Rooting out the Rumor Culprit in Online Social Networks

Chee Wei Tan
City University of Hong Kong

The Fourth International Workshop on M2M Technology
30 August, 2014
Motivation

Rumors

@petershankman
Peter Shankman

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

Source: Google
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

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

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

Motivation

Growth and Expansion of Online Social Networks

- Smartphone revolution 2007
- Twitter 2006
- Tencent Weixin 2012
Motivation

Reach of Online Social Networks

WORLD MAP OF SOCIAL NETWORKS

December 2010

credits: Vincenzo Cosenza www.vincos.it
license: CC-BY-NC
source: Google Trends for Websites / Alexa
Motivation
National Grand Challenges

NAE
GRAND
CHALLENGES FOR
ENGINEERING

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
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?
Outline

- Related Work and Spreading Model
- Rumor Centrality
- Detection in Tree Network
- Detection in General Network
- Cybersecurity Forensics
- Conclusion
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)

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

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

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

$|G_t + 1| = |G_t| + 1$
Outline

- Related Work and Spreading Model
- Rumor Centrality
- Detection in Tree Network
- Detection in General Network
- Cybersecurity Forensics
- Conclusion
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\}, ...
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

\[
\begin{align*}
\sigma_1 &= v_1 \rightarrow v_2 \rightarrow v_3 \rightarrow v_4 \\
\sigma_2 &= v_1 \rightarrow v_2 \rightarrow v_4 \rightarrow v_3 \\
\sigma_3 &= v_1 \rightarrow v_3 \rightarrow v_2 \rightarrow v_4 \\
\sigma_4 &= v_1 \rightarrow v_3 \rightarrow v_4 \rightarrow v_2 \\
\sigma_5 &= v_1 \rightarrow v_4 \rightarrow v_3 \rightarrow v_2 \\
\sigma_6 &= v_1 \rightarrow v_4 \rightarrow v_2 \rightarrow v_3
\end{align*}
\]
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

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

■ Related Work and Spreading Model
■ Rumor Centrality
■ Detection in Tree Network
■ Detection in General Network
■ Cybersecurity Forensics
■ Conclusion
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 Symp. on Information Theory 2013

Zhao, Dong, Zhang and Tan
ACM SIGMETRICS 2014
Detectability

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**
  \[
  \hat{s} = \arg \max_{s \in \{S \cap G_n\}} \sum_{s \in S} R(s) \cap G_n
  \]
  - optimal MAP estimator
  - focus on regular trees

- **For general tree-type networks**
  \[
  \hat{s} = \arg \max_{s \in \{S \cap G_n\}} \sum_{s \in S} R(s) \cap G_n
  \]
  - Analyze the correct detection probability upon observing \( n \) infected nodes

- **For general networks**
  \[
  \hat{s} = \arg \max_{s \in \{S \cap G_n\}} \sum_{s \in S} R(s, T_{bfs}(s))
  \]
  \[
  P_c(n) = \text{Prob} \left[ \hat{s} = s^* \right]
  \]
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
  \[ \lim_{n \to \infty} P_G \left[ \bigcap_{j=1}^{\delta} \left( \frac{X_j}{n} = y_j \right) \right] \]
  \[ \lim_{n \to \infty} P_G \left[ \frac{X_1}{n} = y_1 \right] \]
Equivalence to Pólya’s Urn Model

- **Rumor spreading process**

- **Ball drawing process**

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

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

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

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

- The detection probability is asymptotically upper bounded by 0.307.
Minimum Detectability

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

validation experiment
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 $\delta = 3$
  \[
  P_c(n) = \frac{k+1}{2k} + \frac{k-1}{k} \frac{1}{\left\lfloor \frac{n}{2} \right\rfloor + 2} \sim \frac{k+1}{2k} + O(1/n) \geq \frac{1}{k} + \frac{k-1}{2k}
  \]

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

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

- The performance is significantly improved and reliable detection can be achieved.
Connected Suspects

validation experiment

Theoretical prediction (k=2)
Numerical simulation (k=2)
Theoretical prediction (k=3)
Numerical simulation (k=3)
Theoretical prediction (k=5)
Numerical simulation (k=5)
Theoretical prediction (k=10)
Numerical simulation (k=10)
Connected Suspects
----with vs. without prior knowledge

