Optimization and Inference for Cyber-Security in Complex Engineered Networks

Chee Wei Tan
City University of Hong Kong
28 August, 2014
2014 IEEE SPS Summer School on IoT and M2M
National Taiwan University
Motivation

Rumors

Source: Google

Dear CNN: Morgan Freeman is still busy living. He's yet to get busy dying. Please confirm first.

Peter Shankman, Twitter
Motivation

Rumors

Source: Bloomberg
Motivation

1854 London Cholera Epidemic

Outbreak source


Dr. John Snow
Motivation

2003 SARS (severe acute respiratory syndrome)

Source: The University of Arizona Artificial Intelligence Lab
Motivation

2003 SARS Rumor

Rumor and panic spread faster than virus. Nothing spreads like fear!

Epicenter of Hong Kong SARS
329 Infected
42 killed


- Rumor: "Entire City was poised to be quarantined"
- 14 year-old boy arrested for creating fake news page
Motivation
2009 Online Flu Map Goes Viral

- Created by Pittsburgh biochemist Harry Niman on April 21, 2009
- Surpassing 290,000 web views and 3000 comments within 9 days

Motivation
Growth and Expansion of Online Social Networks

- Smartphone revolution 2007
- Twitter 2006
- Tencent Weixin 2012
Motivation
Reach of Online Social Networks
Motivation
National Grand Challenges

NAE
GRAND
CHALLENGES FOR ENGINEERING

The usefulness of these approaches depends on numerous variables — how infectious and how deadly the virus is, the availability of antiviral drugs and vaccines, and the degree of public compliance with quarantines or travel restrictions. Again, understanding the mathematics of networks will come into play, as response systems must take into account how people interact. Such models may have to consider the “small world” phenomenon, in which interpersonal connections are distributed in a way that assists rapid transmission of the virus through a population, just as people in distant parts of the world are linked by just a few intermediate friends.

- Challenge No. 7: Advance health informatics
Motivation
National Grand Challenges

All engineering approaches to achieving security must be accompanied by methods of monitoring and quickly detecting any security compromises. And then once problems are detected, technologies for taking countermeasures and for repair and recovery must be in place as well. Part of that process should be new forensics for finding and catching criminals who commit cybercrime or cyberterrorism.

- Challenge No. 11: Secure Cyberspace
Motivation

DARPA Mathematical Challenges


- Challenge No. 2: The Dynamics of Networks
- Challenge No. 14: An Information Theory for Virus Evolution

  - can Shannon’s theory shed light on this fundamental area?

Source: https://www.fbo.gov/download/9bc/9bce380aafb19f9ad3bda188bfc1ab20/DARPA-BAA-08-65.doc
Rumor Spreading in Online Social Network

- Outbreak of infectious virus
- Diffusion of viral Information in Network
- Cause of outbreak

- Epidemic-like information flow = rumor spreading in a network.
Who is the culprit?

- Spread of computer virus
- Tweeting and Retweeting in Twitter Network

- A rumor, originating from a suspect set, spreads on a network.
- We only know the prior suspect set and infected nodes.
- Can we find the single rumor source?
Literature

----Research on epidemic outbreak/rumor spreading

• understand impacts of network structure and infection/cure rates
  [Moore—PRE’00, Pastor-Satorras—PRL’01, Newman—PRE’02]

• learn network parameters and predict propagation characteristics
  [Streftaris—IWSM’02, Okamura—ISSRE’07, Gomez-Rodriguez—SIGKDD’10]

• extract influential source nodes
  [Kempe—SIGKDD’03, Chen—SIGKDD’09, Dong—Allerton’12]

– Rumor source estimation problem has only been recently studied.
Literature

---estimation of rumor source

• identification of single rumor source using SI model
  [Shah—TIT’11, Shah—SIGMETRICS’12]

• geometric trees, random graphs
  [Shah—arXiv’11, Shah—TIT’11]

• identification of multiple rumor sources (SI model),
  identification of single rumor source (SIR model)
  [Luo—TSP’13, Zhu—ITA’13]

• noisy estimation of a single rumor source
  [Pinto—PRL’12]

--- Important features such as suspects, no. of observations, topology has not been considered.
SI Spreading Model
(Kermack & McKendrick 1927)

