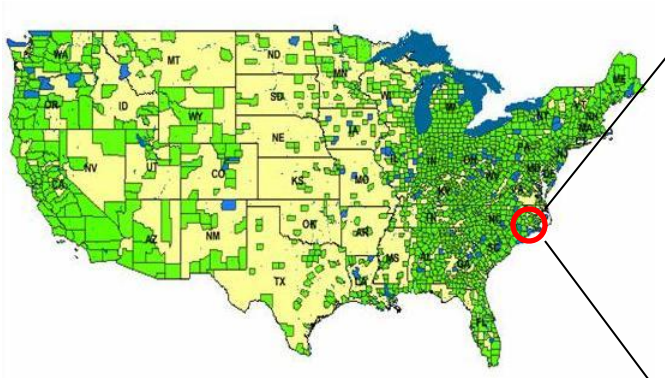
Fast Multidimensional Subset Scan for Outbreak Detection and Characterization

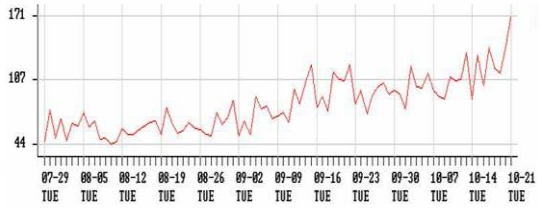
Daniel B. Neill* and Tarun Kumar
Event and Pattern Detection Laboratory
Carnegie Mellon University
neill@cs.cmu.edu

This project was partially supported by NSF grants IIS-0916345, IIS-0911032, and IIS-0953330, and a UPMC Healthcare Technology Innovation grant.

Multivariate outbreak detection



Spatial time series data from spatial locations s_i (e.g. zip codes)



Time series of counts $c_{i,m}{}^t$ for each zip code s_i for each data stream d_m.

d_1 = respiratory ED

d_2 = constitutional ED

d_3 = OTC cough/cold

d_4 = OTC anti-fever

(etc.)

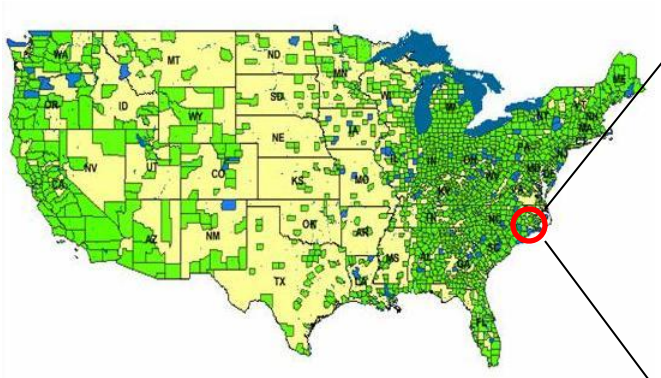
Typical goals: **detect** any emerging disease outbreaks, **pinpoint** the affected space-time region, and **characterize** the outbreak.

Identify affected time window and proximity-constrained **subset** of spatial locations.

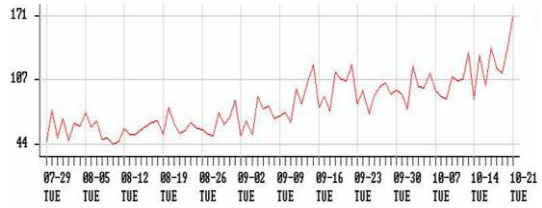
Determine which **subset** of the monitored data streams are likely to have been affected by the outbreak.

Solution: maximize a **likelihood ratio statistic** F(D,S,W) over all subsets of streams D, all proximity-constrained subsets of locations S, and time windows W.

Multidimensional outbreak detection



Spatial time series data from spatial locations s_i (e.g. zip codes)



Time series of counts $c_{i,m}{}^t$ for each zip code s_i for each data stream d_m.

d_1 = respiratory ED

d_2 = constitutional ED

d_3 = OTC cough/cold

d_4 = OTC anti-fever

(etc.)

