Rumor Source Detection in the SIR Model: A Sample Path Approach

Kai Zhu, Lei Ying

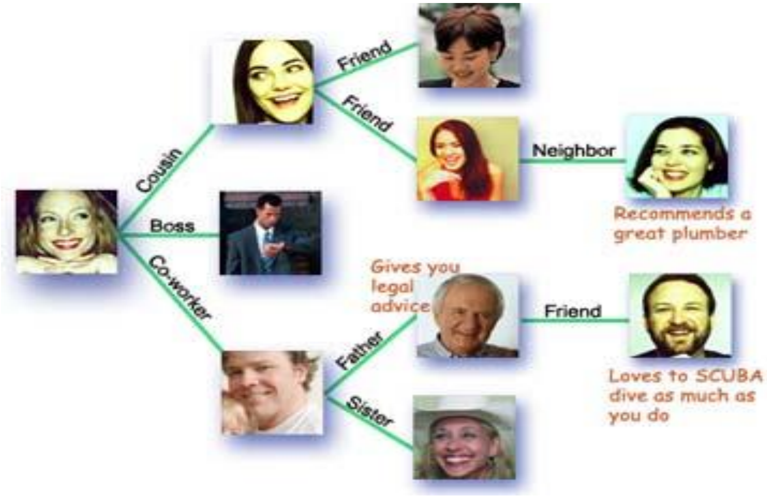
Arizona State University

Presented by Bao Yuanyuan

- Kai Zhu, Lei Ying. Information Source Detection in the SIR Model: A Sample Path Based Approach. Information Theory and Application Workshop (**ITA 2013**).
- Kai Zhu, Lei Ying. A Robust Information Source Estimator with Sparse Observations. **IEEE INFOCOM 2014**.

Background

- **Social networks**



- **Rumor**
  - **Top 100 hottest events on Sina Weibo of 2012.1-2013.1: 1/3 are rumors.**

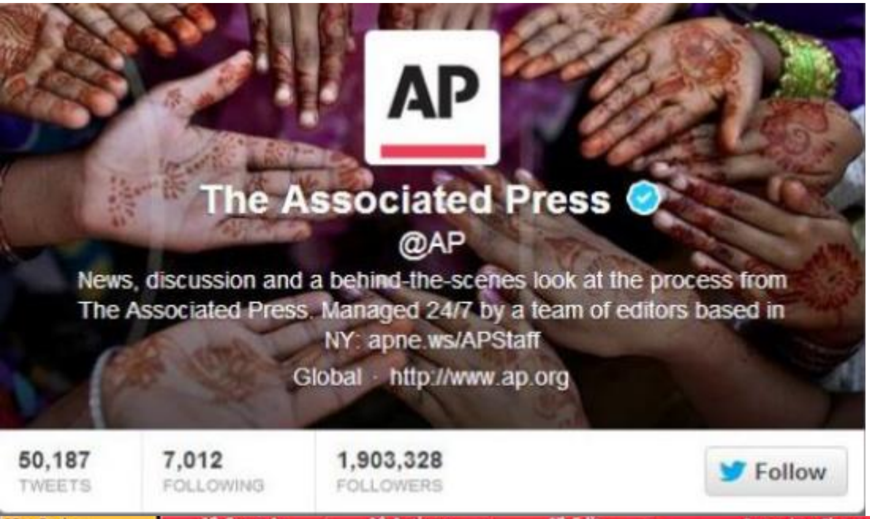
Background

When Hurricane Sandy came, rumors about “confirmed flooding” of the **New York Stock Exchange**, failure of the Old Bridge Township **water system** and **bodies of victims** been found in Seaside Heights circulated on Twitter and resulted in **social panics**.

Background

It said that the president of Syria is dead, which hit twitter greatly and was circulated fast among population, leading to a sharp, quick increase in the price of oil.

Background



AP

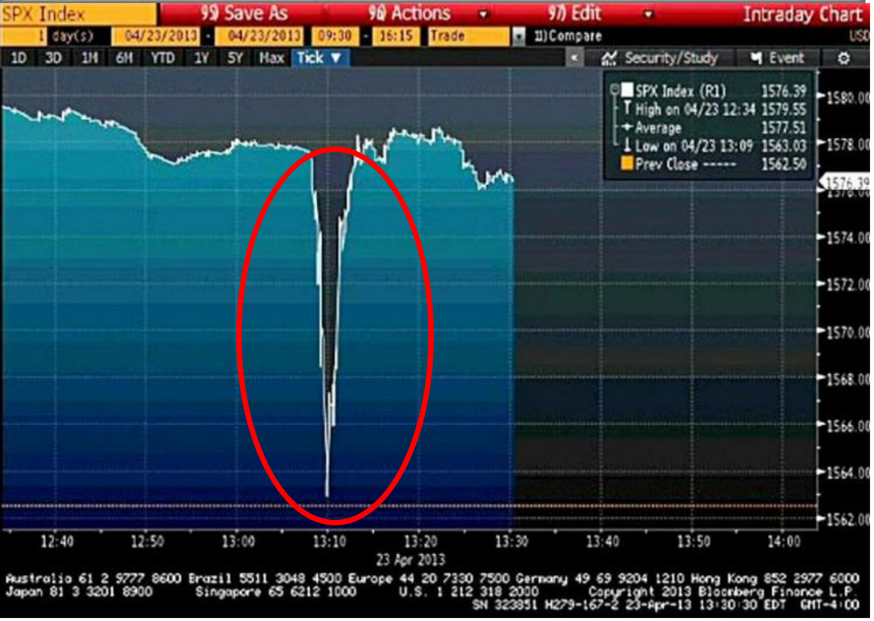
The Associated Press

@AP

News, discussion and a behind-the-scenes look at the process from The Associated Press. Managed 24/7 by a team of editors based in NY: apne.ws/APStaff