- **Fixed Population** $N$ (*only one infected at each $t$*)
- **Susceptible Set at time $t$** $S(t)$
- **Infected Set at time $t$** $I(t)$

$S_t$ denote $|S(t)|$ and $I_t$ denote $|I(t)|$

$$S_t + 1 = S_t + 1$$

$$I_t + 1 = I_t + 1$$

$S_0 = N, I_0 = 0.$
SI Spreading Model

• Vertex of a graph $G$ to model the susceptible and the infected node (person)
• An edge in $G$ models the relationship between two nodes
  • Two persons connected as Facebook Friends or Twitter Follower

Let $G_t$ be a subgraph of order $t$ of $G$.
$G_t$ is composed of $t$ infected vertices
$G_1$ rumor source
$|G_t + 1| = |G_t| + 1$
Random SI Model for Rumor Spreading

- SI (susceptible-infectious) model
  - An infected node keeps the rumor forever

- Uniform probability of any node in a prior suspect set being source
  - \( P_s(\text{source} = s) = \frac{1}{|S|}, \ s \in S \) \( S \) consists of suspect nodes

- Time to infect neighbor is independent and exponentially distributed with rate \( \lambda \)

[Shah—TIT’11, Shah—SIGMETRICS’12]
Tan—ISIT’13, SIGMETRICS’14]
Deterministic SI Model

- Time to infect neighbor is assumed to be a fixed time-slot
- A susceptible node is infected by each of its neighbors with probability $q$
- Assuming the number of infected neighbors of a susceptible node is $n$, the probability that the node becomes infected is $\hat{1} - (1 - q)^n$
- Time-slot model used by researchers for network inference [Luo—TSP’13, Zhu—ITA’13]
Inference with Single Snapshot Observation

Toy Example
Toy Example

Counts the number of permitted permutation to spread a rumor
Toy Example

• **Suppose Source 1**
  – Two permutations: \{1,2,3,4\}, \{1,2,4,3\}

• **Suppose Source 2**
  – Six permutations: \{2,1,3,4\}, \{2,1,4,3\}, \{2,3,1,4\}, …

[Shah—TIT’11]
Inference in Tree

- Let $T$ be a tree:
- let $G_n$ be the subtree of $T$ at time $n$
- $P(G_n|v^*)$ is the probability that view $v^*$ as the source
- Let $\sigma_i$ be the possible infecting order
- $S(v^*, G_n)$ be the collection of all $\sigma_i$ where $v^*$ is viewed as the source

$$P(G_n|v^*) = \sum_{\sigma_i \in S(v^*, G_n)} P(\sigma_i | v^*).$$
Toy Example
Toy Example

Suppose Node 1 is Rumor

\[ \sigma_1 = v_1 \rightarrow v_2 \rightarrow v_3 \rightarrow v_4 \]
\[ \sigma_2 = v_1 \rightarrow v_2 \rightarrow v_4 \rightarrow v_3 \]
\[ \sigma_3 = v_1 \rightarrow v_3 \rightarrow v_2 \rightarrow v_4 \]
\[ \sigma_4 = v_1 \rightarrow v_3 \rightarrow v_4 \rightarrow v_2 \]
\[ \sigma_5 = v_1 \rightarrow v_4 \rightarrow v_3 \rightarrow v_2 \]
\[ \sigma_6 = v_1 \rightarrow v_4 \rightarrow v_2 \rightarrow v_3 \]
Toy Example

Suppose Node 1 is Rumor

Let’s calculate the probability of $\sigma_1$

$$P(\sigma_1|v_1) = \frac{1}{3} \cdot \frac{1}{4} \cdot \frac{1}{5}$$
Toy Example

Suppose Node 1 is Rumor

\[ P(\sigma_i|v_1) = \frac{1}{3} \cdot \frac{1}{4} \cdot \frac{1}{5} \quad \text{for } i = 1, 2, 3, 4, 5, 6 \]
General Tree

\[ P(\sigma_i | v_1) = \prod_{k=1}^{n-1} \frac{1}{\sum_{v_i \in V(G_k)} d(v_i) - 2(k - 1)} \]

where \( G_k \) is a subgraph of \( G_n \) and it represents the infected subgraph at \( k_{th} \) time step along with the infecting order \( \sigma_i \).
General Degree-Regular Tree

\[ P(\sigma_i | v_1) = \prod_{k=1}^{n-1} \frac{1}{\sum_{v_i \in V(G_k)} d(v_i) - 2(k-1)} \]

where \( G_k \) is a subgraph of \( G_n \) and it represents the infected subgraph at \( k_{th} \) time step along with the infecting order \( \sigma_i \).

