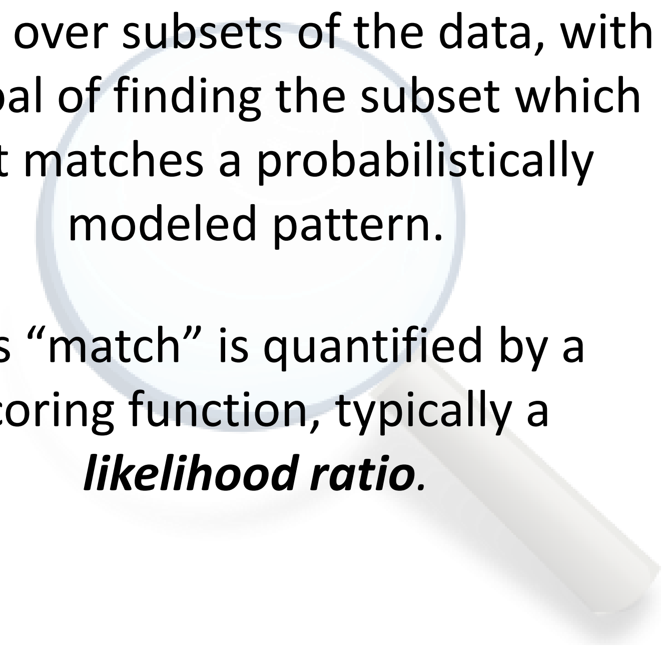
Efficient Subset Scanning with Soft Constraints

Daniel B. Neill
Event and Pattern Detection Laboratory
H.J. Heinz III College, Carnegie Mellon University
neill@cs.cmu.edu

Joint work with Skyler Speakman, Edward McFowland III, and Sriram Somanchi.

This work was partially supported by NSF grants IIS-0916345, IIS-0911032, and IIS-0953330.

Pattern Detection as Subset Scanning



Pattern Detection can be framed as a **search** over subsets of the data, with the goal of finding the subset which best matches a probabilistically modeled pattern.

This “match” is quantified by a scoring function, typically a ***likelihood ratio***.

Computational Problems: Infeasible to perform exhaustive search for more than 30 data records → 2^{30} subsets

Linear-time Subset Scanning (LTSS)

property allows for exact, efficient identification of “highest scoring” subset without an exhaustive search.

Neill, JRSS-B, 2012

GraphScan extended LTSS to only consider ***connected*** subsets. Increases power to detect patterns that affect a subgraph of a larger network.

Speakman & Neill, Proc. ISDS 2010

Subset Scanning with Soft Constraints

Most previous work assumes **hard** constraints,
e.g., the cluster must be **connected**, or have **radius** $\leq$ **r**.

Here we provide a framework for incorporating
"soft constraints" (bonuses or penalties) without
violating the properties that allow for efficient search.

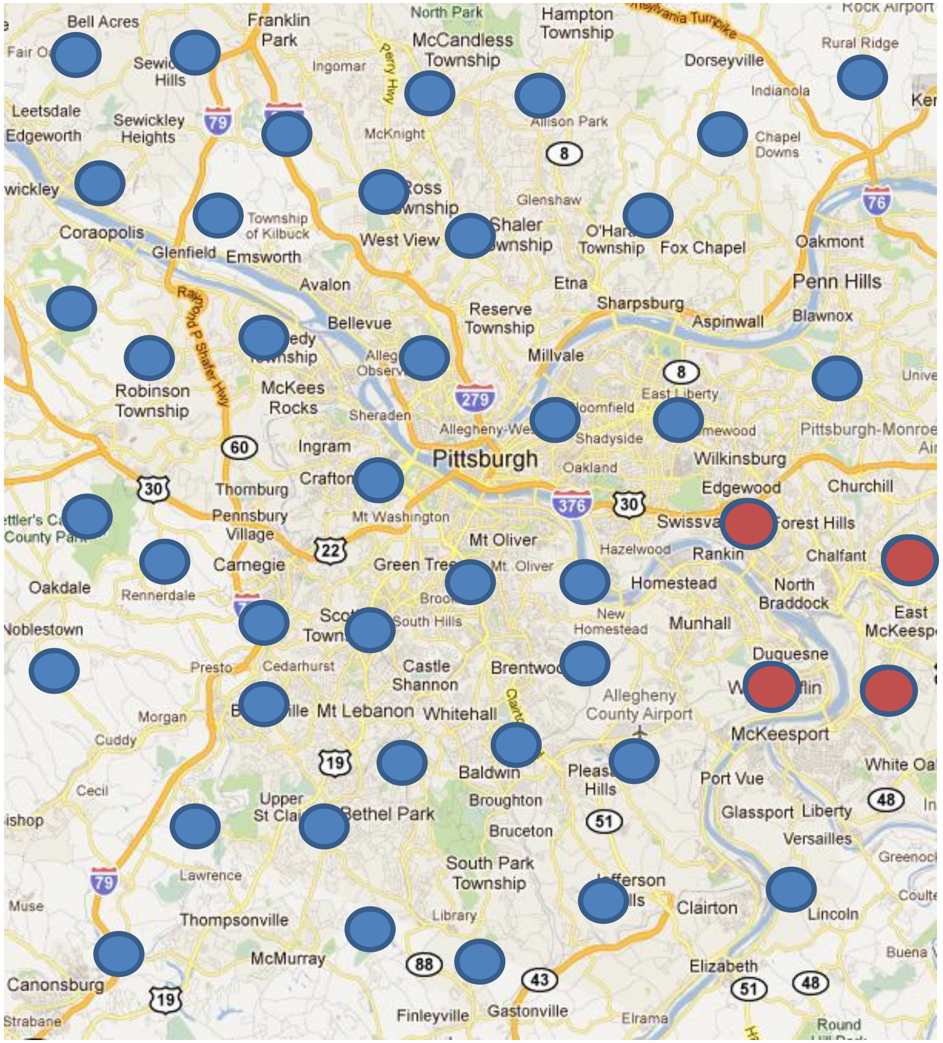
Soft constraints on **compactness**
(prefer more compact spatial clusters)

Example: disease outbreak detection

Soft constraints on **temporal consistency**
(prefer dynamic clusters that change smoothly over time)

Example: detecting spreading contamination in a water distribution network

Example: Detecting Disease Clusters



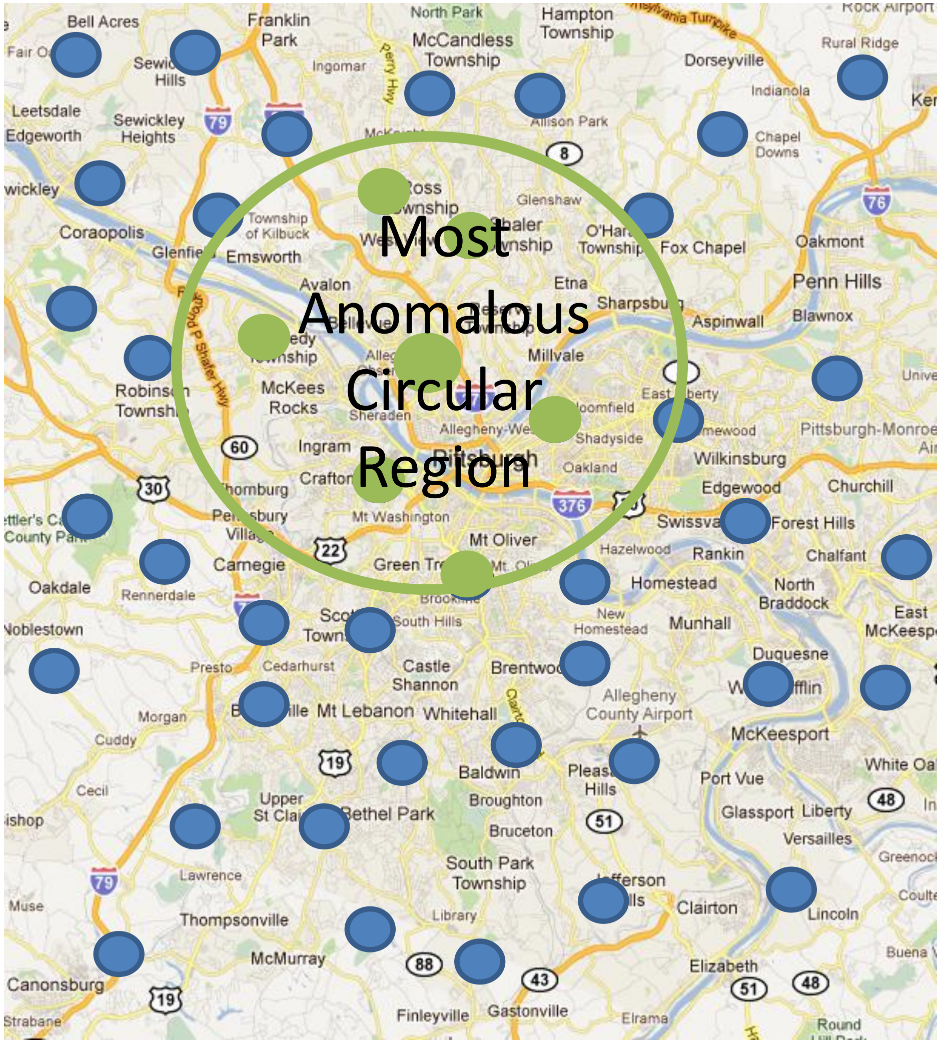
● Location of a monitored data stream

- # of hospital ED visits by zip code
- # of OTC drug sales by zip code

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

An effective detection method should detect an outbreak *early* and have high *spatial accuracy*, while minimizing *false positives*.

