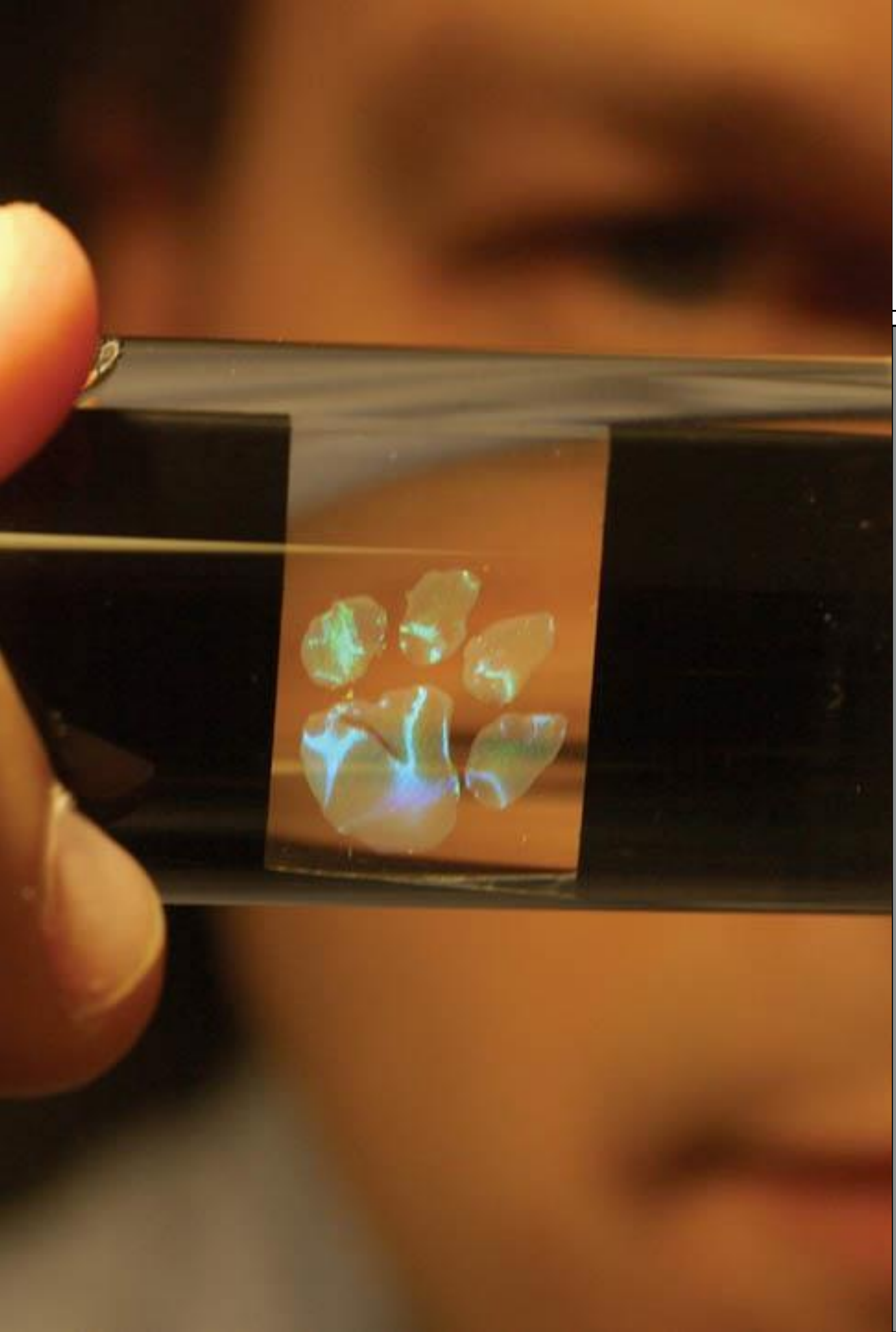


Learning Network Graph of SIR Epidemic Cascades Using Minimal Hitting Set based Approach

Zhuozhao Li and **Haiying Shen**

Dept. of Electrical and Computer Engineering
Clemson University, SC, USA

Kang Chen
Dept. of Electrical and Computer Engineering
Southern Illinois University

