SAND2008-3007C

Distributed Fusion for Water Quality

Mark W. Koch

Sensor Exploitation Applications Department

Sean A. McKenna

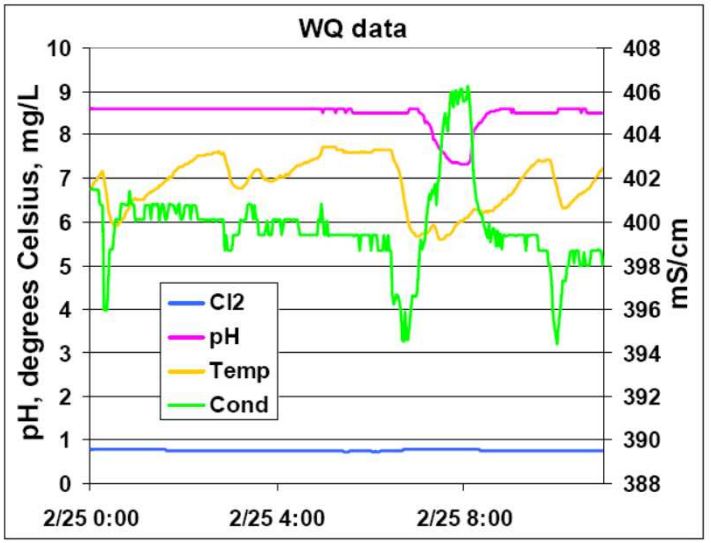
National Security Applications Department

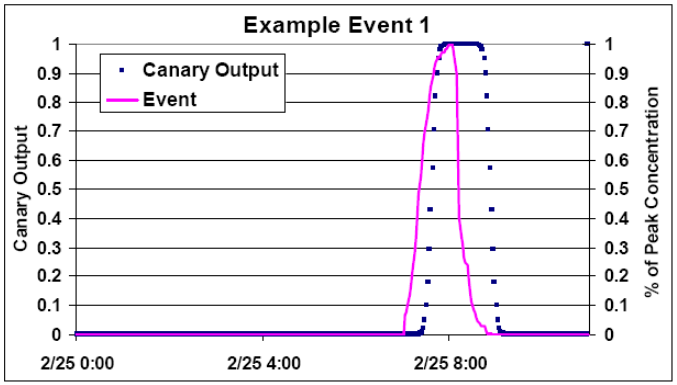
Sandia National Laboratories
Albuquerque, New Mexico

Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

CANARY: Water Quality Change Detection

- Event detection software that provides on-line & automated water quality event detection
- Connects to utility SCADA system to analyze water quality signals from each monitoring location independently
- Multivariate algorithms provide the probability of an anomalous water quality event *P(event)* at every time step



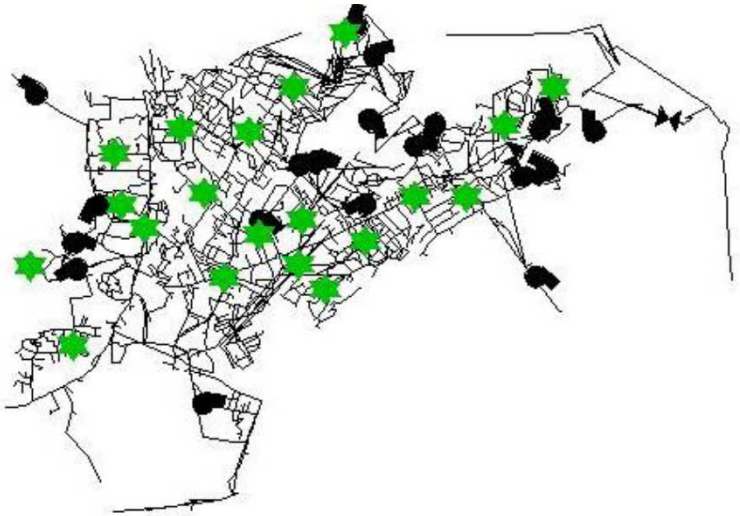


- Distributed Detection
  - Move from independent analysis at each location to integrating information across all locations