For \( d - regular \) tree

\[ P(\sigma_i | v_1) = \prod_{k=1}^{n-1} \frac{1}{dk - 2(k-1)} \]
Toy Example

Suppose Node 1 is Rumor

\[ P(\sigma_i | v_1) = P(\sigma_j | v_1) \text{ for all } \sigma_i, \sigma_j \in S(v, G_n) \]
Toy Example

Suppose Node 1 is Rumor

\[
P(G_n | v^*) = \sum_{\sigma_i \in S(v^*, G_n)} P(\sigma_i | v^*)
\]

\[
= \left| S(v^*, G_n) \right| \cdot P(\sigma | v^*) \quad \forall \sigma_i \in S(v, G_n)
\]

\[
= \left| S(v^*, G_n) \right| \cdot \prod_{k=1}^{n-1} \frac{1}{dk - 2(k - 1)}
\]

\[\propto \left| S(v^*, G_n) \right|.
\]
Toy Example

Suppose Node 1 is Rumor

compute each $P(G_n|v_i)$ by finding out the value $|S(v_i, G_n)|$.

they are of the same value $\frac{1}{3 \cdot 4 \cdot 5} = \frac{1}{60}$

$$P(G_n|v_1) = 3! \cdot \frac{1}{60}$$

$$P(G_n|v_2) = P(G_n|v_3) = P(G_n|v_4) = 1 \cdot 2! \cdot \frac{1}{60}$$

The Node with the Maximum Likelihood: $v_1$
Rumor Centrality

Rumor centrality counts the number of permitted permutation to spread a rumor

\[ R(v, G_n) = |S(v, G_n)| \]

Shah and Zuman
IEEE Transactions on Information Theory 2011
Rumor Centrality

Rumor centrality counts the number of permitted permutation to spread a rumor

\[ R(s, G_n) = n! \prod_{u \in G_n} |T_u^s|^{-1} \]

Shah and Zuman
IEEE Transactions on Information Theory 2011

subtree \(T_u^s\)

\[ |T_2^1| = 3\]
\[ |T_7^1| = 1\]
Rumor Center

- **ML** (maximum likelihood) estimator

\[
\hat{v} = \arg \max_{v \in G_N} P(G_N | v^* = v) = \arg \max_{v \in G_N} R(v, G_N)
\]

- **Most likely source is at the “center” of the network!**

[Shah—TIT’11]
Detection Probability $P_c(n)$

Detecting the most probable source by Bayes Theorem

Let $G$ be a tree and $G_n$ is a subtree of $G$ at time $n$

$v$ is the rumor center of $G_n$

Let the event $v = source$ denoted as $S_v$. $P(S_v|G_n)$ is the probability of the event that the rumor center is exactly the rumor source when given $G_n$

$$P_c(n) = P(S_v|G_n) = \frac{P(G_n|S_v) \cdot P(S_v)}{\sum_{i \in G_n} [P(G_n|S_i) \cdot P(S_i)]}$$
Detection Probability

Detecting the most probable source by Bayes Theorem

The probability of each vertex to be the source are equal, that is

$$P(S_i) = P(S_j) \quad \forall \ i, j \in G_n,$$

and also we have $$P(G_n|S_v) \propto R(v, G_n).$$

$$P(S_v|G_n) = \frac{P(G_n|S_v) \cdot P(S_v)}{\sum_{i \in G_n} [P(G_n|S_i) \cdot P(S_i)]}$$

$$\Rightarrow P(S_v|G_n) = \frac{R(v, G_n)}{\sum_{i \in G_n} R(i, G_n)}$$
Detection Probability

Bound

\[ P(S_v \mid G_n) \leq \frac{1}{2} \]

no matter how large the size is or what shape of \( G_n \) is.

Shah and Zuman

IEEE Transactions on Information Theory 2011
Detection Probability

Bound

\[ P(S_v | G_n) \leq \frac{1}{2} \]

no matter how large the size is or what shape of \( G_n \) is.

