# Spotting Culprits in Epidemics: How many and Which ones?

**B. Aditya Prakash** *Virginia Tech*

**Jilles Vreeken** *University of Antwerp*

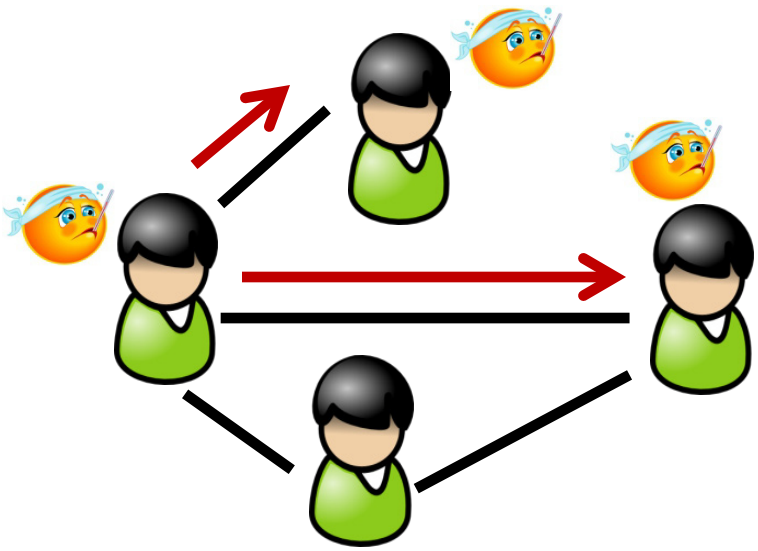
**Christos Faloutsos** *Carnegie Mellon University*

Universiteit Antwerpen

IEEE ICDM Brussels
December 11, 2012

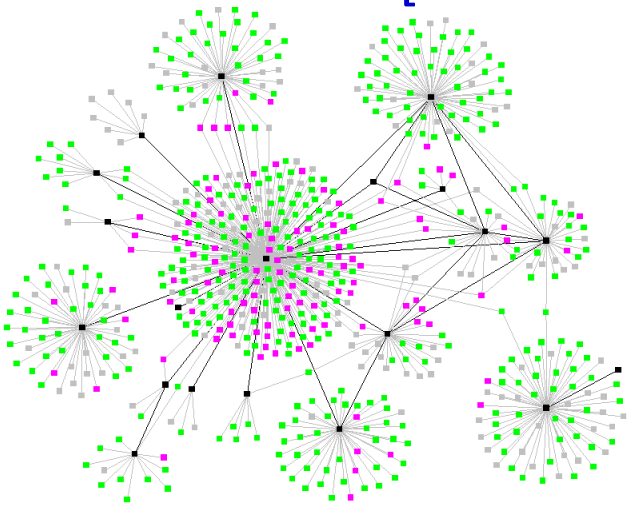
# Virus Propagation

- Susceptible-Infected (SI) Model



*Diseases over contact networks*

[AJPH 2007]



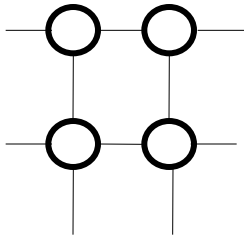
CDC data: Visualization of the first 35 tuberculosis (TB) patients and their 1039 contacts

# Outline

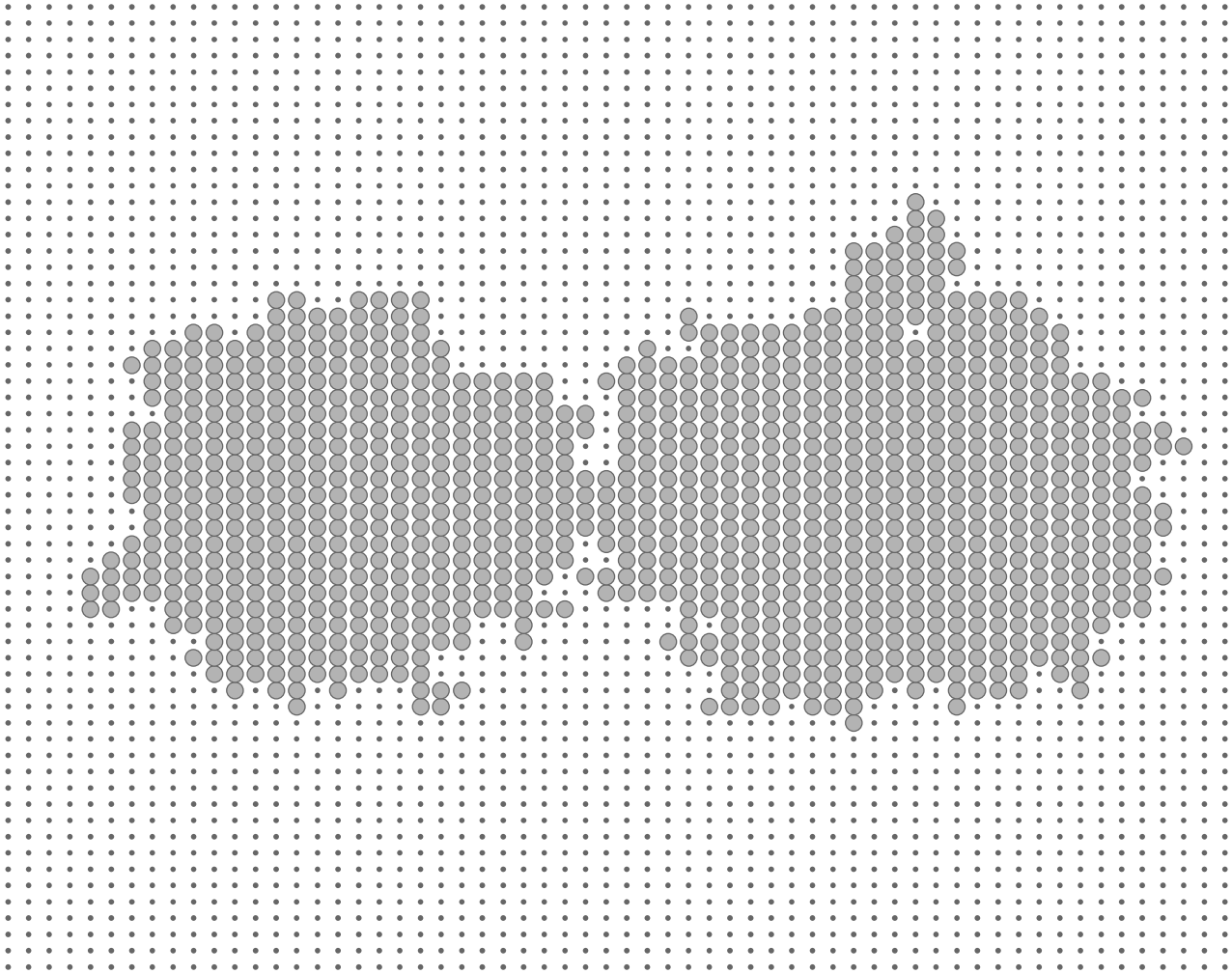
- Motivation---Introduction
- **Problem Definition**
- Intuition
- MDL
- Experiments
- Conclusion