Global · http://www.ap.org

50,187 TWEETS · 7,012 FOLLOWING · 1,903,328 FOLLOWERS



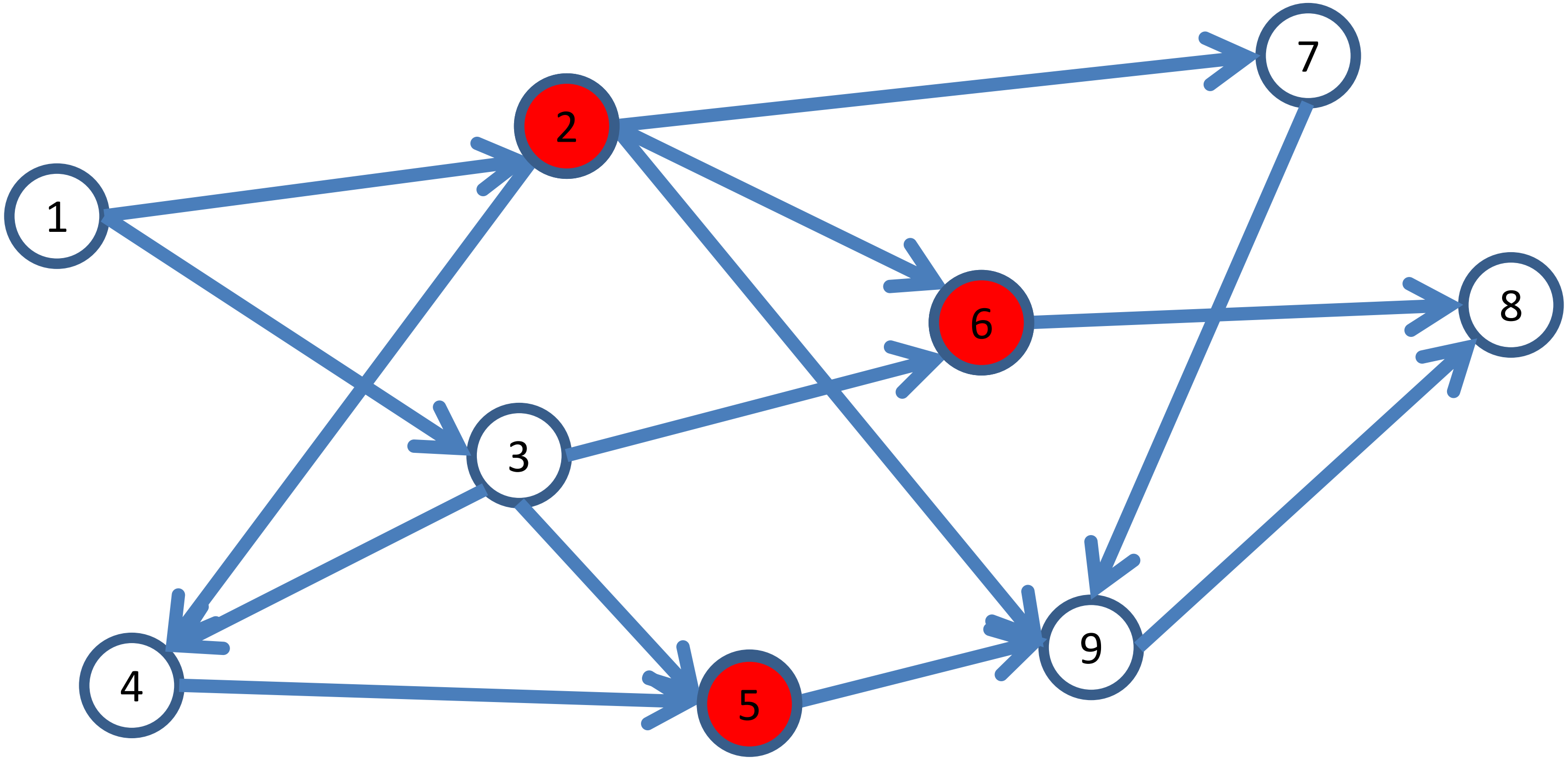
Rumor about explosions at the White House injuring President Obama tweeted by a news agency, made the Dow plunge more than 140 points and the temporary loss of market cap in the S&P 500 alone totaled $136.5 billion.

Here the problem comes!

- Rumor Control
- Rumor Source Detection
- Ideal condition: all tweets in chronological sequence
- Actual condition: only some tweets
- **Rumor source detection problem:**

Given a snapshot of the diffusion process at time t, tell which node is the source of the diffusion.

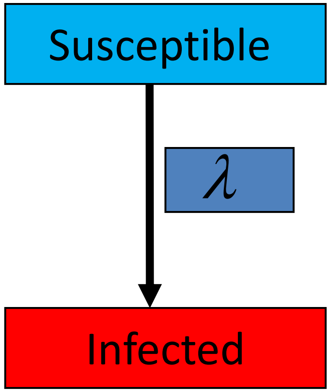
Rumor Source Detection Problem



Given a snapshot of the diffusion process at time t, which node is the source of the diffusion? (Topology is also known.)