Sandia National Laboratories

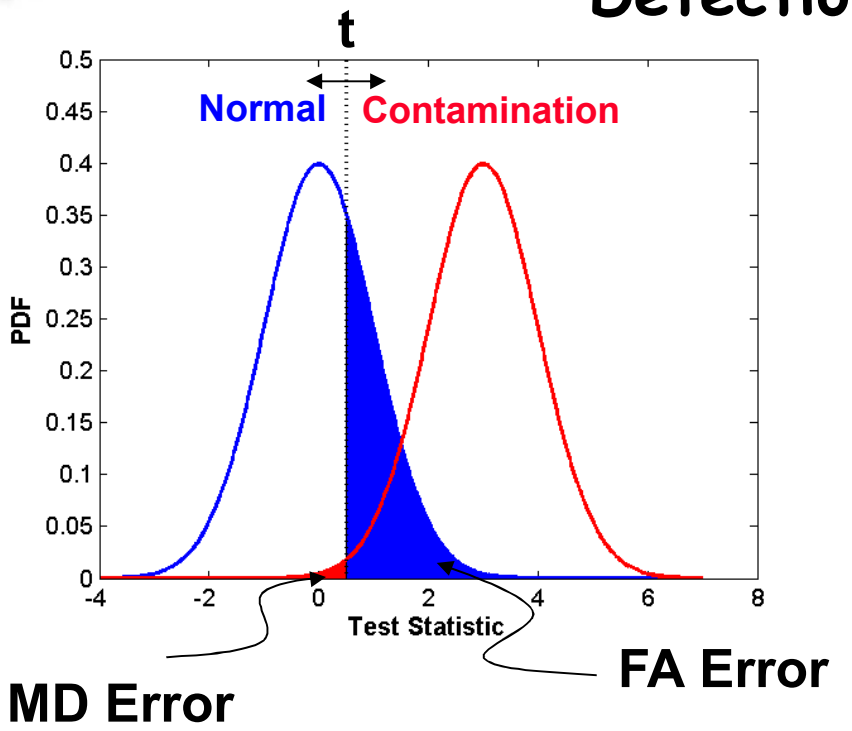
Distributed Network Fusion

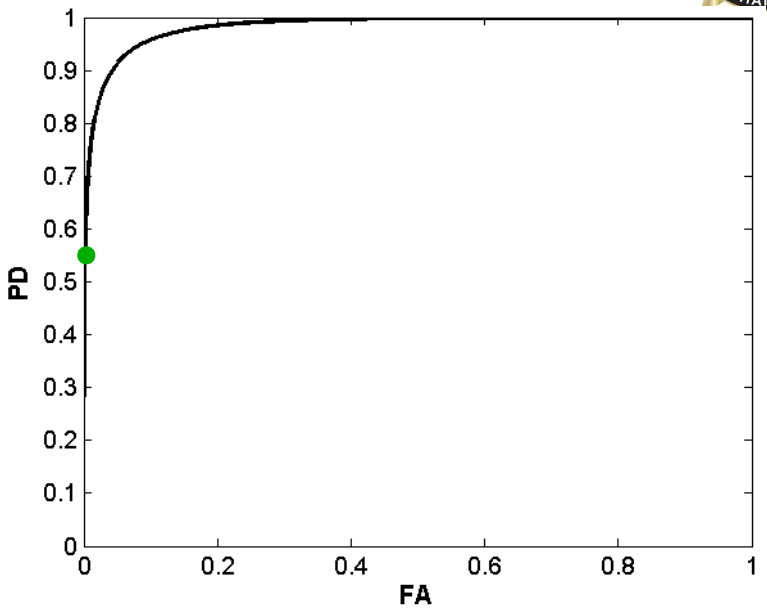
- Currently detection of a water quality event occurs when data from a single station reports an anomaly
- How can the results from individual sensing nodes at multiple locations be combined or fused to get improve detection performance?
- If an individual monitoring station has 1 FA / day
  - 50 sensing nodes will have 2 per hour
  - 100 sensing nodes will have 4 per hour
  - 400 sensing nodes with have 16 per hour
- Goals
  - Improve performance
    - Reduce false alarms
    - Increase probability of detection
    - Faster detection time



Sandia National Laboratories

Measuring Performance of a Water Quality Detection System

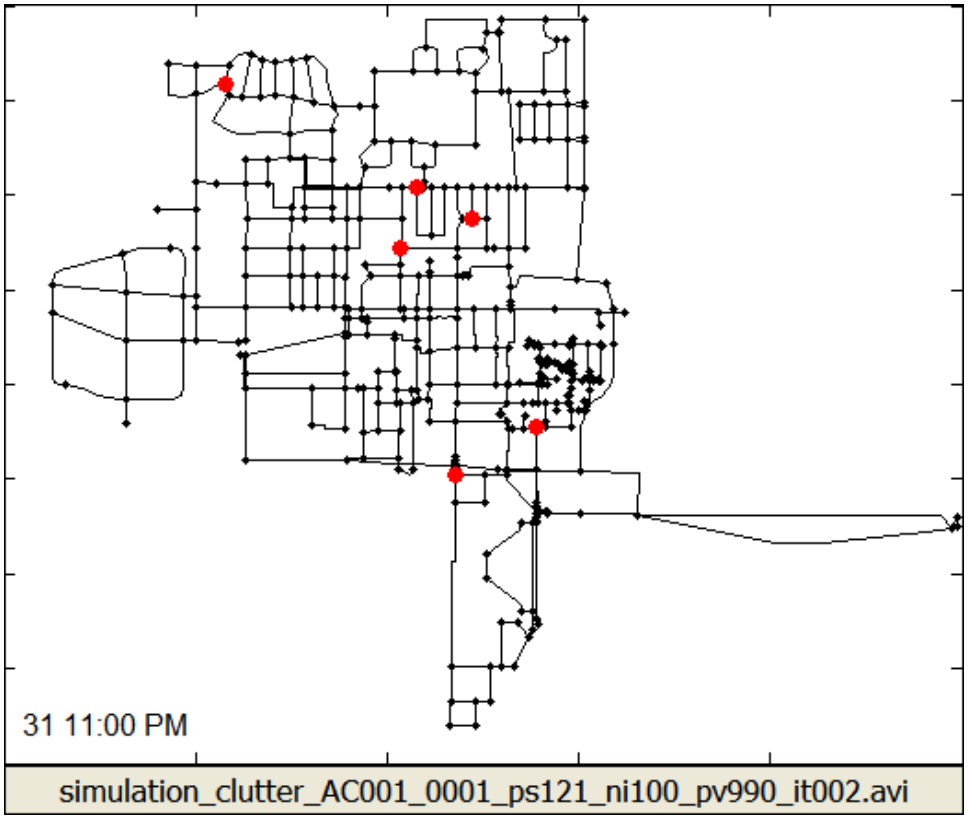




- Trade off between missed detection (MD error) and false alarm (FA) error
- If we want to improve probability of detection (1-MD) then FA error will increase
- Varying the threshold, t, and computing the FA and PD produces a receiver operating characteristic (ROC) curve.