# Culprits: Problem definition

2-d grid



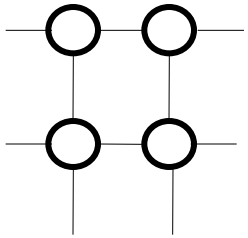
Q: *Who started it?*



Prakash, Vreeken, Faloutsos 2012

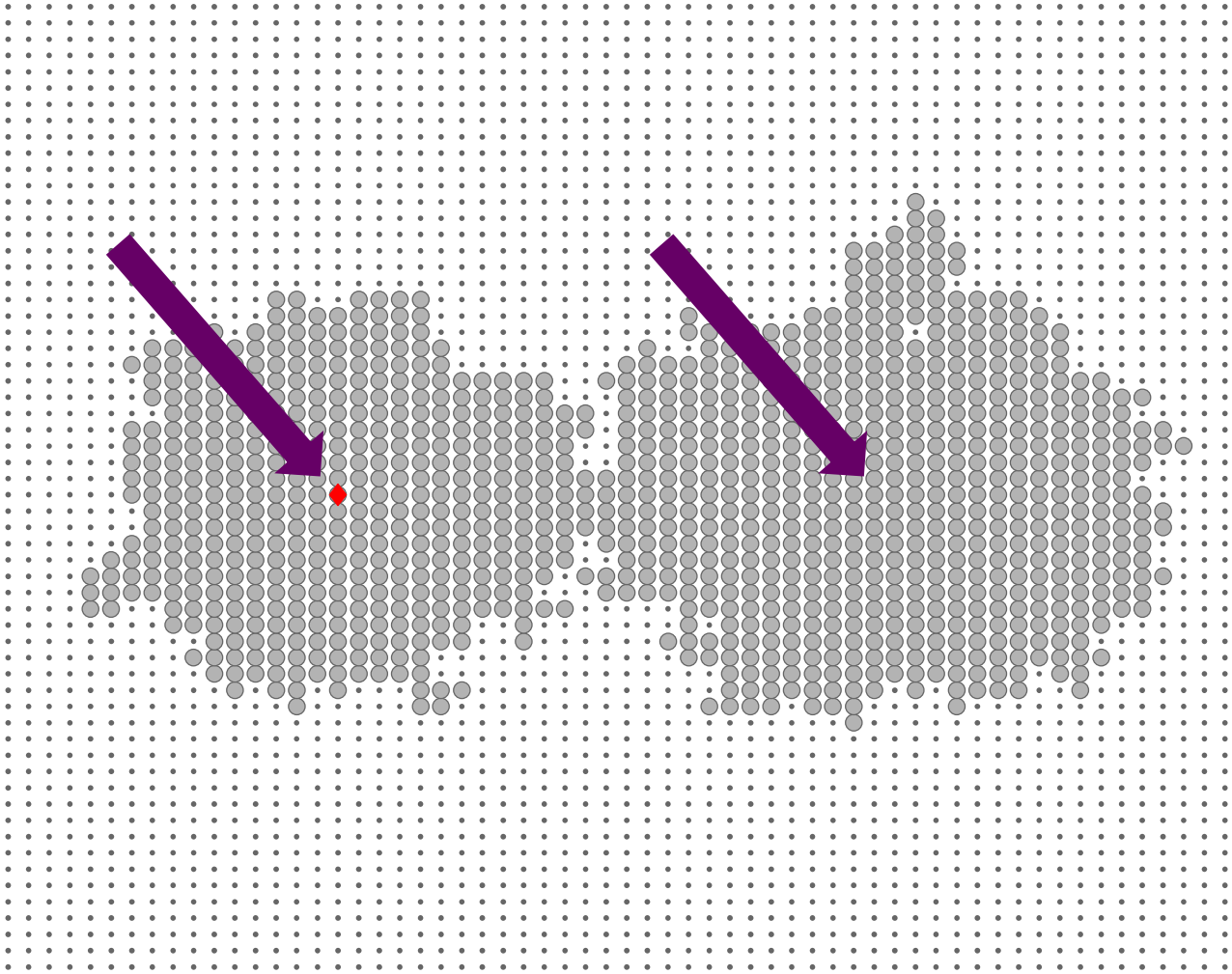
# Culprits: Problem definition

2-d grid



Q: *Who started it?*

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

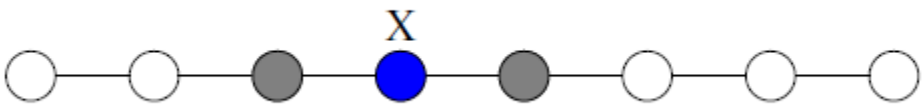


Prakash, Vreeken, Faloutsos 2012

# Outline

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

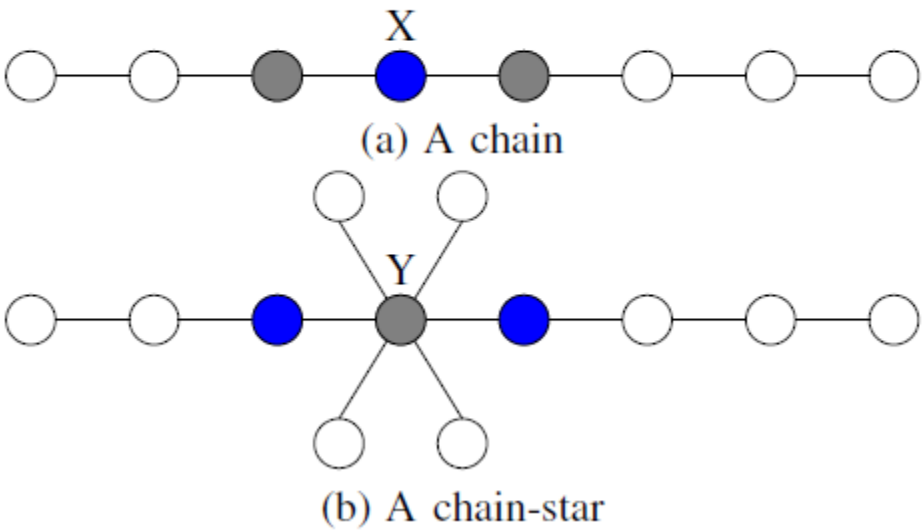