Our additional goal: identify any differentially affected **subpopulations** P of the monitored population.

Gender (male, female, both)

Age groups (children, adults, elderly)

Ethnic or socio-economic groups

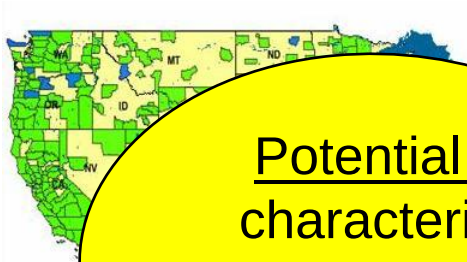
Risk behaviors: e.g. intravenous drug use, multiple sexual partners

[or all of the above…]

More generally, assume that we have a set of additional discrete-valued attributes $A_1..A_J$ observed for each individual case.

We will identify not only the affected streams, locations, and time window, but also a **subset** of values for each attribute.

Multidimensional outbreak detection



Potential advantages: more precise outbreak characterization; improved detection when the outbreak differentially affects some subpopulation.

Huge challenge: **computational infeasibility!**

Assuming M monitored data streams, a spatial neighborhood of size $k \leq N$, and a set of possible values $V_1..V_{Q(j)}$ for each attribute A_j, the total number of subsets to consider is $O(2^{M+k+Q})$, where $Q = \sum_{j=1..J} Q(j)$ is the sum of all attributes' arities.

...y ED

...D

...e have a set

... discrete valued attributes

...erved for each individual case.

Age groups (...

Ethnic ... -economic groups

Risk beha... e.g. intravenous drug ... multiple sexual partners

[or all of the above…]

We will identify not only the affected streams, locations, and time window, but also a **subset** of values for each attribute.

The LTSS property

- In certain cases, we can search over the exponentially many subsets in linear time!
- Many commonly used scan statistics have the property of linear-time subset scanning:
  - Just sort the data records from highest priority to lowest priority according to some criterion…
  - … then search over groups consisting of the top-k highest priority records, for k = 1..N.

The highest scoring subset is guaranteed to be one of these!

Multivariate LTSS

- Neill et al. (2012) extended LTSS from univariate to multivariate data.
  - For each time window and spatial neighborhood (center + k-nearest neighbors), we do the following:

1. Start with a randomly chosen subset of streams.

2. Use LTSS to efficiently find the highest scoring subset of locations for the given streams.

3. Use LTSS to efficiently find the highest scoring subset of streams for the given locations.

4. Iterate steps 2-3 until convergence to a local maximum of the score function, and use multiple restarts to approach the global maximum.

Multidimensional LTSS

- Our **MD-Scan** approach (Neill and Kumar, 2012) extends MLTSS to the multidimensional case:
  - For each time window and spatial neighborhood (center + k-nearest neighbors), we do the following:

1. Start with randomly chosen subsets of **locations** S, **streams** D, and **values** V_j for each attribute A_j (j=1..J).

2. Choose an attribute (randomly or sequentially) and use LTSS to find the highest scoring subset of values, locations, or streams, conditioned on all other attributes.

3. Iterate step 2 until convergence to a local maximum of the score function F(D,S,W, $\{V_j\}$), and use multiple restarts to approach the global maximum.