Related Work

SI Model

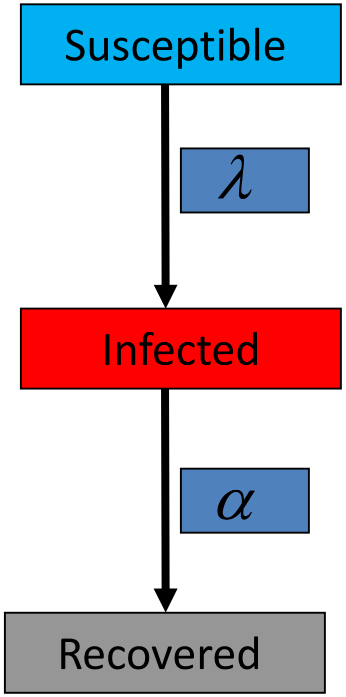


Susceptible

λ

Infected

SIR Model



Susceptible

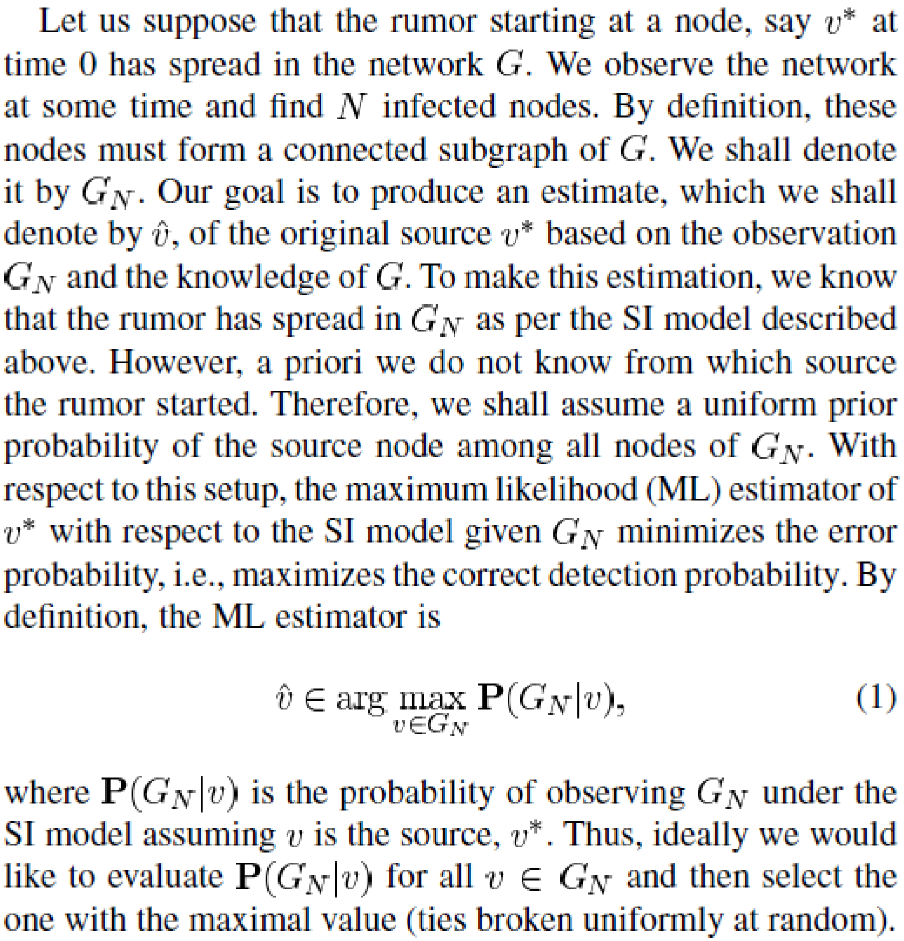
λ

Infected

α

Recovered

Related Work



Let us suppose that the rumor starting at a node, say v^* at time 0 has spread in the network G. We observe the network at some time and find N infected nodes. By definition, these nodes must form a connected subgraph of G. We shall denote it by G_N. Our goal is to produce an estimate, which we shall denote by $\hat{v}$, of the original source v^* based on the observation G_N and the knowledge of G. To make this estimation, we know that the rumor has spread in G_N as per the SI model described above. However, a priori we do not know from which source the rumor started. Therefore, we shall assume a uniform prior probability of the source node among all nodes of G_N. With respect to this setup, the maximum likelihood (ML) estimator of v^* with respect to the SI model given G_N minimizes the error probability, i.e., maximizes the correct detection probability. By definition, the ML estimator is

$$\hat{v} \in \arg\max_{v \in G_N} \mathbf{P}(G_N|v), \qquad (1)$$

where $\mathbf{P}(G_N|v)$ is the probability of observing G_N under the SI model assuming v is the source, v^*. Thus, ideally we would like to evaluate $\mathbf{P}(G_N|v)$ for all $v \in G_N$ and then select the one with the maximal value (ties broken uniformly at random).

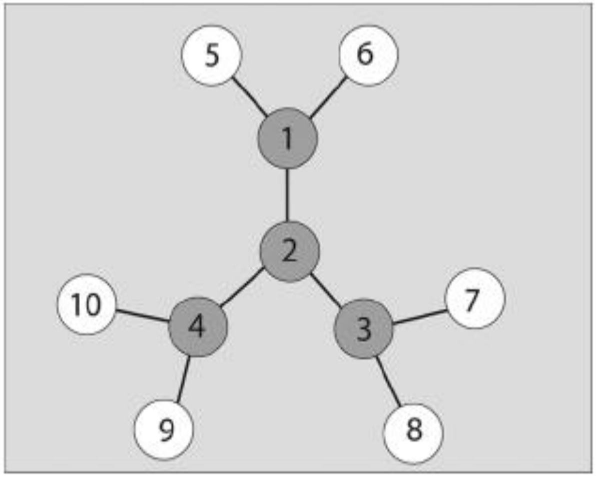


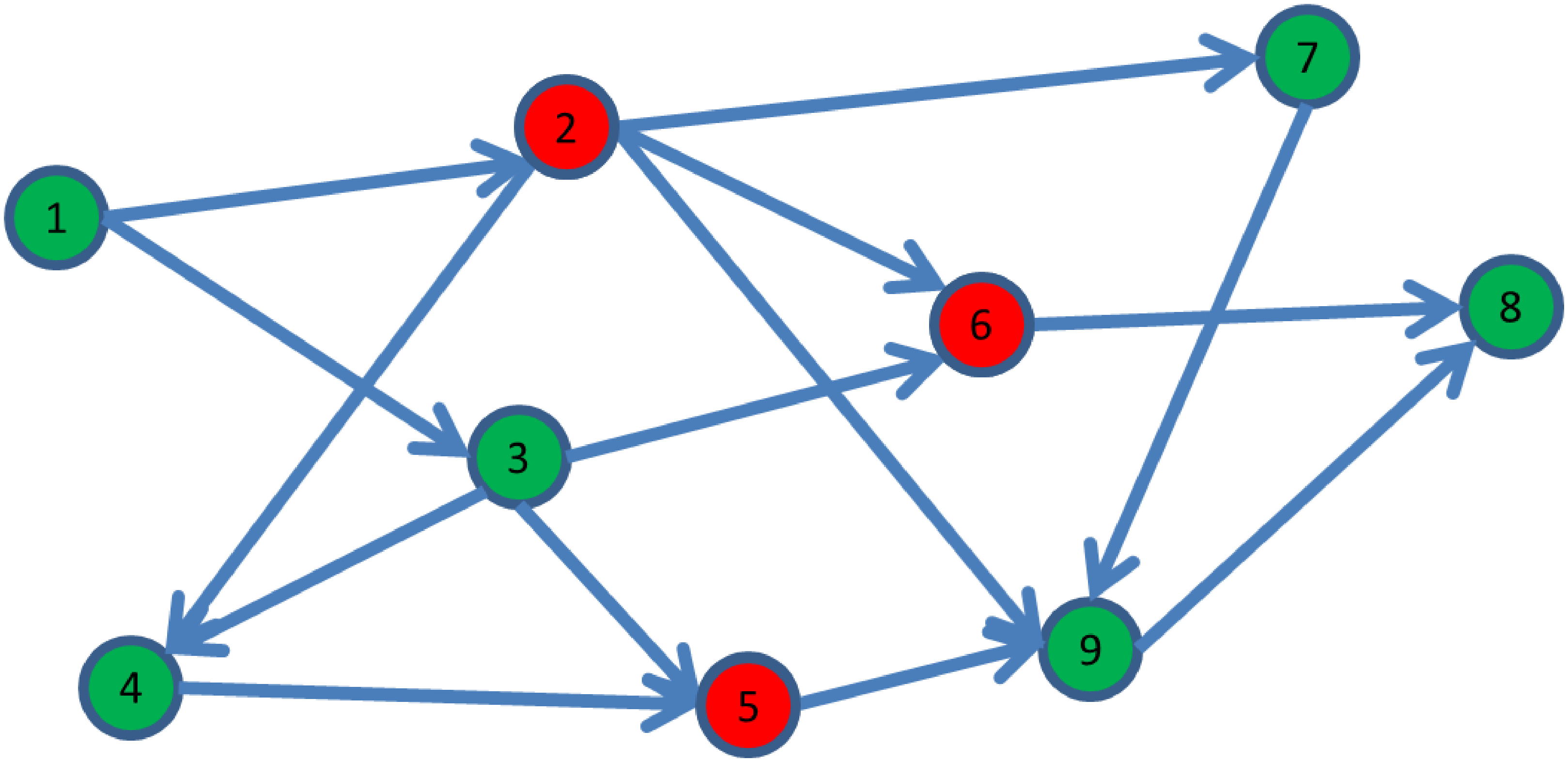
Fig. 1. Example network where the rumor graph has four nodes.