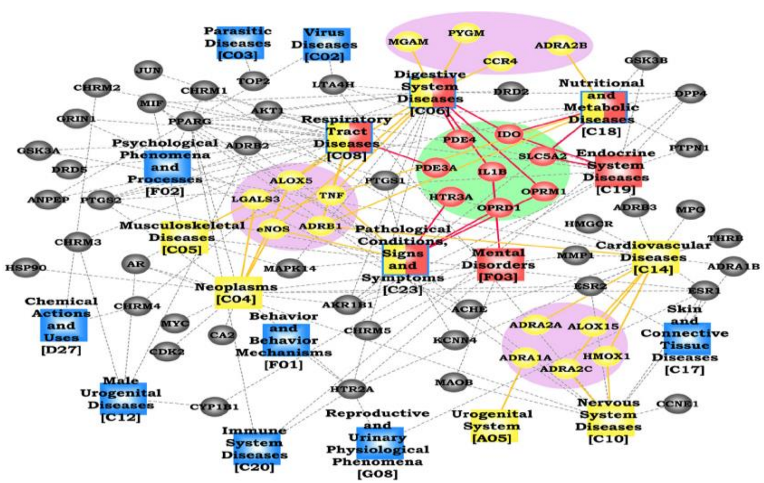
Outline

- Introduction
- Related Work
- System Model
- Main Design
- Performance Evaluation
- Conclusion

Networks are everywhere

Facebook Social Network

The Internet



Human Disease Network

Introduction

- Learning the underlying network structure is very important.

- In this paper, we consider learning the network structure, based on some observations of the network in the context of epidemic spreading.

Epidemic Spreading Model

- Various phenomena can be modeled as epidemic spreading model (epidemic cascade).
  - Biological
    - Diseases via contagion
  - Technological
    - Cascading failures
    - Spread of information
  - Social
    - Rumors, news, new technology
    - Viral marketing
  - Wireless
    - Information dissemination
  - Internet
    - Propagation of worms

Infection Spread Model

- Models of Infection [Easley 10b]:
  - SIS: Susceptible-Infective-Susceptible (e.g., flu)
  - SIR: Susceptible-Infective-Recovered (e.g., chickenpox)

Susceptible, Infected and Recovered

- Susceptible
  - the nodes that can catch disease from a network neighbor
- Infected
  - the nodes that have caught the disease and can pass it on
- Recovered
  - the nodes that have recovered and cannot be infected again