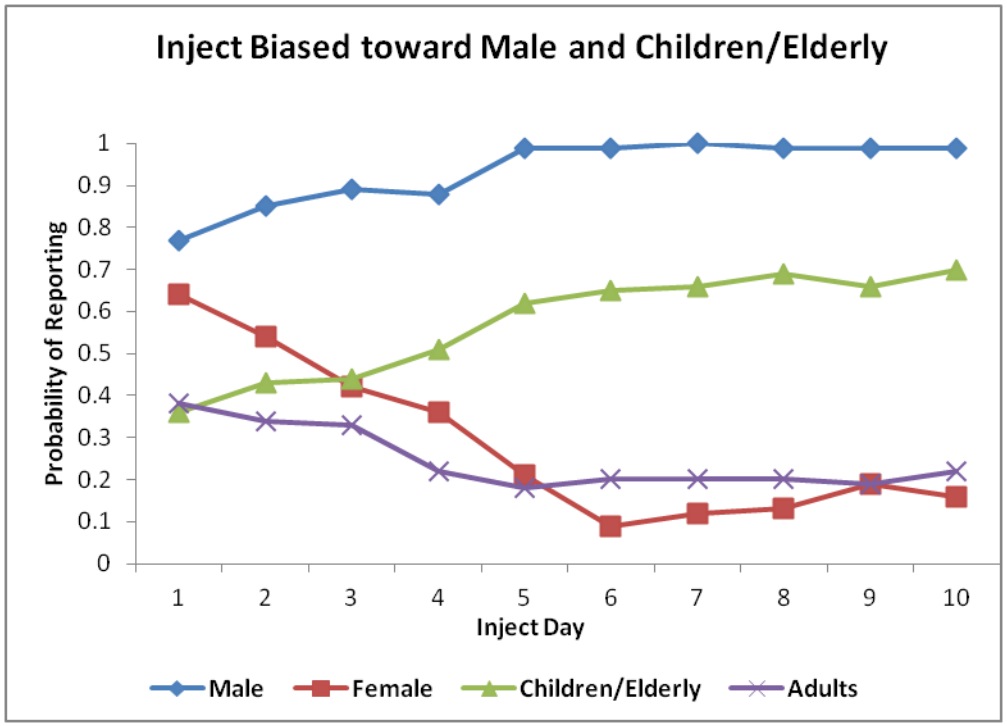
Evaluation

- We compared the detection performance of MD-Scan to MLTSS for multivariate, synthetic outbreaks injected into real-world Emergency Department data from Allegheny County, PA.
- For each case, the data included date, zip code, prodrome, gender, and age decile.
  - MD-Scan included additional search constraints on age decile (must be **connected** in circular graph).
- We considered outbreaks with various types and amounts of age and gender bias.
  - Shown here: biased toward males, children, and the elderly (age deciles 0, 1, 8, 9, 10).

1) Identifying affected subpopulations

By the midpoint of the outbreak, MD-Scan is able to correctly identify the affected gender and age deciles with high probability, without reporting unaffected subpopulations.

