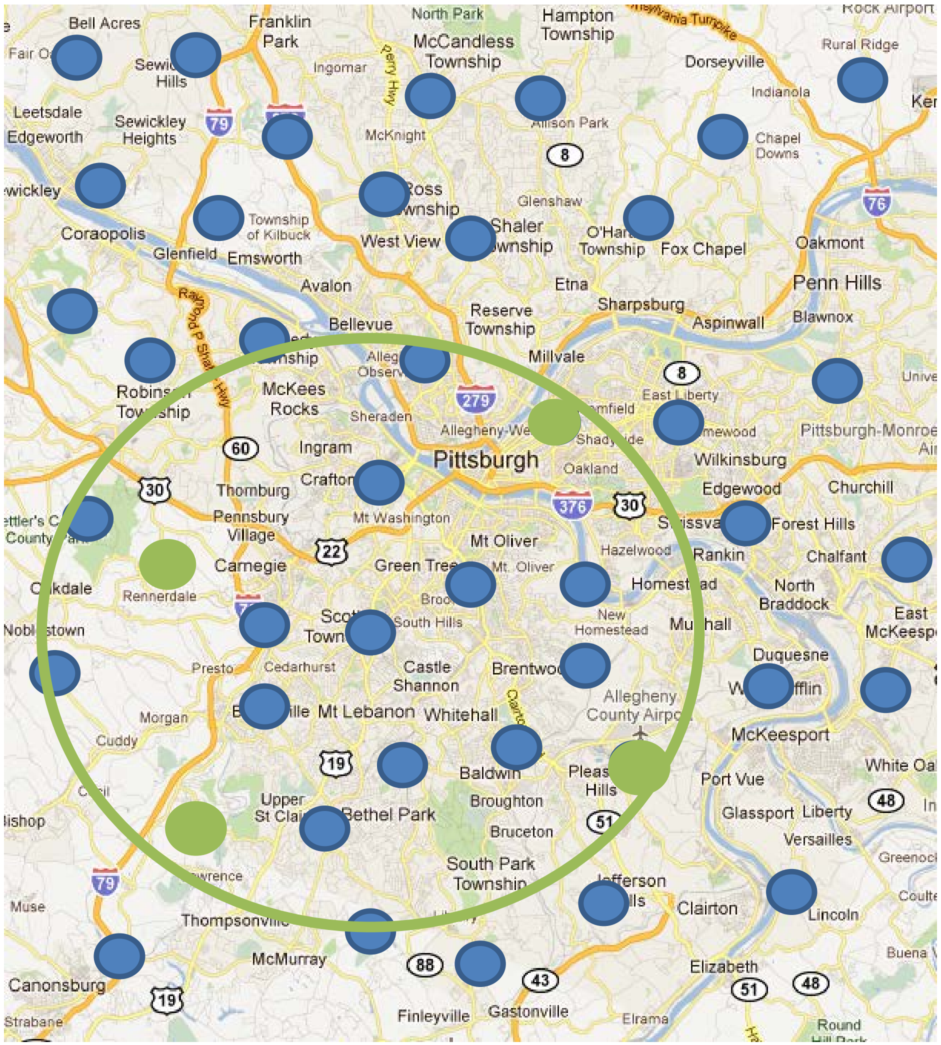
(Neill, 2011)

Fast Subset Scan

Instead of clustering ***ALL locations*** within the region together, only the ***most anomalous subset of locations*** within the region is used

Increases power to detect irregularly shaped disease clusters

Detecting *Irregular* Disease Clusters



(Neill, 2011)

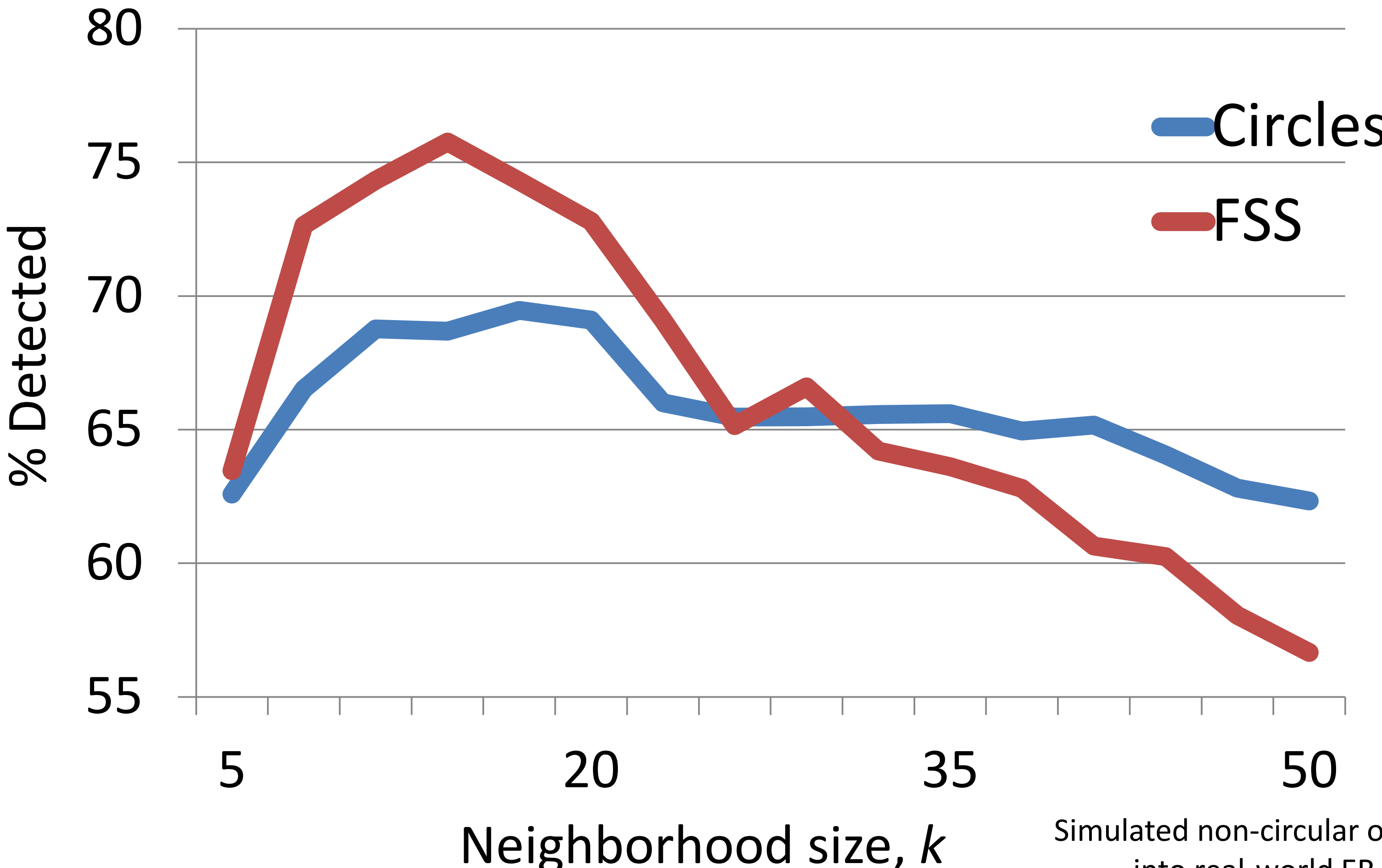
Fast Subset Scan

Instead of clustering ***ALL locations*** within the region together, only the ***most anomalous subset of locations*** within the region is used

Increases power to detect irregularly shaped disease clusters

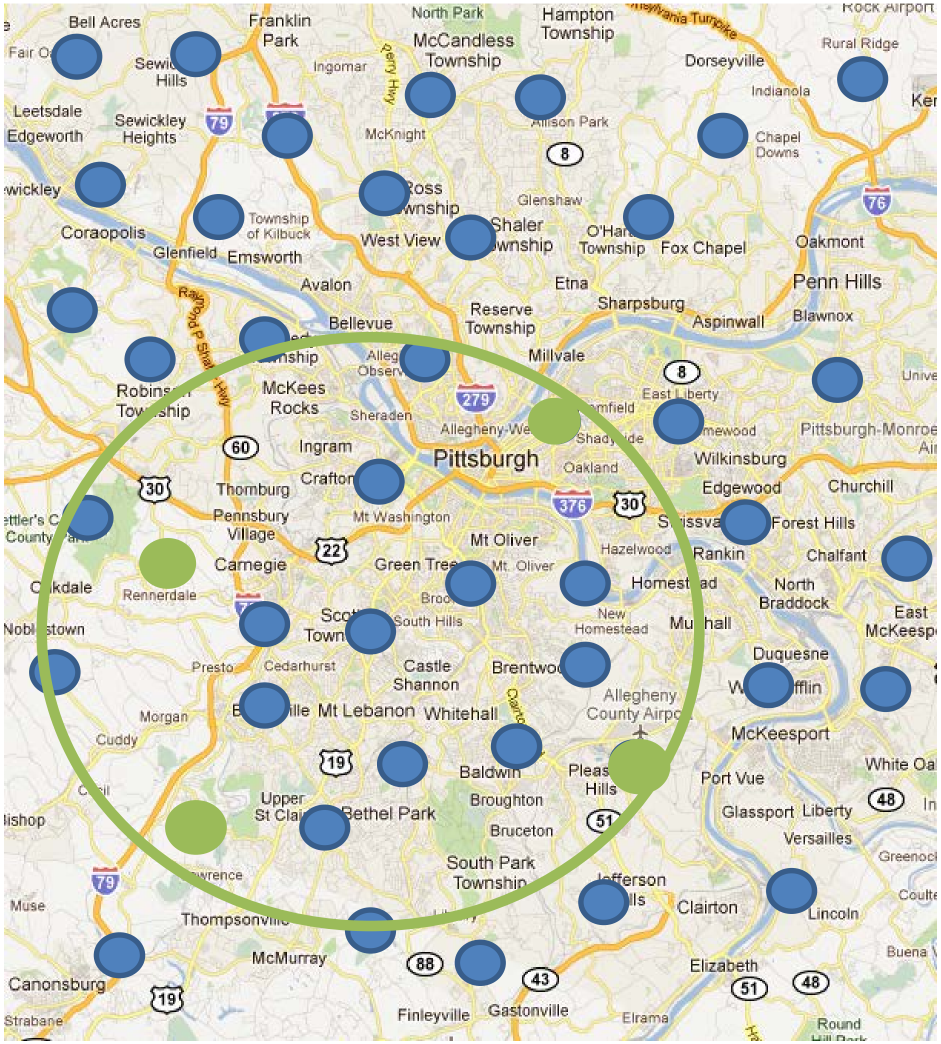
...but may return ***spatially dispersed subsets*** that do not reflect an outbreak of disease

Detection Power for Varying Neighborhood Size



Simulated non-circular outbreaks injected into real-world ER background data. Fixed false positive rate of 1 per year.

Detecting *Irregular* Disease Clusters



(Neill, 2011)

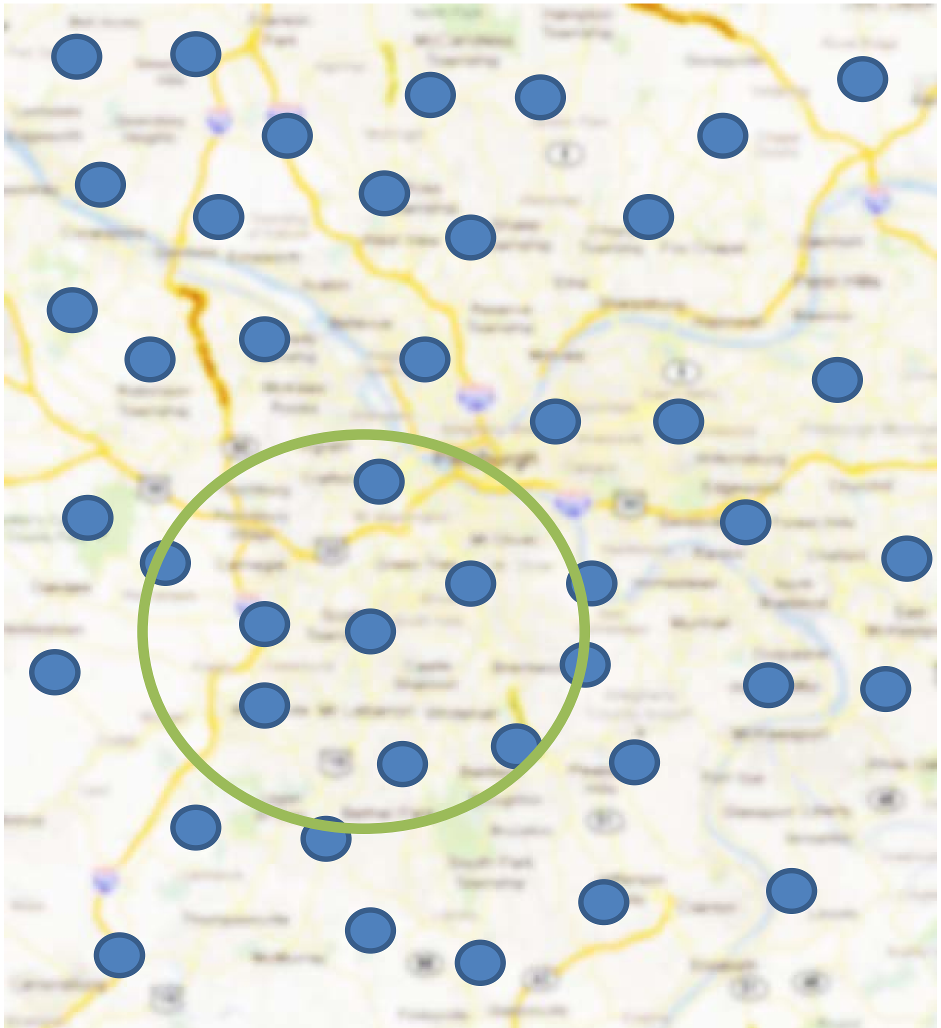
Fast Subset Scan

Instead of clustering ***ALL locations*** within the region together, only the ***most anomalous subset of locations*** within the region is used