Sandia National Laboratories

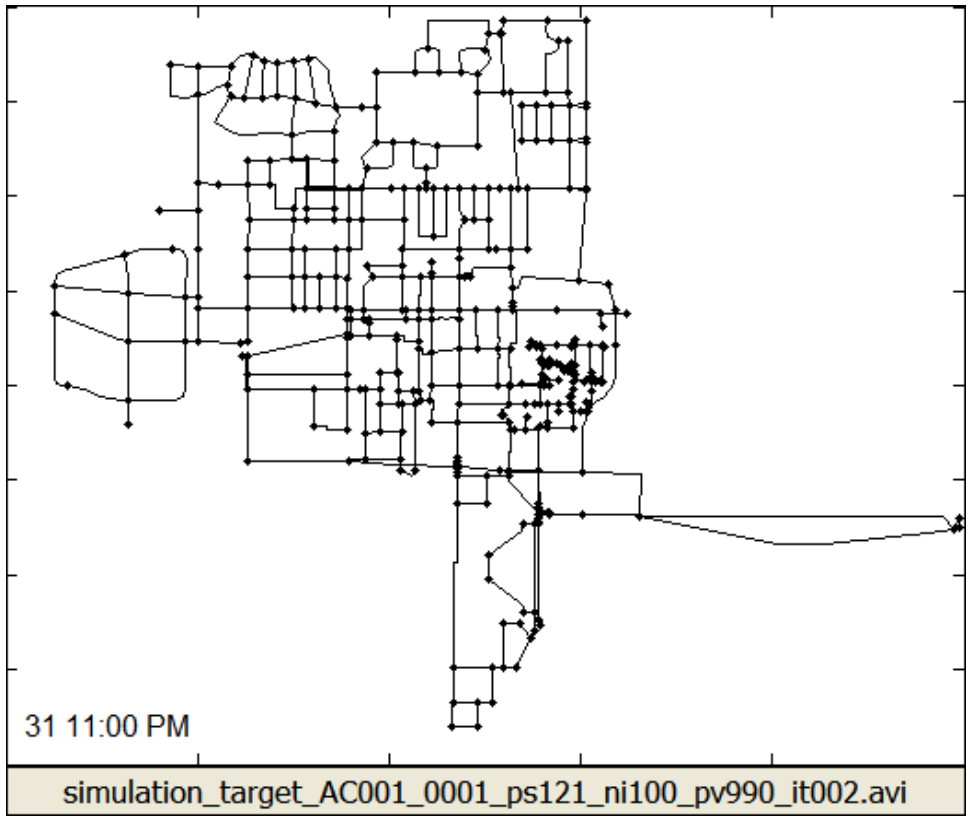
Simulation of Background Clutter



- AnyCity Network
- Parameters
  - 400 sensing nodes
  - 1 FA / day
  - 10 min. sample rate

