Cost-Effective Outbreak Detection in Networks

J. Leskovec et al.

David LaPalomento

February 8, 2010

Motivation

Detecting contamination in municipal water distribution network
Selecting blogs to monitor for influential trends

Goals

- Early detection
- Guaranteeing detection
- “Sensor” cost
- Incident impact

Evaluating Sensor Placement

Definition

Penalty:

- Time before incident is detected by sensor
- Number of nodes exposed to incident before detection

Evaluating Sensor Placement

Minimize penalty over all possible incidents

$$\pi(A) = \sum_i P(i)\pi(T(i, A))$$

- π is the penalty
- A is the set of sensors in the a placement
- P is the incident probability distribution
- $T(i, A)$ is the best detection time across all sensors in the placement A

An Alternative Evaluation Function

Reduction in penalty for a specific incident:

$$R_i = \pi(\infty) - \pi_i(T(i, A))$$

Reduction in penalty across all incidents:

$$R = \sum_i P(i) R_i(A) = \pi(\emptyset) - \pi(A)$$

A Reminder...

Definition

Submodularity: the benefit of adding an element to a smaller set is guaranteed to be equal or greater than the benefit of adding that element to any larger set

Defining Penalty

- Detection likelihood
- Detection time
- Population affected

Multicriterion Optimization

$$\overrightarrow{R} = (R_i(A), R_2(A), ..., R_m(A))$$

Definition

Pareto-optimal: a solution such that no other solution exists which is at least as good in all criteria and strictly better in one

Definition

Scalarization: pick positive weights for all criteria and sum over the products

Unit-Cost Greedy Algorithm

All nodes have equal cost, maximize the marginal benefit at each step
If the cost is actually equal, this algorithm is within 63% of optimal

Variable Cost Greedy Algorithm

Maximize benefit-cost ratio at each step
No longer guaranteed bound against the optimal solution

Cost Effective Forward (CEF) Selection

Compute greedy benefit-cost and greedy unit cost
Select the solution with the better score
Guaranteed bound against the optimal
$O(B|V|)$

"Online" Bound Computation

Observations

Assume outbreaks are "sparse"
Allows for powerful optimizations in conjunction with penalty-reduction formulation

Inverted Index

Index reduction in penalty by sensor index, s

$$R(A) = \sum_{\text{i s.t. i detected by A}} P(i) max_{s \in A} R_i(\{s\})$$

Submodularity!
Reduce the number of marginal-benefit calculations
Store marginal benefits calculated in a priority queue
Evaluate invalidated nodes in decreasing order. More often than not, the highest benefit node stays on the top after re-evaluation

Blog Networks

Dataset drawn from blogs with at least 3 incoming links in the first 6 months of 2006

- 45,000 blogs
- 10.5 million posts
- 16.2 million links, 1 million "internal" links
- 30 GB of data
- 17,589 cascades

Setup

Objectives: Detection Likelihood, Detection Time, Population Affected
Unit-cost model selects big blogs with many posts
Cost-model + number of posts $\Rightarrow$ aggregators

Results

Comparison with Heuristics Friday is the best day to read blogs

Shortcomings

CELF overfit present data
Prevent CELF from selecting particularly small blogs
Degraded results but better generalization

Watere Sensor Networks

Data from Battle of Water Sensor Networks (BWSN)

- 21k nodes
- 25k pipes
- 3.6 million contamination scenarios

Setup

- Detection Likelihood
- Detection Time
- Population Affected

Population affected increases very quickly (sparsity) Detection Likelihood, Detection Time: grow with more sensors

Results

Concentrate in areas with high population density to decrease population affected Spread sensors out to increase detection likelihood and detection time