Example: Detecting Disease Clusters



(Kulldorff, 1997)

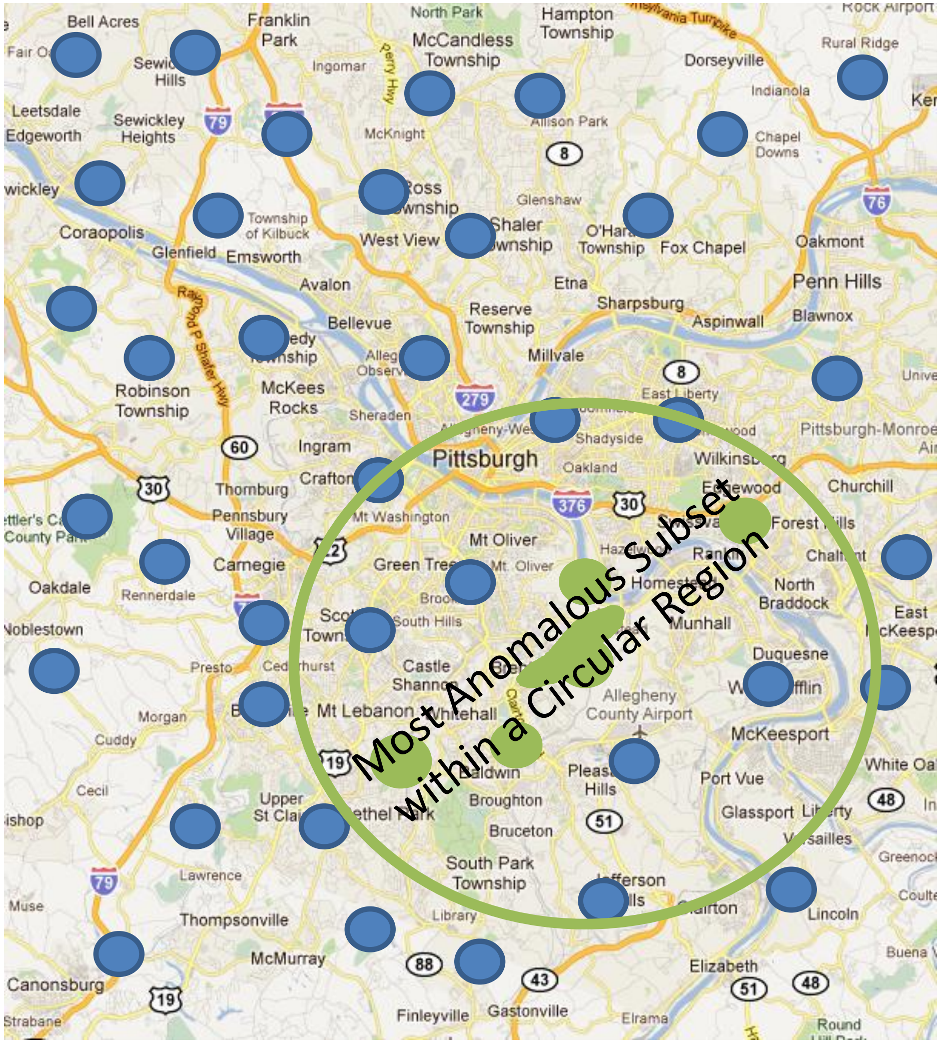
Spatial Scan Statistic
(Circles)

Maximize log-likelihood ratio statistic over circles of varying radius centered at each location.

High power to detect compact clusters (close to circular)

But what about irregular shaped clusters?

Detecting *Irregular* Disease Clusters



(Neill, 2012)

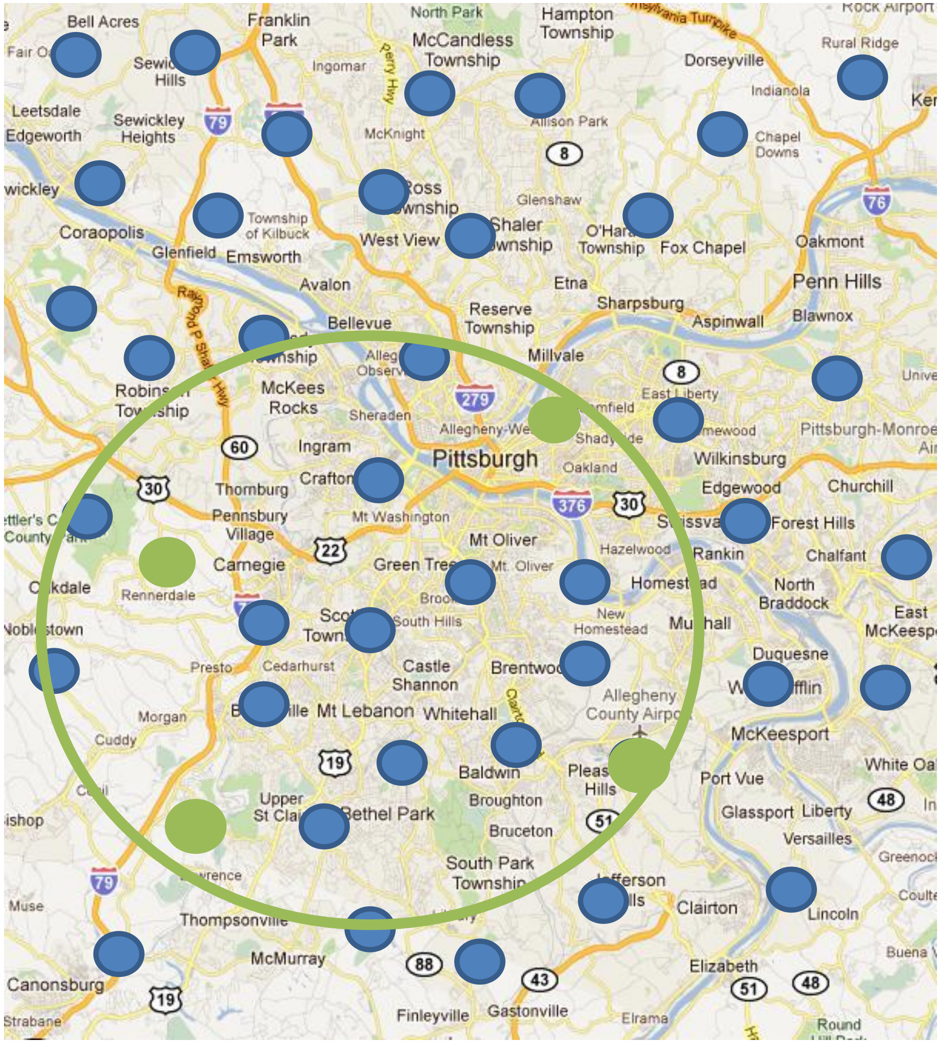
Fast Localized Scan

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 sparse subsets*** that do not reflect an outbreak of disease

Detecting *Irregular* Disease Clusters



(Neill, 2012)

