APDL: A Probabilistic Modeling Language for Anomalous Pattern Detection on Large Graphs

Feng Chen

University at Albany, SUNY

August 12, 2015

Introduction - Applications - 1

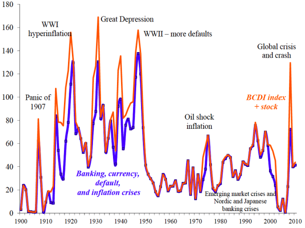
(a) Craigslist Scams

(b) Road Damage

(c) Email Spams

(d) Faked Reviews

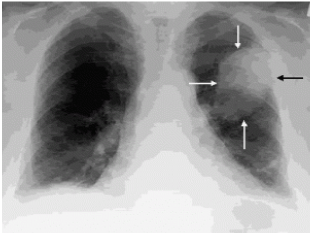
Introduction - Applications - 2



(a) Financial Crisis Events

(b) Scams on Facebook

(c) Fraud Receipts

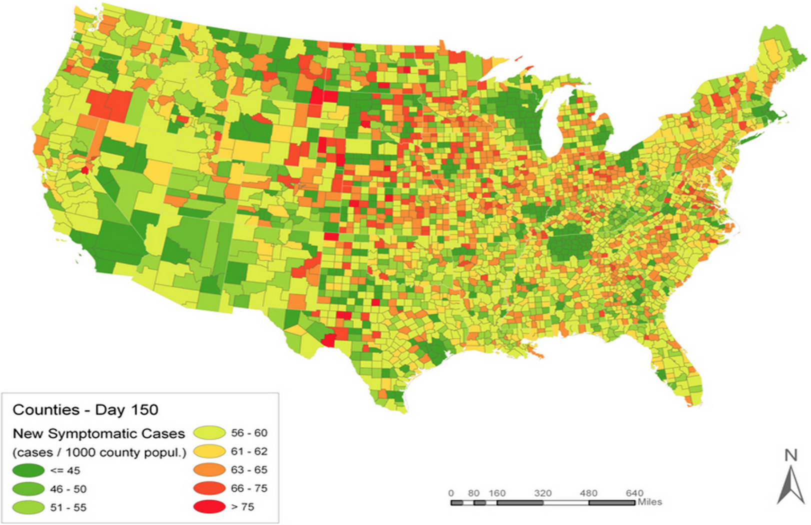


(d) Lung Cancer

How to Detect Anomalous Patterns?

- What are anomalous patterns?
  - A subset or group of data records that are interesting and unexpected
- How to detect anomalous patterns?
  - We model the distribution of normal patterns, and then any patterns (subsets) that deviate significantly from normal patterns are returned as anomalous patterns.
- What are the challenges?
  - Hard problems in general (e.g., exhaustive search takes time $O(2^N)$).
- What are the limitations of existing methods?
  - Most existing methods are designed based on specific assumptions of distributions of normal patterns.
  - The prior knowledge about anomalous patterns is not supported.
  - There is no unified and user-friendly framework that supports the detection of all kinds of anomalous patterns.

Case Study: Disease Outbreak Detection



Suppose Data $= \{d_1, d_2, \cdots, d_N\}$, where $d_i = (c_i^t, b_i^t)$, the pair of reported and expected number of flu cases in a county i on day t.

Hypothesis Testing :

Null hypothesis (H_0):

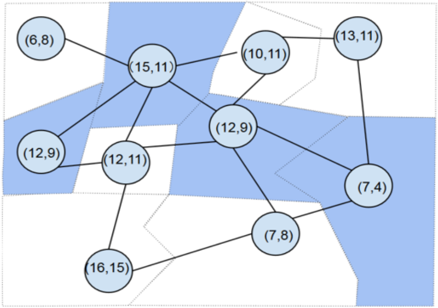
$$c_i^t \sim Poisson(q_{all} \cdot b_i^t)$$