Increases power to detect irregularly shaped disease clusters

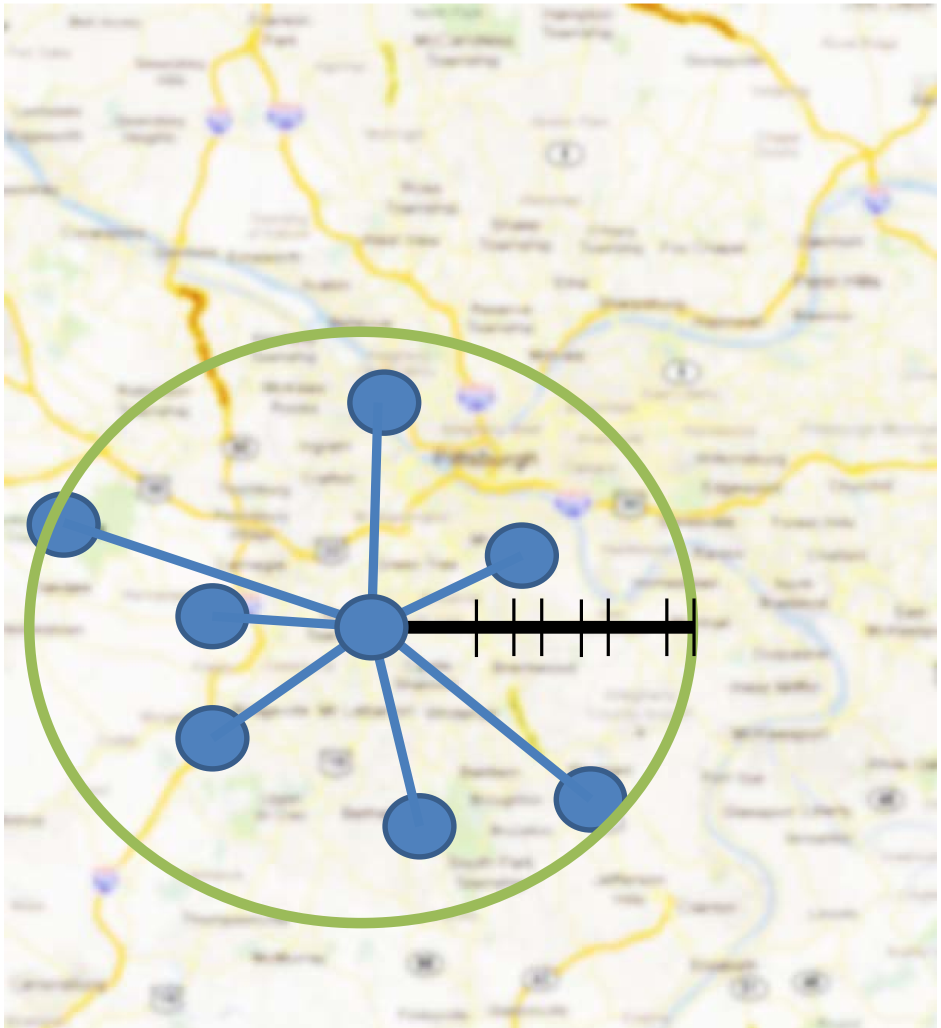
...but may return ***spatially dispersed subsets*** that do not reflect an outbreak of disease

Detecting *Irregular* Disease Clusters



Soft Compactness Constraints

Detecting *Irregular* Disease Clusters

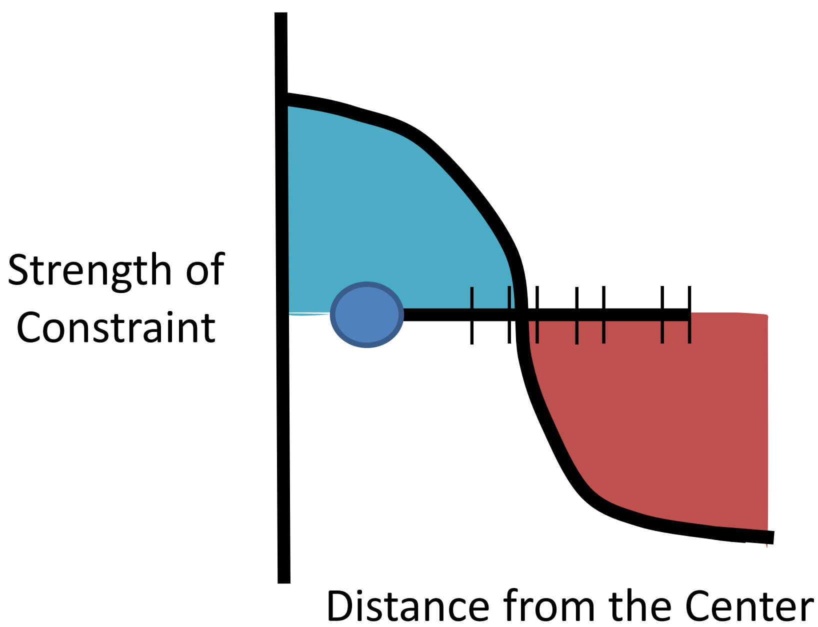


Soft Compactness Constraints

Use the distance of each location from the center as a measure of compactness/sparsity

Detecting *Irregular* Disease Clusters

Soft Compactness Constraints



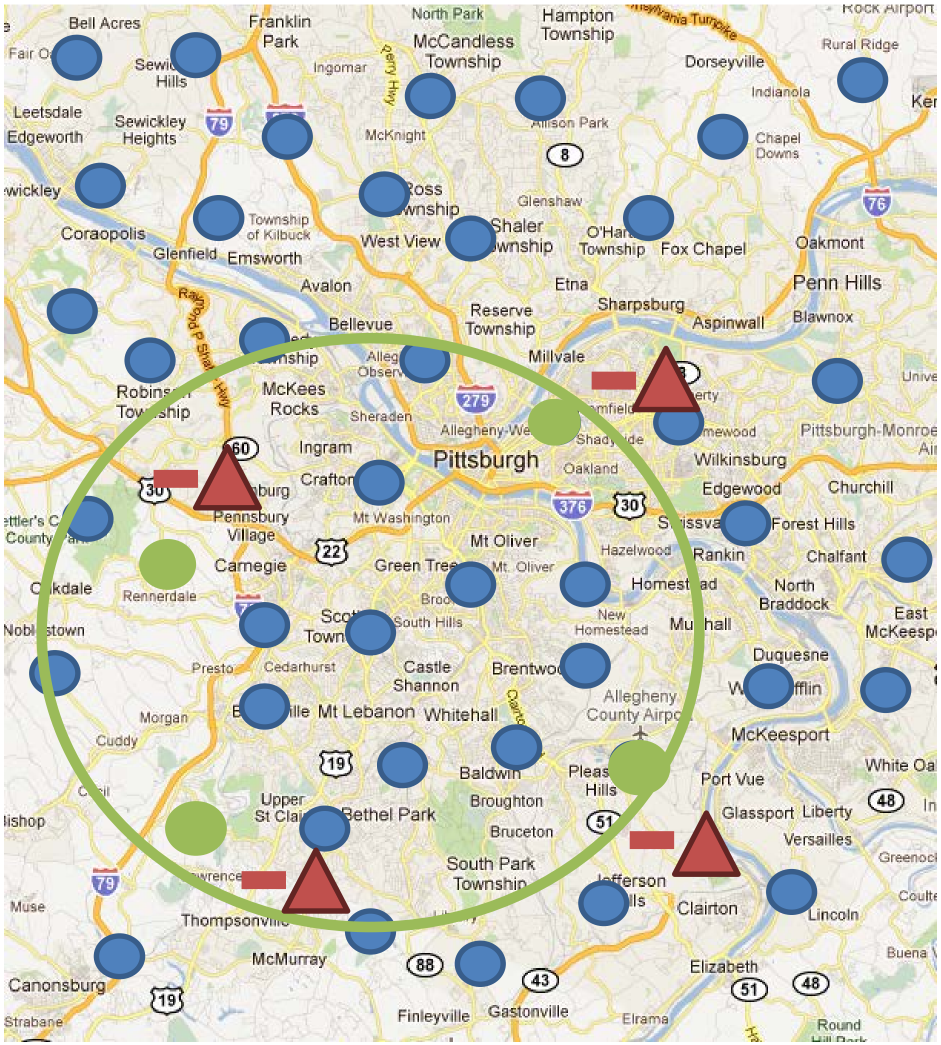
Use the distance of each location from the center as a measure of compactness/sparsity

Reward subsets that contain locations close to the center

and

Penalize subsets that contain locations far from the center

Detecting *Irregular* Disease Clusters



Soft Compactness Constraints

...but may return

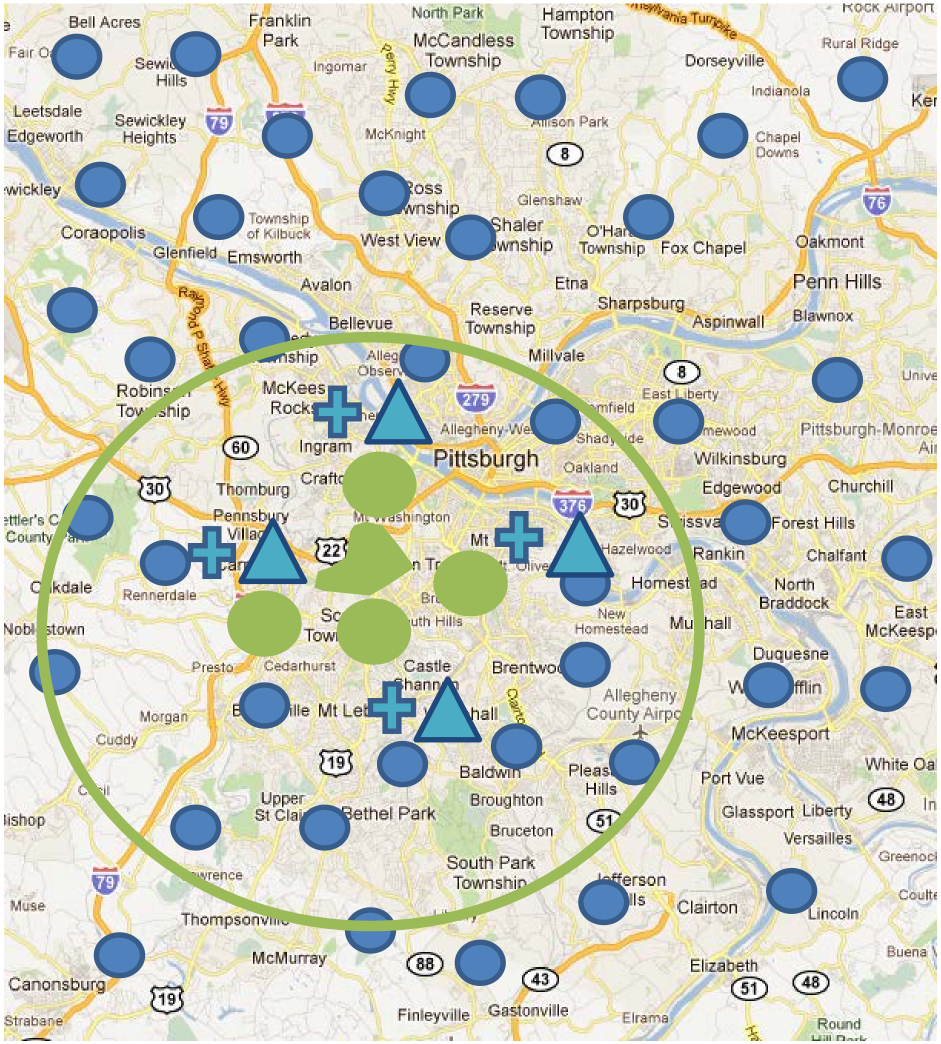
spatially dispersed subsets

that do not reflect an outbreak of disease.

This particular subset would be less likely returned as the optimal one when compactness constraints are used.