Shah and Zuman

IEEE Transactions on Information Theory 2011
Minimum Detectability

- Line Network is undetectable!
- Can multiple observations help?
Minimum Detectability

• For finite $n$

\[
P_c(n) = \frac{1}{2} \sum_{\max\{x_1, x_2\} = n/2} P_G \left( \bigcap_{j=1}^{2} (X_j = x_j) \right) + \sum_{\max\{x_1, x_2\} < n/2} P_G \left( \bigcap_{j=1}^{2} (X_j = x_j) \right),
\]

\[= \frac{1}{2^{n-1}} \left( \left\lfloor \frac{n-1}{2} \right\rfloor \right).\]
Minimum Detectability

• For asymptotically large $n$

As $n \to \infty$, by the Stirling’s formula, we have

\[
P_c(n) \approx \frac{1}{2^n} \cdot \frac{n!}{[(n/2)!]^2} \approx \frac{1}{2^n} \cdot \frac{\sqrt{2\pi n} \cdot \left(\frac{n}{e}\right)^n}{\left[\sqrt{\pi n} \cdot \left(\frac{n}{2e}\right)^{n/2}\right]^2}
\]

\[
= \sqrt{\frac{2}{\pi n}}
\]

\[
= O\left(\frac{1}{\sqrt{n}}\right).
\]

Stirling’s asymptotic formula:

\[
n! \sim \sqrt{2\pi n} \left(\frac{n}{e}\right)^n
\]
Detectability

- Degree-regular tree with degree strictly larger than 2 is asymptotically detectable
- Phase-Transition like phenomenon!
Connection with Graph Convexity in Graph Theory

- Recall that \( R(s, G_n) = n! \prod_{u \in G_n} |T_u^s|^{-1} \)

Shah and Zuman
IEEE Transactions on Information Theory 2011

consider two adjacent vertices \( u, v \) in \( G_n \) and a vertex \( w \in G_n - \{u, v\} \)
Connection with Graph Convexity in Graph Theory

Since

\[ t_u^v = n - t_v^u \text{ and } t_w^v = t_w^u \]

\[ \frac{P(u | G_n)}{P(v | G_n)} = \frac{R(u, G_n)}{R(v, G_n)} = \frac{t_u^v}{n - t_v^u} \]
Alternative Characterization of Rumor Cente

Given an $n$ vertices tree $G_n$. $v \in G_n$ is a rumor center if and only if

$$t_u^v \leq \frac{n}{2}$$

for all $u \in G_n - \{v\}$

Shah and Zuman
IEEE Transactions on Information Theory 2011
Connection with Graph Theory

• Distance Centrality (Social Network Analysis)

\[ D(v, G_n) = \sum_{j \in G_n} d(v, j) \]

where \( d(v, j) \) is the shortest path length between \( v \) and \( j \) (eccentricity)

• Vertex in \( G_n \) with minimum distance centrality is called distance center

Distance from 1 to all shaded nodes:
From Node 1 to Node 2: 1
From Node 1 to Node 3: 2
From Node 1 to Node 4: 2
\[ \Rightarrow \text{Distance Centrality of Node 1} = 1 + 2 + 2 = 5 \]

Distance Centrality of Node 2 is 3, therefore Node 2 is Distance Center
Connection with Graph Theory

• Mass of graph (Graph Theory)

\[
weight(v) = \max_{u \in \text{child}(v)} \{ t_u \}
\]

• Vertex in \( G_n \) with minimum weight is called mass center


Subtree of a child (Node 2) of Node 1 has size of 3
\( \Rightarrow \) Weight of Node 1 = 3

Weight of Node 2 is 1 for its three subtrees, therefore Node 2 is Mass Center
Connection with Graph Theory

Let $G_n$ be an infected subtree of $G$ and $v$ be a vertex in $G$. Then the following statements are equivalent:
1. $v$ is a distance center of $G_n$.
2. $v$ is a rumor center of $G_n$.
3. $v$ is a mass center of $G_n$. 
Connection with Graph Theory

Let $G_n$ be an infected subtree of $G$ and $v$ be a vertex in $G$. Then the following statements are equivalent:

1. $v$ is a distance center of $G_n$.
2. $v$ is a rumor center of $G_n$.
3. $v$ is a mass center of $G_n$.