Alternative hypothesis ($H_1(S)$),

where $S \subset \{1, \cdots, N\}$:

$$c_i^t \sim Poisson(q_{in} \cdot b_i^t)\text{, if } i \in S$$

$$c_i^t \sim Poisson(q_{out} \cdot b_i^t)\text{, otherwise}$$



The shaded region stand for flu outbreak regions

Log Likelihood Ratio $F(S)$:

$$F(S) = C(S) \log\frac{C(S)}{B(S)} + \left(C_{all} - C(S)\right) \log\left(\frac{C_{all} - C(S)}{B_{all} - B(S)}\right) - C_{all} \log\left(\frac{C_{all}}{B_{all}}\right)$$

Problem Formulation:

$$\max_{S \subset \{1, \cdots, N\}} F(S) \text{ s.t } S \text{ is connected}$$

Disease Outbreak Detection: Expectation Scan Statistic

Suppose Data $= \{d_1, d_2, \cdots, d_N\}$, where $d_i = (c_i^t, b_i^t)$, the pair of reported and expected number of flu cases in a county i on day t.

Hypothesis Testing:

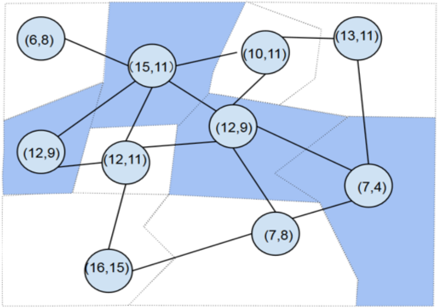
Null hypothesis (H_0):

$$c_i^t \sim Poisson(b_i^t)$$

Alternative hypothesis ($H_1(S)$), where $S \subset \{1, \cdots, N\}$:

$$c_i^t \sim Poisson(q_{in} \cdot b_i^t)\ , \text{ if } i \in S$$

$$c_i^t \sim Poisson(b_i^t)\ , \text{ otherwise}$$



The shaded region stand for flu outbreak regions

Log Likelihood Ratio $F(S)$:

$$F(S) = C(S) \log \frac{C(S)}{B(S)} + B(S) - C(S)$$

Problem Formulation:

$$\max_{S \subset \{1, \cdots, N\}} F(S) \text{ s.t } S \text{ is connected}$$

Example APDL language for Kulldorff's Scan Statistic

Suppose data $= \{d_1, \cdots, d_N\}$, where $d_i = (c_i^t, b_i^t)$, the pair of reported and expected counts of flu cases in a county i on day t.

Denote $V = \{1, \cdots, N\}$, $E \subseteq V \times V$, $C = \{c_1^t, \cdots, c_N^t\}$, $B = \{b_1^t, \cdots, b_N^t\}$

Hypothesis Testing:

- Null hypothesis (H_0):
  - $c_i^t \sim \text{Poisson}(q_{all} \cdot b_i^t)$
- Alternative hypothesis ($H_1(S)$), where $S \subseteq V$:
  - $c_i^t \sim \text{Poisson}(q_{in} \cdot b_i^t)$, if $i \in S$; $c_i^t \sim \text{Poisson}(q_{in} \cdot b_i^t)$, otherwise.

```
real  q_all
real  q_in
real  q_out
constrain(q_all > 0)
constrain(q_in > q_out)
constrain(q_out > 0)
V, E, C, B =
LoadGraphData(FileName)
set S
constrain(S ⊆ V)
constrain(S is connected)
hypothesis = {null, alternative}
if hypothesis == null:
    for v in V:
        C(v) ~ Poisson(q_all * B(v))
else hypothesis == alternative:
    for v in S:
        C(v) ~ Poisson(q_in * B(v))
    for v not in S:
        C(v) ~ Poisson(q_out * B(v))
Infer S
```

Example APDL language for Expectation Scan Statistic

Suppose data $= \{d_1, \cdots, d_N\}$, where $d_i = (c_i^t, b_i^t)$, the pair of reported and expected counts of flu cases in a county i on day t.