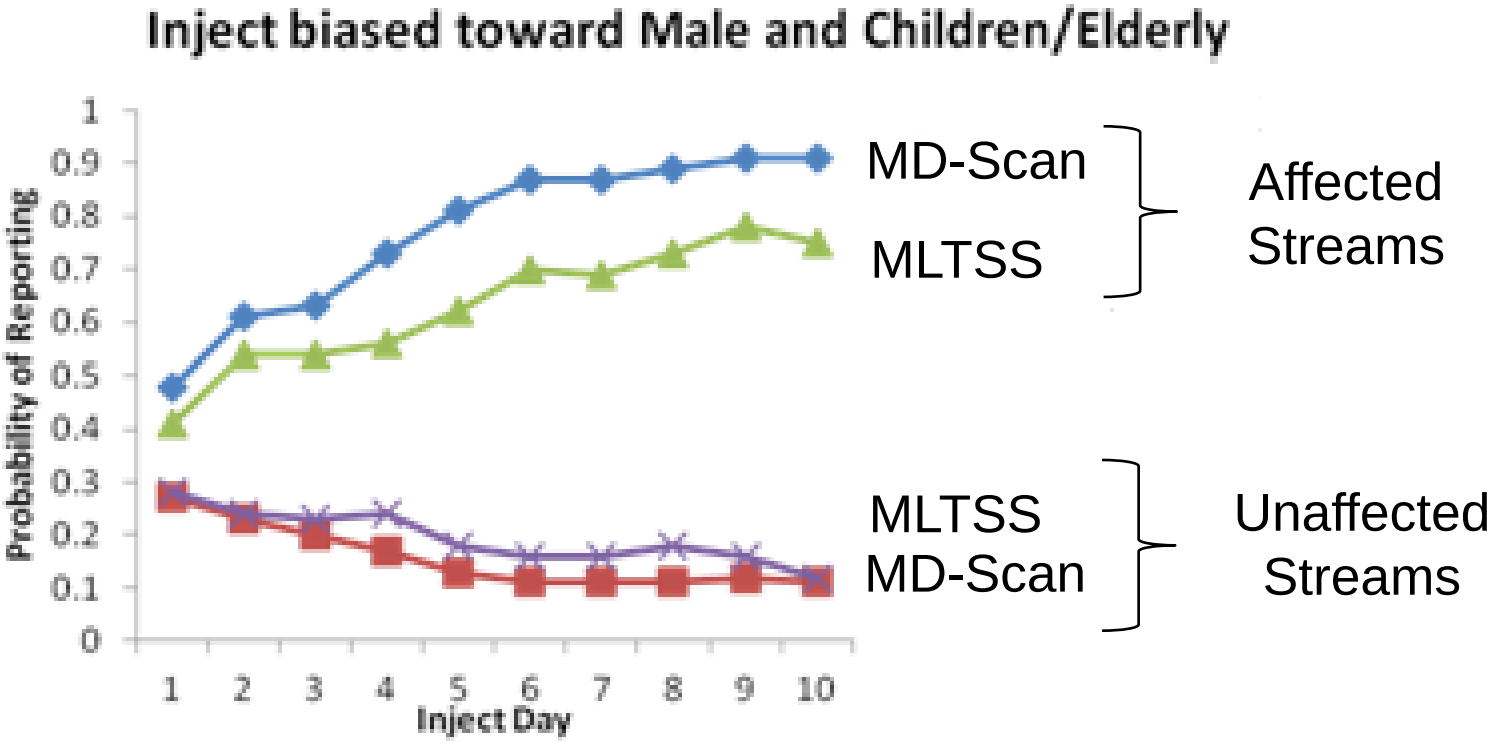
(MLTSS simply ignores the age and gender information, implicitly assuming that all ages and genders are affected.)



2) Characterizing affected streams

As compared to MLTSS, MD-Scan is better able to characterize the affected subset of the monitored streams.