- A tree has either exactly one or two mass centers (graph theory result)
- There are at most two rumor centers when the maximum subtree tree is $n/2$ (Shah’s IT result)
Algorithms

• A myriad of polynomial-time algorithms for computing rumor center
  • Mass center algorithm (graph theory)
  • Distance center algorithm (social network analysis)
• Message passing algorithm (Shah’s IT result)
• ...etc
Detection with Suspects

Why consider suspects?

- Not all infected are suspects
- Spread of infectious disease from cities to cities (frequent travellers)
- Infection of rumors or computer viruses in cyberspace (vulnerable hosts)

- Suspect characteristics significantly affect detectability and add an interesting dimension to identifying the source reliably.
Detection with Suspects

What’s new with suspects?

Finite regime:
- at most vs. at least 0.5 detection probability

Asymptotic regime:
- 0.307 vs. 1 best detection probability

Insightful monotonicity and averaging results

Dong, Zhang and Tan
Proc. of IEEE Symposium on Information Theory 2013

Zhao, Dong, Zhang and Tan
ACM SIGMETRICS 2014
Detectability is graph constrained. Network connectivity matters!
MAP Rumor Source Estimator

- **MAP (maximum a posteriori) estimator**
- A prior suspect set $S$
- An observation of $n$ infected nodes $G_n$

\[
\hat{s} = \arg\max_{s \in \{S \cap G_n\}} P_G \{s \mid G_n \} = \arg\max_{s \in \{S \cap G_n\}} P_G \{G_n \mid s\}
\]
MAP Rumor Source Estimator

- **For regular tree-type networks**
  \[ \hat{s} = \arg \max_{s \in \{S \cap G_n\}} R(s, G_n) \]

- **For general tree-type networks**
  \[ \hat{s} = \arg \max_{s \in \{S \cap G_n\}} R(s, G_n) \]

- **For general networks**
  \[ \hat{s} = \arg \max_{s \in \{S \cap G_n\}} R(s, T_{bfs}(s)) \]

- Optimal MAP estimator
- Focus on regular trees
- Analyze the correct detection probability upon observing \(n\) infected nodes

\[ P_e(n) = \text{Prob} \left[ \hat{s} = s^* \right] \]
Detectability on Tree

- How to detect on a tree?
- Can fewer number of infected nodes help detectability?
- Can a higher degree help detectability?
Pólya’s Urn Model

- Joint distribution
  \[ P_G \left[ \bigcap_{j=1}^{\delta} (X_j = x_j) \right] \]
- Marginal distribution
  \[ P_G [X_1 = x_1] \]
- Limit distributions
  \[ \lim_{n \to \infty} P_G \left[ \bigcap_{j=1}^{\delta} \left( \frac{X_j}{n} = y_j \right) \right] \]
  \[ \lim_{n \to \infty} P_G \left[ \frac{X_1}{n} = y_1 \right] \]

[Johnson—JWS’97]
Equivalence to Pólya’s Urn Model

- **Rumor spreading process**
  - \( t = 0 \)
  - Infected nodes
  - \( t = t_1 \)
  - Newly infected
  - \( t = t_2 \)

- **Ball drawing process**
  - \( t = 0 \)
  - \( t = t_1 \)
  - \( t = t_2 \)

\[
s = \delta - 2
\]
Suspecting all Nodes

--- main results

**Case 1**

- Any infected node might be the rumor source.

[Shah—TIT’11, Shah—SIGMETRICS’12]

**Main results**

- node degree $\delta = 2$
  \[ P_c(n) = \frac{1}{2^{n-1}} \left( \frac{n-1}{(n-1)/2} \right) \sim O(1/\sqrt{n}) \]

- node degree $\delta = 3$
  \[ P_c(n) = \frac{1}{4} + \frac{3}{4} \left( \frac{1}{2 \left\lfloor n/2 \right\rfloor + 1} \right) \sim \frac{1}{4} + O(1/n) \]

- node degree $\delta > 3$
  \[ \lim_{n \to \infty} P_c(n) = 1 - \delta \left( 1 - I_{1/2} \left( \frac{1}{\delta - 2}, \frac{\delta - 1}{\delta - 2} \right) \right) \to 0.307 \]

- Monotonicity: Detection probability increases with degree and decreases with $n$
Minimum Detectability

- Line Network is undetectable!
- Can multiple observations help?
Suspecting all Nodes

validation experiment

Theoretical prediction
Numerical simulation

Correct Detection Probability vs Node Degree
Connected Suspects

---main results

**Case 2**

- All suspect nodes form a connected subgraph.