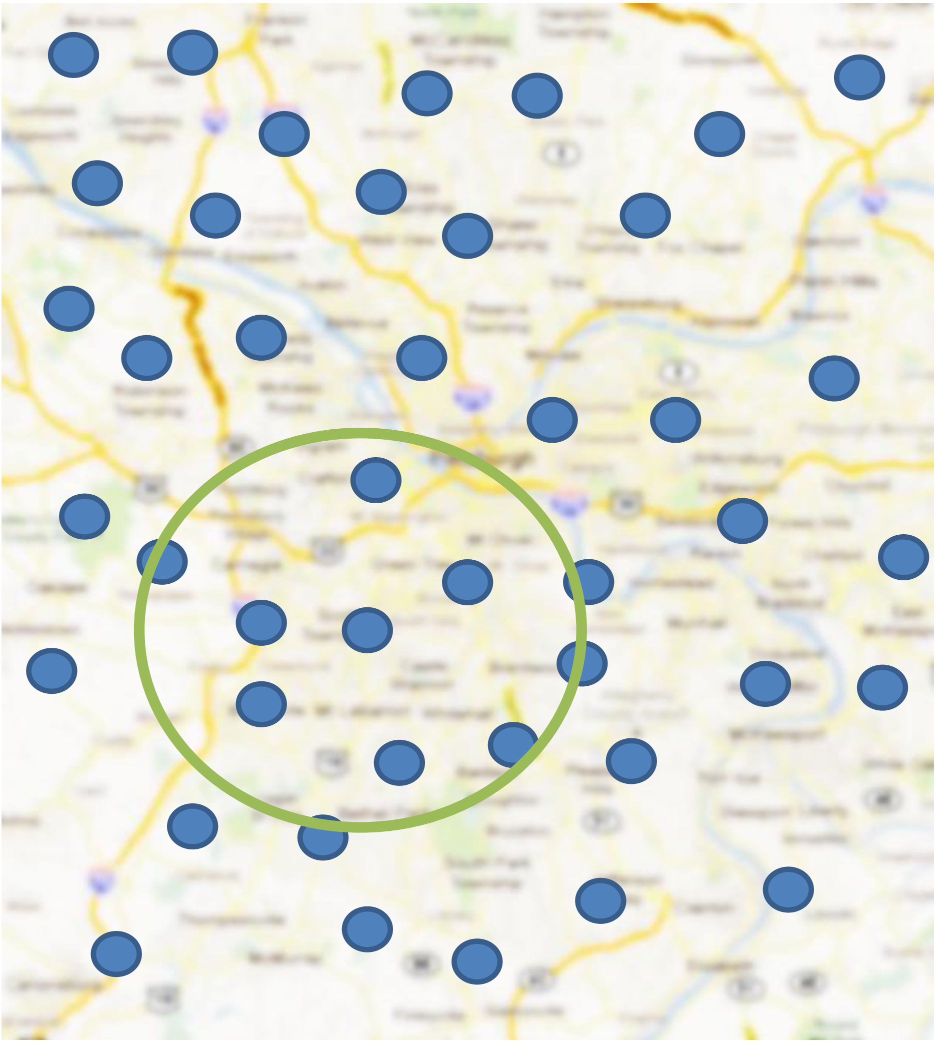
Fast Localized Scan

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

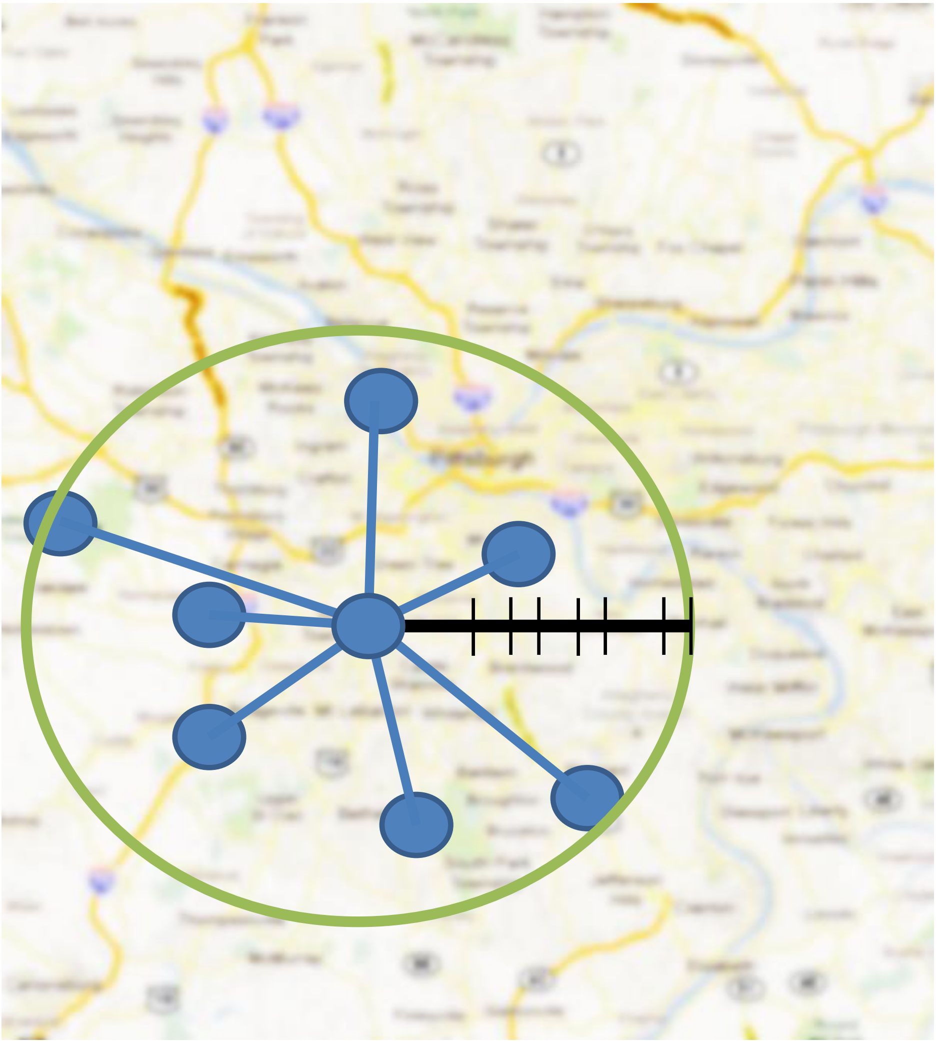
...but may return ***spatially sparse subsets*** that do not reflect an outbreak of disease

Detecting *Irregular* Disease Clusters



Soft Compactness Constraints

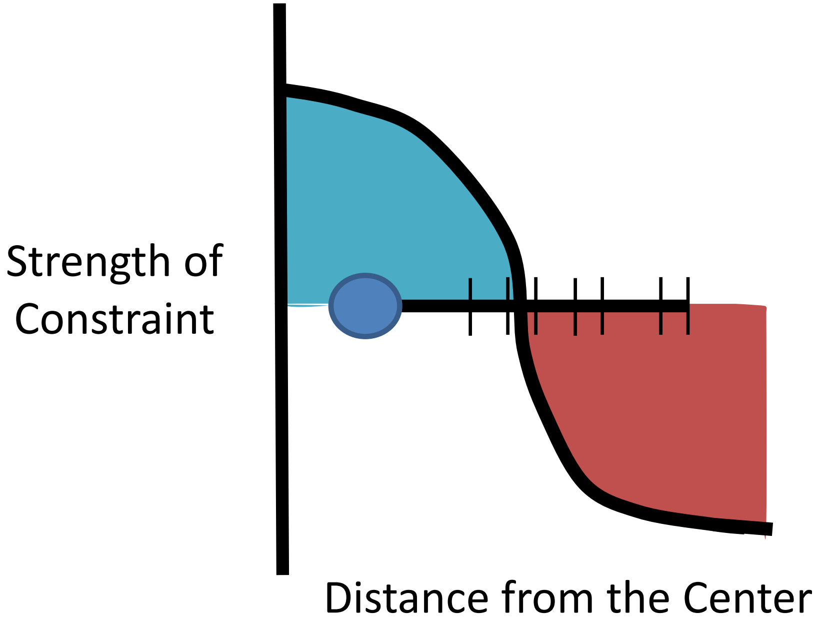
Detecting *Irregular* Disease Clusters



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

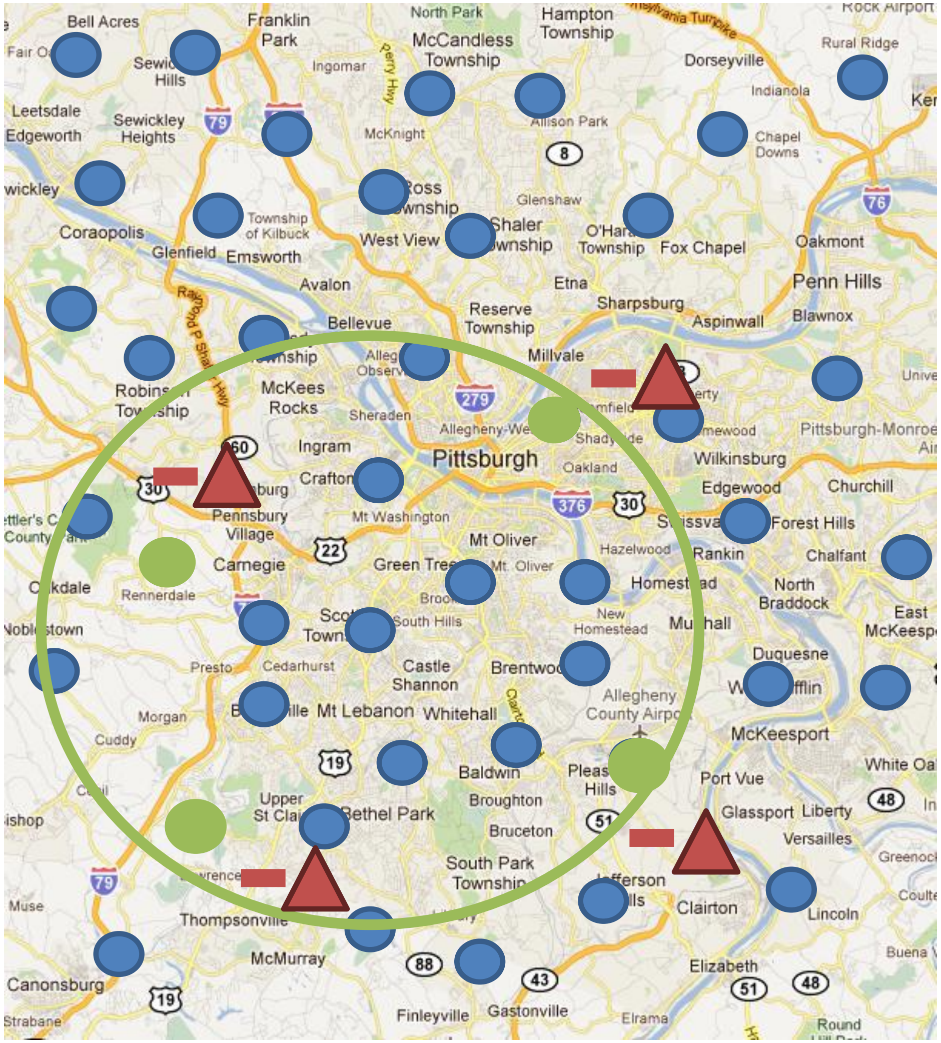
Use the distance of each location from the center as a measure of compactness/sparsity