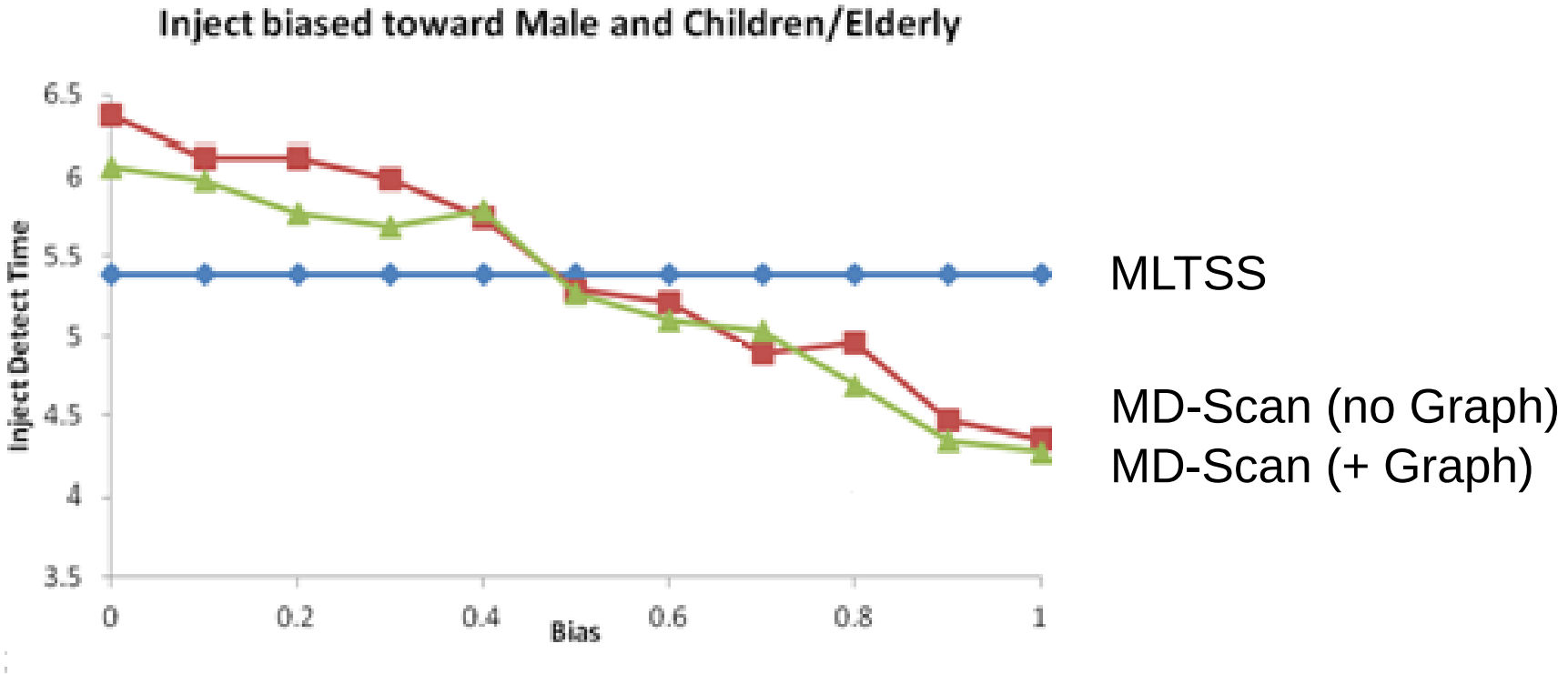
(Counts were injected into three of the eight monitored streams- respiratory, diarrhea, and fever.)



3) Time to detect (1 fp/month)

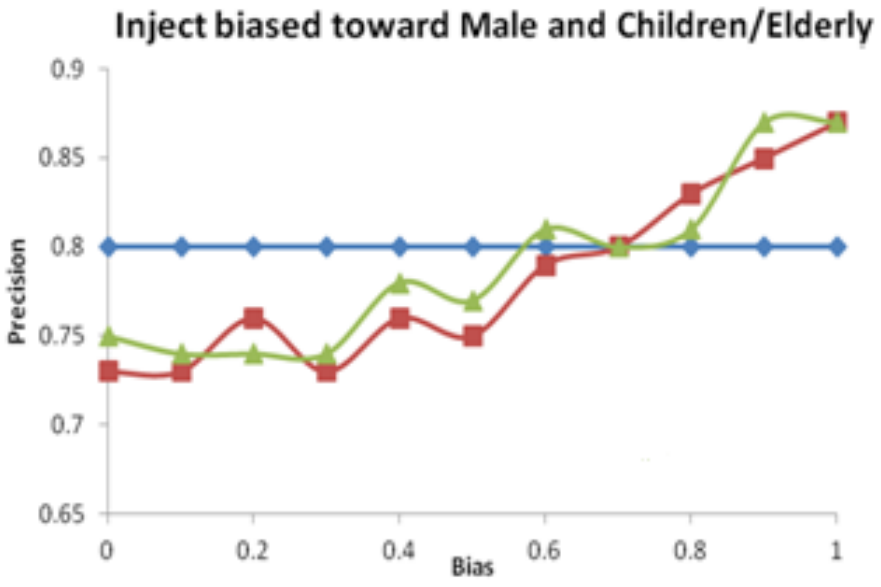
At a fixed false positive rate of 1 per month, MD-Scan achieved faster detection than MLTSS for outbreaks which were sufficiently biased by age and/or gender.