The penalties associated with the distance between the locations and center of the circle would decrease the "score" of the subset

Detecting *Irregular* Disease Clusters



Soft Compactness Constraints

...but may return

spatially dispersed subsets

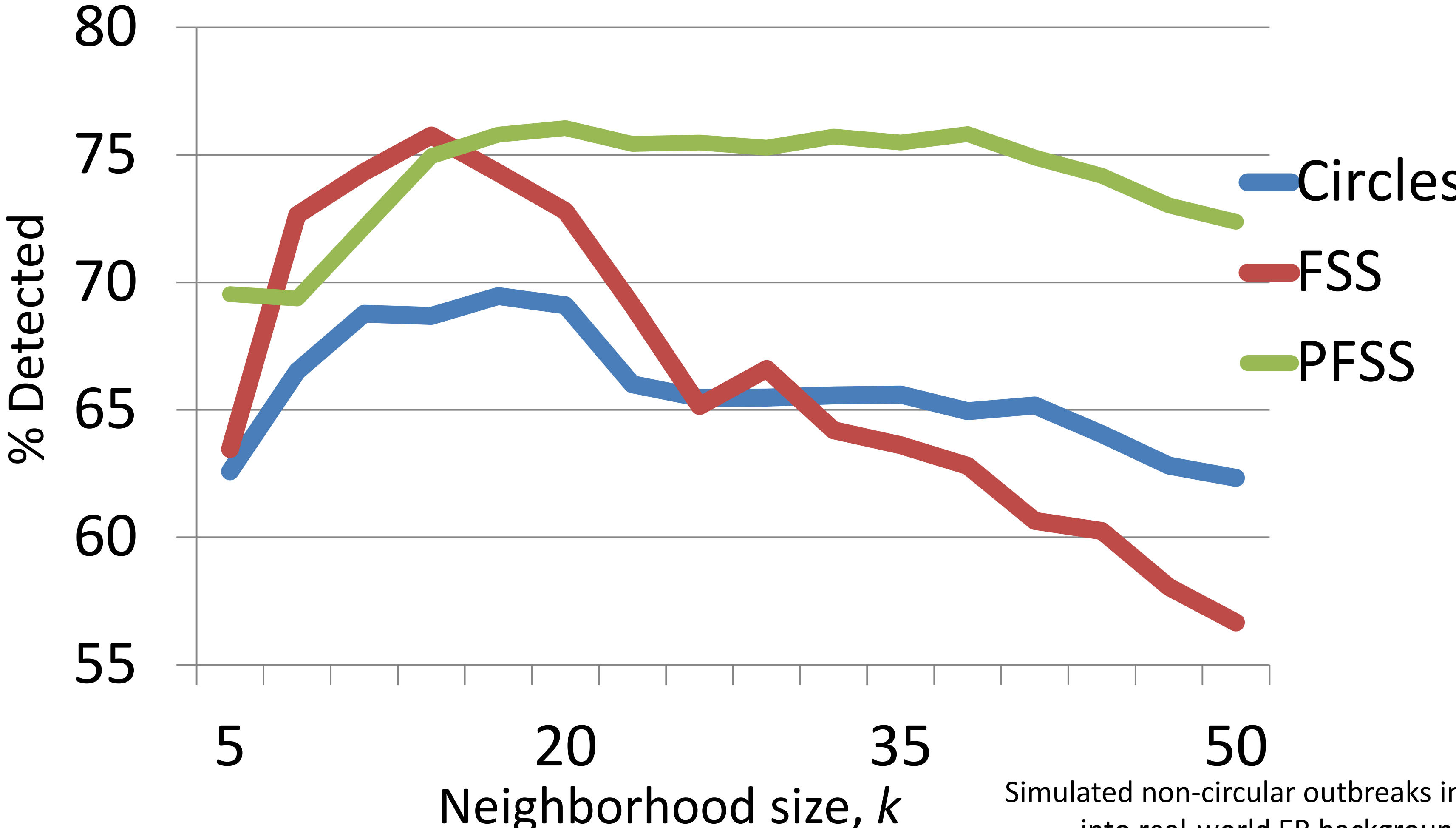
that do not reflect an outbreak of disease.

This particular subset would be less likely returned as the optimal one when compactness constraints are used.

The penalties associated with the distance between the locations and center of the circle would decrease the "score" of the subset

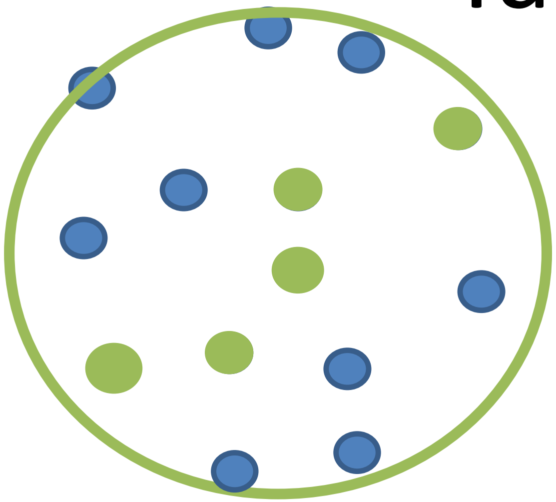
...while increasing the score of compact clusters

Detection Power for Varying Neighborhood Size

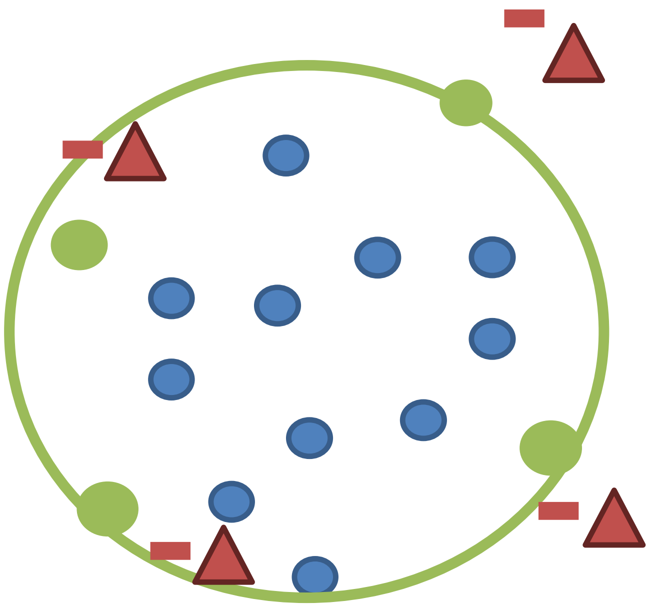


Simulated non-circular outbreaks injected into real-world ER background data. Fixed false positive rate of 1 per year.

Take-Away Message



The subset scanning approach substantially improves detection power of spatial scan statistics for irregular region shapes.



This increased flexibility requires careful choice of neighborhood size, *k*.

Enforcing soft proximity constraints to penalize dispersed subsets addresses this concern and increases overall detection power.

Take-Away Message

Penalized Fast Subset Scanning is very general and provides a framework for incorporating soft constraints into commonly used expectation-based scan statistics.

In the PFSS framework, we demonstrate:

- Exactness: The most anomalous (highest scoring) subset is guaranteed to be identified.
- Efficiency: Only $O(N)$ subsets must be scanned in order to identify the most anomalous penalized subset in a dataset containing N elements (same as the un-penalized scan).
- Interpretability: Soft constraints may be viewed as the prior log-odds for a given record to be included in the most anomalous penalized subset.

Three Contributions

Additive Linear Time Subset Scanning (ALTSS)
property of commonly used
expectation-based scan statistics

Efficient computation of the optimal penalized
subset for functions satisfying ALTSS

One example of penalty terms:
soft proximity constraints

Expectation-based Scan Statistics

$$F(S) = \log \frac{P(Data \mid H_1(S))}{P(Data \mid H_0)}$$

$$H_0 : x_i \sim \mathrm{Dist}(\mu_i)$$

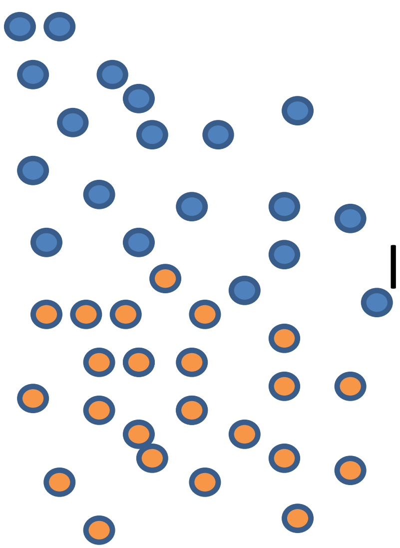
$$H_1(S) : x_i \sim \mathrm{Dist}(q\mu_i) \text{ in } S$$

Expectation-based Scan Statistics

$$F(S) = \max_{q>1} \log \frac{P(Data \mid H_1(S))}{P(Data \mid H_0)}$$

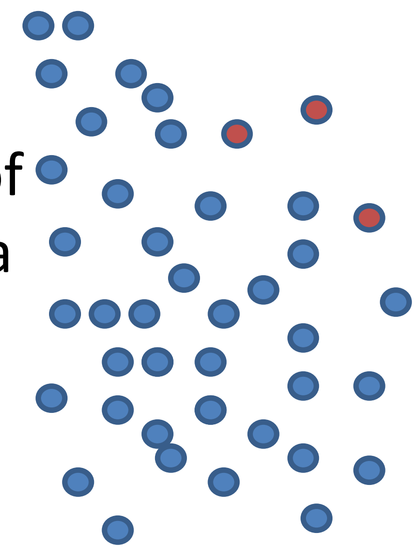
$$H_0 : x_i \sim \text{Dist}(\mu_i)$$

$$H_1(S) : x_i \sim \text{Dist}(q\mu_i) \text{ in } S$$



Large number locations with a moderate risk

Small number of locations with a high risk



Additive Linear Time Subset Scanning

$$F(S) = \max_{q>1} \log \frac{P(Data \mid H_1(S))}{P(Data \mid H_0)}$$

$$H_0 : x_i \sim \mathrm{Dist}(\mu_i)$$
$$H_1(S) : x_i \sim \mathrm{Dist}(q\mu_i) \text{ in } S$$

Definition: For a given dataset *D*, the score function *F(S)* satisfies the Additive Linear Time Subset scanning property if for all $S \subseteq D$ we have

$$F(S) = \max_{q>1} F(S|q) \text{ where } F(S|q) = \textstyle\sum_{s_i \in S} \lambda_i$$

and λ_i depends only on the observed count x_i, expected count μ_i, and the relative risk, *q*.

Additive Linear Time Subset Scanning

$$F(S) = \max_{q>1} \log \frac{P(Data \mid H_1(S))}{P(Data \mid H_0)}$$

$$H_0 : x_i \sim \text{Dist}(\mu_i)$$

$$H_1(S) : x_i \sim \text{Dist}(q\mu_i) \text{ in } S$$

Conditioning ALTSS functions on the relative risk, *q,* allows the function to be written as an ***additive*** set function over the data elements s_i contained in *S*.

Poisson example:

$$F(S) = \max_{q>1} \sum_{s_i \in S} x_i(\log q) + \mu_i(1-q)$$

Additive Linear Time Subset Scanning

Consequence #1: Extremely easy to maximize by including "positive" elements and excluding "negative".

Consequence #2: Additional, element-specific, terms may be added to the scoring function while maintaining the additive property.

$$\boldsymbol{F(S)} = \max_{\boldsymbol{q>1}} \sum_{\boldsymbol{s_i \in S}} \left[x_i(\log q) + \mu_i(1 - q) + \boldsymbol{\Delta_i} \right]$$

Additive Linear Time Subset Scanning

Consequence #1: Extremely easy to maximize by including "positive" elements and excluding "negative".

Consequence #2: Additional, element-specific, terms may be added to the scoring function while maintaining the additive property.

$$F_{penalized}(S) = \max_{q>1} \sum_{s_i \in S} [\, \overbrace{x_i(\log q) + \mu_i(1-q) + \boldsymbol{\Delta_i}}^{\text{“Total Contribution” } \gamma_i \text{ of record } s_i \text{ for fixed risk, } q} \,]$$

Additive Linear Time Subset Scanning

Consequence #1: Extremely easy to maximize by including "positive" elements and excluding "negative".