D. Shah, T. Zaman. Rumors in a Network: Who’s the Culprit?. IEEE Transactions on Information Theory, Vol. 57, No. 8, August 2011.

Limitations

- **SIR** is the natural (somewhat standard) model for viral epidemics.
- It is very important to take **recovery** into consideration.
  - A contraband material uploader may **delete** the file;
  - Anti-virus software **removes** the virus;
  - A user **deletes** the rumor from his/her microblog.

Challenge



Only can identify **infected nodes** and **healthy nodes** (susceptible nodes and recovered nodes). Susceptible nodes and recovered nodes are **indistinguishable**.

PROBLEM FORMATION

- **THE SIR MODEL FOR INFORMATION PROPAGATION**
- **INFORMATION SOURCE DETECTION**
- **MAXIMUM LIKELIHOOD DETECTION**
- **SAMPLE PATH BASED DETECTION**

THE SIR MODEL FOR INFORMATION PROPAGATION

- Undirected graph G={V, E}, where V is the set of nodes and E is the set of edges.
- Each node vЄV has three possible states: susceptible (S), infected (I), and recovered (R).
- Nodes change their states at the beginning of each time slot, and the state of node v in time slot is denoted by $X_v(t)$.
- Initially, all nodes are in state S except node v* which is in state I and is the information source.
- Infected with probability q and recover with probability p.
- **The states of all the nodes at time slot t: $X(t)=\{X_v(t), v\in V\}$ Markov chain**

INFORMATION SOURCE DETECTION

- However, X(t) is not full observable. Only observe Y={Y_v, v∈V} such that

$$Y_v = \begin{cases} 1, & \text{if } v \text{ is in state } I; \\ 0, & \text{if } v \text{ is in state } S \text{ or } R. \end{cases}$$

- The information source detection problem is to identify v* given the graph G and Y.

An Example of Information Propagation

(infection time, recovery time)

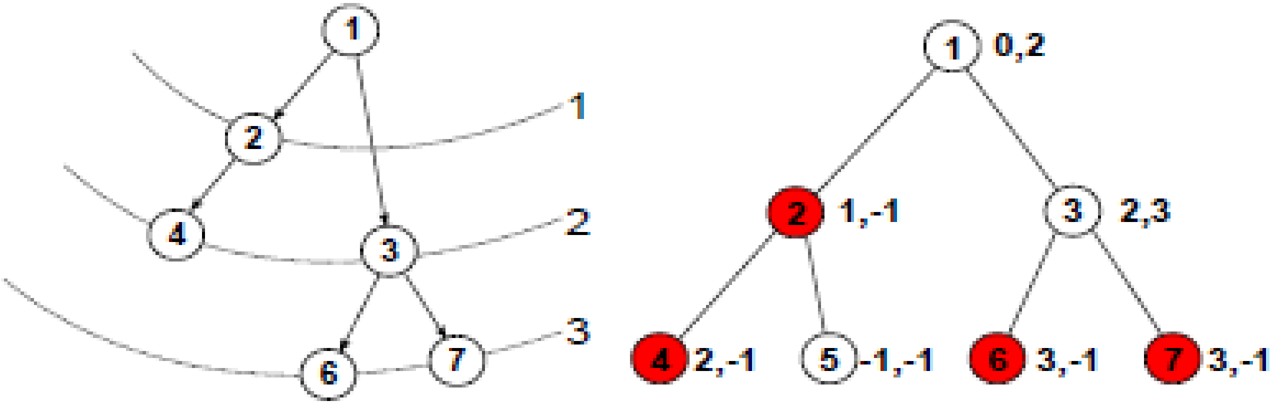


Figure 1. An Example of Information Propagation

If we observe the network at the end of the time slot 3, then the snapshot of the network is Y={0,1,0,1,0,1,1}.

MAXIMUM LIKELIHOOD DETECTION

$$v^{\dagger} \in \arg\max_{v \in \mathcal{V}} \sum_{\mathbf{X}[0,t]:\mathbf{F}(\mathbf{X}(t))=\mathbf{Y}} \Pr(\mathbf{X}[0,t] | v^{*} = v),$$

If source=v_1, exist X(1), X(2),…, X(t); → $\sum \boldsymbol{P_r(X[0,t])}$

If source=v_2, exist X(1), X(2),…, X(t); → $\sum \boldsymbol{P_r(X[0,t])}$

…

If source=v_n, exist X(1), X(2),…, X(t); → $\sum \boldsymbol{P_r(X[0,t])}$

Max $\sum \boldsymbol{P_r(X[0,t])}$

- sample path X[0,t] given the source is node v.

CURSE OF DIMENSIONALITY

$$v^{\dagger} \in \arg\max_{v\in\mathcal{V}} \sum_{\mathbf{X}[0,t]:\mathbf{F}(\mathbf{X}(t))=\mathbf{Y}} \Pr(\mathbf{X}[0,t]|v^{*} = v),$$

If Y_v=1, need to decide the infection time. O(t) possible choices.

If Y_v=0, need to decide the infection time and recovery time. O(t^2) possible choices.

Even for a fixed t, the number of possible sample paths is at lease t^N.