Sandia National Laboratories

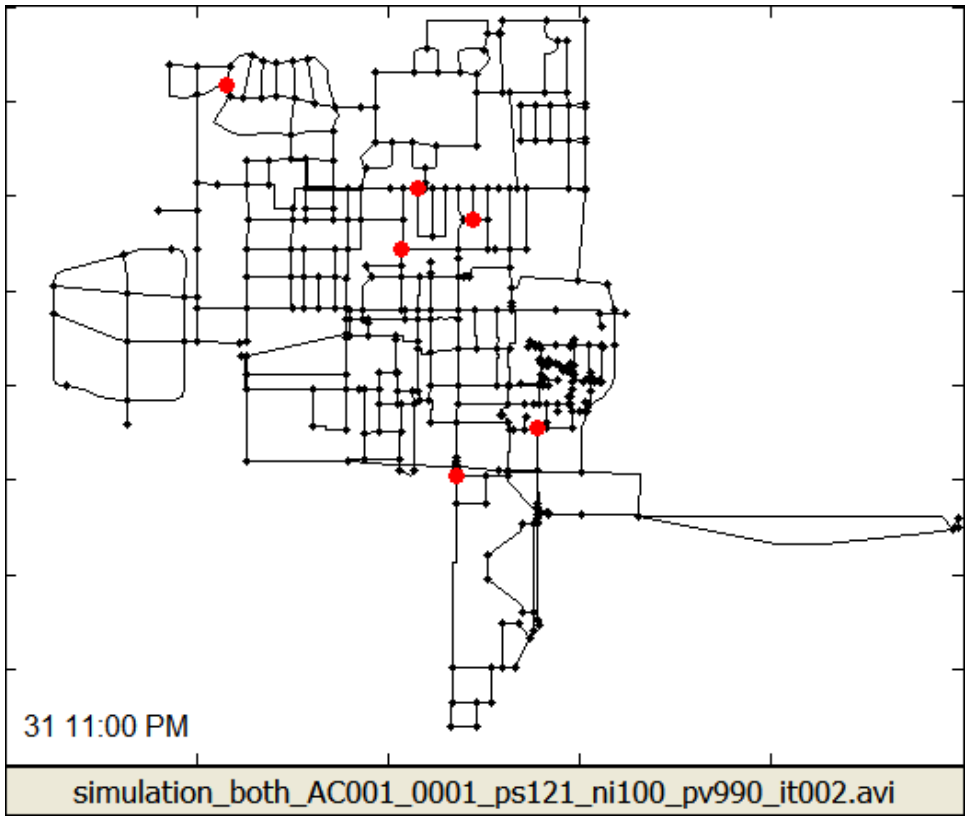
Simulation of Contamination



- Use EPANET to simulate injection of a tracer
- Parameters
  - 30 min injection
  - 50 mg/L
  - 5 mg/L detection threshold
  - 1/100 missed detections

Sandia National Laboratories

Combine Clutter & Contamination



Sandia National Laboratories

Space-Time Clustering

- Kulldorff scan test
  - Originally used to test whether disease is randomly distributed over space and time
  - Evaluate statistical significance of disease clusters
- Count cases in a sliding space-time window or zone
- Two hypotheses:
  - Null hypothesis H_0
    - For all the zones the probability of an event inside the zone is the same as outside
  - Alternative hypothesis H_1
    - There is at least one zone where the probability of an event inside a zone is greater that the probability outside
- Clusters can have different sizes and locations
- Use likelihood ratio
  - Not ad-hoc
- Clearly defined alternative
  - Avoid multiple and dependent tests

Kulldorf's Likelihood Function

$$L(z,p,q) = p^{c_z}(1-p)^{n_z-c_z}q^{C-c_z}(1-q)^{(N-n_z)-(C-c_z)}$$

- Likelihood function
  - *L(z,p,q)*
  - Likelihood that
    - number of events inside zone *z* is c_z and
    - number of events outside zone *z* is C-c_z

Parameters

- *z*: zone (window)
- *p*: probability of detecting an event inside *z*
- *q*: probability of detection an event outside *z*
- *C*: total number of events
- *N*: total number of sensing nodes in area of interest
- c_z: number cases inside *z*
- n_z: number of sensing nodes inside *z*
- N-n_z: number of sensing nodes outside *z*
- C-c_z: number of events outside *z*

Sandia National Laboratories

Kulldorff's Scan Test

- Null hypothesis H_0
  - For all the zones the probability of an event inside the zone is the same as outside

$$H_0 : p = q$$

- Alternative hypothesis H_1
  - There is at least one zone where the probability of an event inside a zone is greater that the probability outside

$$H_1 : z \in Z, p > q$$

$$\frac{L(z)}{L_0} = \frac{\sup_{z \in Z, p>q} L(z,p,q)}{\sup_{p=q} L(z,p,q)}$$

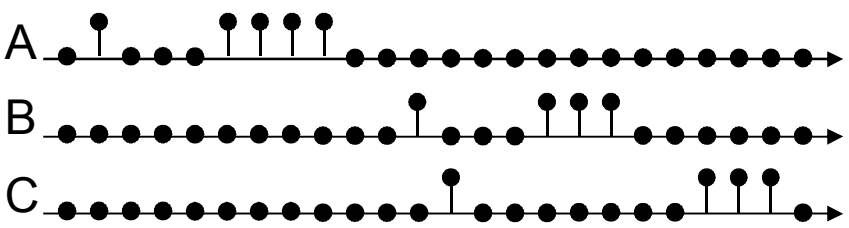
- Use randomization to determine the distribution of the null hypothesis
  - Monte Carlo randomization
    - Use *p=C/N* (=FA error)
  - Permutation testing

$$\lambda = \frac{\max_z L(z)}{L_0}$$

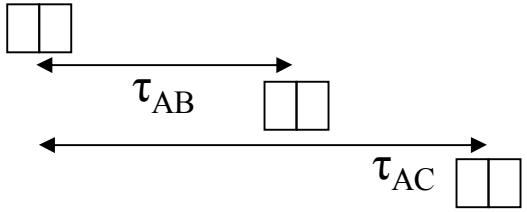
Sandia National Laboratories

Kulldorff's Test for Water Networks

Example time series at 3 sensing nodes



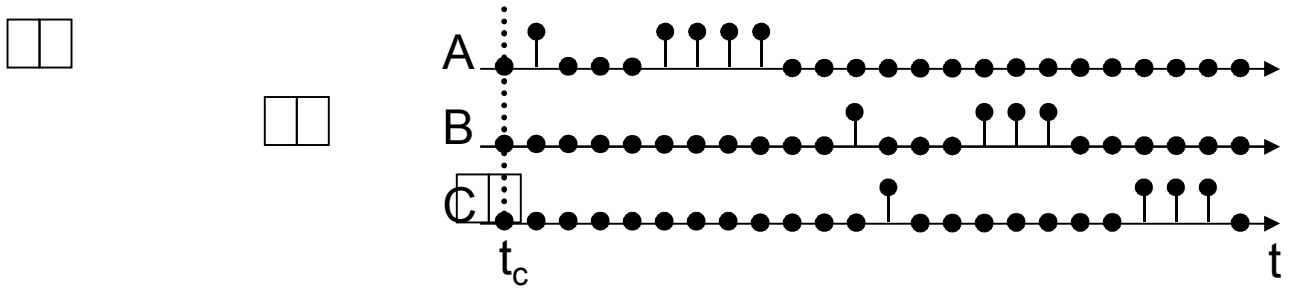
3x2 zone template with node A at the cluster center



