Using Partial Probes to Infer Network States

Pavan Rangudu[°], Bijaya Adhikari[*], B. Aditya Prakash[*], Anil Vullikanti[*][°]

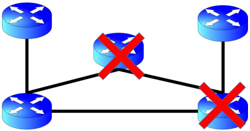
[*]Department of Computer Science, Virginia Tech
[°]NDSSL, Biocomplexity Institute, Virginia Tech

Contact: `badityap@cs.vt.edu`

VirginiaTech
Invent the Future

Motivation

- Network nodes and links fail dynamically
- Networks not known fully because of privacy constraints
- Our focus: if some failed nodes are known, can we infer the states of the remaining nodes?



Node failures in internet

Traffic jam in road network

Prior works fail to the address the problem directly.

Our model

- Graph $G(V, E)$ with set $I \subseteq V$ which have failed
- Goegraphically correlated failure model [Agarwal et al., 2013]
  - Single seed of the failure, with probability $p_s(v)$ of node v being the seed
  - Correlated failure model: $F(u|v)$ denotes the probability that node u fails given that v has failed
    - Assume independence, i.e., $F(u_1, u_2|v) = F(u_1|v) \cdot F(u_2|v)$
  - Motivation: attacks or natural disasters in infrastructure networks
- Probes: subset $Q \subseteq I$ of failed nodes is known
- Objective: find the set $I - Q$

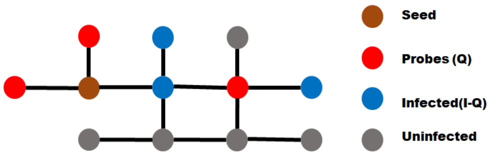


Figure: A toy road network with node failures

Our approach: Minimum Description Length

- Model cost $\mathcal{L}(|\mathcal{Q}|, |I|, I)$ has three components

$$\mathcal{L}(|\mathcal{Q}|, |I|, I) = \mathcal{L}(|\mathcal{Q}|) + \mathcal{L}\big(|I| \,\big|\, |\mathcal{Q}|\big) + \mathcal{L}\big(I \,\big|\, |\mathcal{Q}|, |I|\big).$$

  - $\mathcal{L}(|\mathcal{Q}|) = -\log\Big(Pr(|\mathcal{Q}|)\Big)$ by using the *Shannon-Fano* code
  - $\mathcal{L}\big(|I| \,\big|\, |\mathcal{Q}|\big) = -\log\left(\frac{Pr\left(|\mathcal{Q}| \,\middle|\, |I|\right) Pr(|I|)}{Pr(|\mathcal{Q}|)}\right)$
  - $\mathcal{L}\big(I \,\big|\, |\mathcal{Q}|, |I|\big) = -\log\Big(Pr\big(I \,\big|\, |\mathcal{Q}|, |I|\big)\Big) = -\log\Big(Pr\big(I \,\big|\, |I|\big)\Big)$
- Data cost: description of $\mathcal{Q}^+ = I \setminus \mathcal{Q}$ (assuming no observation errors)
  - $\mathcal{L}(\mathcal{Q}^+|I) = -\log\left(\gamma^{|\mathcal{Q}|}(1-\gamma)^{|\mathcal{Q}^+|}\right) =$ $-|\mathcal{Q}|\log(\gamma) - (|I| - |\mathcal{Q}|)\log(1-\gamma)$

Problem Description

Model Cost

$$\begin{aligned}
\mathcal{L}(|\mathcal{Q}|, |I|, I) =& \mathcal{L}(|\mathcal{Q}|) + \mathcal{L}\big(|I| \,\big|\, |\mathcal{Q}|\big) + \mathcal{L}\big(I \,\big|\, |\mathcal{Q}|, |I|\big) \\
=& -\log \binom{|I|}{|\mathcal{Q}|} - |\mathcal{Q}| \log(\gamma) - (|I| - |\mathcal{Q}|) \log(1-\gamma) \\
& - \log \left(\sum_{s \in V} p_s(s) \prod_{v \in I} F(v \mid s) \prod_{v' \notin I} \big(1 - F(v' \mid s)\big) \right)
\end{aligned}$$