Consequence #2: Additional, element-specific, terms may be added to the scoring function while maintaining the additive property.

$$F_{penalized}(S) = \max_{q>1} \sum_{s_i \in S} \overbrace{[\quad \lambda_i \quad + \quad \Delta_i \quad]}^{\text{“Total Contribution” } \gamma_i \text{ of record } s_i \text{ for fixed risk, } q}$$

Additive Linear Time Subset Scanning

Distribution	$\lambda_i(q)$
Poisson	$x_i \log q + \mu_i(1-q)$
Gaussian	$x_i\mu_i \frac{(q-1)}{\sigma_i^2} + \mu_i^2\left(\frac{1-q^2}{2\sigma_i^2}\right)$
exponential	$\frac{x_i}{\mu_i}(1-\frac{1}{q}) - \log q$
binomial	$x_i \log(q) + (n_i - x_i)\log\left(\frac{n_i - q\mu_i}{n_i - \mu_i}\right)$
negative binomial	$x_i \log(q) + (r_i + x_i)\log\left(\frac{r_i + \mu_i}{r_i + q\mu_i}\right)$

Three Contributions

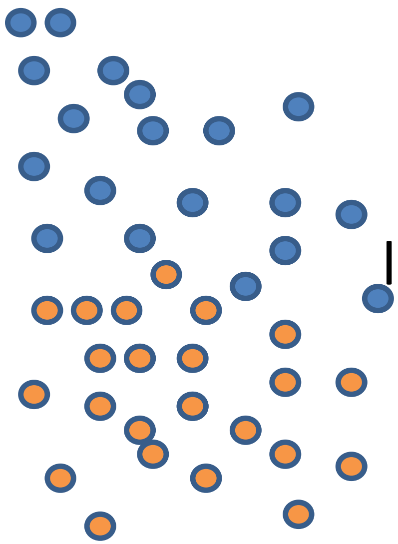
Additive Linear Time Subset Scanning (ALTSS) property of commonly used expectation-based scan statistics

Efficient computation of the optimal penalized subset for functions satisfying ALTSS

One example of penalty terms: soft proximity constraints

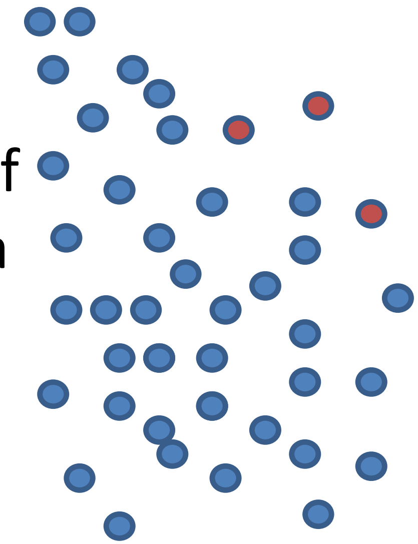
Penalized Fast Subset Scanning

$$F(S) = \max_{q>1} \log \frac{P(Data \mid H_1(S))}{P(Data \mid H_0)}$$



Large number locations with a moderate risk

Small number of locations with a high risk

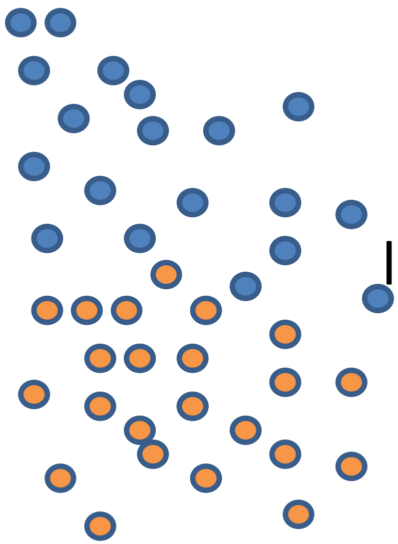


… but the ALTSS property requires evaluating the function at a *fixed* risk.

How do we optimize over the entire range *q > 1* ?

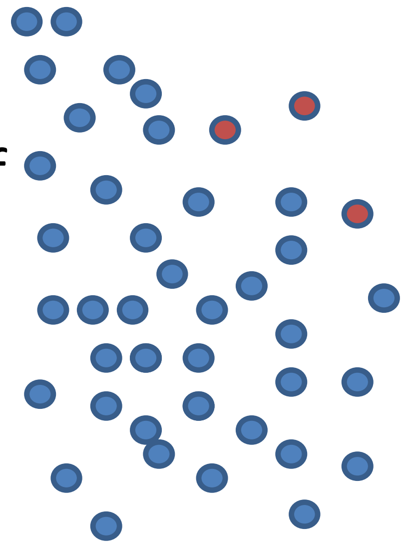
Penalized Fast Subset Scanning

$$F(S) = \max_{q>1} \log \frac{P(Data \mid H_1(S))}{P(Data \mid H_0)}$$



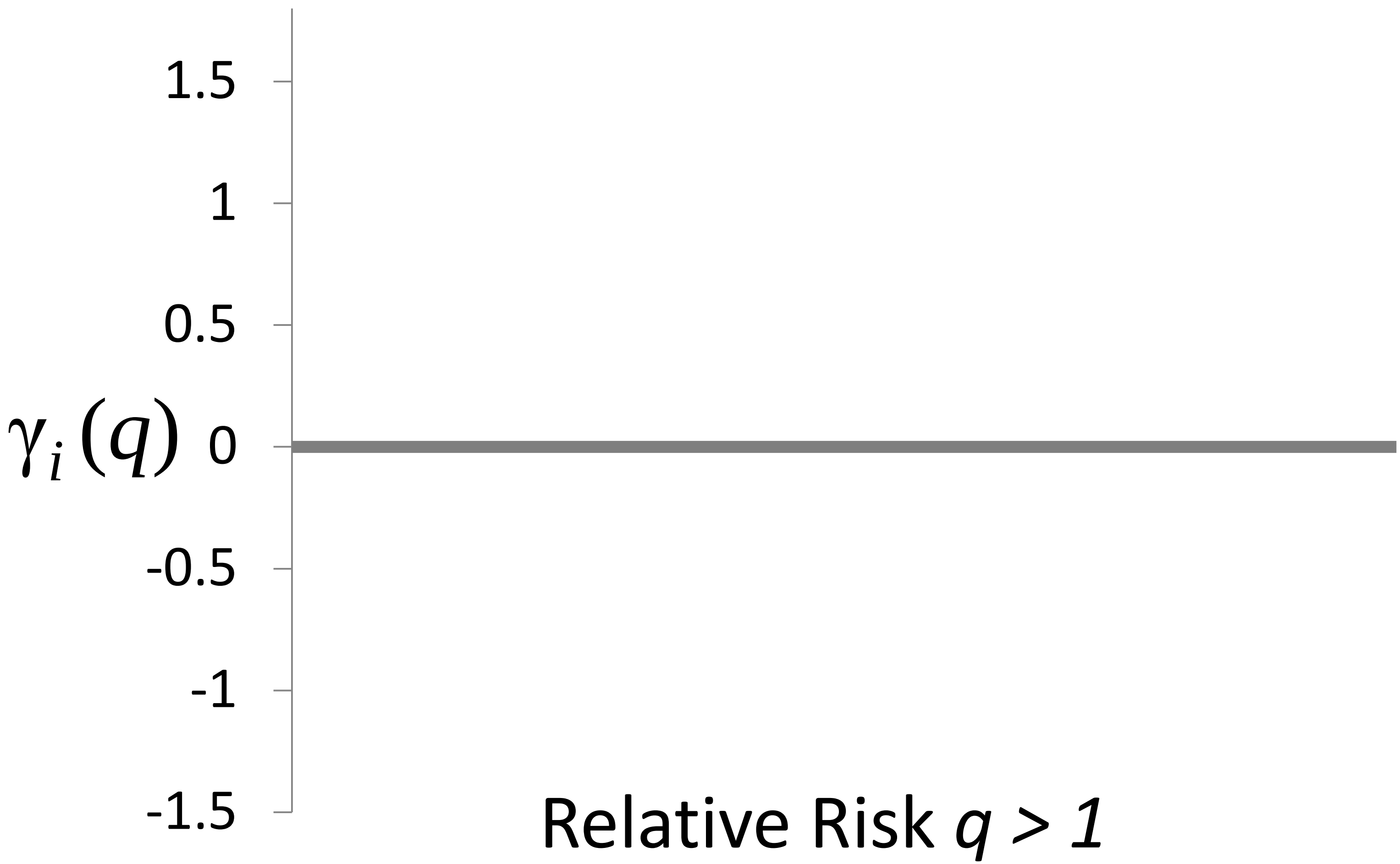
Large number locations with a moderate risk

Small number of locations with a high risk



Theorem: The optimal subset $S^* = \arg\max_S F_{pen}(S)$ maximizing a penalized expectation-based scan statistic satisfying the ALTSS property may be found be evaluating only *O(N)* subsets, where *N* is the total number of data elements.