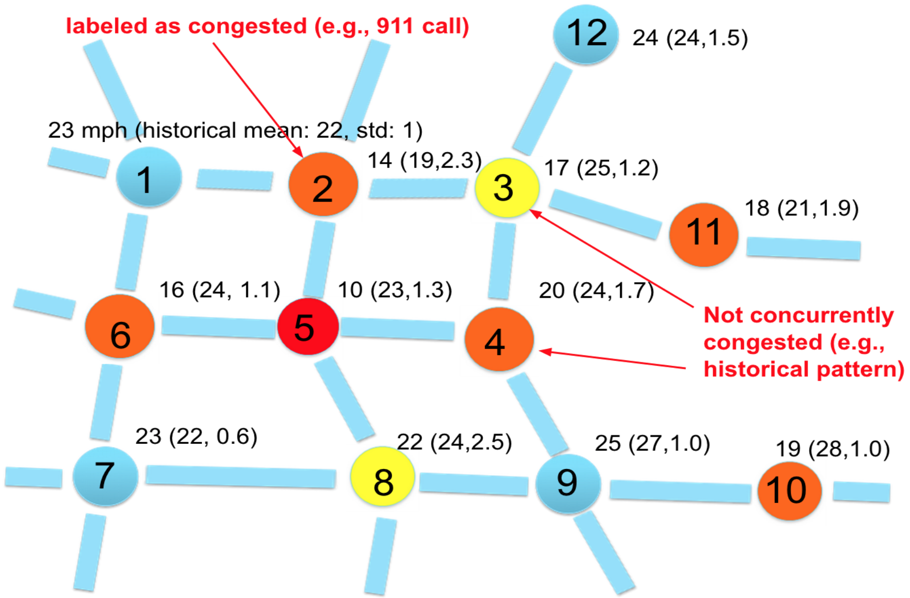
Denote $V = \{1, \cdots, N\}$, $E \subseteq V \times V$, $C = \{c_1^t, \cdots, c_N^t\}$, $B = \{b_1^t, \cdots, b_N^t\}$

Hypothesis Testing:

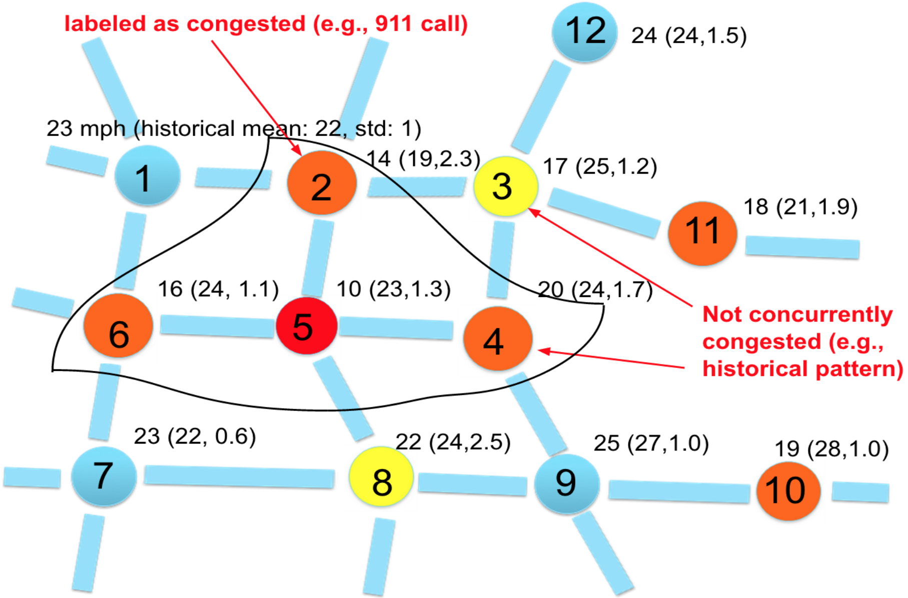
- Null hypothesis (H_0):
  - $c_i^t \sim \text{Poisson}(b_i^t)$
- Alternative hypothesis ($H_1(S)$), where $S \subset V$:
  - $c_i^t \sim \text{Poisson}(q_{in} \cdot b_i^t)$, if $i \in S$; $c_i^t \sim \text{Poisson}(b_i^t)$, otherwise.

```
real q_in
constrain(q_in > 1)
V, E, C, B =
LoadGraphData(FileName)
set S
constrain(S ⊆ V)
constrain(S is connected)
hypothesis = {null, alternative}
if hypothesis == null:
    for v in V:
        C(v) ~ Poisson(B(v))
else hypothesis == alternative:
    for v in S:
        C(v) ~ Poisson(q_in * B(v))
    for v not in S:
        C(v) ~ Poisson(B(v))
Infer S
```