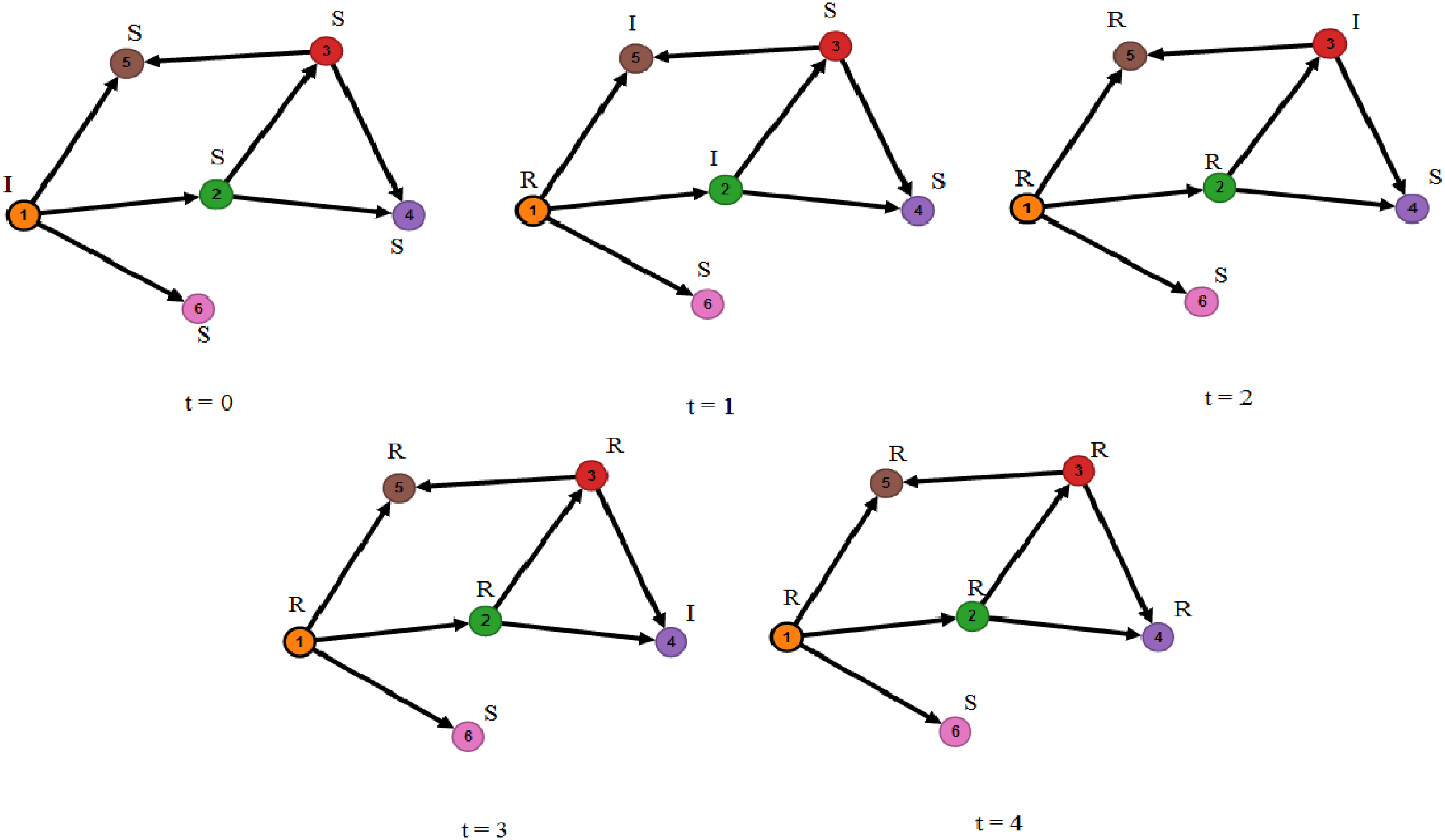
Susceptible, Infected and Recovered

- How is the epidemic spread in SIR model?

- Process
  - Initially, each node independent of other nodes gets infected with probability p_{init}. These infected nodes are seeds.
  - Each node in the *infected* state remains infectious for one time step.
  - Each node i in the *infected* state can infect each of its susceptible neighbors j, with a probability of p_{ij}.
  - After one time step, each infected node is no longer infectious or susceptible and enters the *recovered* state.

- This SIR model is suitable for modeling a disease that each individual can only catch once during his/her life time.

Example SIR Epidemic in One Cascade



Observation Model

- We model the network as a Graph, $G = (V, E)$.
- We denote $V_i = \{j: j\ is\ the\ parent\ of\ node\ i\}$.
- Consider a number of cascades.
- In each cascade u, seeds start to spread the infection.
- We observe the time that each node $i \in V$ gets infected, denoted by t_i^u.
- We set $t_i^u = 0$ for the seeds, and $t_i^u = \infty$ for the nodes that never get infected.

Goal

What is the smallest number of cascades (i.e., sample complexity) to recover a correct network structure with high probability?

Related Work

- Maximum Likelihood (ML) [SIGMETRICS'12]
  - ML guarantees to detect a subset of the parents of a node rather than its exact parental set
  - ML requires a relatively high sample complexity for reliable graph recovery when the graph is dense (i.e., nodes have high degrees)
  - To achieve high performance, ML requires an appropriately predefined threshold parameter

Our proposal

We propose to use a minimal hitting set approach to recover network graph G.

Minimal Hitting Set

- Hitting set of a collection of sets
  - a set that intersects all of the sets in the collection
- Minimal hitting set of a collection of sets
  - if and only if no proper subset of it is a hitting set for this collection
  - Consider a collection of sets $\{\{1,2\},\{1,3\},\{1,2,4\},\{1,3,5\}\}$. $\{2,3\}$ is a minimal hitting set of the collection

Minimal Hitting Set Algorithm

- Process
  - We have m epidemic cascades.
  - Recovering the network structure is equivalent to recovering the parental nodes of every node j, $\forall j \in V$.
  - In each cascade u, we denote t_j^u as the infection time of the node j.
  - We denote S_j^u as the set of nodes i that have infection time $= t_j^u - 1$ in observation u, i.e., $S_j^u = \{i: t_i^u = t_j^u - 1\}$
  - Therefore, for m cascades, we have a collection of S_j^u sets.
  - The parent set of node j is the minimal hitting set of the collection of S_j^u sets from all the cascades.

Minimal Hitting Set Algorithm

- Theorem: as the sample complexity $m \to \infty$, the parent nodes of node j is the minimal hitting set of the collection of S_j^u sets.

- For details of the proof, please refer to the paper.

- Simple rationale: the parent nodes of node j is definitely a minimal hitting set of the collection of S_j^u sets. A minimal hitting set must include all the parents of node j. Otherwise, as $m \to \infty$, there must exist one cascade that node j is infected by the parent that are not in the minimal hitting set.