**Main results**

- **node degree \( \delta = 2 \)**
  
  \[
  P_c(n) = \frac{1}{k} \left( 1 + \frac{k-1}{2^{n-1}} \left( \left\lfloor \frac{n-1}{2} \right\rfloor \right) \right) \sim \frac{1}{k} + O(1/\sqrt{n})
  \]

- **node degree \( \delta = 3 \)**
  
  \[
  P_c(n) = \frac{k+1}{2k} + \frac{k-1}{k} \frac{1}{4\left\lfloor n/2 \right\rfloor + 2} \sim \frac{k+1}{2k} + O(1/n) \geq \frac{1}{k} + \frac{k-1}{2k}
  \]

- **node degree \( \delta > 3 \)**
  
  \[
  \lim_{n \to \infty} P_c(n) = 1 - \frac{2k-2}{k} \left( 1 - I_{1/2} \left( \frac{1}{\delta-2}, \frac{\delta-1}{\delta-2} \right) \right) > \frac{1}{k} + \frac{k-1}{2k} \to 1
  \]

- **Monotonicity:** Detection probability increases with degree and decreases with \( n \)
Connected Suspects

validation experiment

The diagram shows the relationship between node degree and correct detection probability. The x-axis represents the node degree, while the y-axis represents the correct detection probability. The graph includes lines and markers for theoretical predictions and numerical simulations for different values of k (2, 3, 5, 10). The markers and lines are color-coded to differentiate between the different values of k.
Connected Suspects
----with vs. without prior knowledge

**Closer-up look at case 2**
- node degree $\delta > 2$
- exceed prior probability
- at least 0.5-detection
- achieve reliable detection

$$\lim_{n \to \infty} P_c(n) = 1 - \frac{2k-2}{k} \left( 1 - I_{1/2} \left( \frac{1}{\delta-2}, \frac{\delta-1}{\delta-2} \right) \right)$$

$$P_c(n) \geq \frac{1}{k} + \frac{k-1}{2k}$$

$$P_c(n) \geq 2I_{1/2} \left( \frac{1}{\delta-2}, \frac{\delta-1}{\delta-2} \right) - 1 \geq 0.5$$

**Comparison with case 1**
- node degree $\delta > 2$
- non-trivial positive value
- at most 0.5-detection
- upper-bounded by 0.307

$$\lim_{n \to \infty} P_c(n) = 1 - \delta \left( 1 - I_{1/2} \left( \frac{1}{\delta-2}, \frac{\delta-1}{\delta-2} \right) \right)$$

$$P_c(n) > 0$$

$$P_c(n) \leq 0.5$$

$$\lim_{n \to \infty} \lim_{\delta \to \infty} P_c(n) = 0.307$$

— Suspect characteristics (connectivity) bring about new ingredients.
Connected Suspects

---no degeneration with large $k$

**Averaged by Bayes’ Rule**

$$P_c(n) = \sum_{i=1}^{k} P_s(s_i) P_c(n | s_i)$$

$$= \frac{1}{k} \sum_{s^* \in S} P_c(n | s^*)$$

- Suspect nodes nearby the subgraph boundary are easier to identify.
- Suspect nodes inside the connected subgraph are harder to identify.

**Explanation**

- The results in case 2 don’t degenerate to case 1 with large $k$.
- infinite regular-tree network
Connected Suspects

validation experiment

![Graph showing correct detection probability vs. suspect size for different theoretical predictions and numerical simulations. The graph includes lines and markers for theoretical predictions and numerical simulations with varying parameters.]
Two Suspects

--- main results

### Case 3

- Two suspect nodes is separated by $d$.

- Identifying the rumor source is more difficult if the two suspects are closer.