Reward subsets that contain locations close to the center

and

Penalize subsets that contain locations far from the center

Detecting *Irregular* Disease Clusters



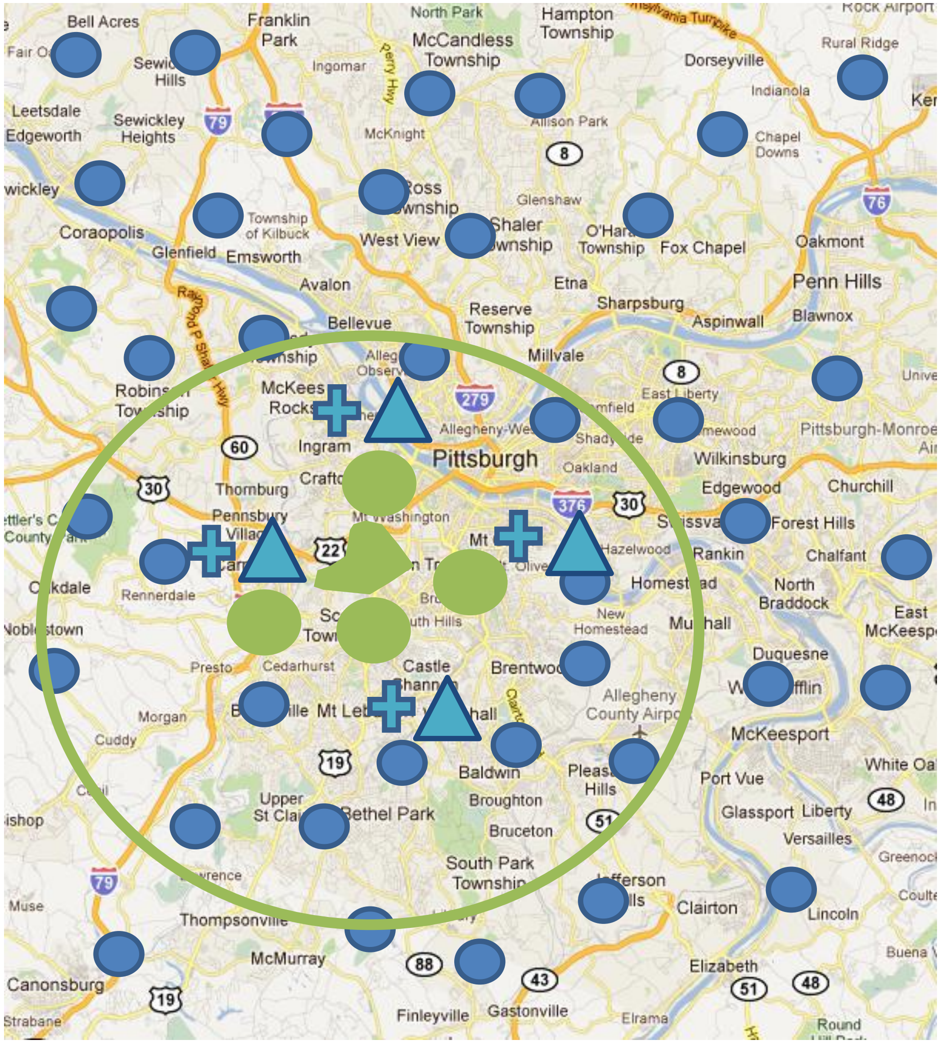
Soft Compactness Constraints

"...but may return
spatially sparse subsets
that do not reflect an outbreak..."

This particular subset would be less likely to be returned as optimal 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 sparse subsets

that do not reflect an outbreak..."

This particular subset would be less likely to be returned as optimal 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

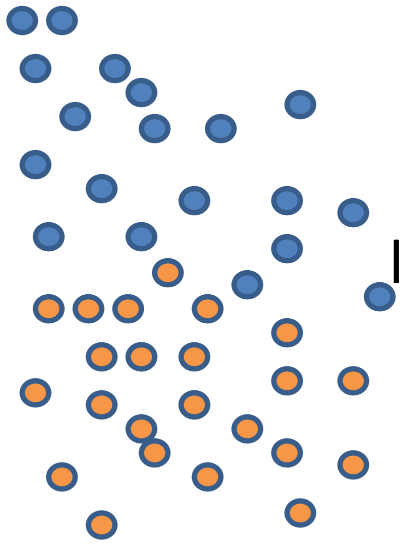
Score Function: Expectation-Based Poisson

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

$$H_0 : c_i \sim Poisson(b_i)$$

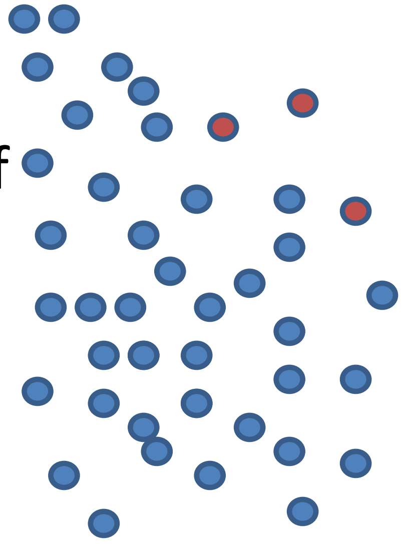
$$H_1 : c_i \sim Poisson(qb_i) \qquad q > 1$$

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



Large number locations with a moderate risk

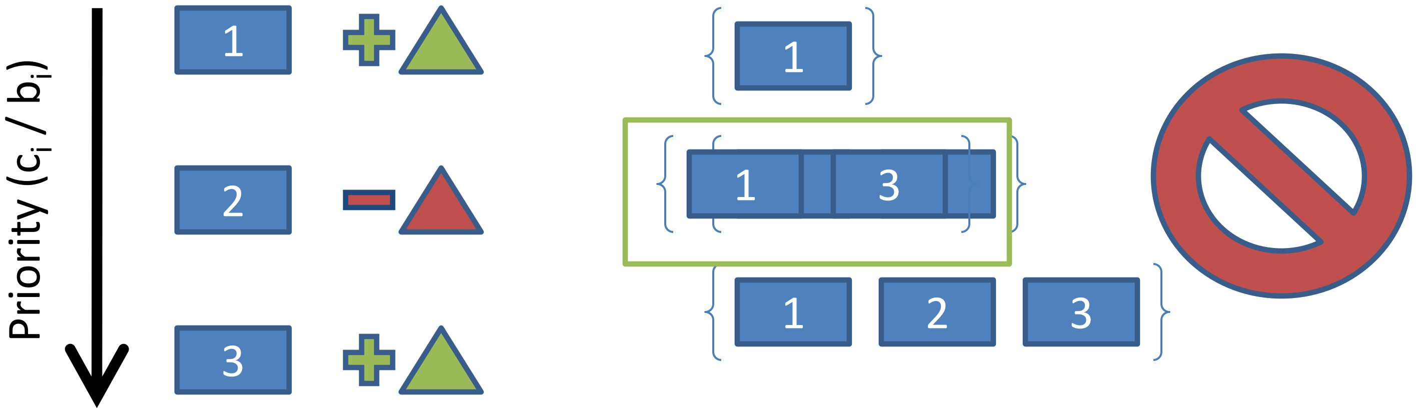
Small number of locations with a high risk



(Neill, 2012)

Linear-Time Subset Scanning

For EBP and any other score function satisfying the LTSS property, the highest scoring subset is guaranteed to be one of the following:



Decreases the search space from 2^N to N

Naively altering the scoring function to enforce soft constraints violates the LTSS property!

Adding Soft Constraints to the Scoring Function

$$F(S) + \sum_{s_i \in S} \Delta_i$$ 🚫

SOLUTION: Interpret the scoring function as a ***sum*** of ***contributions*** from each record in the subset.

Maximizing the scoring function is then equivalent to selecting all records that are making a ***positive contribution***.

$$F(S) = \max_q \sum_{s_i \in S} F(s_i \mid q)$$

INSIGHT: When treated as an additive function, ***further terms*** (i.e. soft constraints) may be introduced without interfering with the maximization step.

$$F(S) = \max_q \sum_{s_i \in S} \left[F(s_i \mid q) + \Delta_i \right]$$ 🙂

Demonstration with Expectation-based Poisson

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

$$F(S) = \max_{q>1} \log \prod_{s_i \in S} \frac{e^{-qb_i}(qb_i)^{c_i} / c_i!}{e^{-b_i}(b_i)^{c_i} / c_i!} = \max_{q>1} \log \prod_{s_i \in S} e^{(1-q)b_i} q^{c_i}$$

$$F(S \mid q) = \sum_{s_i \in S} \overbrace{\left[\underbrace{(1-q)b_i + c_i \log q}_{\text{Log-likelihood } F(s_i|q)} + \underbrace{\Delta_i}_{\text{Reward /Penalty from constraints}} \right]}^{\text{Contribution from each location, for a fixed q}}$$