- "Space" is defined by the water distribution network with units of travel time between nodes
  - τ_{AB}-Estimated travel time between node A & B
  - τ_{AC}-Estimated travel time between node A & C
  - Estimated using EPANET by taking the median of the travel times for each link over 24 hours
    - Driven by the stochastic demand models developed by Steve Buchberger and students
    - No assumptions on travel time direction
    - More refined information could be pulled from network model

Sandia National Laboratories

Searching for Clusters



Count: 00000000000110111113564

- t_c - current time
- Slide window through time to search for clusters
- Count events in template
- Apply scan test
- Multiple templates
  - Different cluster sizes
  - Each sensing node as a possible source

Sandia National Laboratories

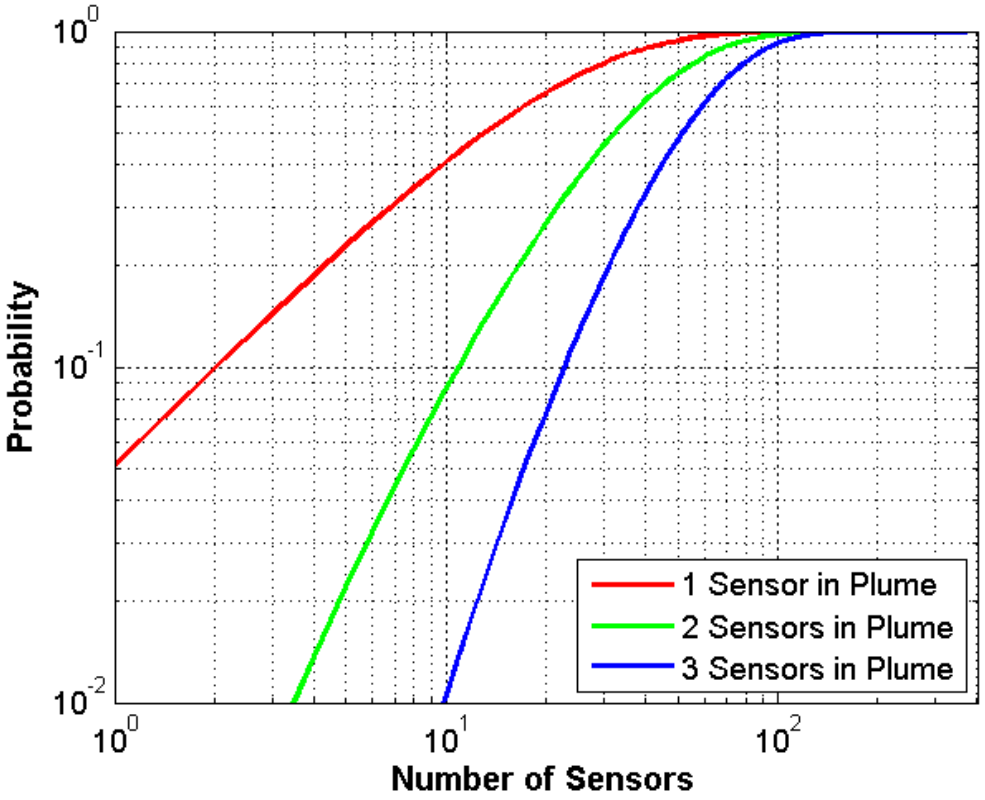
How Many Sensors?

$$\Pr(x \mid m) = \sum_{i=x}^{X} \binom{X}{i} \binom{M-m}{X-i} \Big/ \binom{M}{m}$$

- *x* – number of sensors in plume
- *m* – number of sensors nodes
- Pr(*x*|*m*) – probability of having at least *x* sensors in plume given *m* of sensing nodes
- *X* – smallest plume size (junctions)
- *M* – number of network junctions
- Hypergeometric Distribution