Proof by Picture

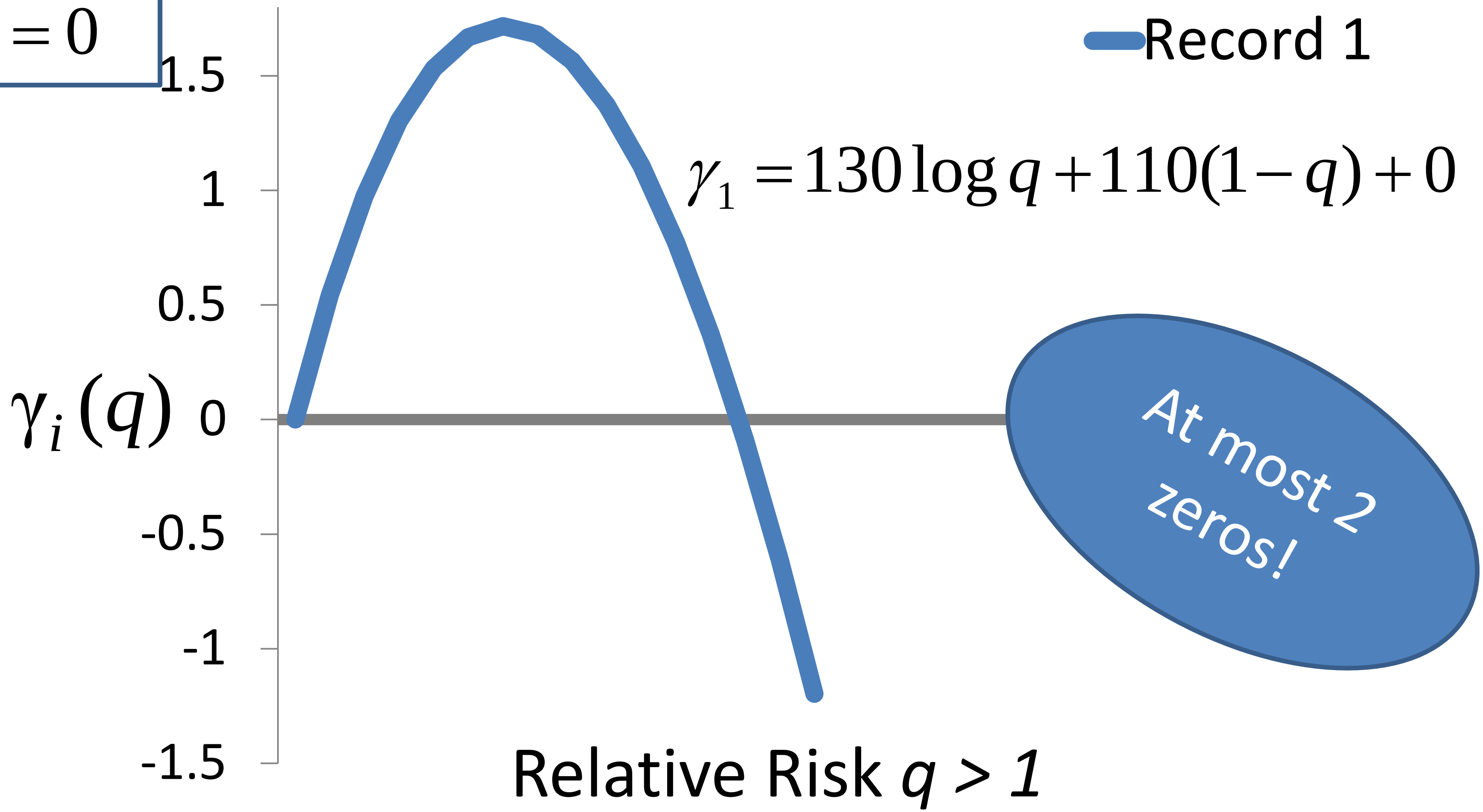


$x_1 = 130$

$\mu_1 = 110$

$\Delta_1 = 0$

Proof by Picture

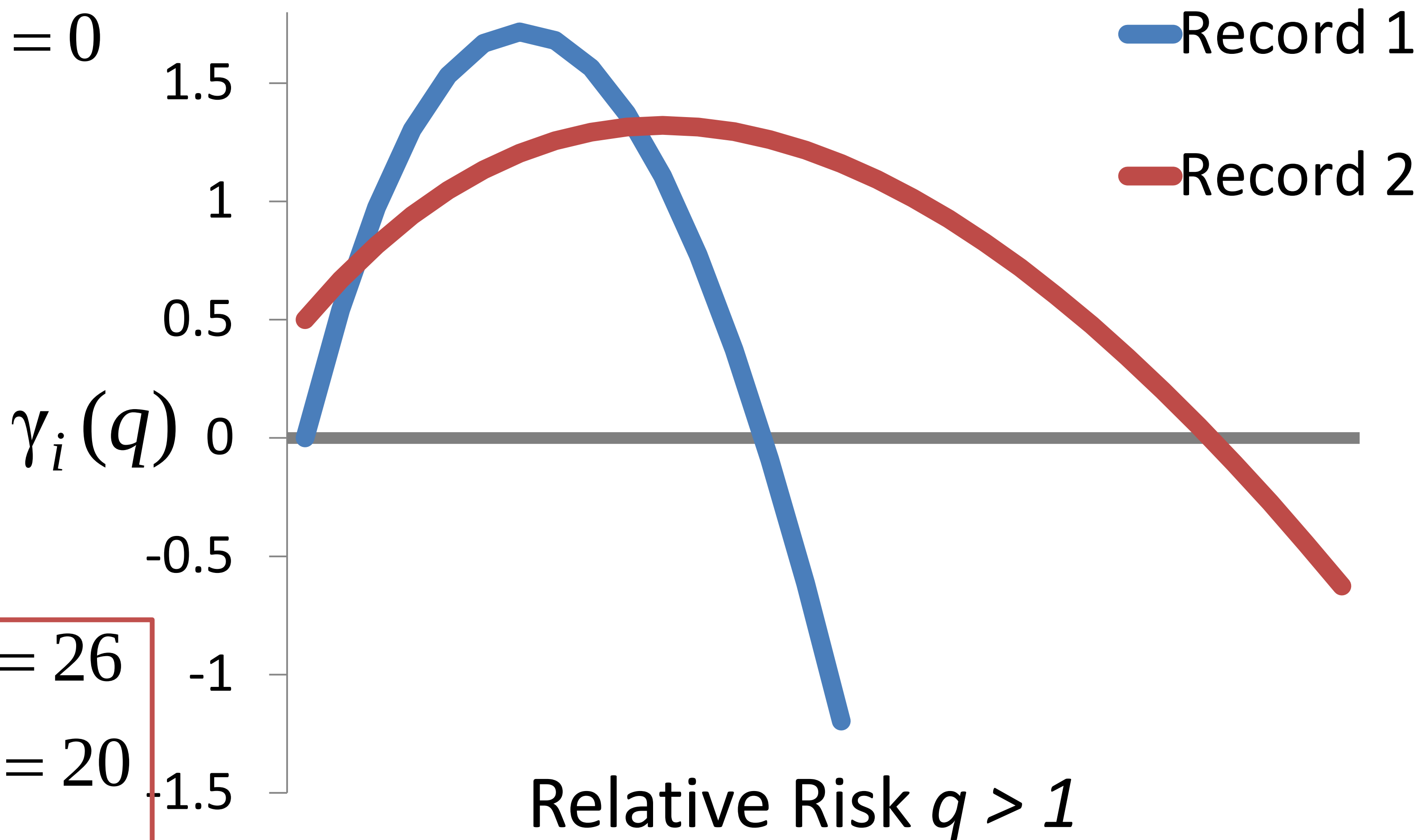


$x_1 = 130$

$\mu_1 = 110$

$\Delta_1 = 0$

Proof by Picture



$x_2 = 26$

$\mu_2 = 20$

$\Delta_2 = 0.5$

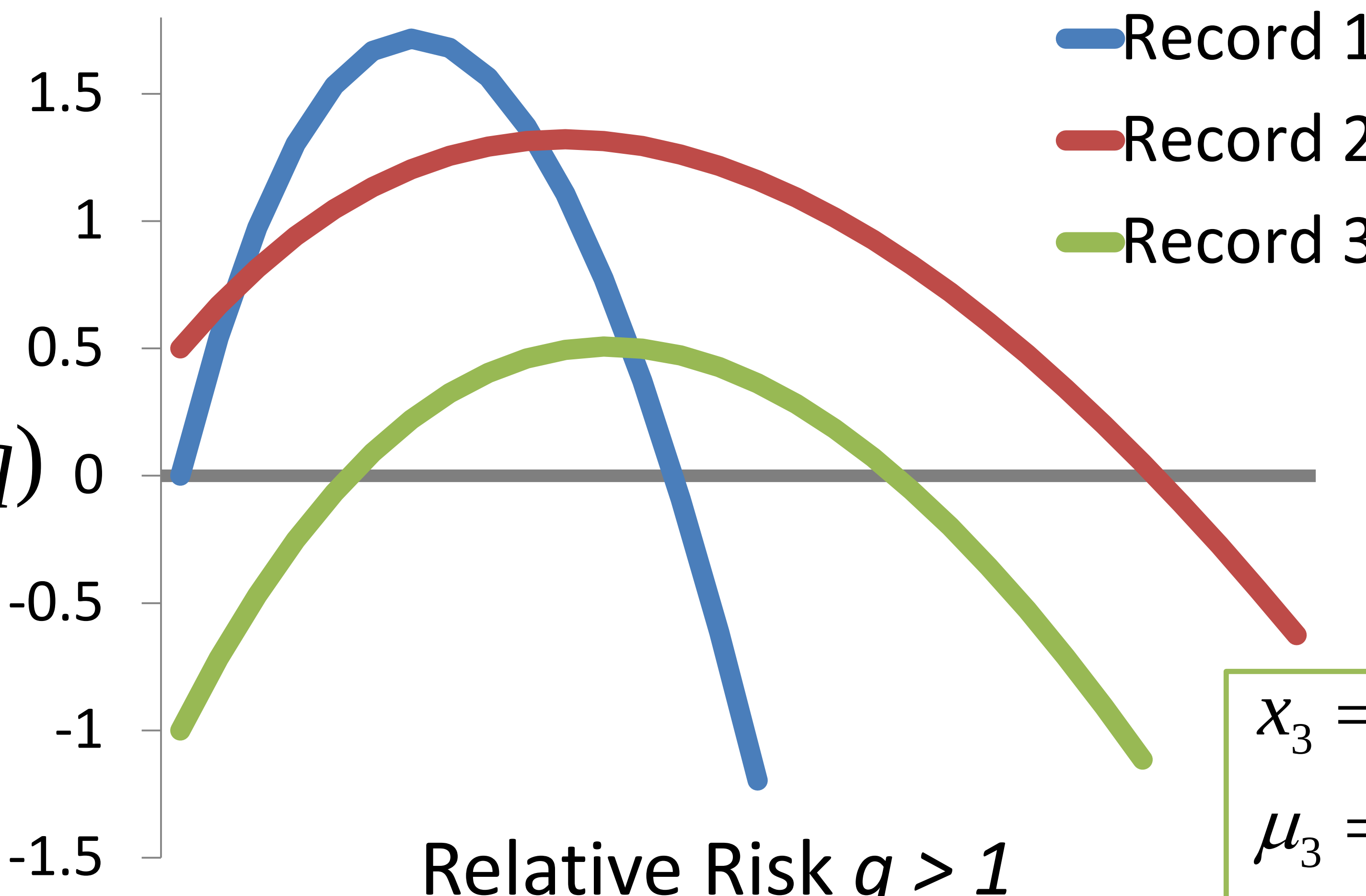
Proof by Picture

$x_1 = 130$

$\mu_1 = 110$

$\Delta_1 = 0$

$\gamma_i(q)$



$x_2 = 26$

$\mu_2 = 20$

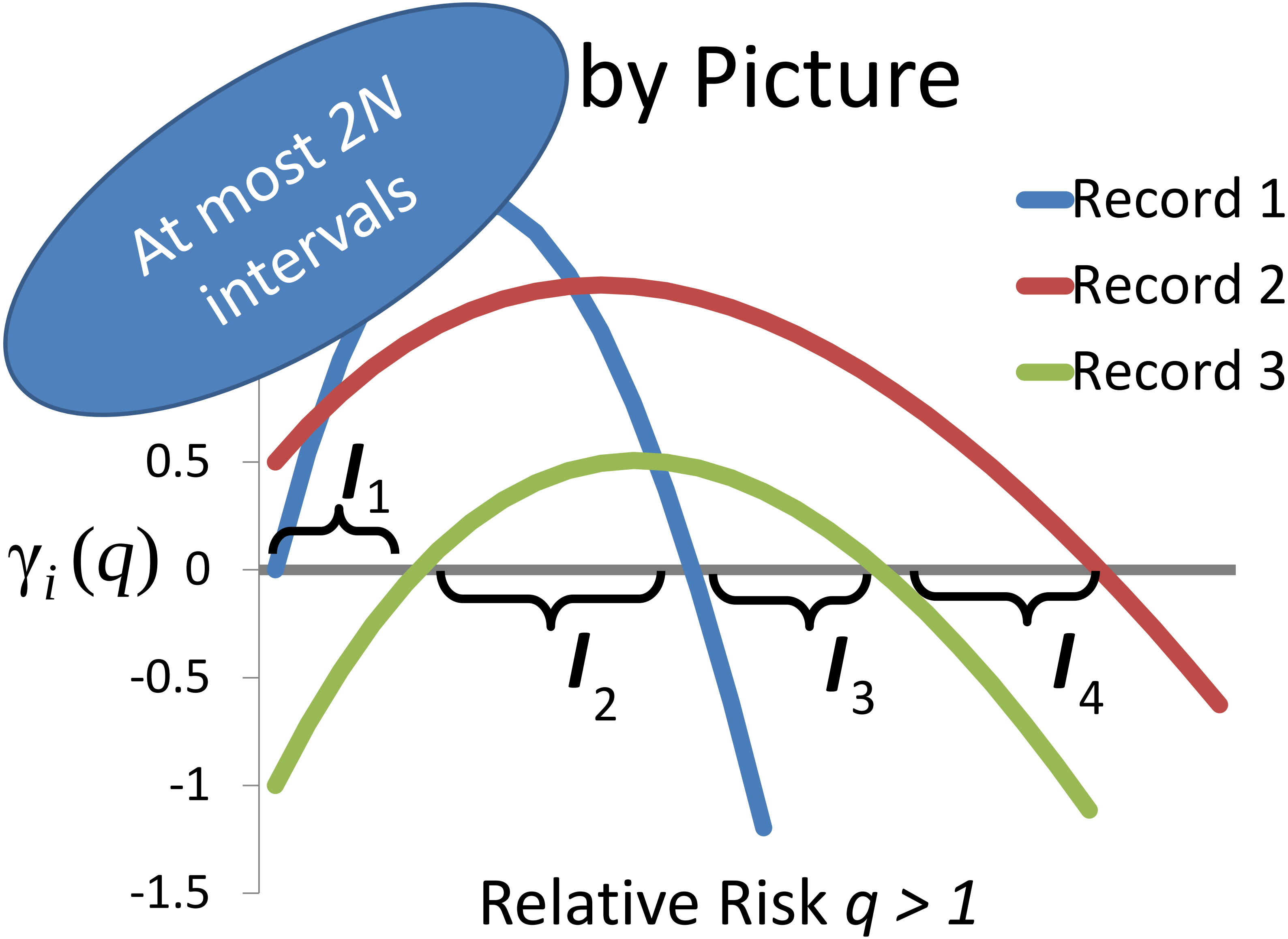
$\Delta_2 = 0.5$

$x_3 = 40$

$\mu_3 = 30$

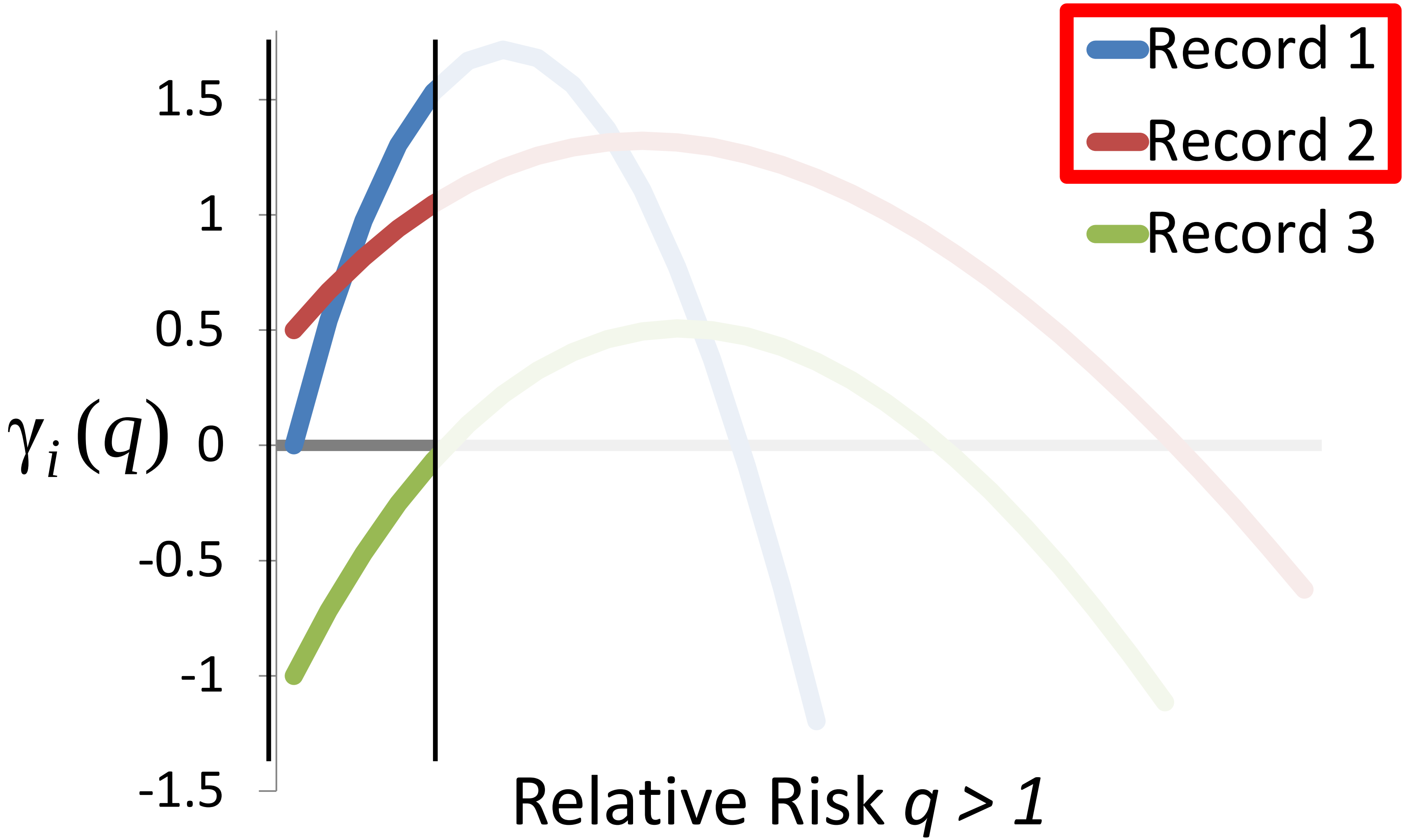
$\Delta_3 = -1$

by Picture

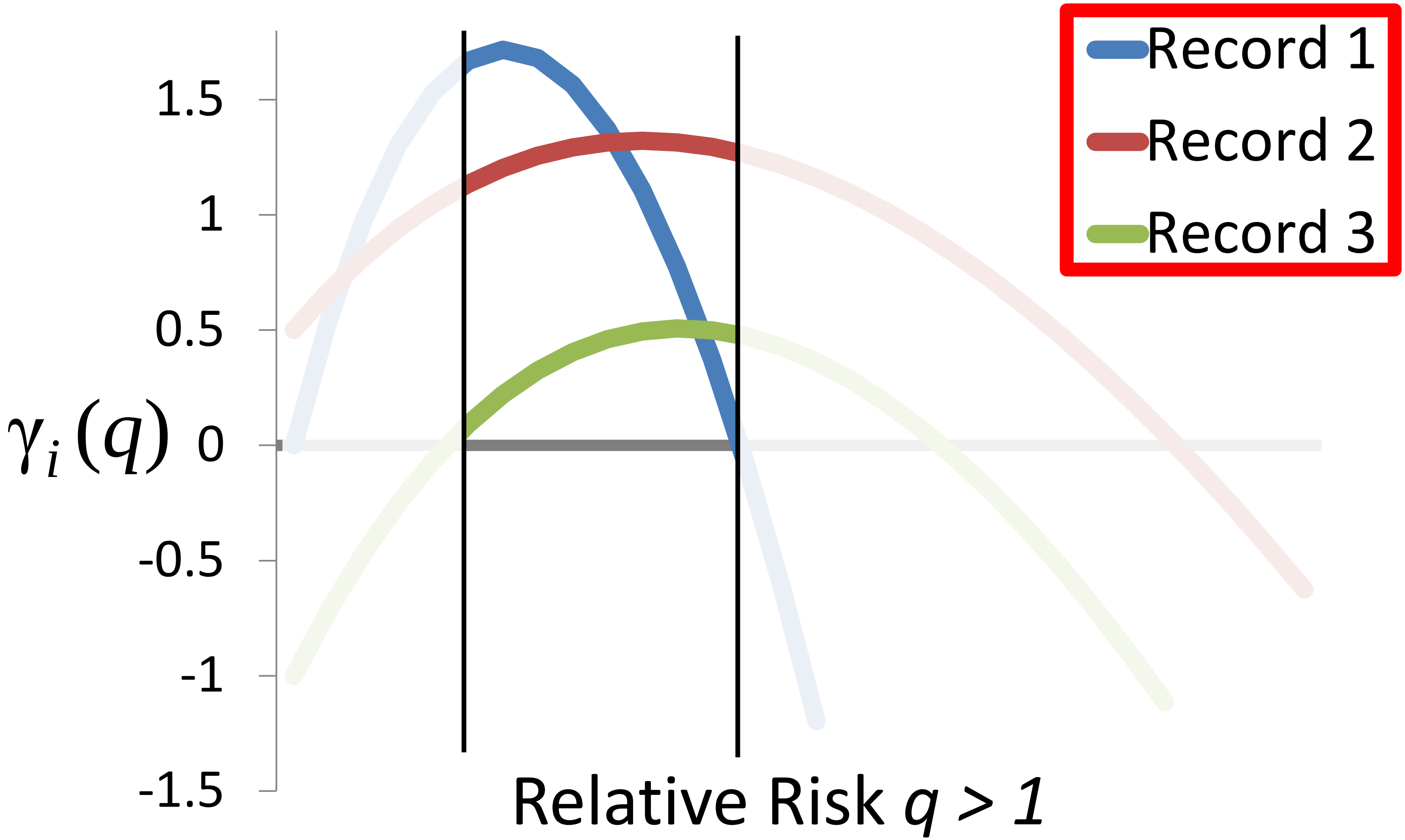


At most 2N intervals
Record 1
Record 2
Record 3
0.5
0
-0.5
-1
-1.5
$\gamma_i(q)$
I_1
I_2
I_3