SAMPLE PATH BASED DETECTION

MLE: $$v^{\dagger} \in \arg\max_{v\in\mathcal{V}} \sum_{\mathbf{X}[0,t]:\mathbf{F}(\mathbf{X}(t))=\mathbf{Y}} \Pr(\mathbf{X}[0,t]|v^{*}=v),$$

To identify the sample path X*[0,t*] that most likely leads to Y:

$$\mathbf{X}^{*}[0,t^{*}] = \arg\max_{t,\mathbf{X}[0,t]\in\mathcal{X}(t)} \Pr\left(\mathbf{X}[0,t]\right),$$

Where $\mathcal{X}(t) = \{\mathbf{X}[0,t]|\mathbf{F}(\mathbf{X}(t)) = \mathbf{Y}\}$.

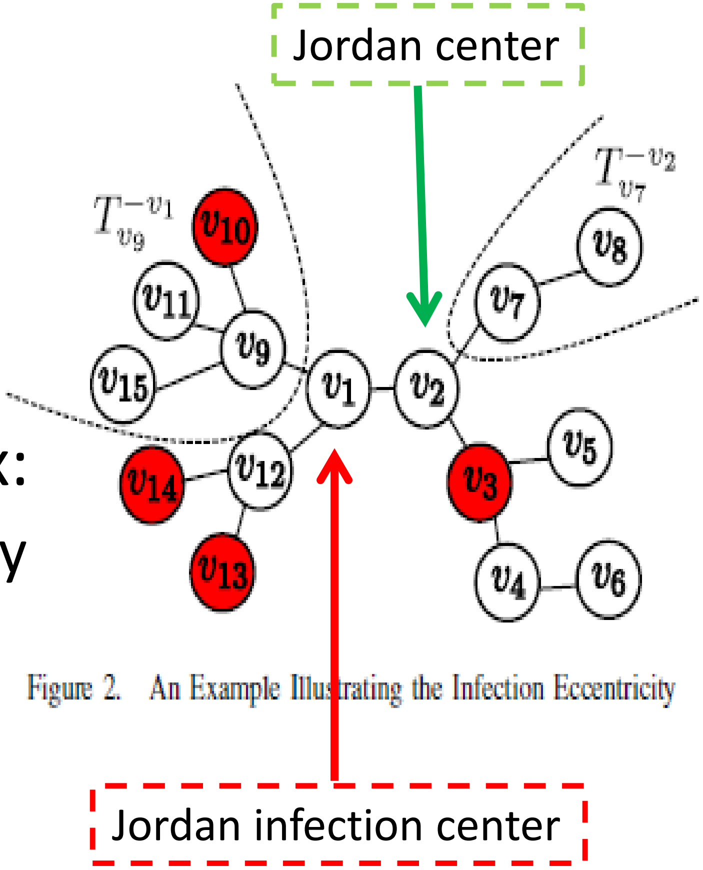
The source node associated with X*[0,t*] is then viewed as the information source.

SAMPLE PATH BASED DETECTION ON TREE NETWORKS

- The optimal sample paths for **general graphs** are still difficult to obtain.
- Focus on **tree networks** and derive structure properties of the optimal sample paths.

Infection Eccentricity

- Eccentricity e(v) of a vertex:
  - maximum distance between v and other vertex in the graph.
- Jordan centers:
  - the nodes having the minimum eccentricity.
- Infection eccentricity ẽ(v) of a vertex:
  - Maximum distance between v and any infected nodes
- Jordan infection centers
  - Nodes with the minimum infection eccentricity.



Jordan center

Figure 2. An Example Illustrating the Infection Eccentricity

Jordan infection center

SAMPLE PATH BASED DETECTION ON TREE NETWORKS

- **The source associated with the optimal sample path=Node with the minimum infection eccentricity.**

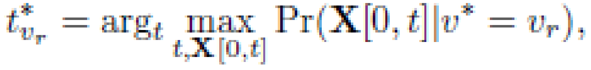
I. Time duration of the optimal sample path equals to the infection eccentricity of node v_r.

II. The optimal sample path starting from a node with a smaller infection eccentricity is more likely to occur. (the optimal sample path rooted at a node with smaller infection eccentricity occurs with a higher probability.)

III. The source of optimal sample path must be a Jordan infection center.

I. Time duration of the optimal sample path equals to the infection eccentricity of node v_r.

Assuming the information source is v_r, analyze **time duration** of the optimal sample path such that