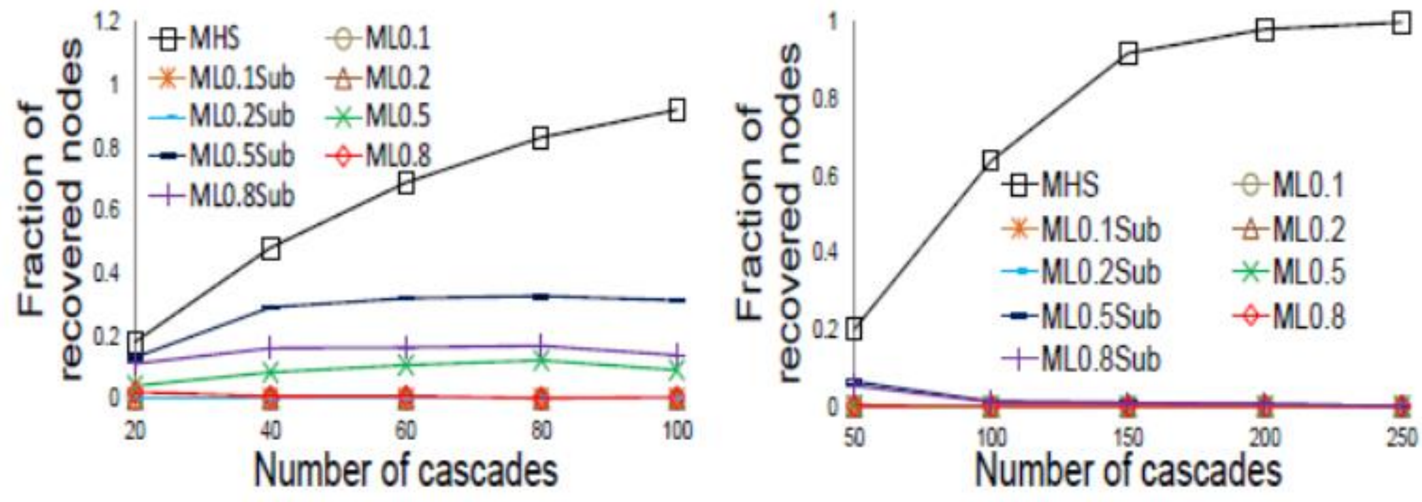
Minimal Hitting Set Algorithm

- Lower bounds (sufficient condition)
  - Recover the network structure with high probability
    - find the sample complexity m, $\forall\delta > 0$
    - guarantee that the network structure is recovered with probability at least $1-\delta,\ \delta > 0$
  - Sample complexity
    - $m \geq \frac{\log\delta - 2\log n}{\log(1-p_{init} * p_{min})} = O(\log n)$
    - p_{min} is the minimum probability for the edge propagation

Evaluation

- Trace-driven simulation
- Comparing method
  - maximum likelihood (ML)
- p_{init}=0.3, $p_{ij} = 0.8$
- The predefined parameter (infection probability threshold) of ML
  - set to x=0.1, 0.2, 0.5, 0.8, denote as MLx
- Recover only a subset of the parents for ML
  - denoted as MLxSub

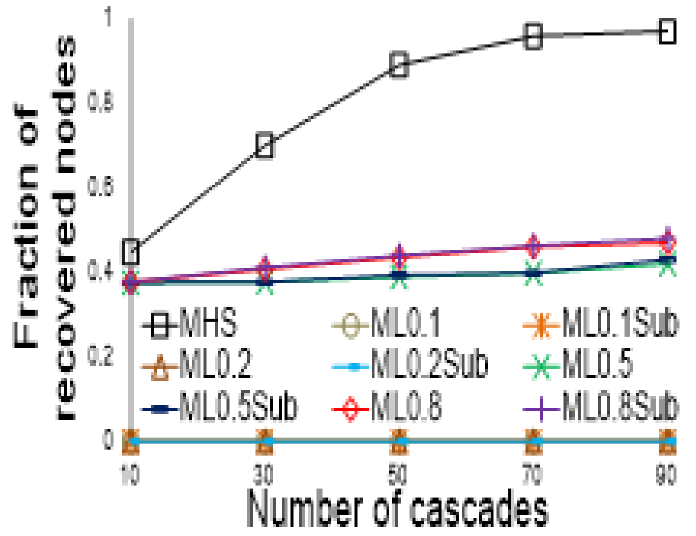
Evaluation



(a) Recovery of 5×5 grids (b) Recovery of 10×10 grids

- Grid Graph. We see the superior performance of our proposed minimal hitting set algorithm.

Evaluation



- A sub-graph of the Google+ network with 500 users.
- We see the superior performance of our proposed minimal hitting set algorithm.

Conclusion

- We consider learning the underlying graph structure of an epidemic cascade based on infection times of nodes.
- We propose a minimal hitting set algorithm to recover the network structure.
- We demonstrate the effectiveness of minimal hitting set algorithm by trace-driven simulation.

Thank you!

Questions & Comments?

Haiying Shen

shenh@clemson.edu

Associate Professor

Pervasive Communication Laboratory

Clemson University