# Culprits: Exoneration




(a) A chain

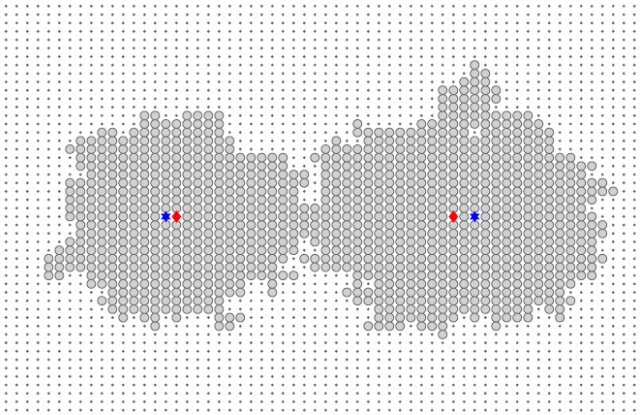
Prakash, Vreeken, Faloutsos 2012

# Culprits: Exoneration




Prakash, Vreeken, Faloutsos 2012

# Who are the culprits



- Two-part solution
  - use MDL for *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**
  - Opitimization
- Experiments
- Conclusion

# Modeling using MDL

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

$$\mathcal{L}(\mathcal{S}) = \mathcal{L}_{\mathbb{N}}(|\mathcal{S}|) + \log \binom{N}{|\mathcal{S}|}$$