### Main results

- **node degree $\delta = 2$**
  
  \[
  P_c(n) = \begin{cases} 
  \frac{1}{2} \left(1 - \frac{(n+d+1)^2}{z}ight), & (n-d) \text{ is odd} \\
  \frac{1}{2} - \sum_{z=(n-d)/2}^{(n+d+1)/2} \left(\frac{n-1}{z}\right), & (n-d) \text{ is even} 
  \end{cases}
  \]

- **node degree $\delta > 2$**
  
  \[
  \lim_{n \to \infty} P_c(n) = I_{1/2} \left(\frac{1}{\delta - 2}, \frac{\delta - 1}{\delta - 2}\right) \geq 0.75, d = 1 \to 1
  \]

- Monotonicity: Detection probability increases with $d$
Two Suspects

validation experiment
SIR Spreading Model

• Fixed Population $N$ (only one infected at each $t$)
• Susceptible Set at time $t$ $S(t)$
• Infected Set at time $t$ $I(t)$
• Recovery Set at time $t$
• Different parameter configuration with infectious rate and recovery rate

Maximum Likelihood Estimator in SIR Spreading

- Sample-path analysis
- Jordan center (graph theory) of a graph is the set of all vertices of minimum eccentricity

\[ \gamma_s \in \arg \min_{b \in \mathcal{V}} \left( \max_{a \in \mathcal{V}_1^{(s)}} d(b, a) \right) \]

- Used for single-source and multiple-source detection [Luo—TSP’13, Zhu—ITA’13]
Detection for General Tree

• Still an open problem
  • Each permitted permutation does not have equal probability

\[ P(\sigma_i|v_1) = \prod_{k=1}^{n-1} \frac{1}{\sum_{v_i \in V(G_k)} d(v_i) - 2(k - 1)} \]

• Breadth-First-Search (BFS) Heuristic Algorithm
  • Rumor centrality algorithm on BFS tree
Detection for General Graph

• Still an open problem
  • How to deal with loops?

• Breadth-First-Search (BFS) Heuristic Algorithm
  • BFS tree approximates *diffusion tree*
  • Rumor centrality algorithm on BFS tree
Detection for General Graph

• Maximum Likelihood Estimator:

\[ \hat{s} \in \arg \max_{s \in G^k_N} P(G^k_N | s) \]

\[ P(G^k_N | s) = \sum_{m=1}^{M} P(\sigma_m | s) \]

- \( I_k(\sigma | s) \) is the number of infected neighboring nodes of the \( k \)th node in \( \sigma(s) \) at time \( k-1 \)
- \( S_k(\sigma | s) \) is the total number of uninfected neighboring nodes for all the infected nodes in \( G^k_N(\sigma | s) \).

\[ P(\sigma_m | s) = \prod_{k=2}^{K} \frac{I_k(\sigma_m | s)}{S_{k-1}(\sigma_m | s)} \]

\[ P(G^k_N | s) = \sum_{m=1}^{M} \left( \prod_{k=2}^{K} \frac{I_k(\sigma_m | s)}{S_{k-1}(\sigma_m | s)} \right) \]
Detection for General Graph

• Toy Example for Regular Lattice Graph
Detection for General Graph

- Toy Example for Regular Lattice Graph

\[ \sigma_1 : s_1 \rightarrow s_2 \rightarrow s_3 \rightarrow s_4, \quad \sigma_2 : s_1 \rightarrow s_2 \rightarrow s_4 \rightarrow s_3, \]
\[ \sigma_3 : s_1 \rightarrow s_3 \rightarrow s_2 \rightarrow s_4, \quad \sigma_4 : s_1 \rightarrow s_3 \rightarrow s_4 \rightarrow s_2. \]

\[ P(\sigma_1 | s_1) = \frac{1}{4} \cdot \frac{2}{6} \cdot \frac{2}{6} = \frac{1}{36}, \quad P(\sigma_2 | s_1) = \frac{1}{4} \cdot \frac{1}{6} \cdot \frac{3}{8} = \frac{3}{192}, \]
\[ P(\sigma_3 | s_1) = \frac{1}{4} \cdot \frac{2}{6} \cdot \frac{2}{6} = \frac{1}{36}, \quad P(\sigma_4 | s_1) = \frac{1}{4} \cdot \frac{1}{6} \cdot \frac{3}{8} = \frac{3}{192}. \]

The exact likelihood of Node 1 is
\[ P(G_8^4 | s_1) = P(\sigma_1 | s_1) + P(\sigma_2 | s_1) + P(\sigma_3 | s_1) + P(\sigma_4 | s_1) = \frac{25}{288}. \]
Detection for General Graph

• Toy Example for Regular Lattice Graph

For a $d$-regular lattice network, every nodes have the same degree $d$
so we can replace $S_{k-1}(\sigma \mid s)$ by $d(k-1) - d(G_N^{k-1}(\sigma))$
where $d(G_N^{k-1}(\sigma))$ is the total number of degree for the subgraph $G_N^{k-1}(\sigma)$
with the first $k-1$ nodes in $\sigma$ infected.