I_4
Relative Risk $q > 1$

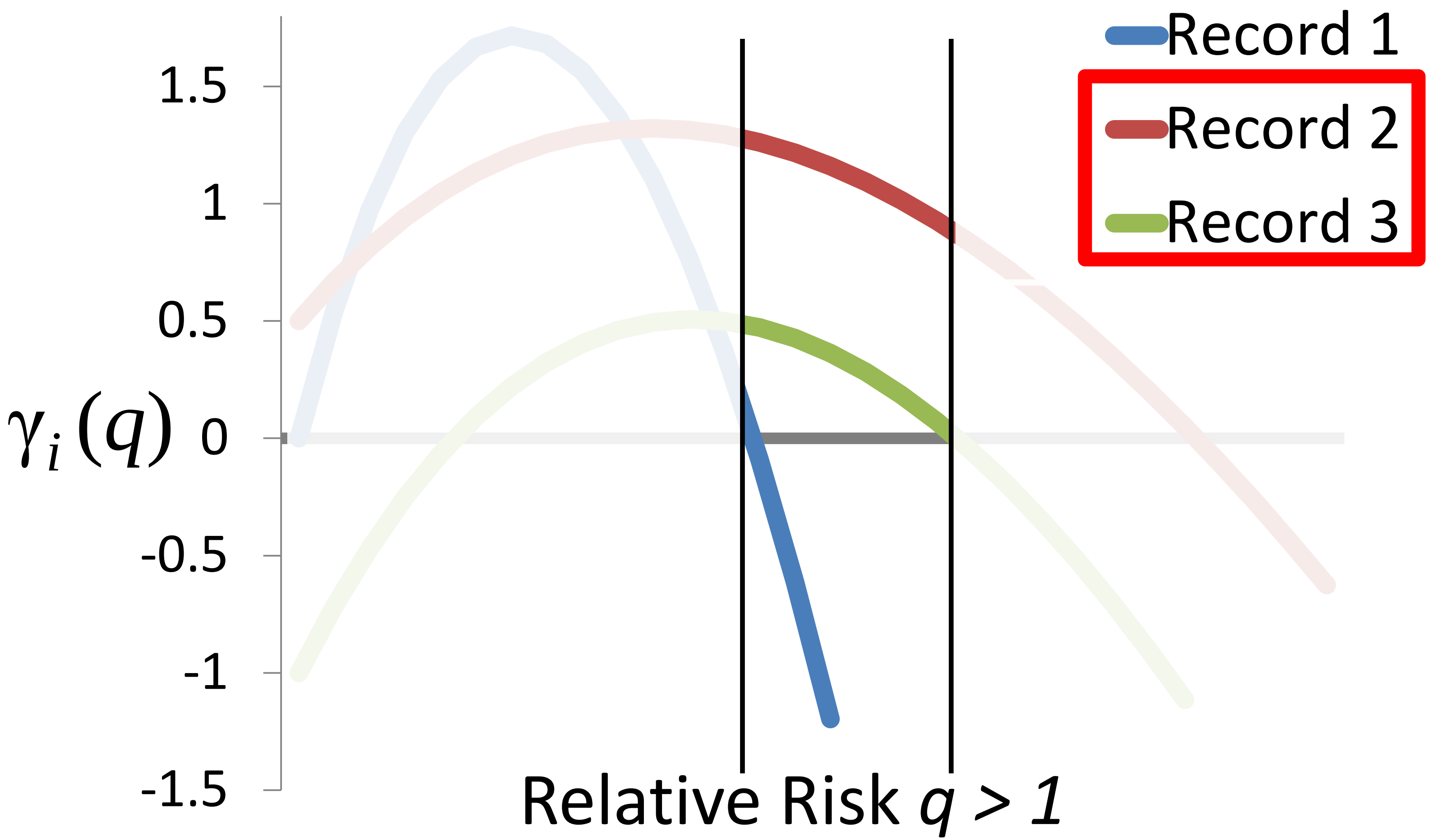
Proof by Picture



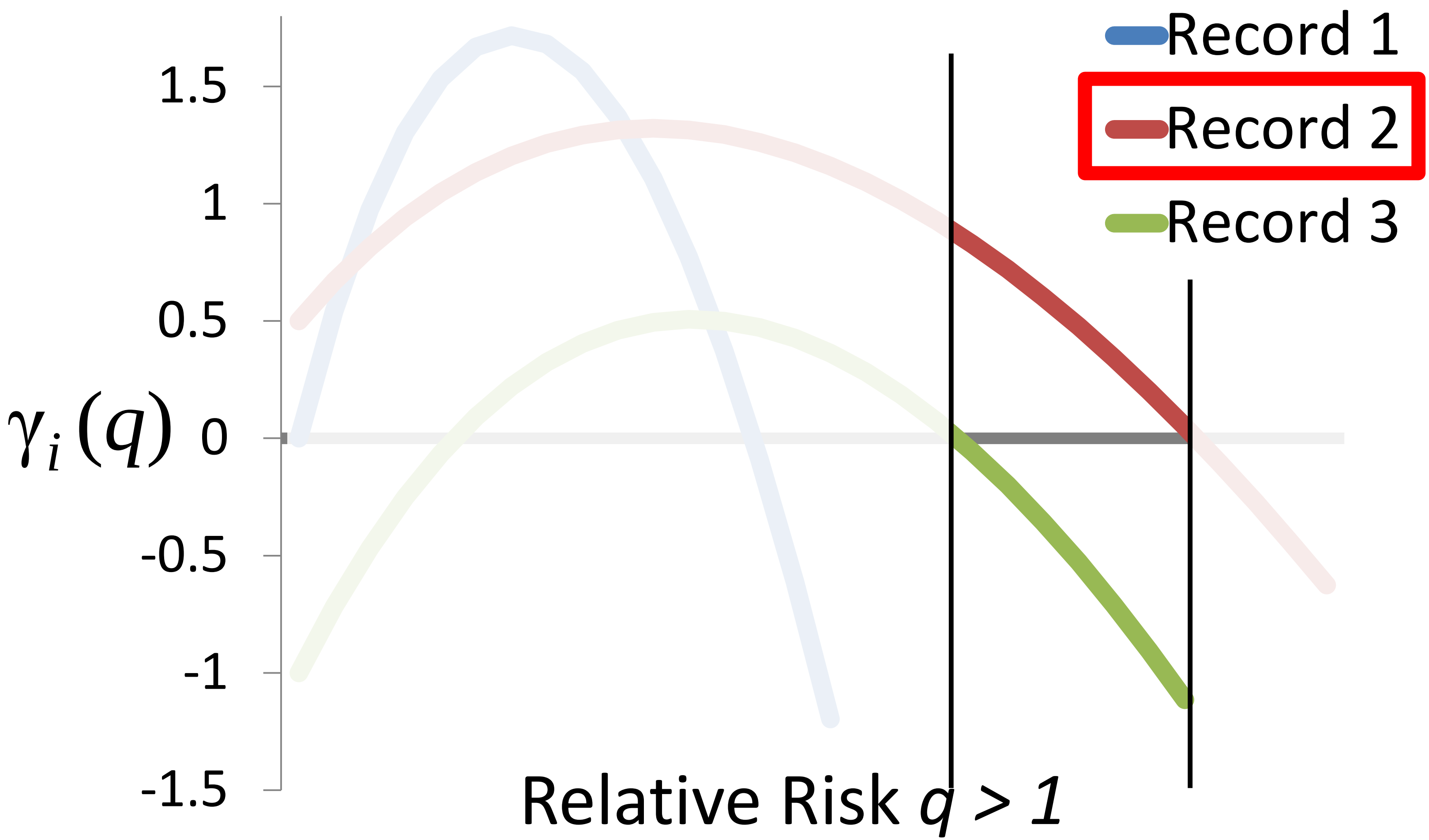
Proof by Picture



Proof by Picture



Proof by Picture



Three Contributions

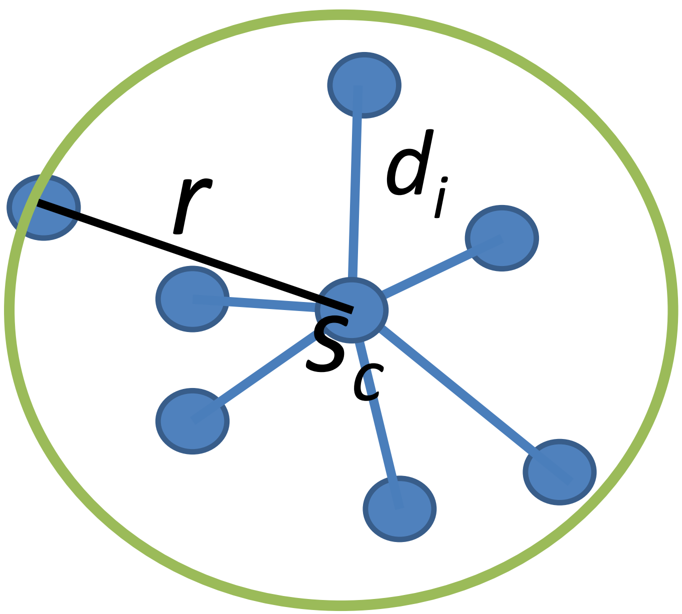
Additive Linear Time Subset Scanning (ALTSS) property of commonly used expectation-based scan statistics

Efficient computation of the optimal penalized subset for functions satisfying ALTSS

One example of penalty terms: soft proximity constraints

Soft Proximity Constraints

Penalized Fast Subset Scanning allows additional spatial information to be included; rewarding spatial compactness and penalizing dispersed subsets within a local neighborhood.