Case Study: Traffic Congestion Detection in Road Network



Case Study: Traffic Congestion Detection in Road Network



Case Study: Traffic Congestion Detection in Road Network

Suppose data $= \{d_1, \cdots, d_N\}$, where $d_i = (x_i^t, \mu_i^t, \sigma_i^t)$, the tuple of reported speed, expected mean and standard deviation of normal speed in a road link i and hour t.

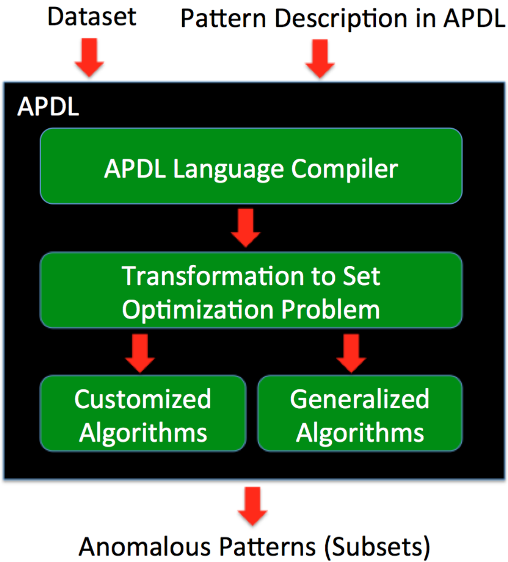
Hypothesis Testing:

- Null hypothesis (H_0):
  - $x_i^t \sim \mathcal{N}(\mu_i^t, \sigma_i^t)$
- Alternative hypothesis ($H_1(S)$):
  - $c_i^t \sim \mathcal{N}(q_{in} \cdot \mu_i^t, \sigma_i^t)$, if $i \in S$; $c_i^t \sim \mathcal{N}(\mu_i^t, \sigma_i^t)$, otherwise.

Prior Knowledge: 1) Road link 5 is currently congested (from 911 calls); 2) road links 3 and 4 are not congested concurrently (patterns from historical data).

```
real q_in
constrain(q_in > 1)
V, E, X, μ, σ =
LoadGraphData(FileName)
set S
constrain(S ⊆ V)
constrain(S is connected)
set S_0 = {5}
constrain(S_0 ⊆ S)
set S_1 = {3, 4}
constrain(|S_1 ∩ S| ≤ 1)
hypothesis = {null, alternative}
if hypothesis == null:
    for v in V:
        X(v) ~ N(μ(v), σ(v))
else hypothesis == alternative:
    for v in S:
        X(v) ~ N(q_in · μ(v), σ(v))
    for v not in S:
        X(v) ~ N(μ(v), σ(v))
Infer S
```