En-coding integer |S|

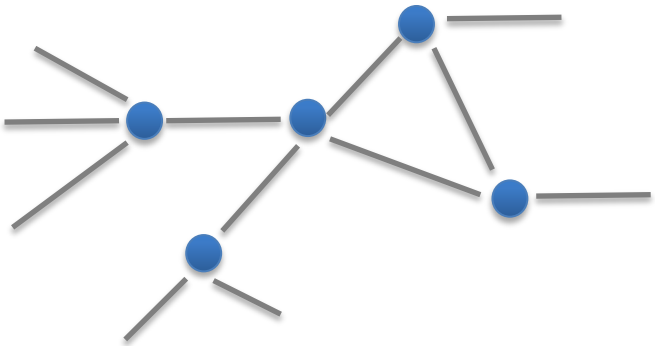
Number of possible |S|-sized sets

Prakash, Vreeken, Faloutsos 2012

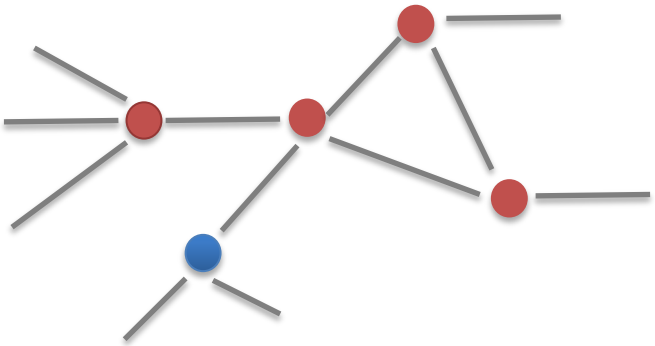
# Modeling using MDL

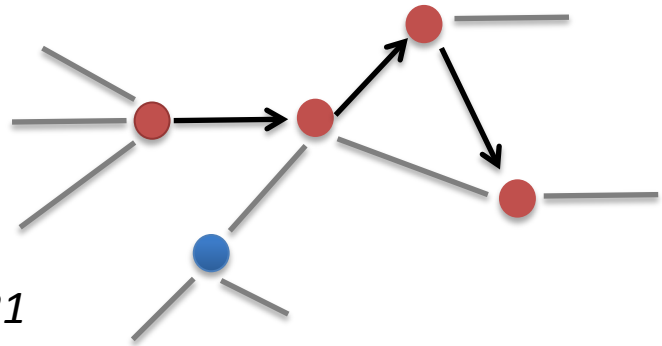
- Data: Propagation Ripples

Original Graph

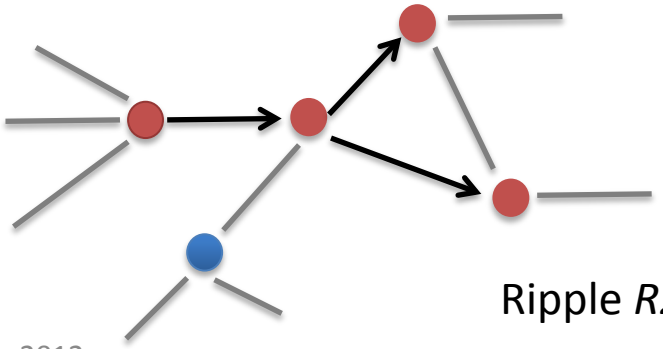


Infected Snapshot





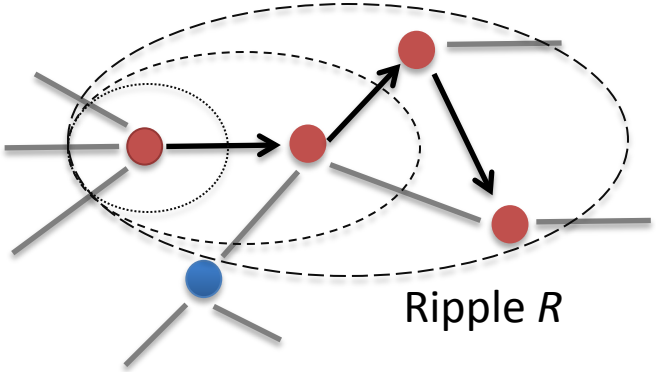
Ripple *R1*



Ripple *R2*

Prakash, Vreeken, Faloutsos 2012

# Modeling using MDL




- Ripple cost

$$\mathcal{L}(R \mid \mathcal{S}) = \mathcal{L}_{\mathbb{N}}(T) + \sum_{t}^{T} \mathcal{L}(\mathcal{F}^t)$$

How long is the ripple

How the 'frontier' advances

- Total MDL cost

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

Prakash, Vreeken, Faloutsos 2012

# Outline

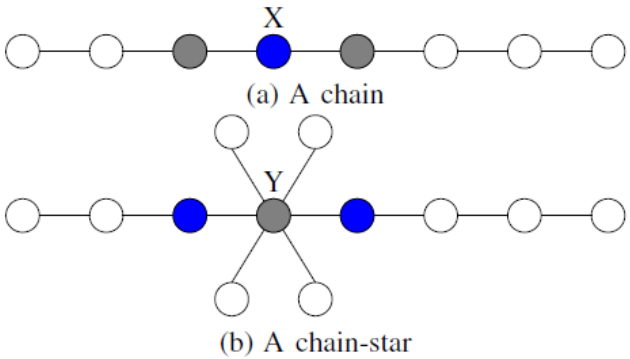
- Motivation---Introduction
- Problem Definition
- Intuition
- **MDL**
  - Construction
  - **Opitimization**
- Experiments
- Conclusion

# 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

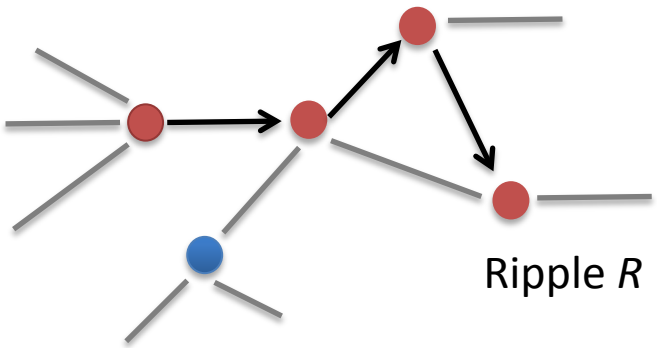
- High-quality *k*-seed-set
  - Exoneration




- Best single seed:
  - *Smallest* eigenvector of Laplacian *sub-matrix*
  - Analyze a *Constrained* SI epidemic
- Exonerate neighbors
- Repeat