Center location and its *k-1* nearest neighbors

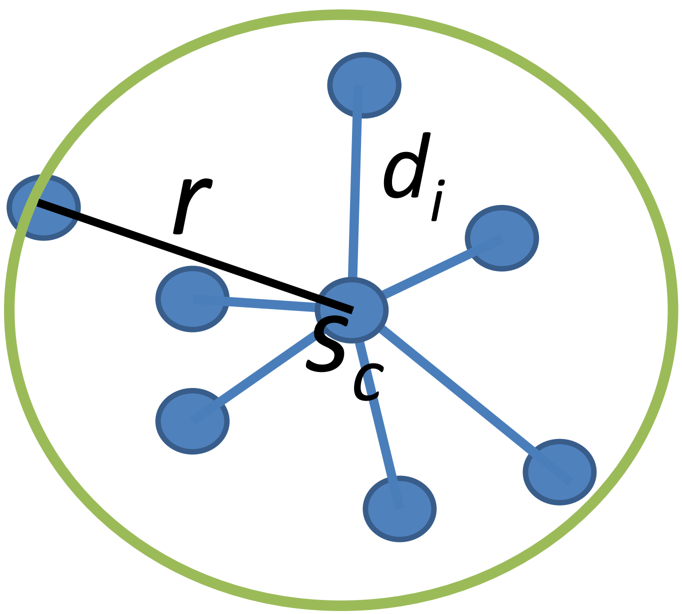
$$\Delta_i = h\left(1 - \frac{2d_i}{r}\right)$$

h is the strength of the constraint

$$\Delta_i \in [-h \ldots h]$$

Soft Proximity Constraints

Penalty terms may be interpreted as prior log-odds for a location to be included in the subset.



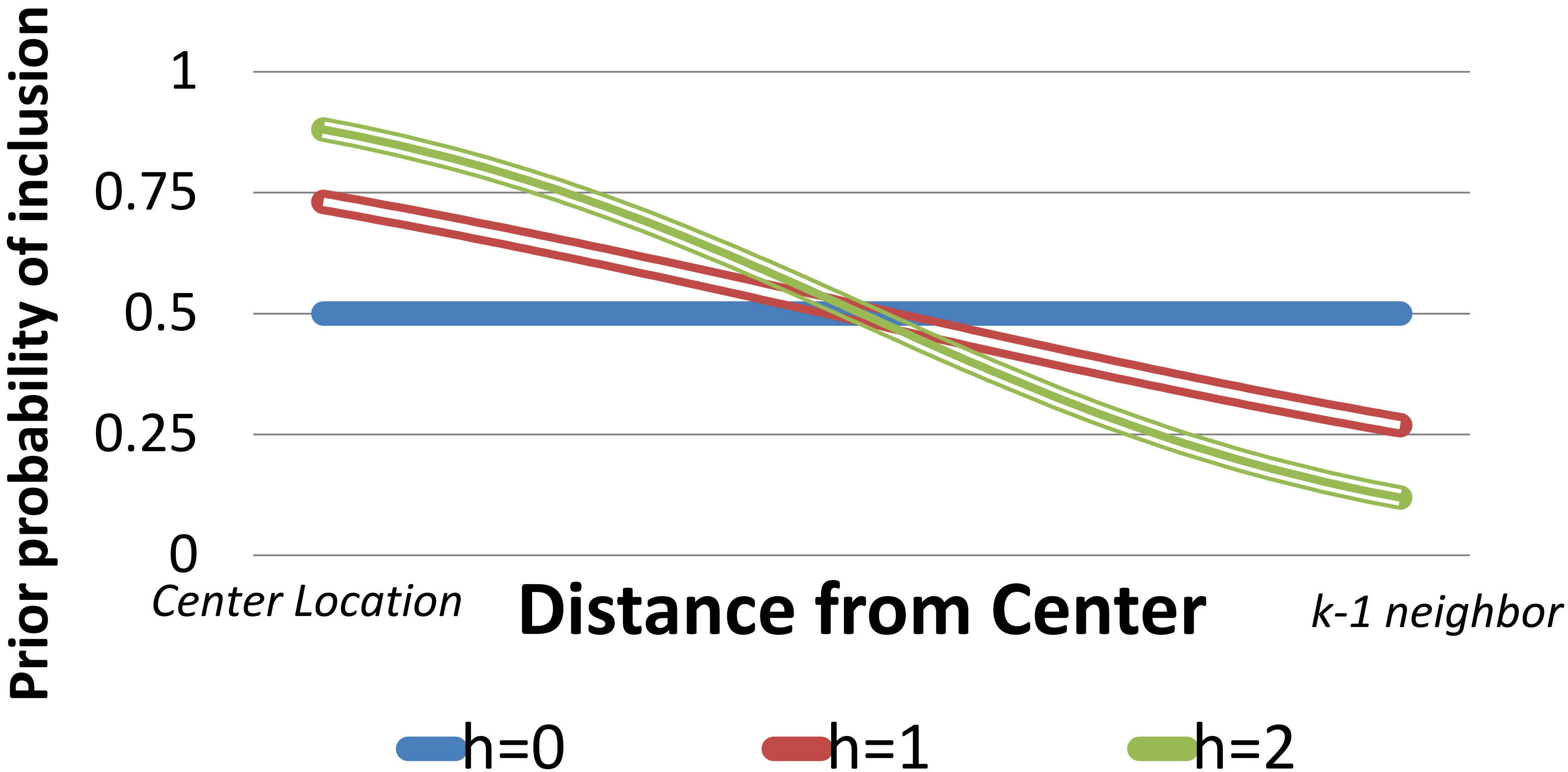
Center location and its *k-1* nearest neighbors

$$\log\left(\frac{p_i}{1-p_i}\right)=\Delta_i$$

The center location is e^h times more likely to be included in the optimal subset than the *k-1* nearest neighbor.

Soft Proximity Constraints

Penalty terms may be interpreted as prior log-odds for a location to be included in the subset.



Evaluation: Emergency Department Data

Two years of admissions from Allegheny County Emergency Departments

The patient's home zip code is used to tally the counts at each location

Centroids of 97 Zip Codes were used as locations

Bayesian Aerosol Release Detector (BARD)

Hogan et al; 2007

Simulates anthrax spores released over a city

Two models drive the simulator:

Dispersion

Which areas will be affected?

Weather data

Gaussian plumes

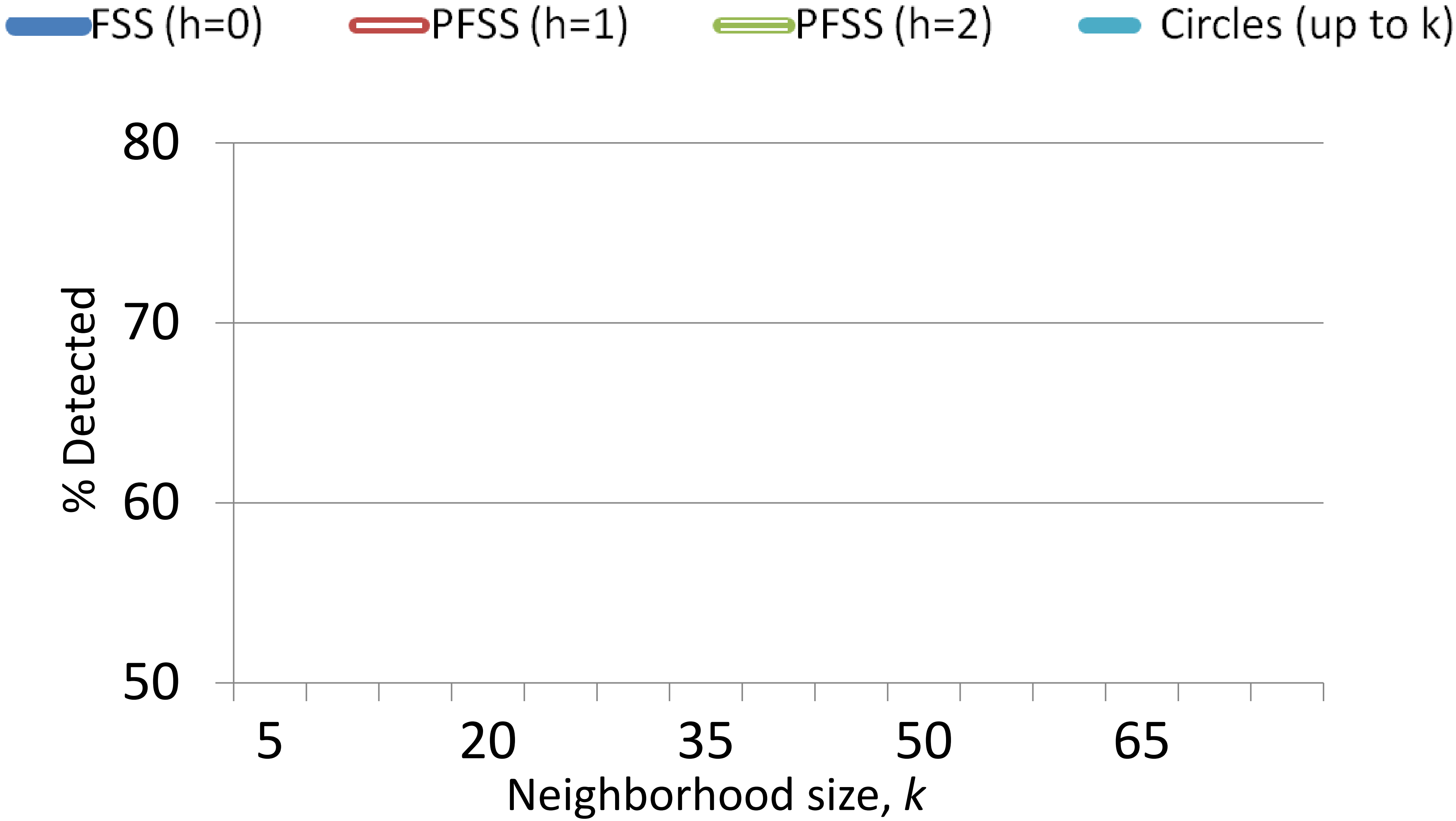
Infection

How many infected people in an area?