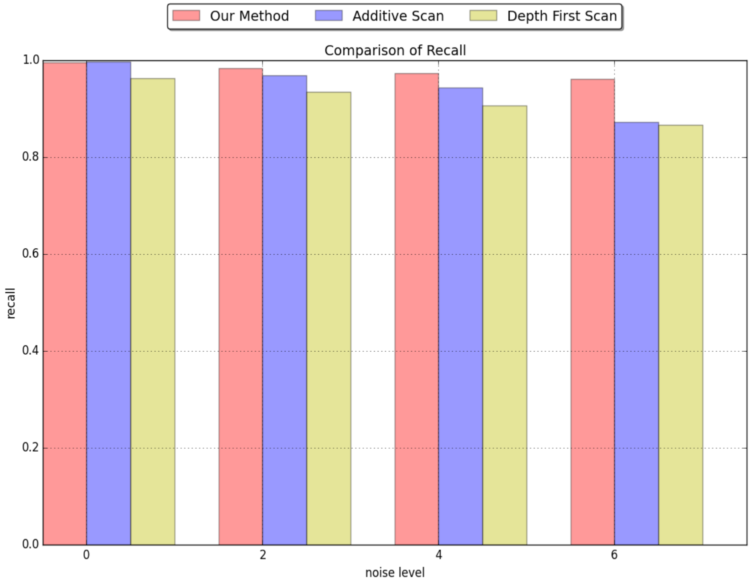
APDL Architecture



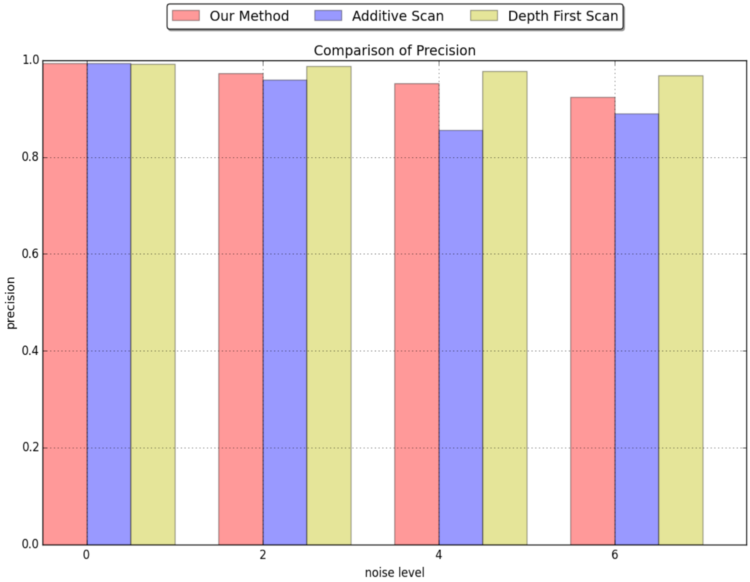
Experiments: Pollution Detection in Water Sensor Network

- Water pollution data set: The "Battle of the Water Sensor Networks" (BWSN) provides a realworld network of 12,527 nodes, and 25 nodes with chemical contaminant plumes that are distributed in four different areas.
- The spreads of these contaminant plumes on graph were simulated using the water network simulator EPANET that was used in BWSN for a period of 8 hours.
- The task of anomalous pattern detection is to detect the infected nodes by chemical contaminant plumes.
- Competitive methods:
  - Depth First Search Graph Scan [Speakman et.al., Journal of Computational and Graphical Statistics, 2015]
  - Additive Graph Scan [Speakman et.al., Proc. 13th IEEE International Conference on Data Mining, 2013]

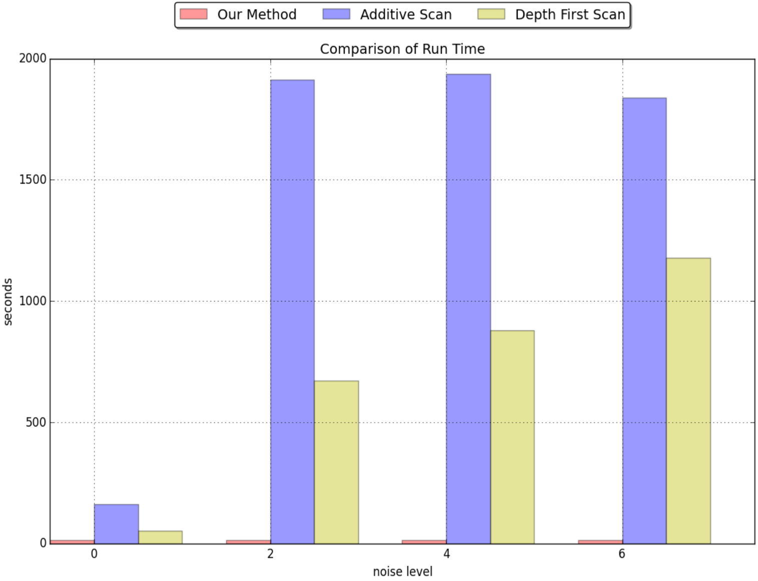
Comparison on Recall



Comparison on Recall



Comparison on Run-Time



The End