# 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 $O(k^*(E_I + E_F + V_I))$

# Outline

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

# Experiments

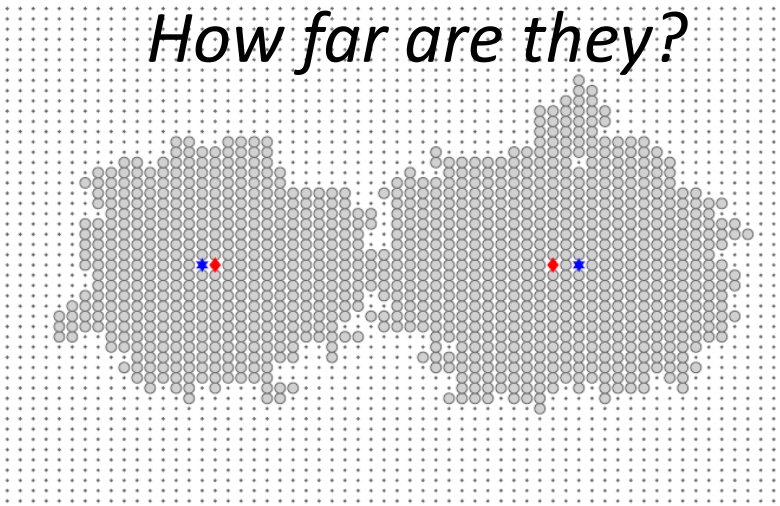
- Evaluation functions:
  - MDL based

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

  - Overlap based

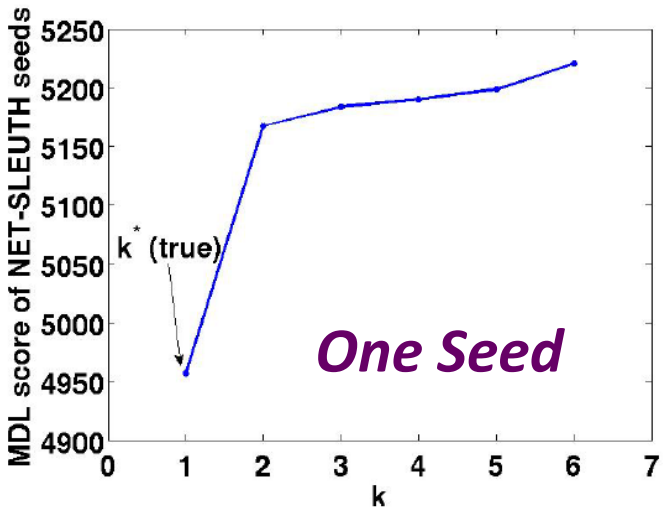
$$Q_{\mathrm{JD}} = \frac{\mathbb{E}[JD_{\mathcal{S}}(\mathcal{V}_I)]}{\mathbb{E}[JD_{\mathcal{S}^*}(\mathcal{V}_I)]}$$

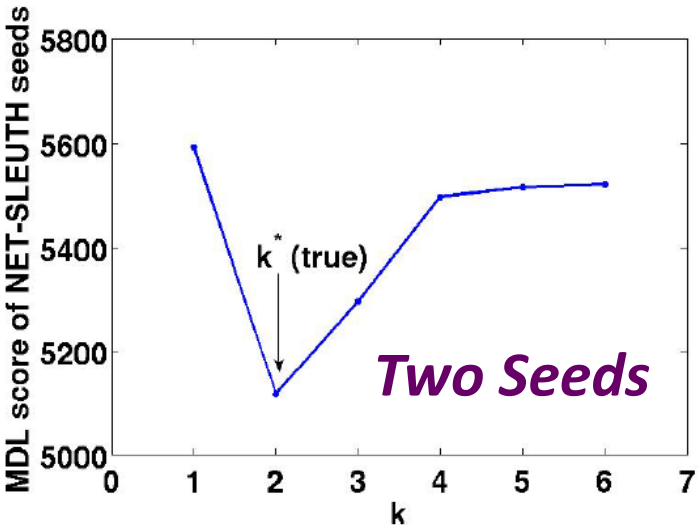
(*JD* == Jaccard distance)

