Spotting Culprits in Epidemics: How many and Which ones?

B. Aditya Prakash  Virginia Tech
Jilles Vreeken  University of Antwerp
Christos Faloutsos  Carnegie Mellon University
Virus Propagation

- Susceptible-Infected (SI) Model

Diseases over contact networks

CDC data: Visualization of the first 35 tuberculosis (TB) patients and their 1039 contacts

Prakash, Vreeken, Faloutsos 2012
Outline

• Motivation---Introduction
• Problem Definition
• Intuition
• MDL
• Experiments
• Conclusion

Prakash, Vreeken, Faloutsos 2012
Culprits: Problem definition

2-d grid

Q: Who started it?

Prakash, Vreeken, Faloutsos 2012
Culprits: Problem definition

Q: Who started it?

Prior work:
[Lappas et al. 2010, Shah et al. 2011]
Outline

• Motivation---Introduction
• Problem Definition
• Intuition
• MDL
• Experiments
• Conclusion

Prakash, Vreeken, Faloutsos 2012
Culprits: Exoneration

(a) A chain
Culprits: Exoneration

(a) A chain

(b) A chain-star
Who are the culprits

- Two-part solution
  - use MDL for \textit{number} of seeds
  - for a given number:
    - exoneration = centrality + penalty

- Running time = $O(k^* (E_I + E_F + V_I))$
  - linear! (in edges and nodes)

NetSleuth

Prakash, Vreeken, Faloutsos 2012
Outline

• Motivation---Introduction
• Problem Definition
• Intuition
• MDL
  – Construction
  – Optimization
• Experiments
• Conclusion

Prakash, Vreeken, Faloutsos 2012
Modeling using MDL

- Minimum Description Length Principle == Induction by compression
- Related to Bayesian approaches
- MDL = Model + Data
- Model
  - Scoring the seed-set

\[ \mathcal{L}(S) = \mathcal{L}_N(|S|) + \log \left( \frac{N}{|S|} \right) \]

- Encoding integer \(|S|\)
- Number of possible \(|S|\)-sized sets

Prakash, Vreeken, Faloutsos 2012
Modeling using MDL

• Data: Propagation Ripples

Original Graph

Infected Snapshot

Ripple R1

Ripple R2

Prakash, Vreeken, Faloutsos 2012
Modeling using MDL

- Ripple cost

\[ \mathcal{L}(R \mid S) = \mathcal{L}_N(T) + \sum_{t=1}^{T} \mathcal{L}(F^t) \]

How long is the ripple

How the ‘frontier’ advances

- Total MDL cost

\[ \mathcal{L}(G_I, S, R) = \mathcal{L}(S) + \mathcal{L}(R \mid S) \]

Prakash, Vreeken, Faloutsos 2012
Outline

• Motivation---Introduction
• Problem Definition
• Intuition
• **MDL**
  – Construction
  – *Optimization*
• Experiments
• Conclusion

Prakash, Vreeken, Faloutsos 2012
How to optimize the score?

• Two-step process
  – Given $k$, quickly identify high-quality set
  – Given these nodes, optimize the ripple $R$
Optimizing the score

• High-quality $k$-seed-set
  – Exoneration

• Best single seed:
  – Smallest eigenvector of Laplacian sub-matrix
  – Analyze a Constrained SI epidemic

• Exonerate neighbors

• Repeat

Prakash, Vreeken, Faloutsos 2012
Optimizing the score

• Optimizing $R$
  – Get the MLE ripple!

• Finally use MDL score to tell us the best set

• NetSleuth: Linear running time in nodes and edges $O(k^* (E_I + E_F + V_I))$

Prakash, Vreeken, Faloutsos 2012
Outline

• Motivation---Introduction
• Problem Definition
• Intuition
• MDL
• Experiments
• Conclusion
Experiments

• Evaluation functions:
  – MDL based
    \[ Q_{MDL} = \frac{\mathcal{L}(G_I, S, R)}{\mathcal{L}(G_I, S^*, R^*)} \]
  – Overlap based
    \[ Q_{JD} = \frac{\mathbb{E}[JD_S(V_I)]}{\mathbb{E}[JD_{S^*}(V_I)]} \]
    \((JD == \text{Jaccard distance})\)

Prakash, Vreeken, Faloutsos 2012
Experiments: # of Seeds

One Seed

Two Seeds

Three Seeds
Experiments: Quality (MDL and JD)

\[ Q_{MDL} = \frac{\mathcal{L}(G_I, S, R)}{\mathcal{L}(G_I, S^*, R^*)} \]

\[ Q_{JD} = \frac{\mathbb{E}[JD_{S(V_I)}]}{\mathbb{E}[JD_{S^*(V_I)}]} \]

Ideal = 1

Prakash, Vreeken, Faloutsos 2012
Experiments: Quality (Jaccard Scores)

Closer to diagonal, the better
Experiments: Scalability

![Graph showing scalability]

Prakash, Vreeken, Faloutsos 2012
Outline

• Motivation---Introduction
• Problem Definition
• Intuition
• MDL
• Experiments
• Conclusion
Conclusion

- **Given**: Graph and Infections
- **Find**: Best ‘Culprits’

- **Two-part** solution
  - use **MDL** for *number* of seeds
  - for a given number:
    \[ \text{exoneration} = \text{centrality} + \text{penalty} \]

- **NetSleuth**:
  - Linear running time in nodes and edges
  \[ O(k^* (E_I + E_F + V_I)) \]
Any Questions?

B. Aditya Prakash
http://www.cs.vt.edu/~badityap

Culprits: Problem definition
2-d grid

Q: Who started it?

Who are the culprits

- Two-part solution
  - use MDL for number of seeds
  - for a given number:
    * exoneration = centrality + penalty

- Running time = $O(k^*(\mathcal{E}_I + \mathcal{E}_F + \mathcal{V}_t))$
  - linear! (in edges and nodes)

Experiments: # of Seeds

Prakash, Vreeken, Faloutsos 2012