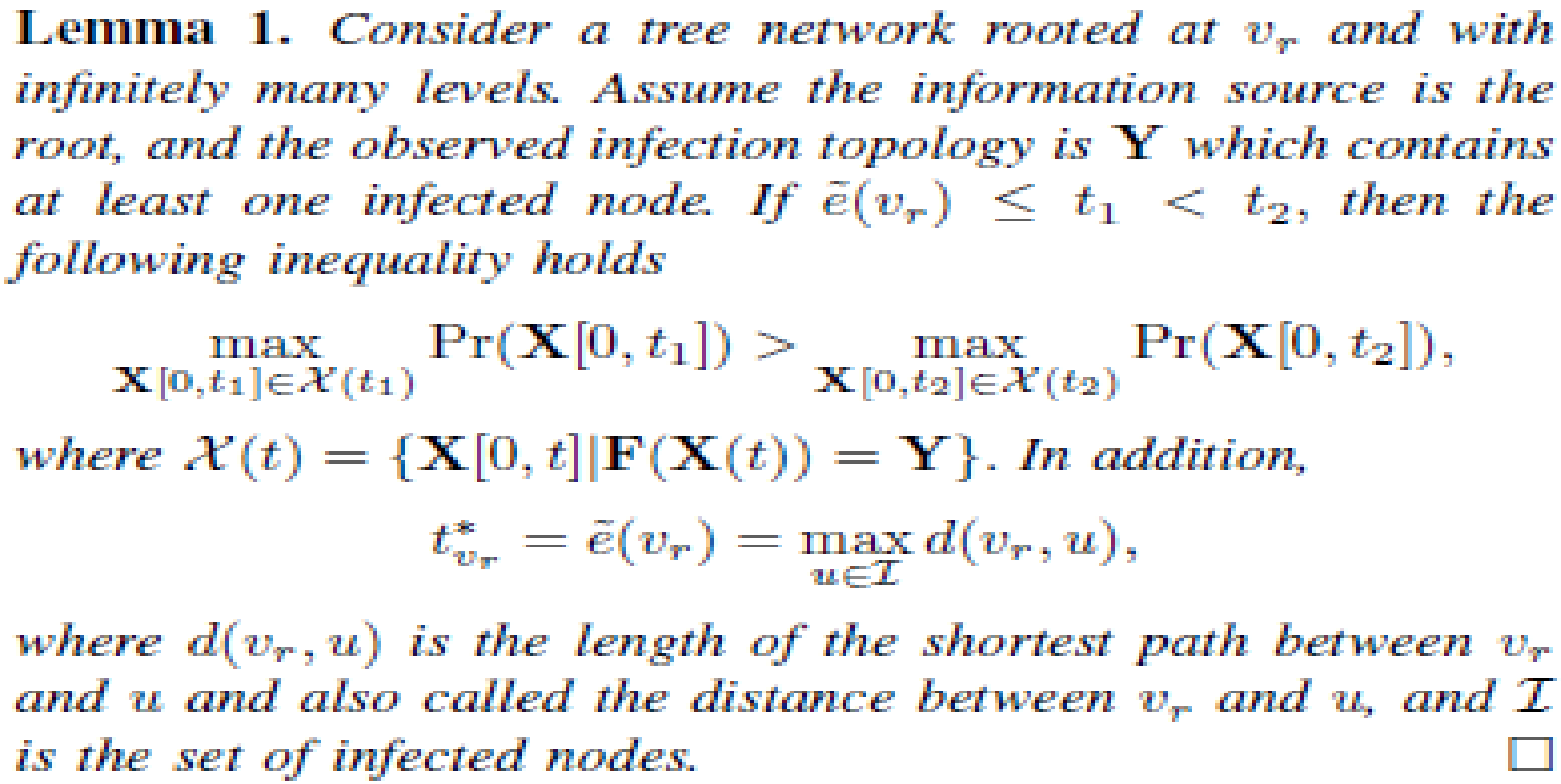


$$t^*_{v_r} = \arg_t \max_{t, \mathbf{X}[0,t]} \Pr(\mathbf{X}[0,t] | v^* = v_r),$$

${t_{v_r}}^*$ is the time duration of the optimal sample path in which v_r is the source.

Time duration of the optimal sample path equals to the infection eccentricity of node v_r.

I. Time duration of the optimal sample path equals to the infection eccentricity of node v_r.



Lemma 1. *Consider a tree network rooted at v_r and with infinitely many levels. Assume the information source is the root, and the observed infection topology is $\mathbf{Y}$ which contains at least one infected node. If $\tilde{e}(v_r) \leq t_1 < t_2$, then the following inequality holds*

$$\max_{\mathbf{X}[0,t_1]\in\mathcal{X}(t_1)} \Pr(\mathbf{X}[0,t_1]) > \max_{\mathbf{X}[0,t_2]\in\mathcal{X}(t_2)} \Pr(\mathbf{X}[0,t_2]),$$

where $\mathcal{X}(t) = \{\mathbf{X}[0,t] | \mathbf{F}(\mathbf{X}(t)) = \mathbf{Y}\}$. In addition,

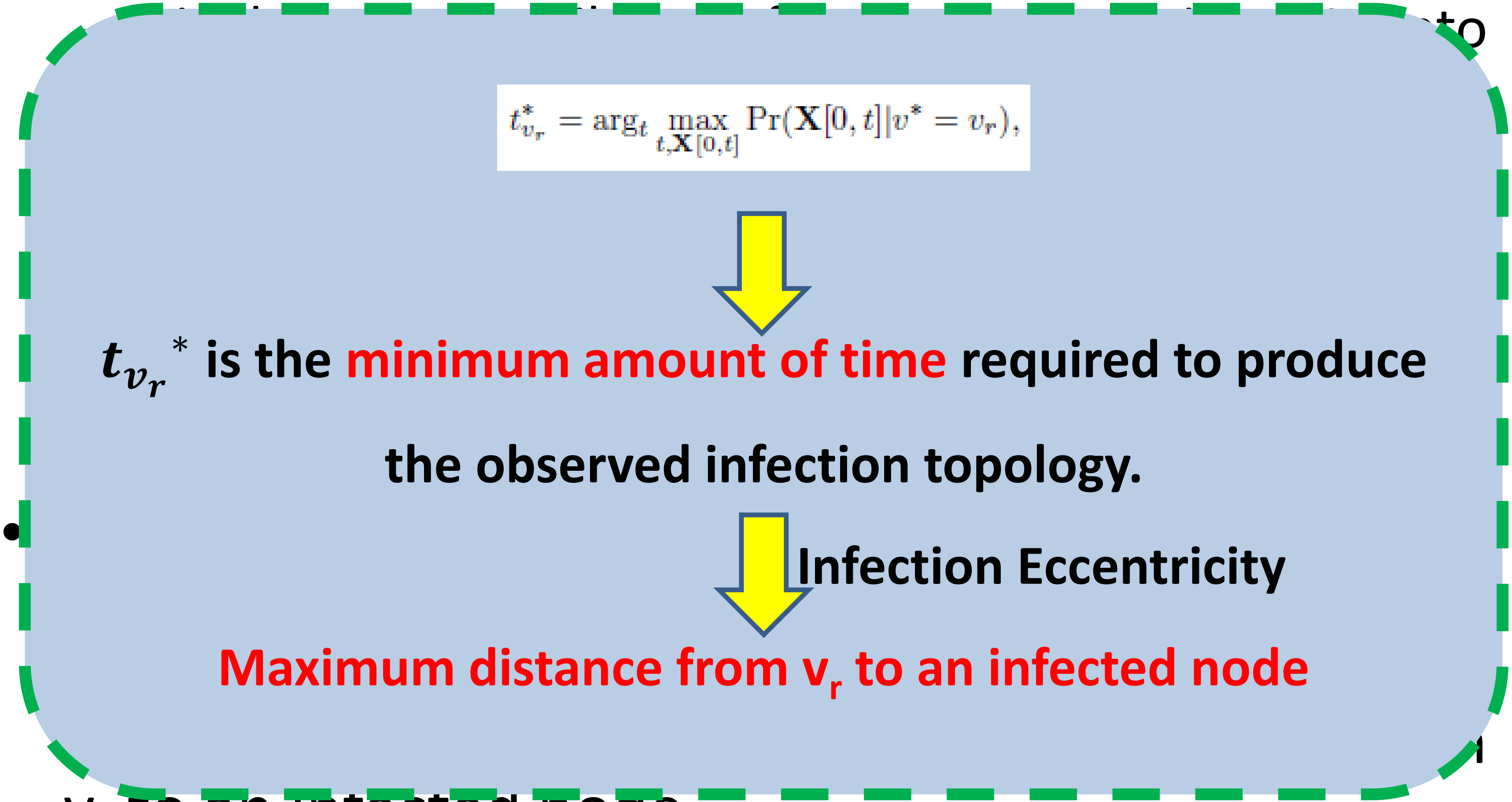
$$t^*_{v_r} = \tilde{e}(v_r) = \max_{u\in\mathcal{I}} d(v_r, u),$$

where $d(v_r, u)$ is the length of the shortest path between v_r and u and also called the distance between v_r and u, and $\mathcal{I}$ is the set of infected nodes. □