Sandia National Laboratories

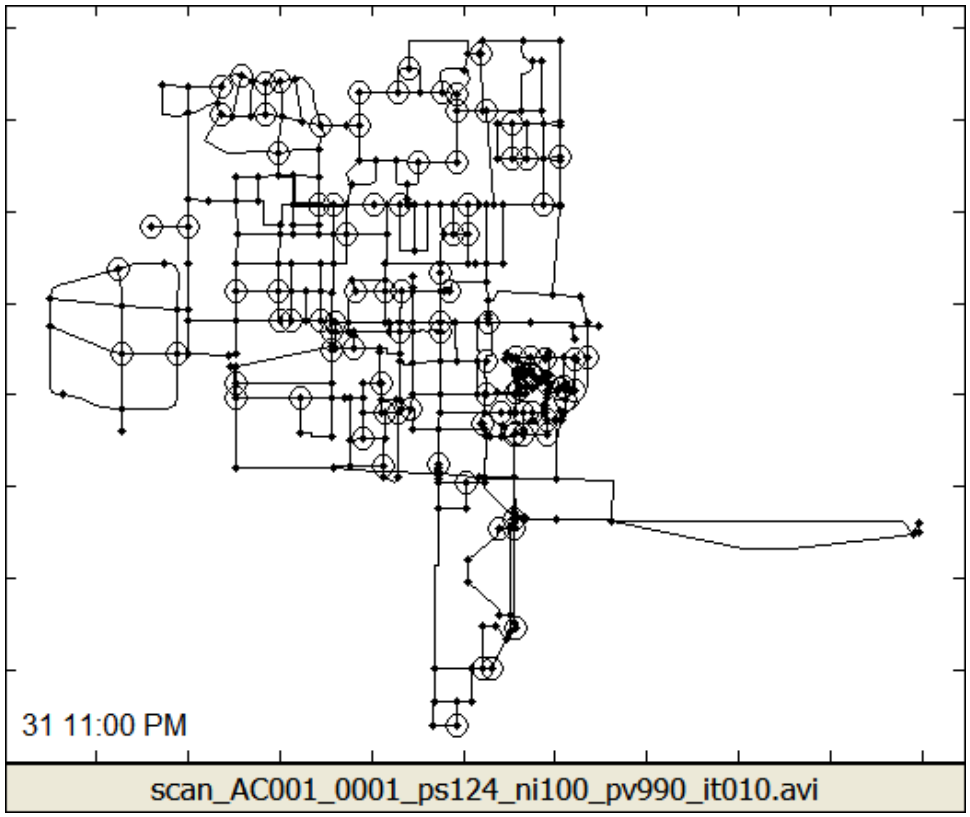
Number of Sensors: Example



- Network size: 396 junctions
- Min plume size: 20 junctions
- 1 sensor for a 0.05 probability of having at least 1 sensor in the plume
- 10 sensors for a 0.41 probability of having at least 1 sensor in the plume
- 80 sensors for a 0.99 probability of having at least 1 sensor in the plume

Sandia National Laboratories

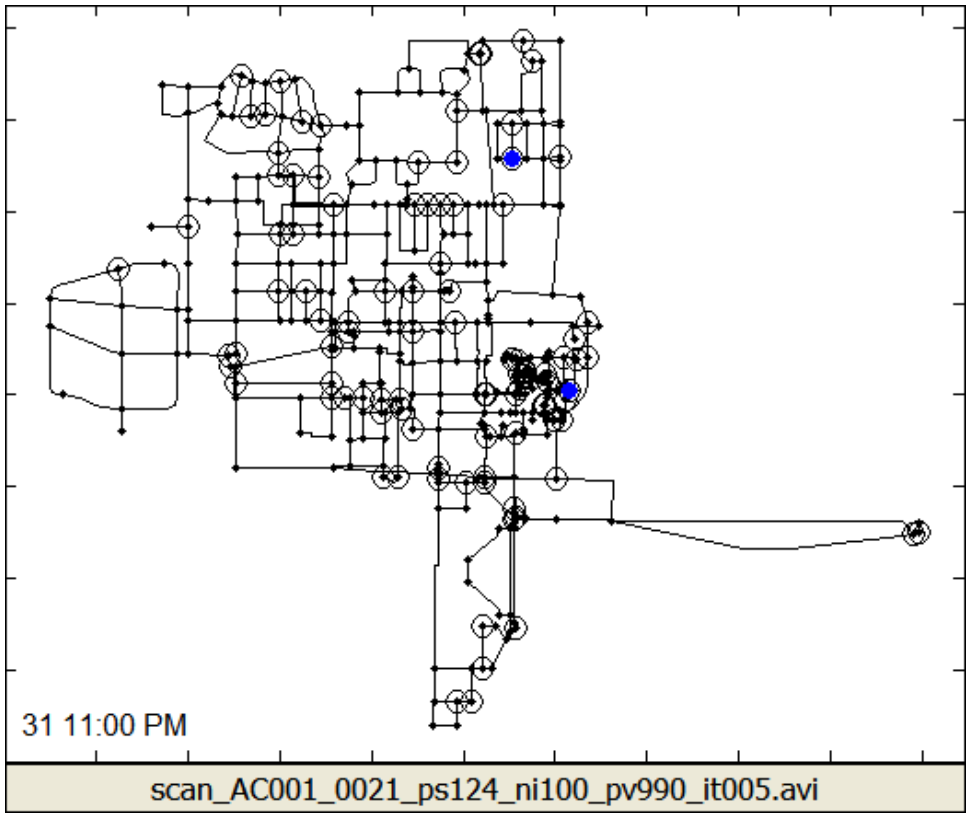
Kulldorff's Test with One Source



- Use EPANET to simulate injection of a tracer
- Parameters
  - 100 randomly placed sensing nodes
  - 10 min. sample rate
  - Background clutter
    - 1 FA / day
  - Contamination
    - 30 min injection
    - 50 mg/L
    - 5 mg/L detection threshold
    - 1/100 missed detections
  - Cluster
    - Threshold: 1 FA / 100 days
    - Space sizes: (1,3,6,12) (nodes)
    - Time sizes: (1,3,6) (time steps)