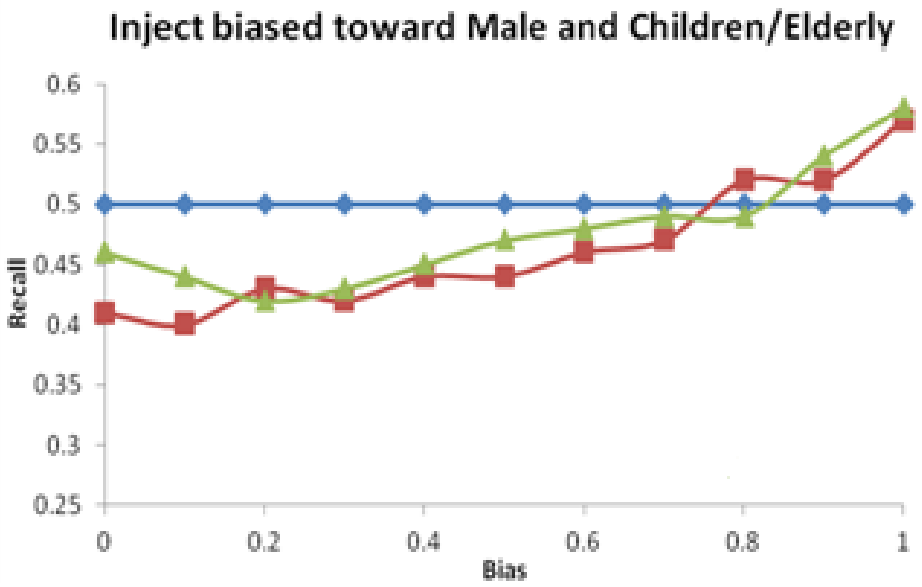
(Bias is linearly scaled, from 0 = same age/gender distribution as background data to 1 = only males and children/elderly affected.)



4) Spatial precision and recall

Similarly, spatial accuracy (precision and recall for identifying the affected subset of locations) was improved for outbreaks which had differential effects based on age and gender.





$$\text{Precision} = \frac{|\text{Detected} \cap \text{Affected}|}{|\text{Detected}|}$$

$$\text{Recall} = \frac{|\text{Detected} \cap \text{Affected}|}{|\text{Affected}|}$$

5) Run time

Run time of MD-Scan is about an order of magnitude slower than MLTSS, but still extremely fast:

For our experiments, MD-Scan required an average of 4.15 seconds to analyze each day of Emergency Department data.

Conclusions

MD-Scan is able to use **individual case data**, rather than aggregate counts, to more accurately identify not only the region, but also the **subpopulation** affected by an outbreak.

Subset of values for each monitored attribute of the population (demographics, behaviors, etc.)

MD-Scan substantially improves timeliness and accuracy of detection for outbreaks which differentially affect a subset of the monitored population (e.g. high-risk demographics or behaviors)

Detection performance can be enhanced by incorporating additional constraints such as spatial proximity and graph connectivity into each conditional optimization step.

Planned future work includes monitoring of sexually transmitted illness (in collaboration with Chicago DPH) to identify changing trends among neighborhoods, demographic groups, high-risk behaviors, prevention and treatment options, etc.

Thanks!!!

References:

D.B. Neill (2012). Fast subset scan for spatial pattern detection. *Journal of the Royal Statistical Society (Series B: Statistical Methodology)* 74(2): 337-360.

D.B. Neill, E. McFowland III, and H. Zheng (2012). Fast subset scan for multivariate event detection. *Statistics in Medicine*, published online 11/22/2012, DOI: 10.1002/sim.5675.

D.B. Neill and T. Kumar (2012). Fast multidimensional subset scan for event detection and characterization. Submitted for publication.

For more information, contact:

Daniel B. Neill, neill@cs.cmu.edu

http://www.cs.cmu.edu/~neill