- Start from the case where the time difference of two sample path is one.

$$\max_{\mathbf{X}[0,t]\in\mathcal{X}(t)} \Pr(\mathbf{X}[0,t]) > \max_{\mathbf{X}[0,t+1]\in\mathcal{X}(t+1)} \Pr(\mathbf{X}[0,t+1]). \tag{2}$$

to



$$t^*_{v_r} = \arg_t \max_{t,\mathbf{X}[0,t]} \Pr(\mathbf{X}[0,t]|v^* = v_r),$$

$t_{v_r}{}^*$ is the minimum amount of time required to produce the observed infection topology.

Infection Eccentricity

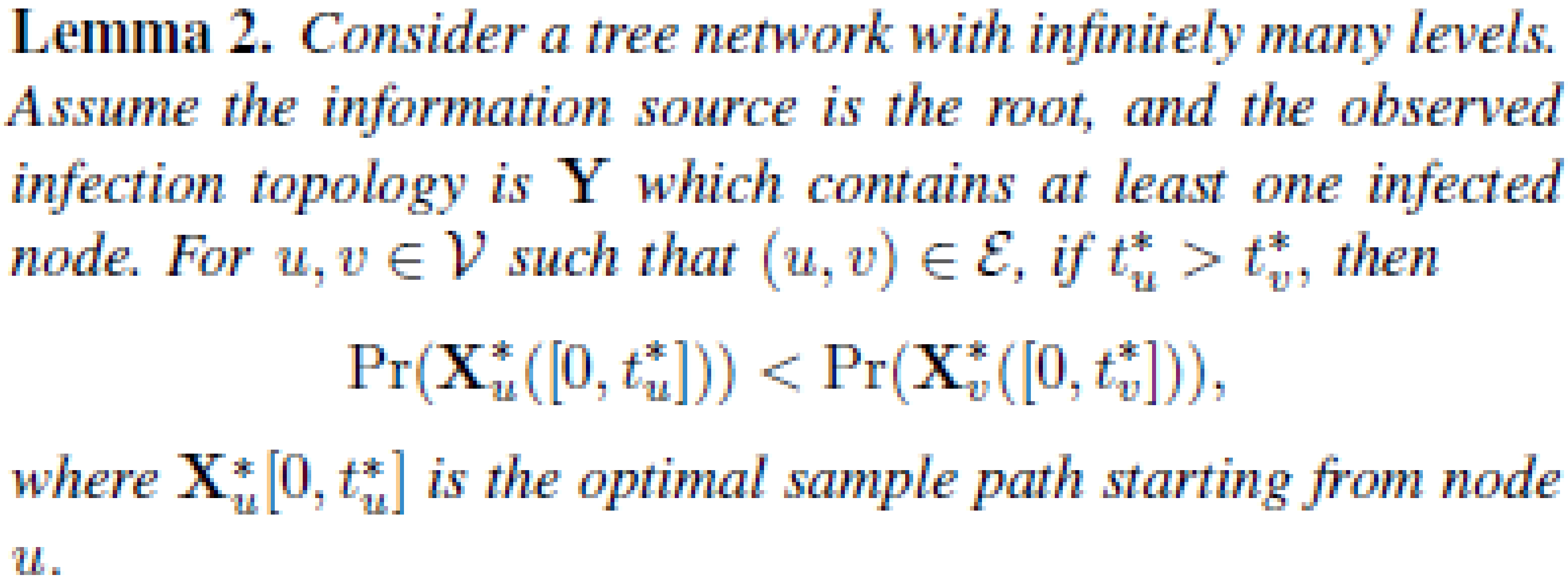
Maximum distance from v_r to an infected node

-

d

v_r to an infected node.

II. The optimal sample path starting from a node with smaller infection eccentricity is more likely to occur.



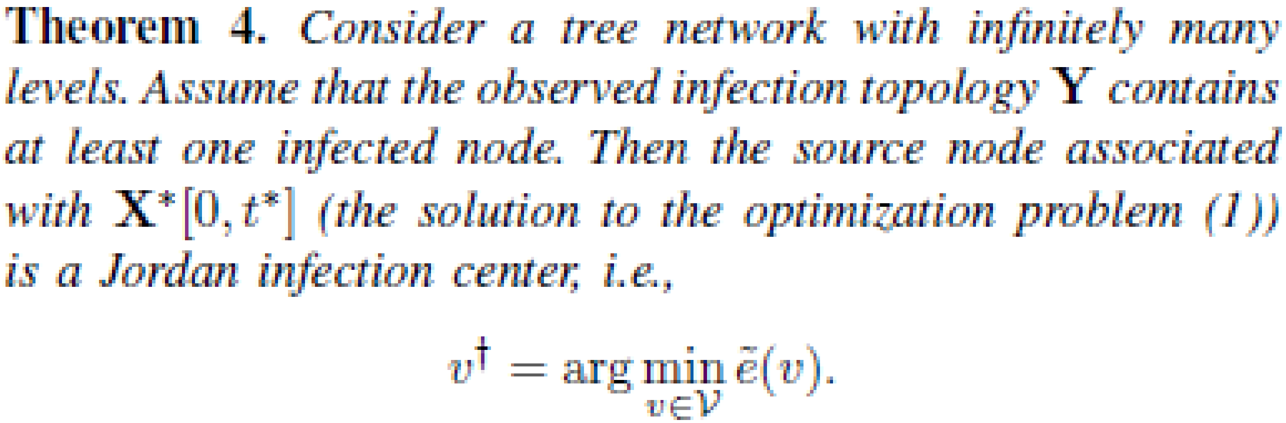
Lemma 2. *Consider a tree network with infinitely many levels. Assume the information source is the root, and the observed infection topology is* $\mathbf{Y}$ *which contains at least one infected node. For* $u, v \in \mathcal{V}$ *such that* $(u, v) \in \mathcal{E}$, *if* $t_u^* > t_v^*$, *then*

$$\Pr(\mathbf{X}_u^*([0, t_u^*])) < \Pr(\mathbf{X}_v^*([0, t_v^*])),$$

where $\mathbf{X}_u^*[0, t_u^*]$ *is the optimal sample path starting from node* u.