*after algebra

Problem Formulation

Given G, p_s, $F(\cdot)$, Q, find I that minimizes the total MDL cost:

$$\mathcal{L}\big(|\mathcal{Q}|, |I|, I, \mathcal{Q}\big) = -\log \binom{|I|}{|\mathcal{Q}|} - \log \left(\sum_{s \in V} p_s(s) \prod_{v \in I} F(v \mid s) \prod_{v' \notin I} \big(1 - F(v' \mid s)\big) \right)$$

$$-2|\mathcal{Q}| \log(\gamma) - 2(|I| - |\mathcal{Q}|) \log(1-\gamma)$$

Algorithm Greedy

Input: Instance $(V, \mathcal{Q}, p, P, \gamma)$
Output: Solution $\hat{I}$ that minimizes $\mathcal{L}(|\mathcal{Q}|, |\hat{I}|, \hat{I}, \mathcal{Q})$

```
for each s ∈ V do
    for each k ∈ [|Q|, |V|] do
        I_s(k) ← Top k − |Q| nodes in V \ Q with highest weight f(s, v)
        I_s(k) ← I_s(k) ∪ Q
    end for
end for
S ← {I_s(k) : ∀s ∈ V & k ∈ [|Q|, |V|]}
Î ← argmin_{I∈S} L(|Q|, |I|, I, Q)
Return Î
```

- **for** each $s \in V$ **do**
  - **for** each $k \in [|\mathcal{Q}|, |V|]$ **do**
    - $I_s(k) \leftarrow$ Top $k - |\mathcal{Q}|$ nodes in $V \setminus \mathcal{Q}$ with highest weight $f(s, v)$
    - $I_s(k) \leftarrow I_s(k) \cup \mathcal{Q}$
  - **end for**
- **end for**
- $\mathcal{S} \leftarrow \{I_s(k) : \forall s \in V \& k \in [|\mathcal{Q}|, |V|]\}$
- $\hat{I} \leftarrow \arg\min_{I \in \mathcal{S}} \mathcal{L}(|\mathcal{Q}|, |I|, I, \mathcal{Q})$
- Return $\hat{I}$

Analysis of GREEDY

Theorem: (Additive Approximation)

Let I^* be the set minimizing the MDL cost, and let I denote the solution computed by Algorithm GREEDY. Then, $\mathcal{L}(|\mathcal{Q}|, |I|, I, \mathcal{Q}) \leq \mathcal{L}(|\mathcal{Q}|, |I^*|, I^*, \mathcal{Q}) + \log(n)$, where n is the number of seed nodes.

Running time

Algorithm GREEDY runs in $O(|V|^3)$ time

Experiments

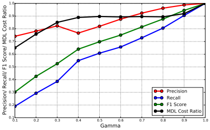
- Baseline: local improvement algorithm LocalSearch
- Datasets
  - Synthetic grid
    - 60×60 grid
    - Uniform seed probability $p_s(\cdot)$
    - Conditional failure probability distribution using model of [Agarwal et al., 2013]: $F(v \mid s) = 1 - d(s, v)$, where $d(\cdot)$ is (normalized) distance
  - Real datasets: Seed and conditional failure probability distributions computed from data
    - JAM data from WAZE for Boston: road network with 2650 nodes.
    - WEATHER data from WAZE for Boston: road network with 1520 nodes.
    - POWER-GRID: network of 24 nodes from Electric disturbance events

WAZE dataset

Visualization of Waze dataset. Partitions in the 119 × 78 grid represent nodes in our network.

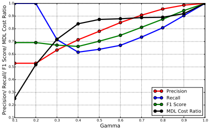
Takeaways

Results for JAM dataset



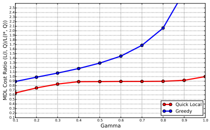
Algori

LocalSearch



Algori

Greedy



Comp

of the MDL costs

- Our MDL based approach helps identify missing failures
- Promising approach for other problems with missing information