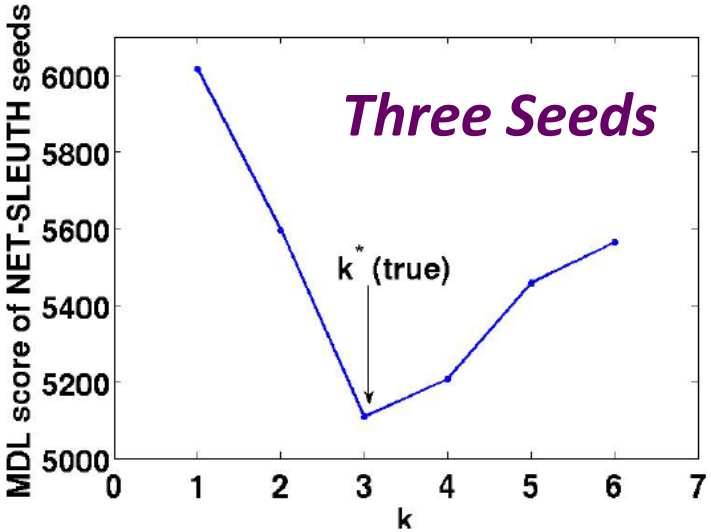



Closer to 1 the better

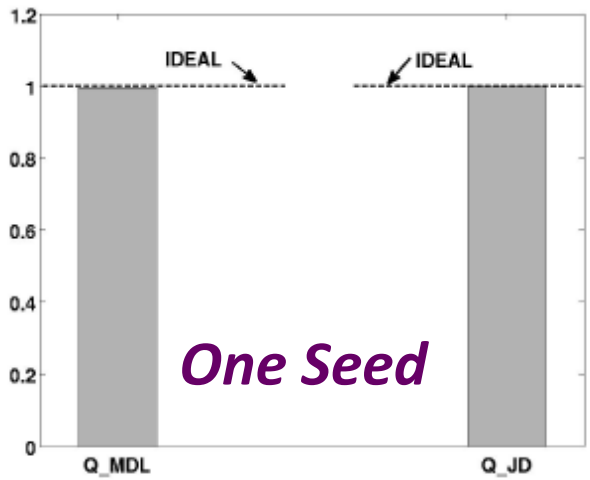
# Experiments: # of Seeds

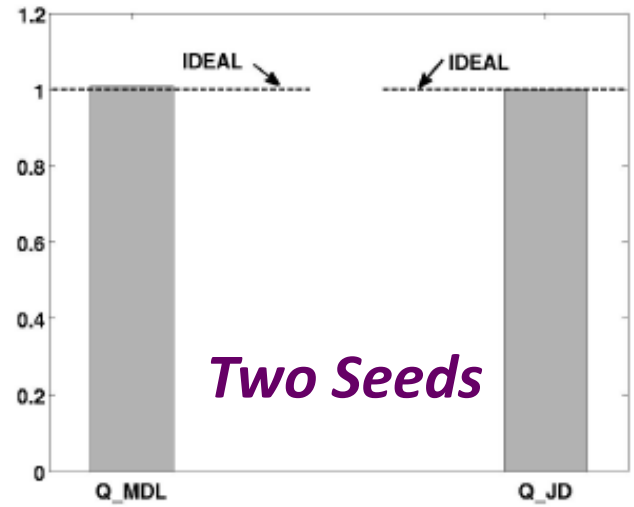


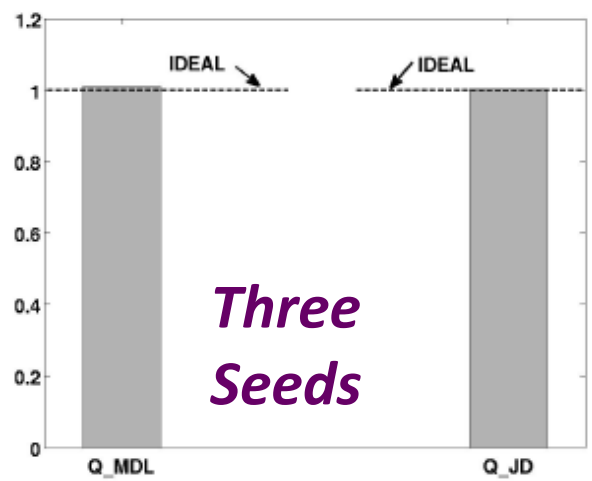
# Experiments: Quality (MDL and JD)




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

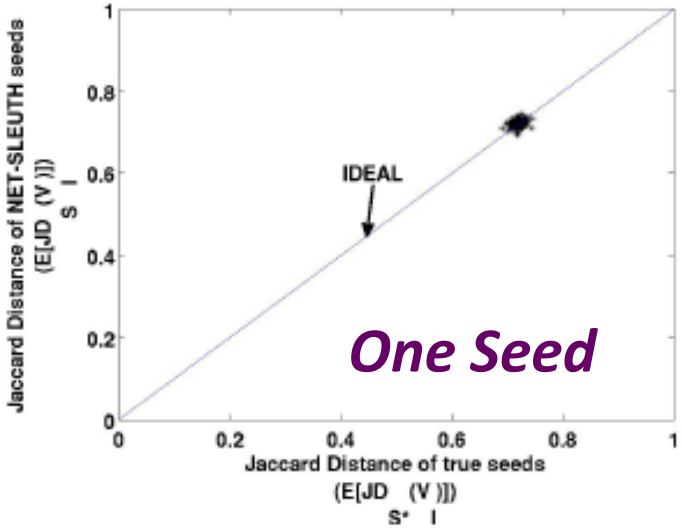
$$Q_{\mathrm{JD}} = \frac{\mathbb{E}[JD_{\mathcal{S}}(\mathcal{V}_I)]}{\mathbb{E}[JD_{\mathcal{S}^*}(\mathcal{V}_I)]}$$