• Special case (Fully connected graph)

the degree of every nodes in the underlying graph $G_N$ and the rumor graph $G_N^K$
are respectively $N - 1$ and $K - 1$.

$$P(\sigma \mid s) = \prod_{k=2}^{K} \frac{k-1}{(K-1)(k-1)-2C_{k-1}^2} = \frac{1}{(K-1)!}$$

$$P(G_N^K \mid s) = (K-1)! P(\sigma \mid s) = 1.$$
Detection for General Graph

- Erdos-Renyi random graph \((N=50, K=20, p=0.4)\)

1) Start with \(N\) isolated nodes;
2) Add an edge between two nodes with probability \(p\).
Detection for General Graph

- Erdos-Renyi random graph
Detection for General Graph

- Newman-Watts small world ($N=50$, $K=20$, $m0=4$, $p=0.4$)

1) Start with a ring-shaped network with $N$ nodes, in which each node is connected to its $2m_0$ neighbors, where $m_0 > 0$ is a (small) positive integer.
2) Add an edge between two nodes with probability $p$. 
Detection for General Graph

- Newman-Watts small world graph
Detection for General Graph

- Barabasi-Albert scale-free graph ($N=50$, $K=20$, $m_0=4, m=2$)

1) Growth: start with a small fully-connected network having $m_0 \geq 1$ nodes, and add one new node to the network each time by connecting to $m$ existing nodes, where ($m \leq m_0$).

2) Preferential attachment: The new node is connected to an existing node $i$ of degree $d_i$ according to the following probability:

$$\Pi_i = \frac{d_i}{\sum_{j=1}^{N} d_j}.$$
Detection for General Graph

• Barabasi-Albert scale free graph
Extensions

- Detection with Multiple Observations
  - How many observations to take?
  - 2 Observations more than double the detectability performance of a single observation

Zhao, Dong, Zhang and Tan
ACM SIGMETRICS 2014

- Open Questions
  - General graph detection
  - How good or how bad is the Breadth-first Search Heuristic?
Extensions

- Centrality in Social Network Analysis
  - Degree
  - Distance
  - Betweenness
  - Eigenvector
  - ...etc

- Center in Graph Theory
  - Mass, Jordan

- General unifying link still missing
  - Jordan center not equal to mass center in general
Data Mining and Network Forensics

- Forensics in Online social networks
Data Mining and Network Forensics

• Many interactions over online social networks that are recorded and available from API service to glimpse social relationship of users
• Provide clues for rumor source detection and other cyber-security forensics algorithms
• How to even obtain a single snapshot observation of the graph?
• Bridge deep gulf between theory and practice
  – There is nothing more practical than a good theory!
Facebook Graph

- Enables application to read/write on Facebook social graph
Facebook Graph

• Example API call (in Javascript)

```javascript
FB.api(
  "/me/friends?fields=id",
  function (response) {
    if (response && !response.error) {
      /* handle the result */
    }
  }
);```

Facebook Graph

- API calls require access token
  - For identification of users, apps etc,
  - For granting permission, web login etc

- Facebook query language
  - Query data from the Graph API (SQL-style interface)

  e.g.

  ```
  SELECT uid1, uid2
  FROM friend  where
  uid1 in ' + user1 + '
  AND uid2 in ' + user2 + ' AND uid1 < uid2';
  ```
Data Mining

• Rate-constrained data scraping
• Access control for privacy and security
• How to use semantics to infer possessing a rumor?
• How to link social graph with technological graph?
Conclusion

- **Rumor Centrality**
  - Center of a Network

- **Network features: Suspects, Connectivity, Observations**

- **Detectability and Detection**
  - Statistical inference, probability theory, graph theory, Information theory
  - Scalable algorithms

- **Numerous Open Issues:**
  - Heterogeneous connectivity and spreading models
  - Real-world data traces
  - Practical network forensics protocol in online social networks
Thank You

cheewtan@cityu.edu.hk

www.cs.cityu.edu.hk/~cheewtan
SI Spreading Model