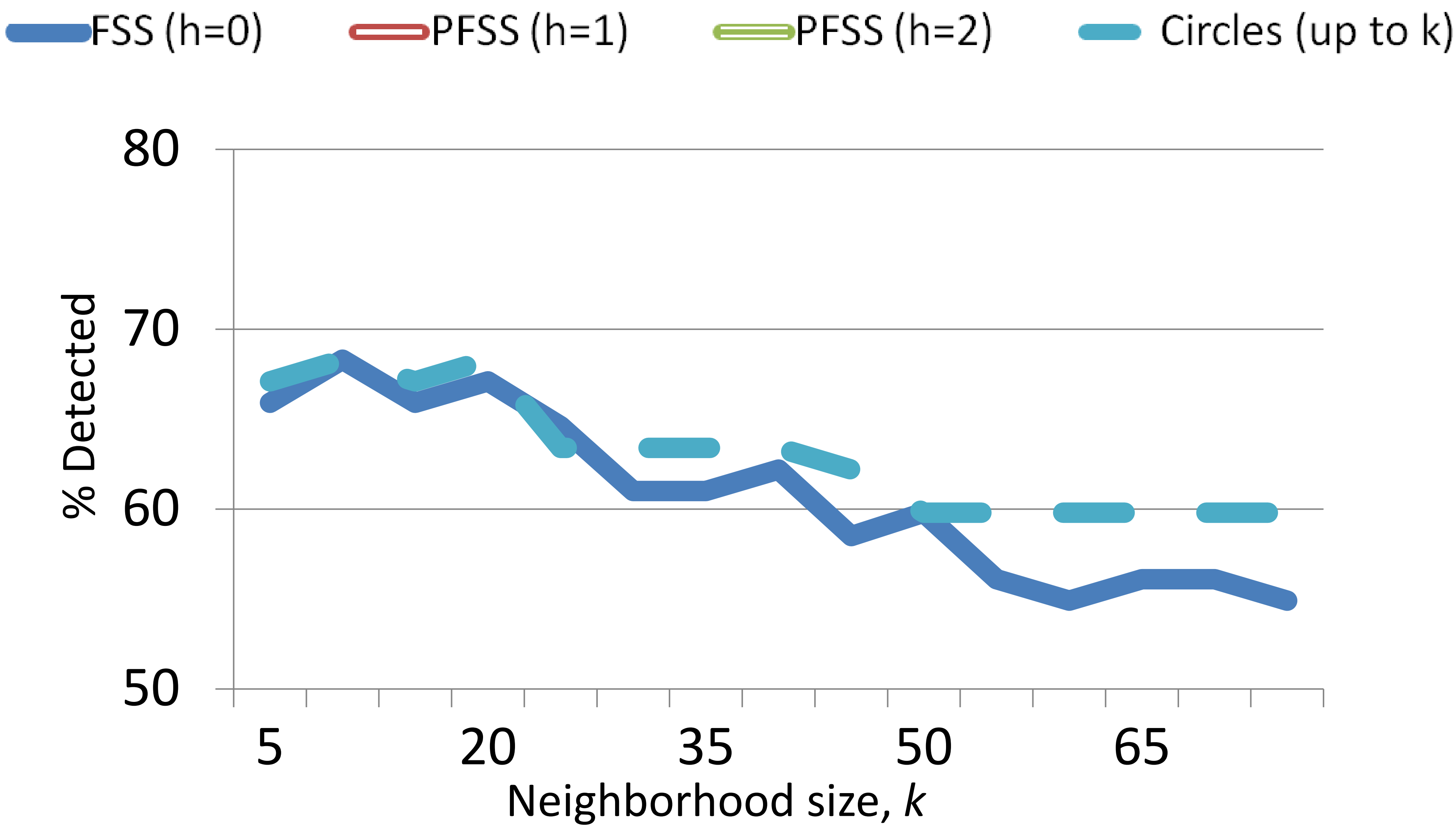
Demographic data

Increased ER visits with respiratory complaints

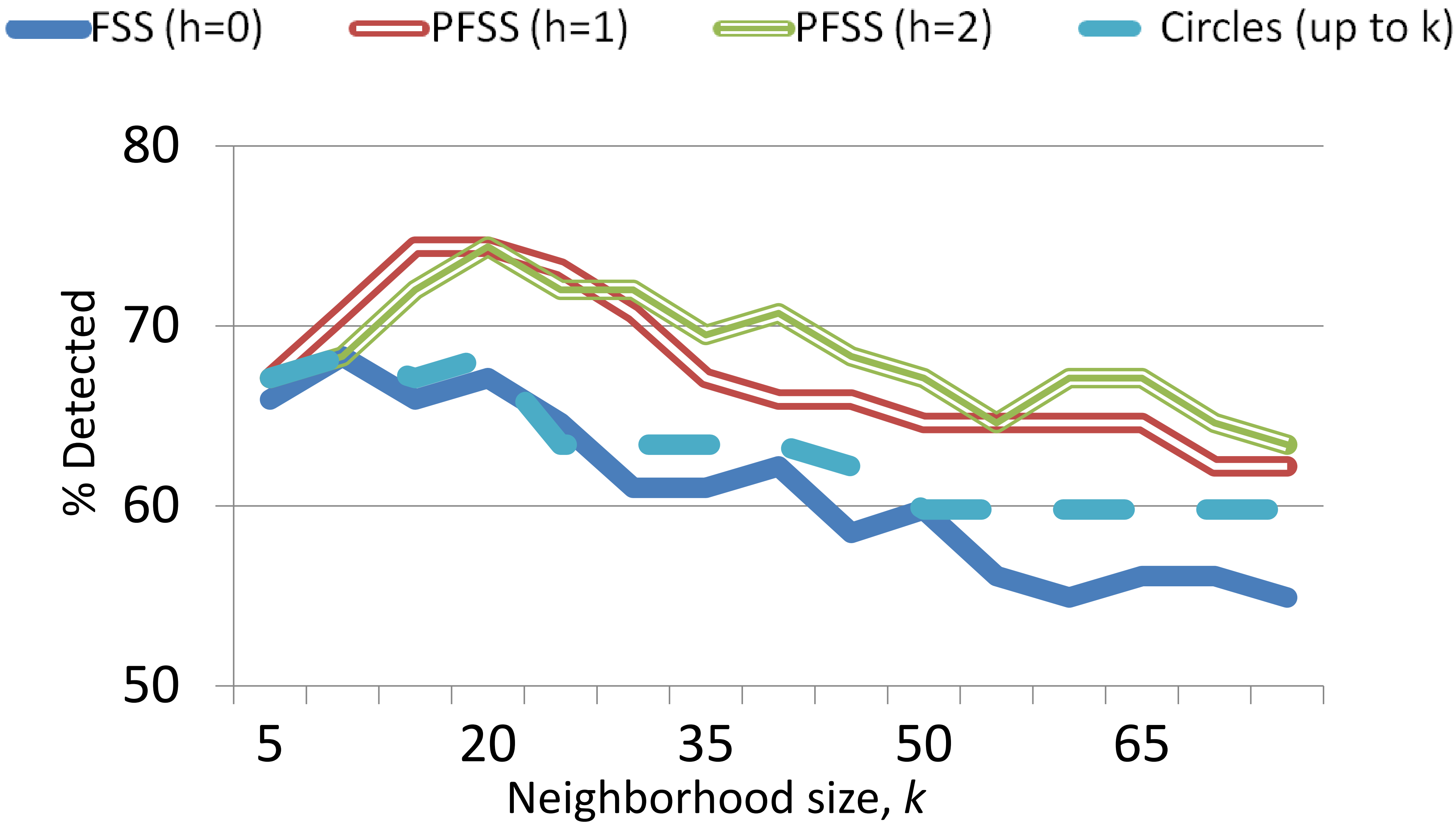
Comparison of Detection Power for BARD Simulated Attacks



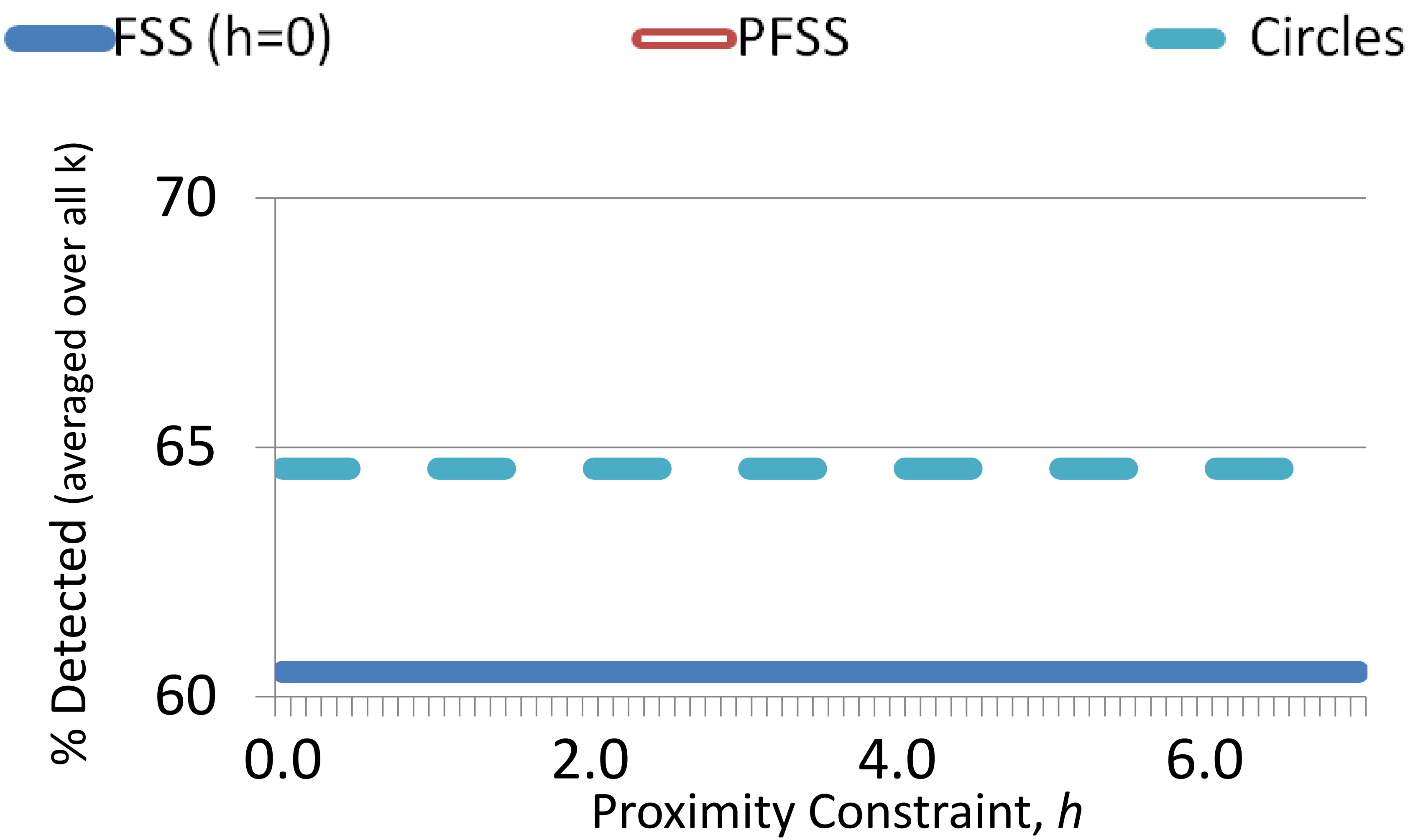
Comparison of Detection Power for BARD Simulated Attacks



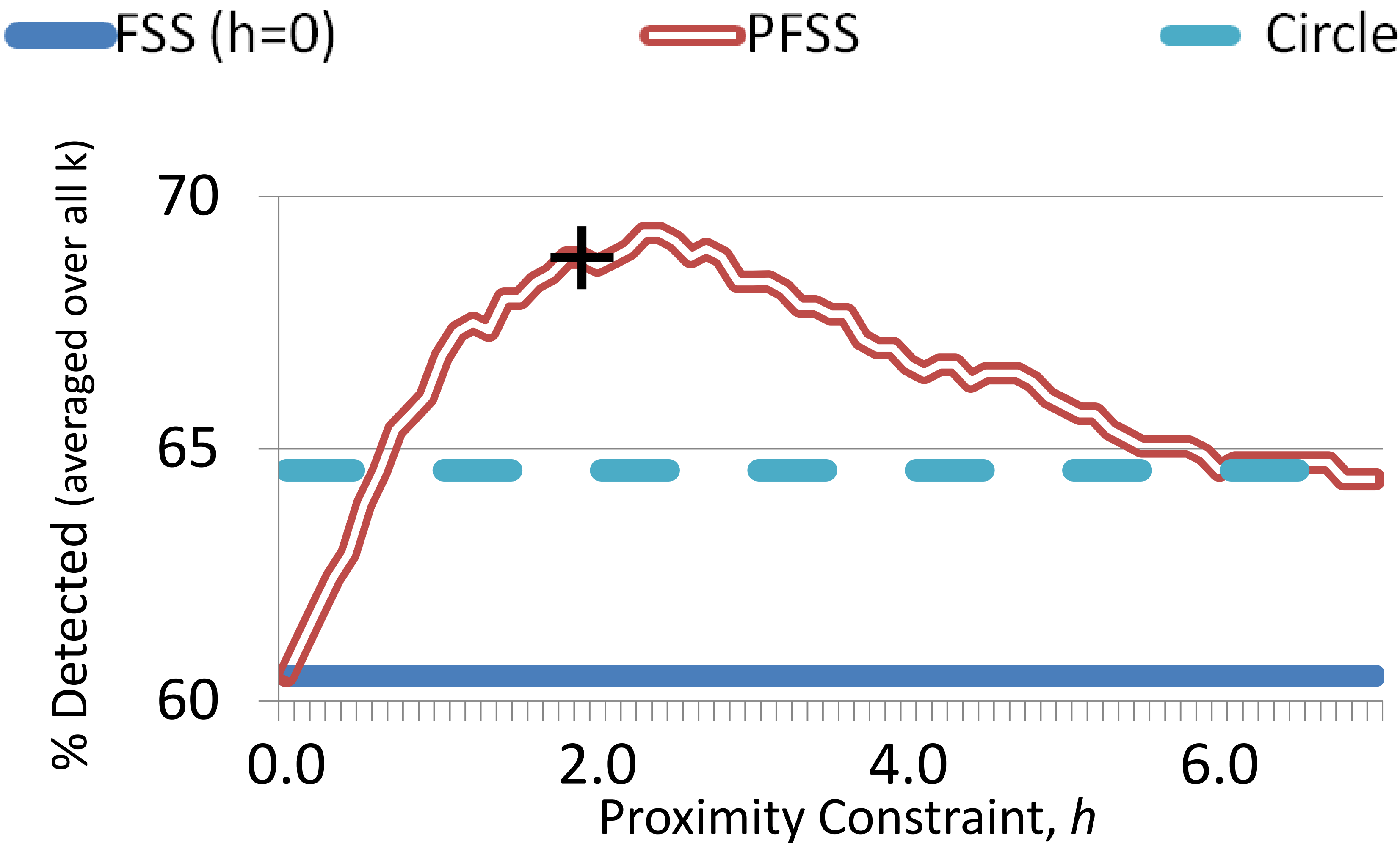
Comparison of Detection Power for BARD Simulated Attacks



Average Detection Power for Varying Proximity Constraint Strength



Average Detection Power for Varying Proximity Constraint Strength



Conclusions

PFSS is very general and provides a framework for incorporating soft constraints into commonly used expectation-based scan statistics.

- **Exact**: The most anomalous (highest scoring) subset is guaranteed to be identified.
- **Efficient**: Only *O(N)* subsets must be scanned in order to identify the most anomalous ***penalized*** subset in a dataset containing *N* elements.
- **Interpretable**: Soft constraints may be viewed as the prior log-odds for a given record to be included in the most anomalous penalized subset.

Conclusions

PFSS is very gene… framework for
incorporati… …nly used

--BONUS!--
Applies to a **larger** class of expectation-based scan statistics (exponential family), as compared to un-penalized Fast Subset Scan.

- **Exact**: Th… …set is guarante…
- **Efficient**… …order to identify t… …n a dataset containing…
- **Interpretabl**… …wed as the prior log-odds for a given … …d in the most anomalous penalized subset.

Interested?

More details on our web page:
http://epdlab.heinz.cmu.edu

Or e-mail me at:
neill@cs.cmu.edu