Sandia National Laboratories

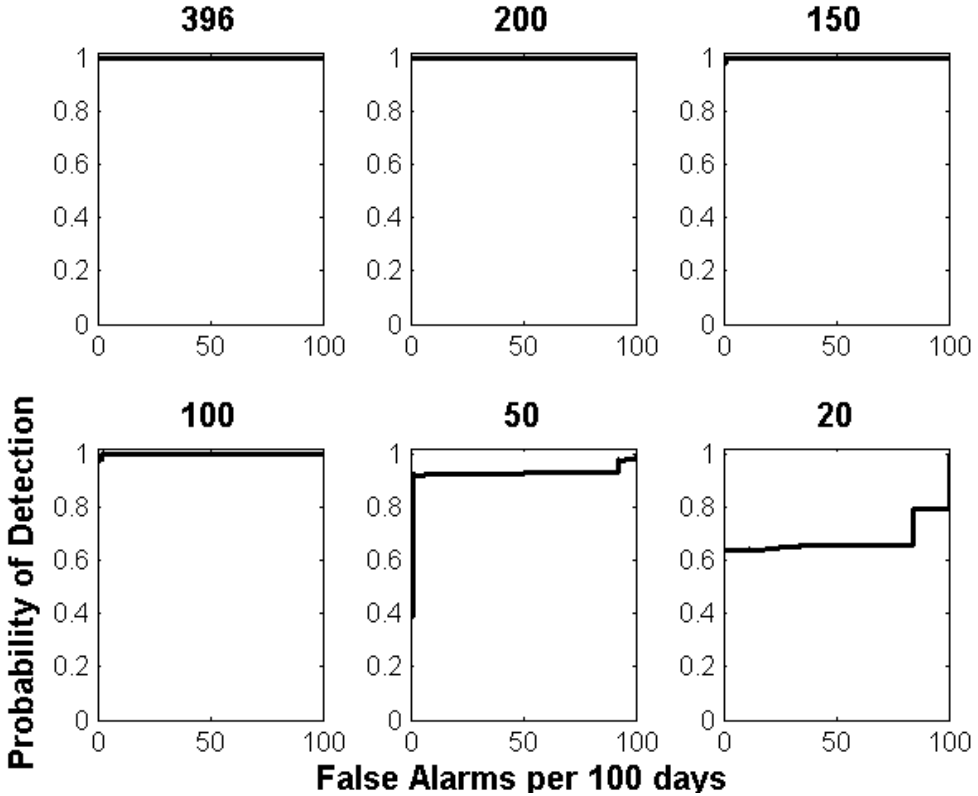
Kulldorff's Test with Two Sources



- Use EPANET to simulate injection of a tracer
- Parameters
  - 2 sources starting at same time
  - 100 randomly placed sensing nodes
  - 10 min. sample rate
  - Background clutter
    - 1 FA / day
  - Contamination
    - 30 min injection
    - 50 mg/L
    - 5 mg/L detection threshold
    - 1/100 missed detections
  - Cluster threshold
    - 1 FA / 100 days
    - Space - (1,3,6,12) (nodes)
    - Time - (1,3,6) (time steps)

Sandia National Laboratories

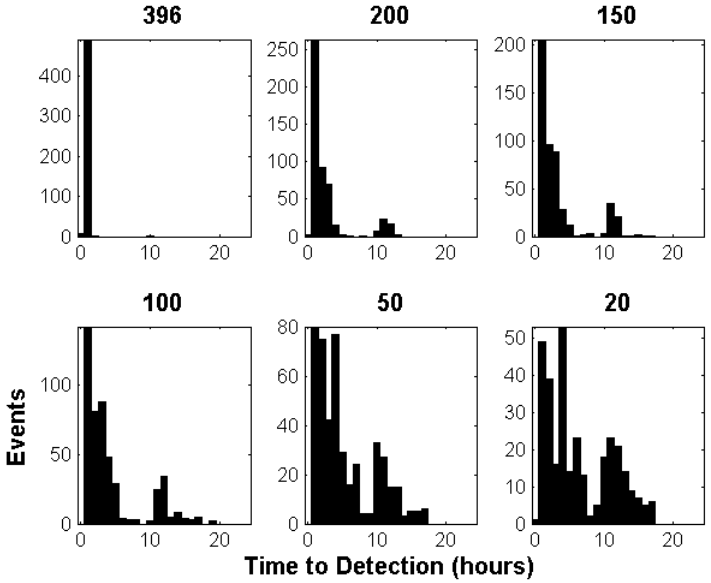
Operating Characteristics



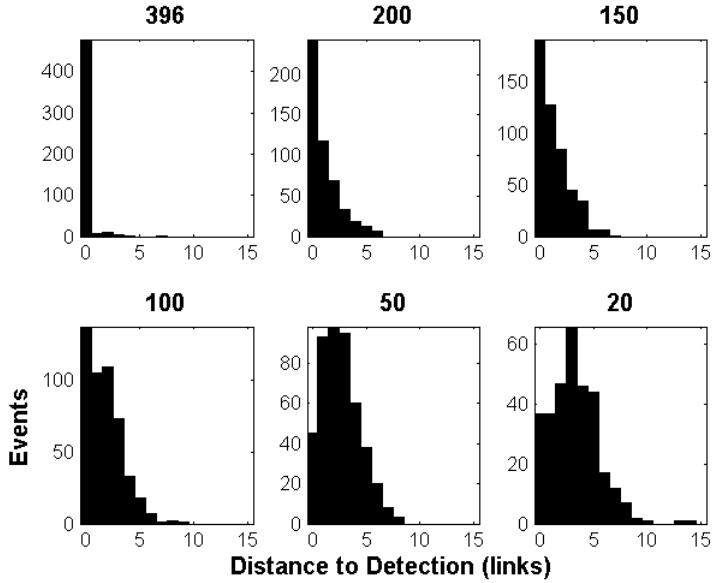
- Results based on single plume scenarios, but should extend to multiple plumes
- Need 80 independent sensors for 0.99 probability of having one sensor in the plume, but have will have 3 FA per hour
- Fusion reduces the false-alarms / day
- Flatness is due to the small number possible of scan-test values