■ Closer-up look at case 2
- node degree $\delta > 2$
  \[ \lim P_c(n) = 1 - \frac{2k-2}{k} \left(1 - I_{1/2} \left( \frac{1}{\delta - 2}, \frac{\delta - 1}{\delta - 2} \right) \right) \]
- exceed prior probability
  \[ P_c(n) \geq \frac{1}{k} + \frac{k - 1}{2k} \]
- at least 0.5-detection
  \[ P_c(n) \geq 2I_{1/2} \left( \frac{1}{\delta - 2}, \frac{\delta - 1}{\delta - 2} \right) - 1 \geq 0.5 \]
- achieve reliable detection
  \[ \lim_{\delta \to \infty} \lim_{n \to \infty} P_c(n) = 1 \]

■ Comparison with case 1
- node degree $\delta > 2$
  \[ \lim P_c(n) = 1 - \delta \left(1 - I_{1/2} \left( \frac{1}{\delta - 2}, \frac{\delta - 1}{\delta - 2} \right) \right) \]
- non-trivial positive value
  \[ P_c(n) > 0 \]
- at most 0.5-detection
  \[ P_c(n) \leq 0.5 \]
- upper-bounded by 0.307
  \[ \lim \lim P_c(n) = 0.307 \]

– Suspect characteristics (connectivity) bring about new ingredients.
Two Suspects

---main results

**Case 3**

- Two suspect nodes is separated by $d$. $d(s_1, s_2) = d$

**Main results**

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

- **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 $d$

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

validation experiment

Theoretical prediction (d=1)
Numerical simulation (d=1)
Numerical simulation (d=2)
Numerical simulation (d=3)
Numerical simulation (d=4)
Detection in General Network

• 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
Outline

- Related Work and Spreading Model
- Rumor Centrality
- Detection in Tree Network
- Detection in General Network
- Cybersecurity Forensics
- Conclusion
Detection in General Network

- 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 in General Network

- Erdos-Renyi random graph
Detection in General Network

- Newman-Watts small world \((N=50, K=20, m_0=4, p=0.4)\)

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 in General Network

- Newman-Watts small world graph

![Graph with actual rumor source and rumor center]
Detection in General Network

• Barabasi-Albert scale-free graph \((N=50, K=20, m0=4, m=2)\)

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:

\[
\Pi_i = \frac{d_i}{N \sum_{j=1}^{N} d_j}
\]
Detection in General Network

- Barabasi-Albert scale free graph
Detection with Multiple Observations

- A source may initiate multiple instances, e.g., email spam, recurring malcodes.
- Can diversity help? How many observations to take?

Zhao, Dong, Zhang and Tan, ACM SIGMETRICS 2014
Performance results

➢ For regular tree

• **Case 2**

  $\delta = 3, \ K = 2$  Given $G_{n_1}, G_{n_2}$

  - (1) $n_1 = n, n_2 = qn \ (q \in \mathbb{Z}^+)$
    
    \[ P_c = \frac{qn + q + 2}{2(qn + 1)} \]

  - (2) $n_1 = n, n_2 = qn + 1 \ (q \in \mathbb{Z}^+)$
    
    \[ P_c = \frac{qn + q + 2}{2(qn + 1)} \]

  - (3) $n_1 = n, n_2 = qn + t \ (q \in \mathbb{Z}^+), t < n$
    
    \[ P_c = \frac{qn + q + 2}{2(qn + 1)} + \Delta P_c \quad \text{with} \quad \Delta P_c < \frac{1}{2(qn + 1)}. \]

As $n \to +\infty$, \( \lim_{n \to +\infty} P_c = \frac{1}{2} \)
Performance results

Detection performance with multiple instances:

\[ n_1, n_2 \rightarrow \infty \]

Asymptotic correct detection probability \( (16,0.9) \)

Asymptotic correct detection probability \( (3,1/2) \)

\[ n_1,n_2,n_3 \rightarrow \infty \]

Asymptotic correct detection probability \( (6,0.9) \)

Asymptotic correct detection probability \( (3,0.65) \)

K=2

K=3
Performance results

Detection performance with multiple instances:

\[ \phi_k(\delta) \text{ vs } K, \delta = 3 \]
Outline

- Related Work and Spreading Model
- Rumor Centrality
- Detection in Tree Network
- Detection in General Network
- Cybersecurity Forensics
- Conclusion
Forensics in Online Social Networks
Network Cyber Security 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!
Forensics using Facebook Graph

- Rate-constrained data scraping
- Access control for privacy and security
- How to use semantics to infer the possession of 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