Demonstration with Expectation-based Poisson

Here we use $\Delta_i = h(1 - 2d_i/r)$:

d_i is that location's distance from the center

r is the neighborhood radius

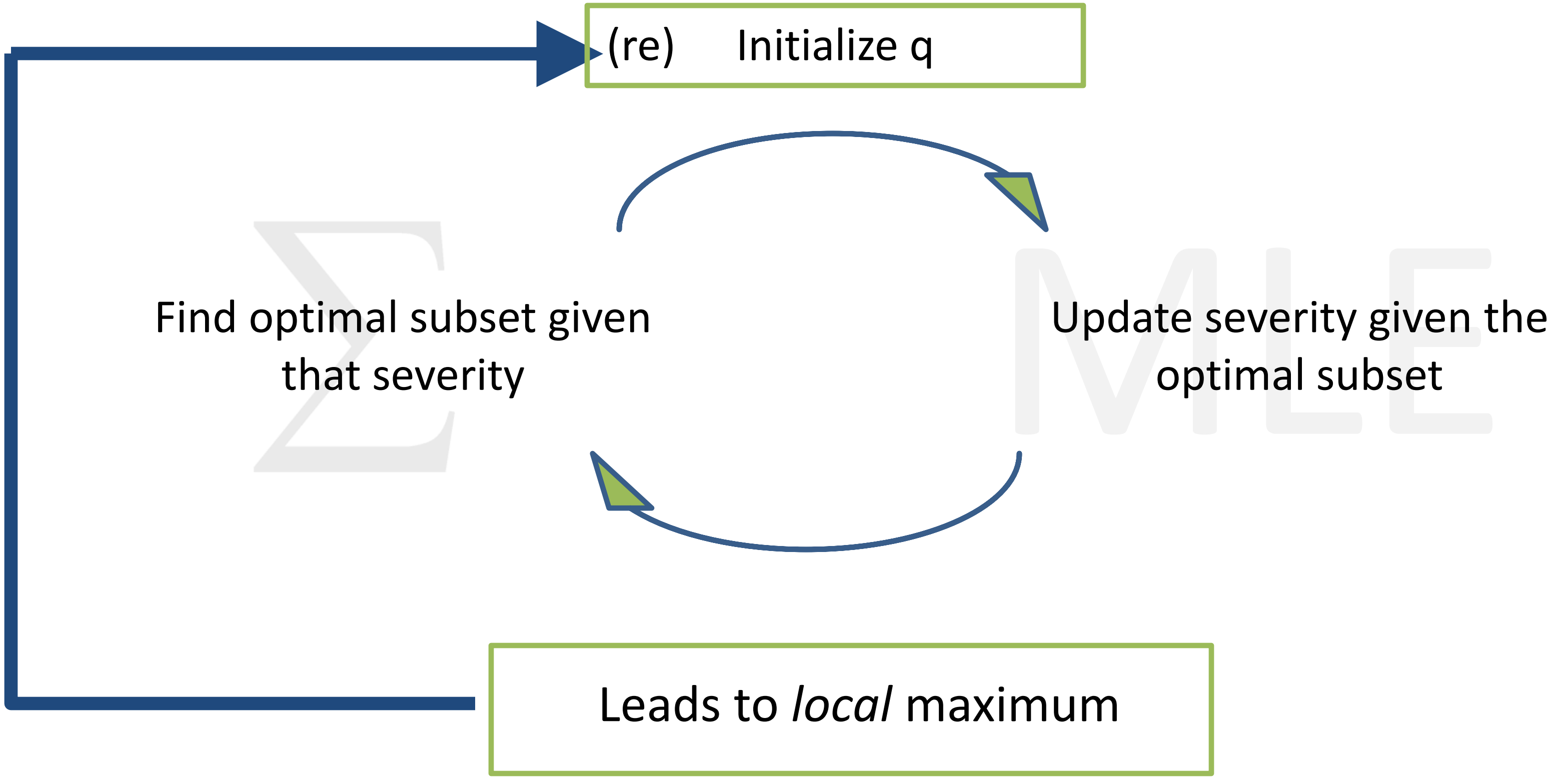
h is a constant representing the strength of the constraint.

Each Δ_i can be interpreted as the *prior log-odds* that s_i will be affected, and thus the center location ($d_i = 0$, $\Delta_i = h$) is e^h times as likely as its (k-1)th nearest neighbor ($d_i = r$, $\Delta_i = -h$).

$$F(S \mid q) = \sum_{s_i \in S} \left[(1-q)b_i + c_i \log q \underbrace{+ \Delta_i}_{\text{Reward /Penalty from constraints}} \right]$$

From *Fixed* q to *All* q

Our goal is to maximize F(S) ***over all q***



(re) Initialize q

Find optimal subset given that severity

Update severity given the optimal subset

Leads to *local* maximum

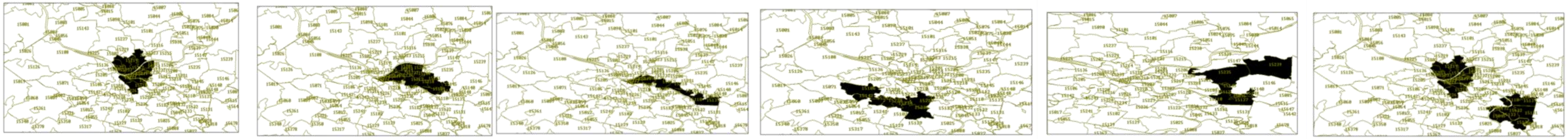
Evaluation: Emergency Department Data

Two years of admissions from 10 different Allegheny County Emergency Departments

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

Centriods of 97 zip codes were used as locations

Semi-Synthetic “injects” were created by artificially increasing the count within various subsets of zip codes: Some compact, some elongated or irregular.



Competing Methods

Circles:
Determines the most anomalous circular region.

Kulldorff, 1997

Circles

Fast Localized Scan:
Determines the most anomalous subset within a circular region.
(This equates to our new method *without additional soft constraints).*

Neill, 2012

h=0

Competing Methods

Weak Compactness Constraints:
Determines the most anomalous subset with weak constraints

h=1

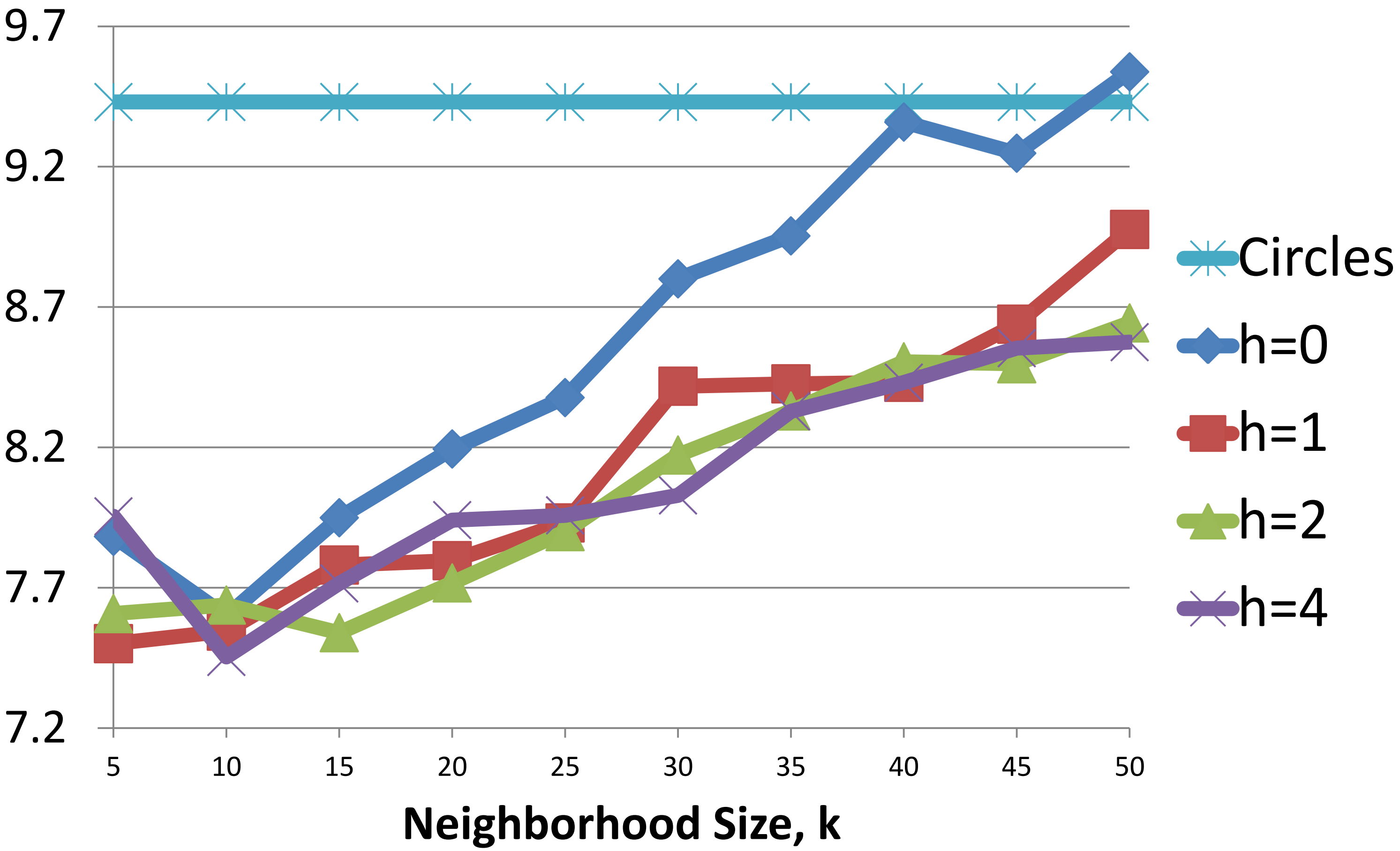
Moderate Compactness Constraints:
Determines the most anomalous subset with moderate constraints

h=2

Strong Compactness Constraints:
Determines the most anomalous subset with strong constraints