Detection Statistics

Time (hours)



Distance (links)



- Results based on single plume scenarios, but should extend to multiple plumes
- Hypothesis: Bimodal time-to-detection histograms come from sensors at edge of plume

Sandia National Laboratories

Conclusions & Future Work

Conclusions

- Fusion improves performance by reducing the number of false alarms
- For the presented simulations
  - 80 independent sensors are needed to have a 0.99 prob. of having a least one sensor in the plume
  - But will give 3 FA's per hour
  - Using 100 nodes and fusion we can have a high PD with only 3 FA's per 100 days
- Center of cluster can help identify contamination source node
- Size of clusters can help identify contamination extent

Future Work

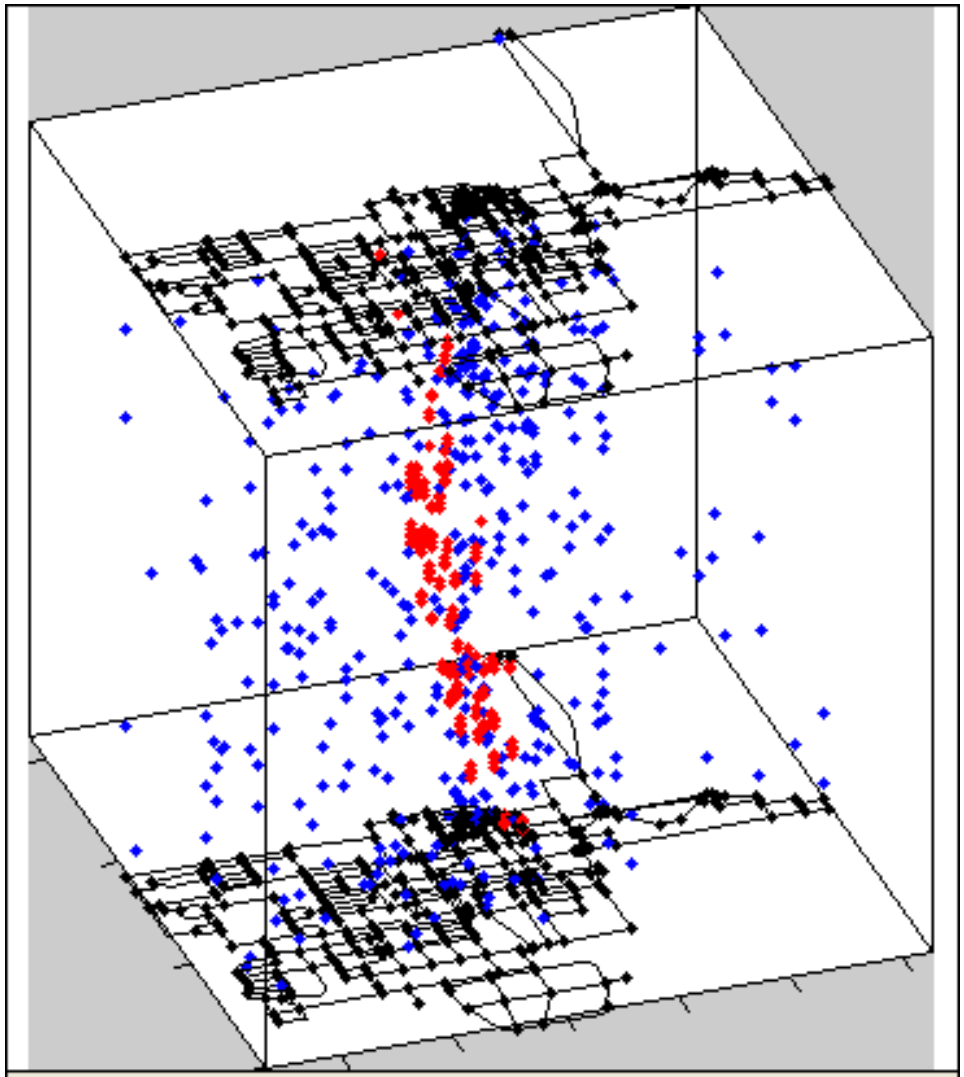
- Investigate larger networks with longer injection times
- Travel times between sensors
- Incorporate constraints of sensor response
- Improved determination of plume extent
  - Currently zone templates are designed to identify the contamination as soon as possible

Sandia National Laboratories

Extra

Sandia National Laboratories

Space Time Cube



cube_AC001_0001_ps121_ni100_pv990_it005.avi

Sandia National Laboratories