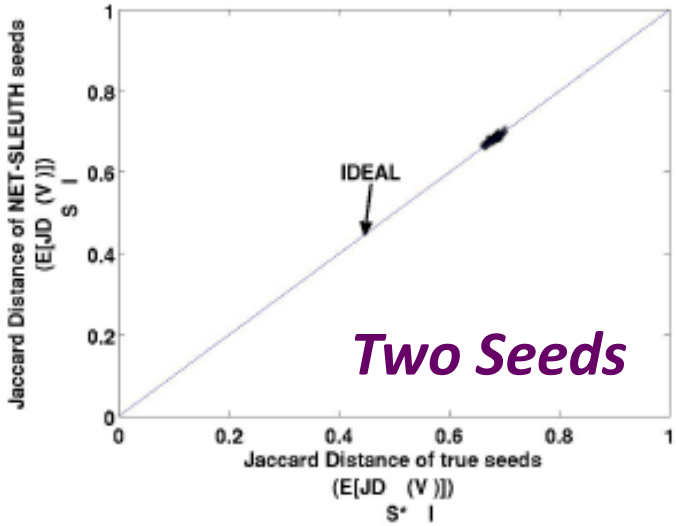


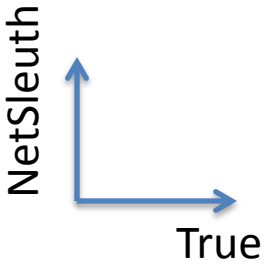

**Ideal = 1**

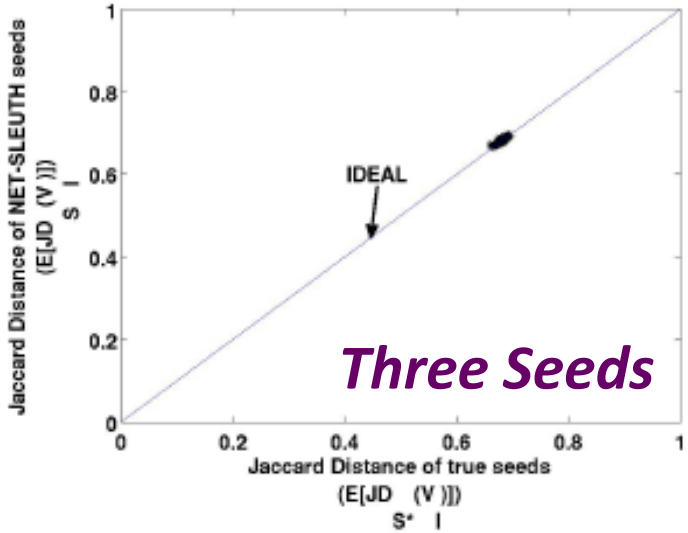
Prakash, Vreeken, Faloutsos 2012

# Experiments: Quality (Jaccard Scores)



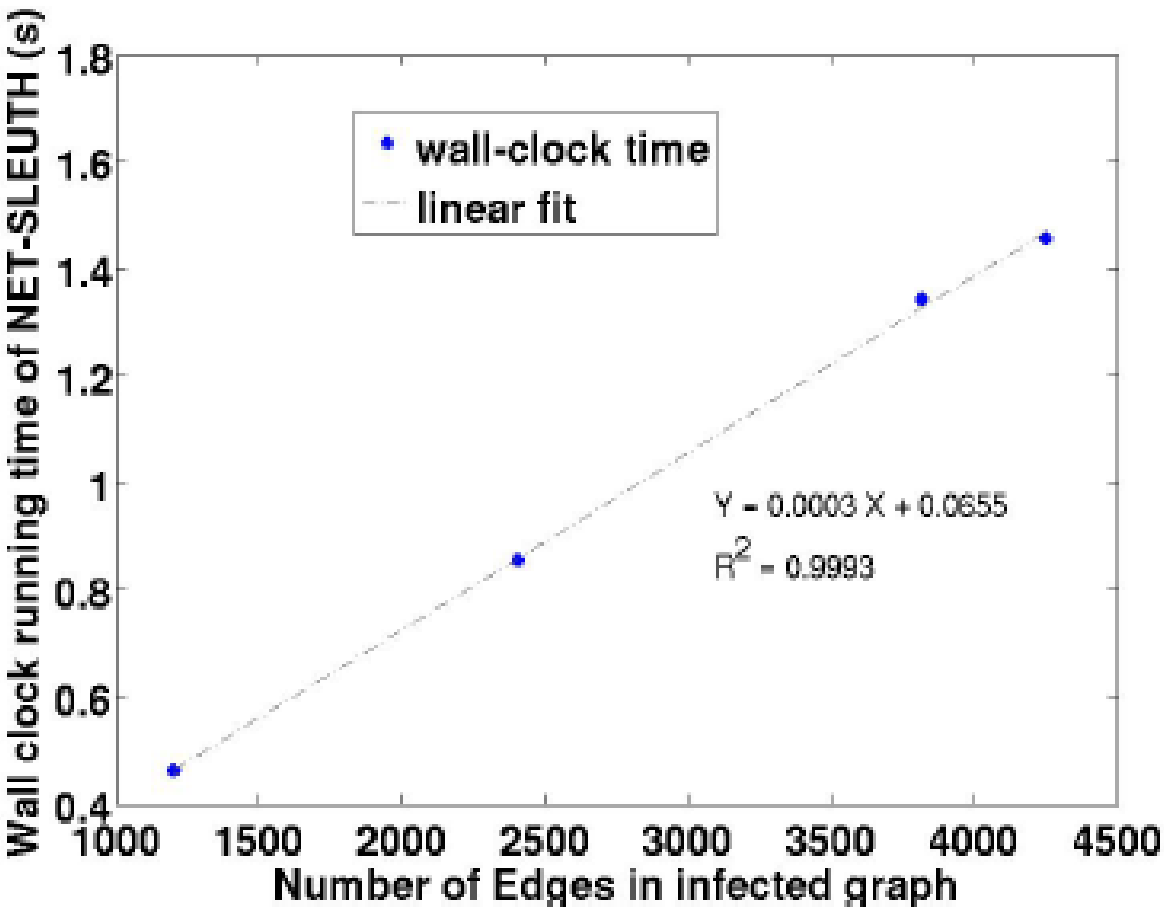




**Closer to diagonal, the better**

# Experiments: Scalability




Prakash, Vreeken, Faloutsos 2012

# Outline

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

# Conclusion

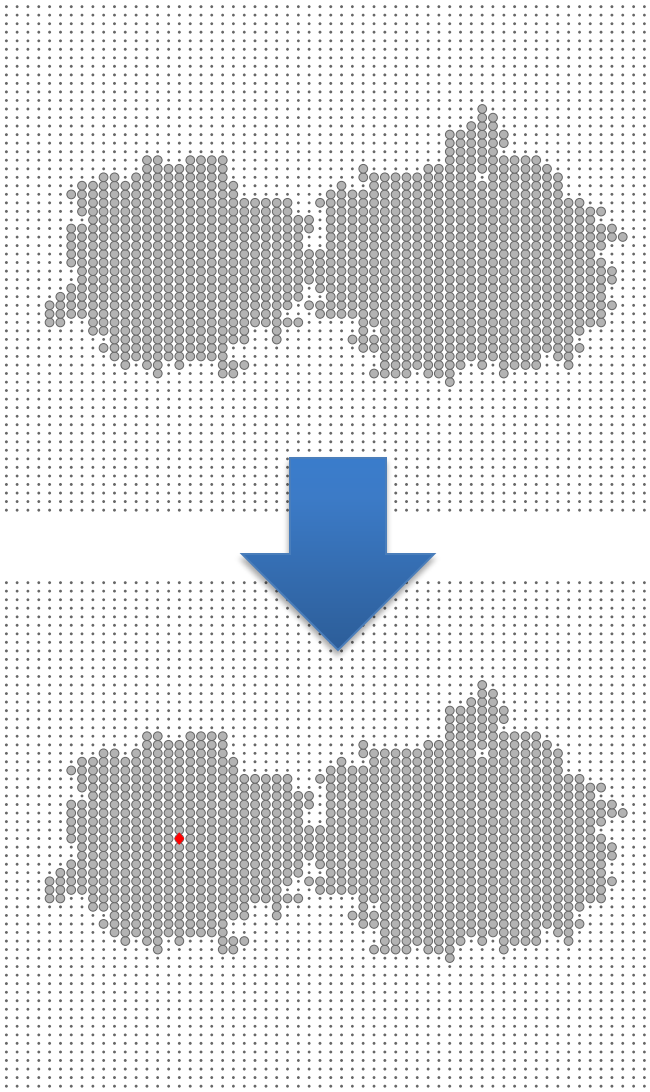