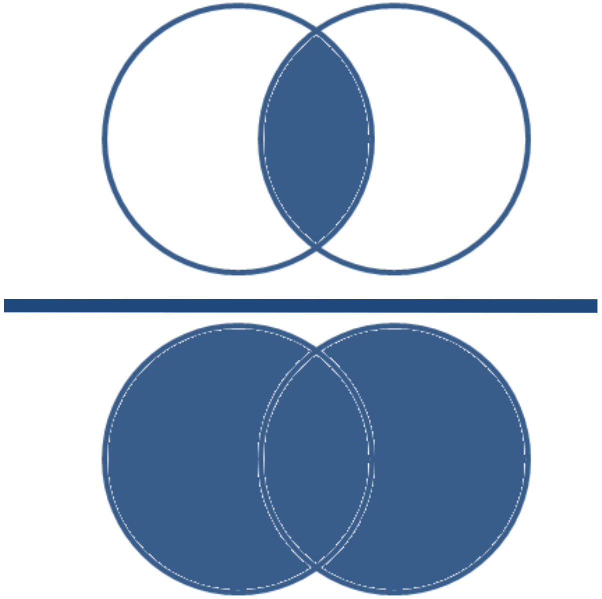
h=4

Results: Time to Detect (Days)



Results: Spatial Overlap

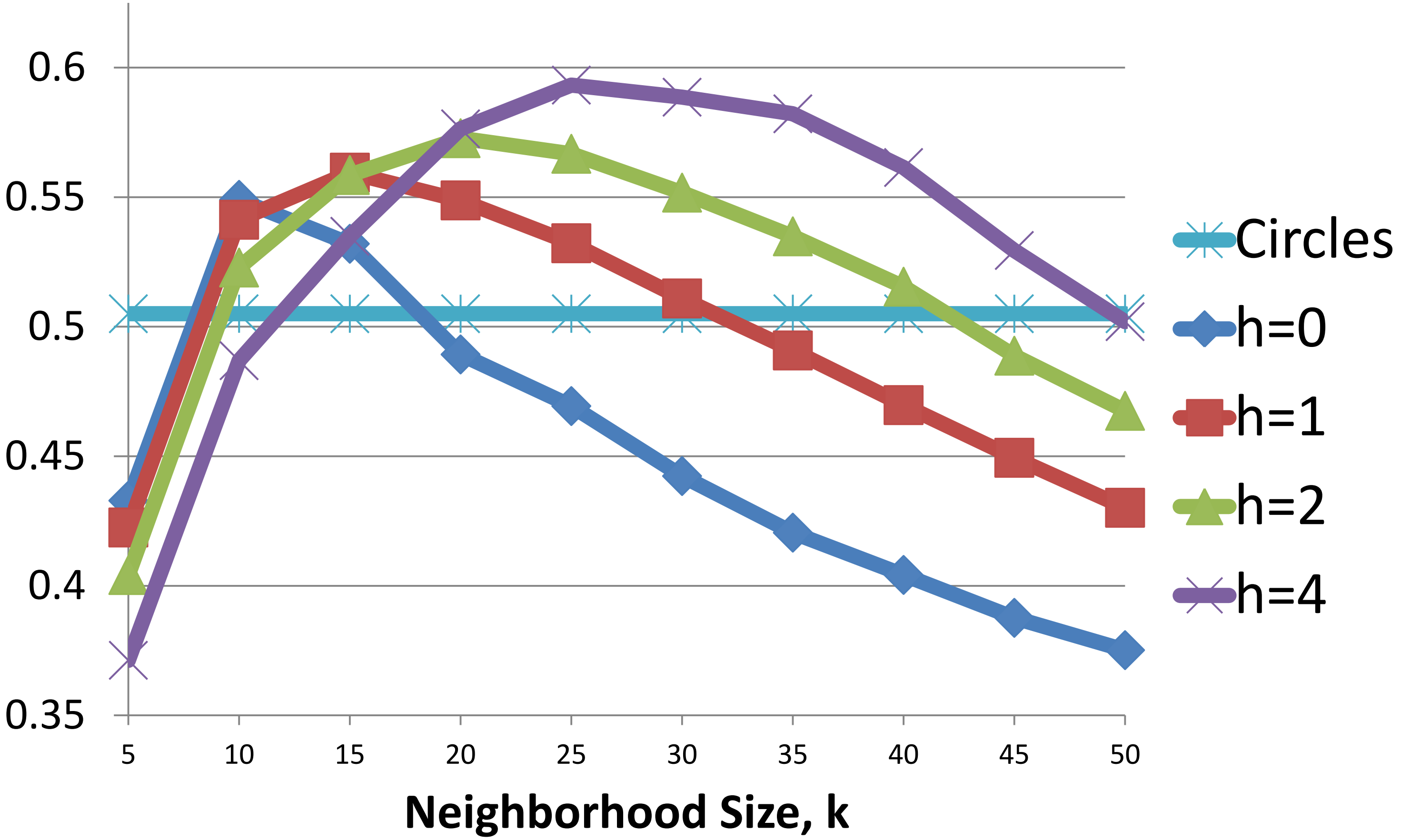
$$Overlap = \frac{A \cap B}{A \cup B} =$$



$Overlap = 1$ Perfect Match

$Overlap = 0$ Completely Disjoint

Results: Spatial Overlap

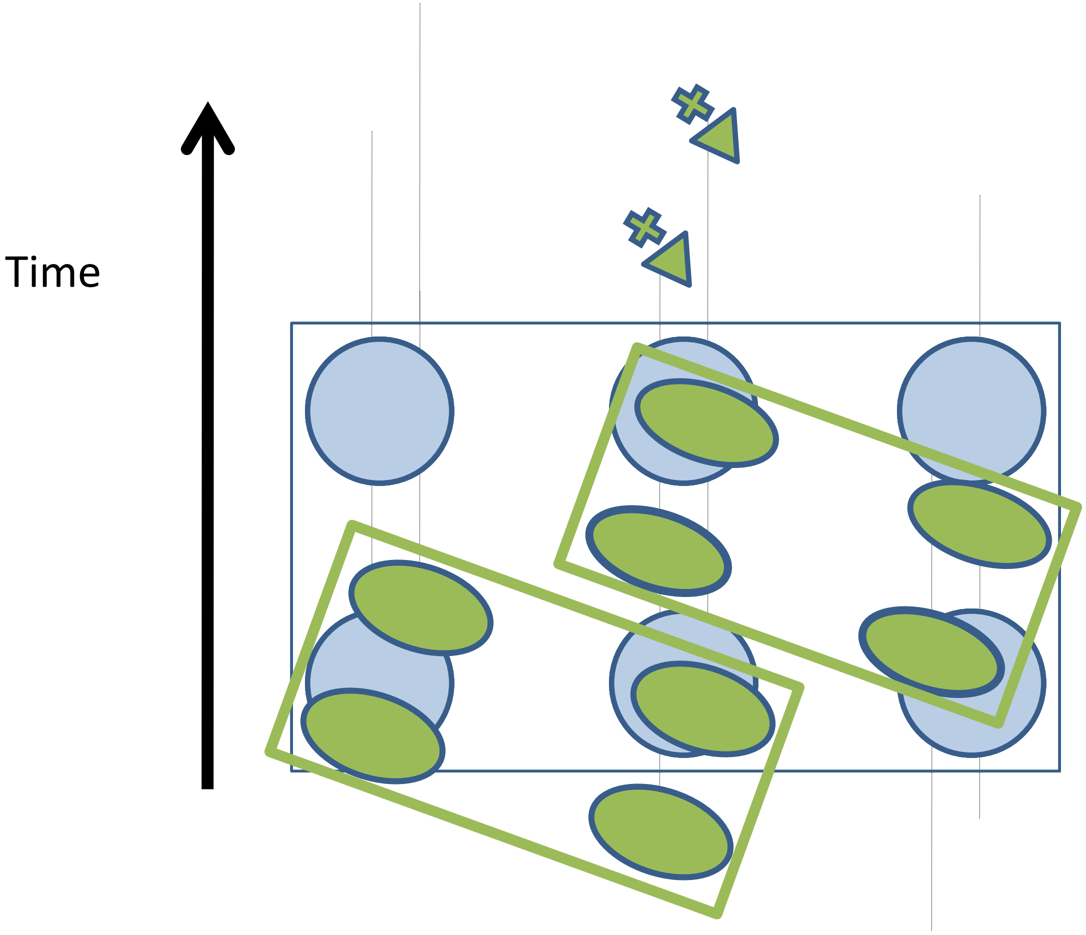


Example 2: Temporal Consistency

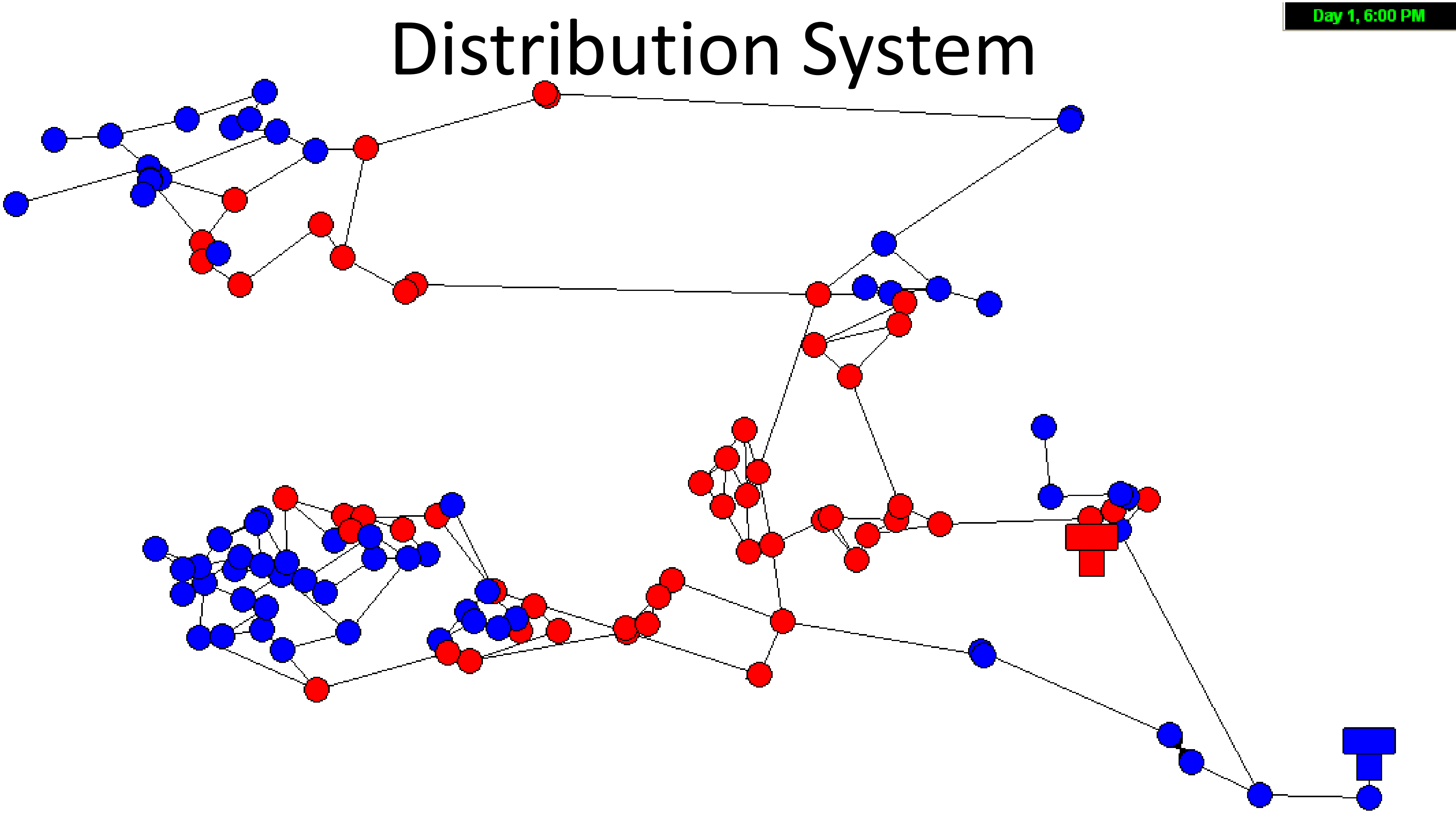
So far, we have naively used temporal information by simply aggregating counts over a temporal window

We can also use temporal information by rewarding locations that were in the optimal subset in previous time steps.

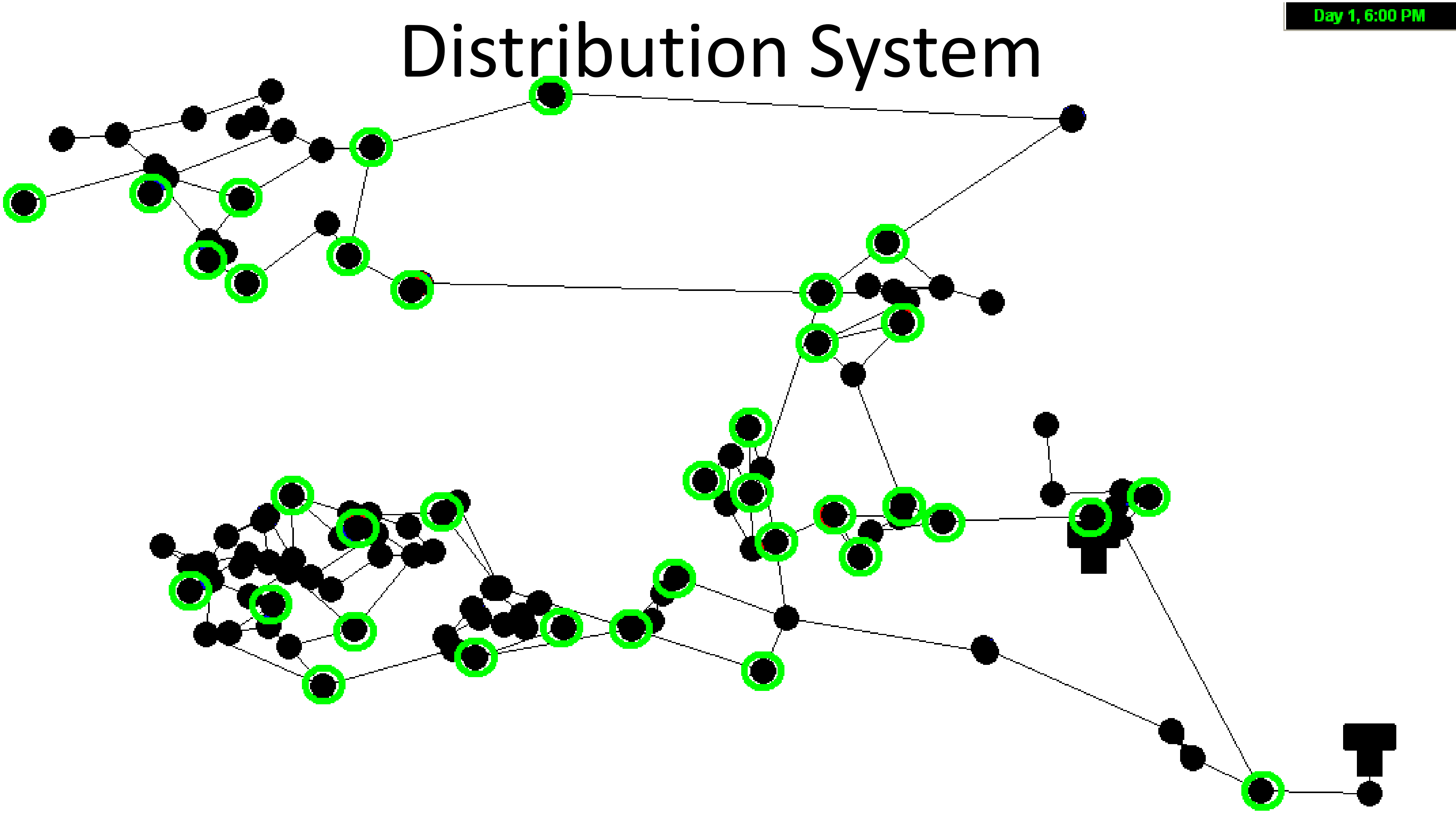
This can increase power to detect **dynamic** patterns that may be changing over time.