- **Given**: Graph and Infections
- **Find**: Best 'Culprits'

- **Two-part** solution
  - use **MDL** for *number* of seeds
  - for a given number:

    **exoneration** = centrality + penalty

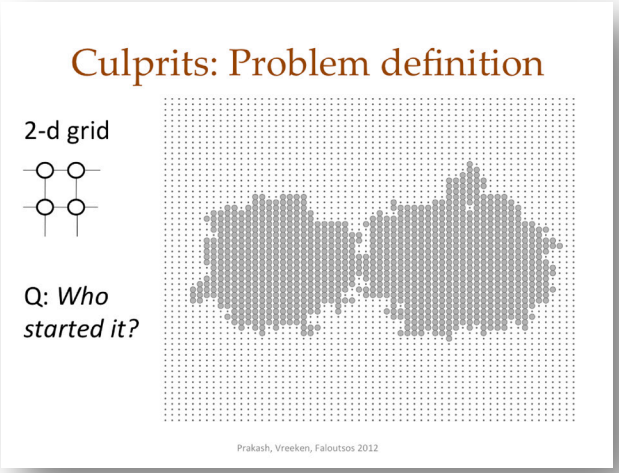
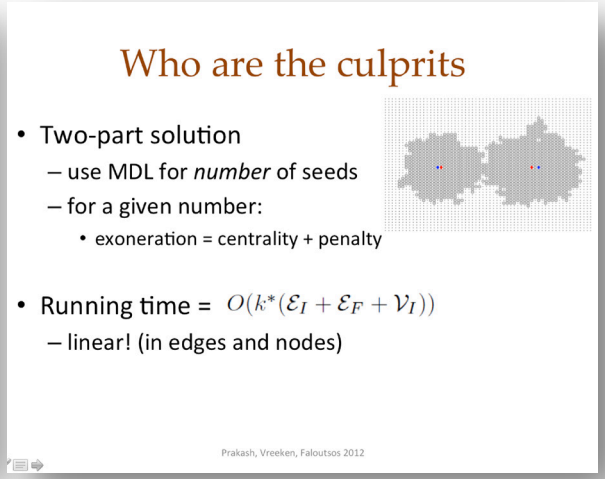
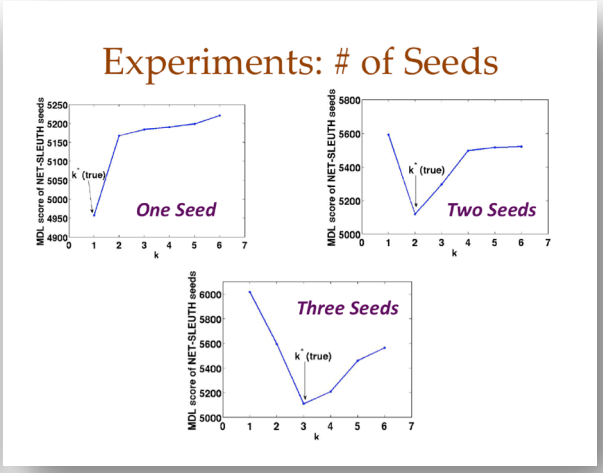
- **NetSleuth**:
  - Linear running time in nodes and edges

$$O(k^*(E_I + E_F + V_I))$$



Prakash, Vreeken, Faloutsos 2012

# Any Questions?


**B. Aditya Prakash**

http://www.cs.vt.edu/~badityap