Spreading Contaminants in a Water Distribution System



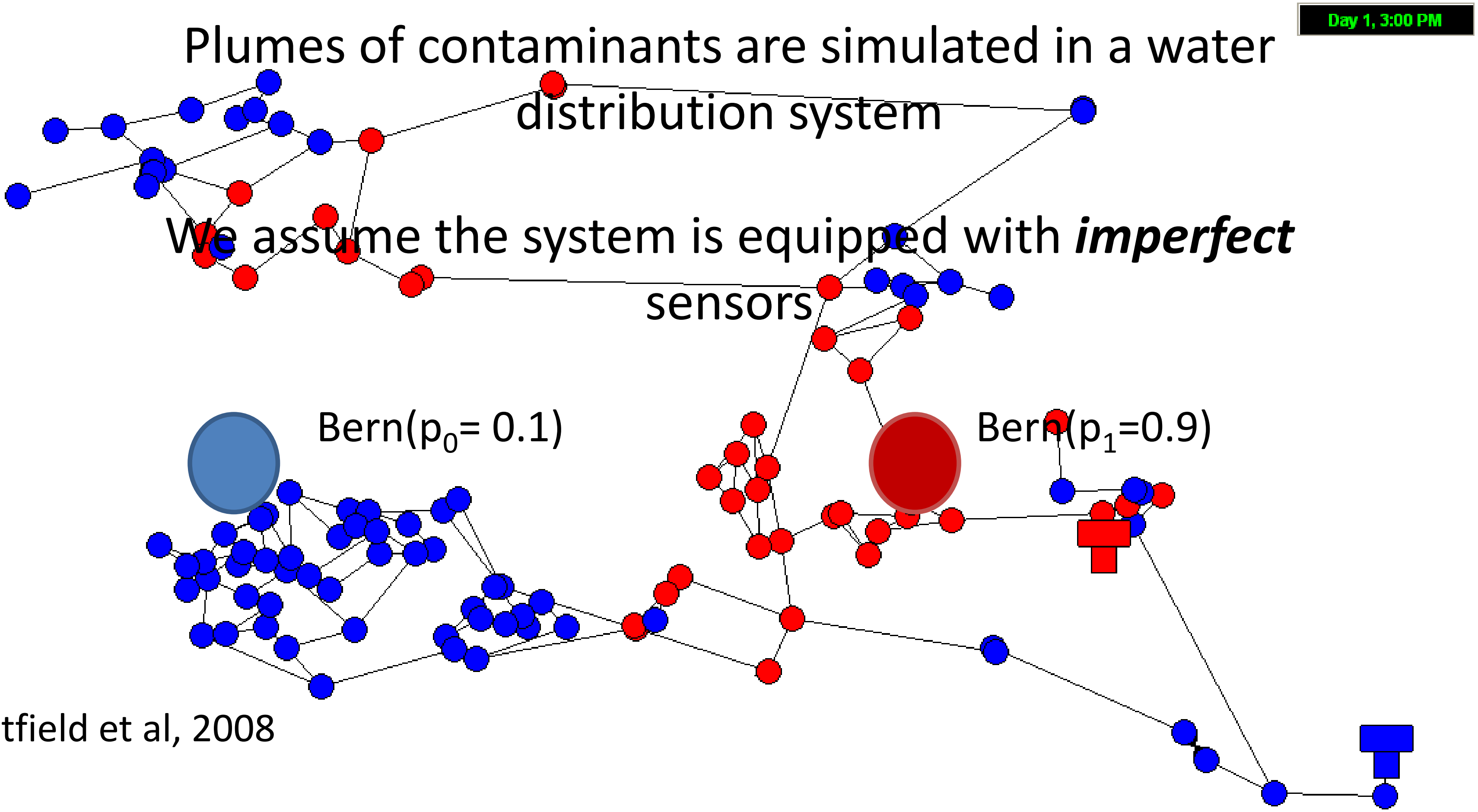
Spreading Contaminants in a Water Distribution System



Data: Battle of the Water Sensor Networks

Plumes of contaminants are simulated in a water distribution system

We assume the system is equipped with ***imperfect*** sensors



● Bern(p_0= 0.1)

● Bern(p_1=0.9)

Ostfield et al, 2008

Competing Methods

Upper Level Sets:
A heuristic that is not guaranteed to find the most anomalous subgraph

ULS

Patil & Taillie, 2004

GraphScan:
Determines the most anomalous subgraph *without further constraints*

GS

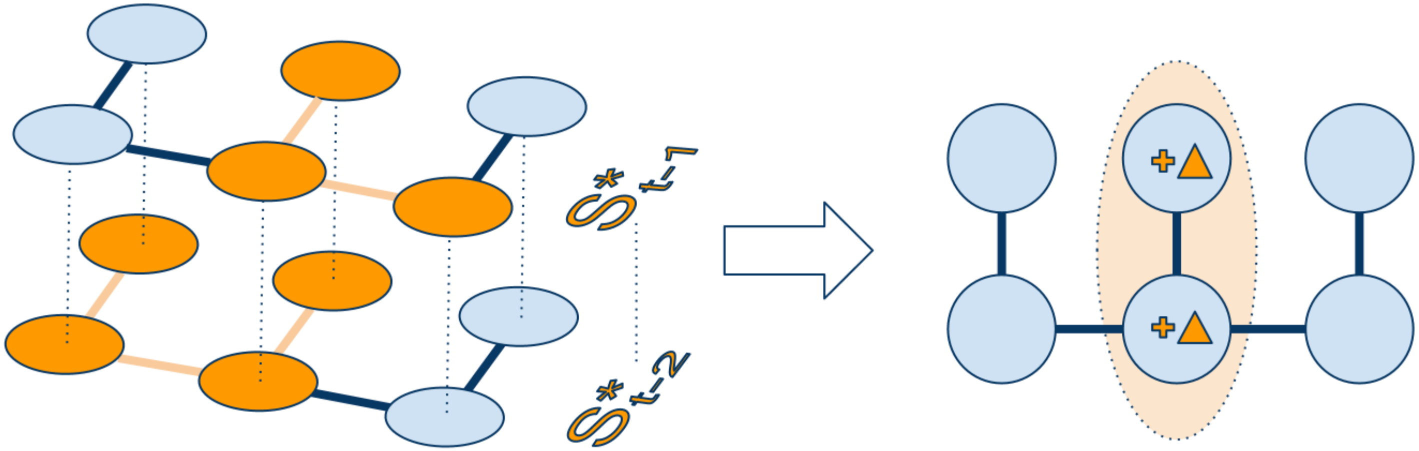
Speakman & Neill, 2010

Competing Methods

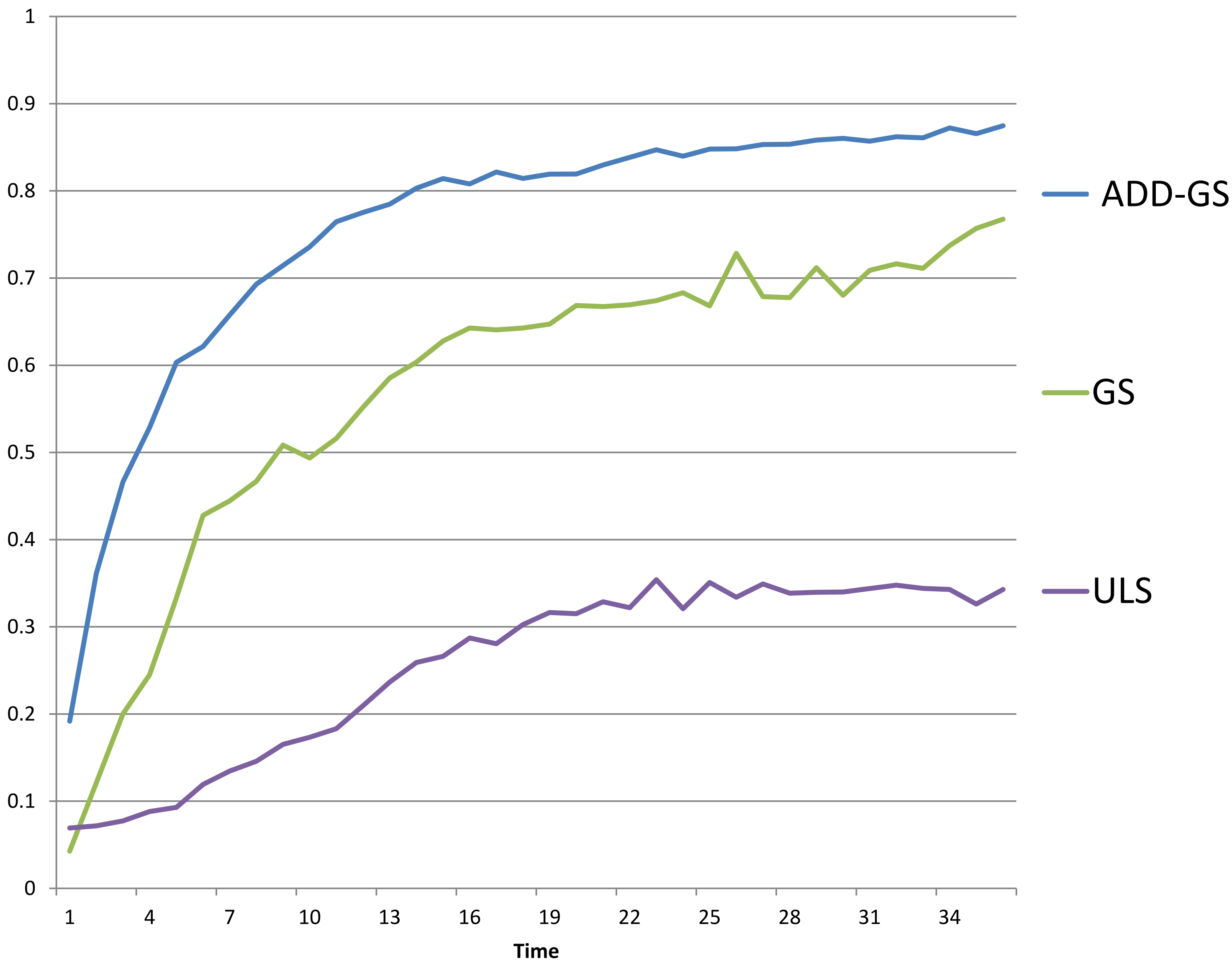
GraphScan with basic temporal consistency

$$F(S) = \max_{q} \sum_{s_i \in S} \left(F(s_i \mid q) + \Delta_i \right) \qquad \Delta_i = \begin{cases} +\Delta \text{ if } s_i \in \Omega \\ 0 \text{ otherwise} \end{cases}$$

ADD-GS



Overlap Coefficient



Hours until Detection	% Detected
7.66	100%
9.65	97.5%
15.4	92.4%

Conclusions

We provided a framework that allows soft constraints to influence the scoring function and give preference to subsets of desired spatial compactness or temporal consistency, while still allowing an efficient search for the highest scoring subset.

We applied **soft proximity constraints** for detecting an increase in ED visits in Allegheny County, PA, and **temporal consistency constraints** to detect dynamic patterns of contamination in a water network.

Empirical results showed that soft constraints ***reduced time to detect*** and ***increased spatial accuracy*** of the methods in each case.