- Step 1: To show $t_u^*=t_v^*+1$;
- Step 2: To prove $t_v^I=1$;
- Step 3: Given sample path $X_u^*=[0, t_u^*]$, construct $X_v=[0, t_v^*]$, which occurs with a higher probability.

III. The source of optimal sample path must be a Jordan infection center.



Theorem 4. *Consider a tree network with infinitely many levels. Assume that the observed infection topology* $\mathbf{Y}$ *contains at least one infected node. Then the source node associated with* $\mathbf{X}^*[0,t^*]$ *(the solution to the optimization problem (1)) is a Jordan infection center, i.e.,*

$$v^\dagger = \arg\min_{v\in\mathcal{V}} \tilde{e}(v).$$

- Step 1-Step 3: If v has the minimum infection eccentricity and u has a larger minimum infection eccentricity, then there exists a path from u to v along which the infection eccentricity monotonically decrease.
- Step 4: Repeatedly applying Lemma 2 along the path from node u to v, can conclude that the optimal sample path rooted at node v is more likely to occur than the optimal sample path rooted at node u.
- Root node associated with the optimal sample path must be a Jordan infection center.

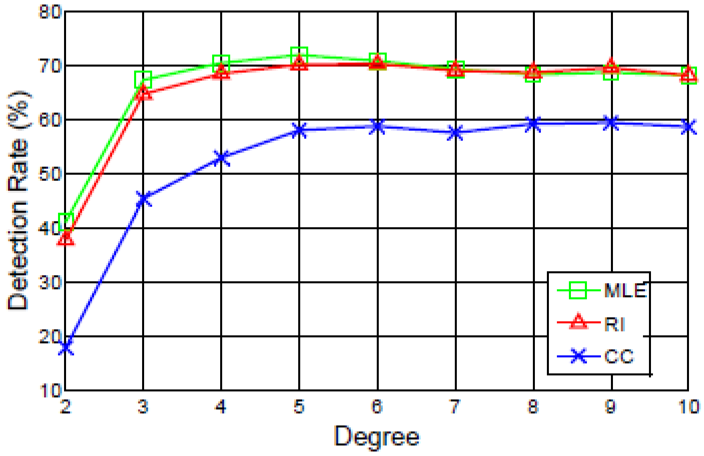
Reverse Infection Algorithm

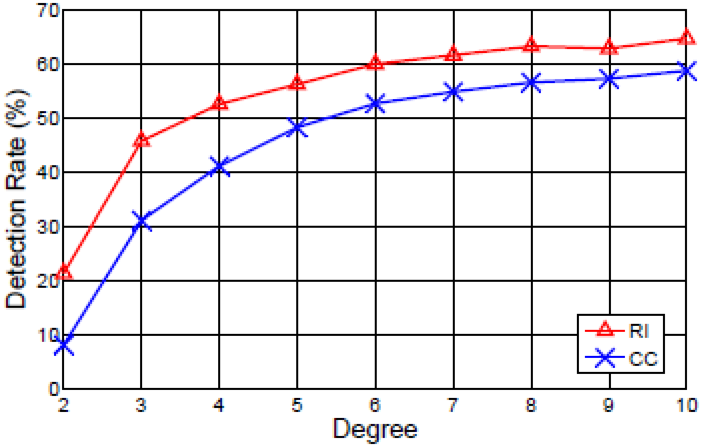
- Let every infected node broadcast a message containing its identity(ID) to its neighbors.
- When a node receives the IDs of all infected nodes, it claims itself as the information source the algorithm terminates.
- Tie-breaking rule: choose the node with the maximum infection closeness(inverse of the sum of distances from a node to all infected nodes)

Performance Analysis

- Demonstrate the effectiveness of the sample path based approach, within a constant distance of from the actual source with a high probability, independent of the number of infected nodes and the time at which the snapshot Y was taken.

Tree network

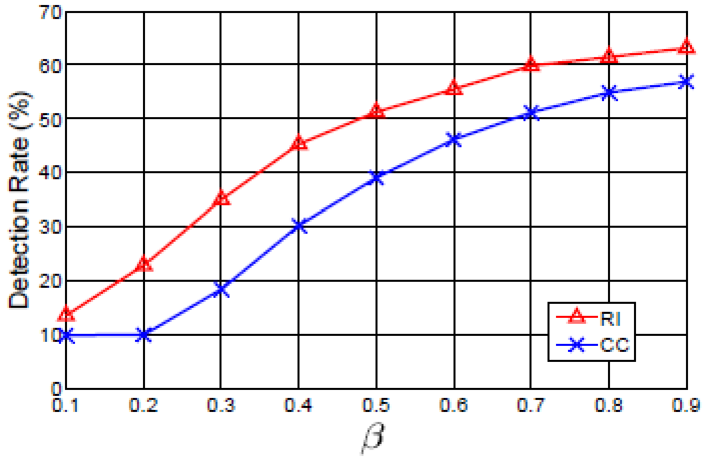




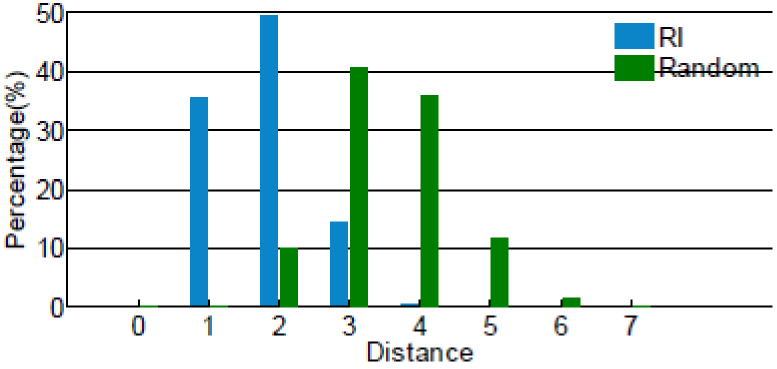
- Small-size tree networks
  - No more than 100
  - Detection rate is almost the same as that of MLE.
  - Higher than that of the closeness centrality 20% when degree is small.
- General g-regular tree networks
  - Higher than 60% when g>6.
  - Higher than that of closeness centrality, average difference is 8.86%.

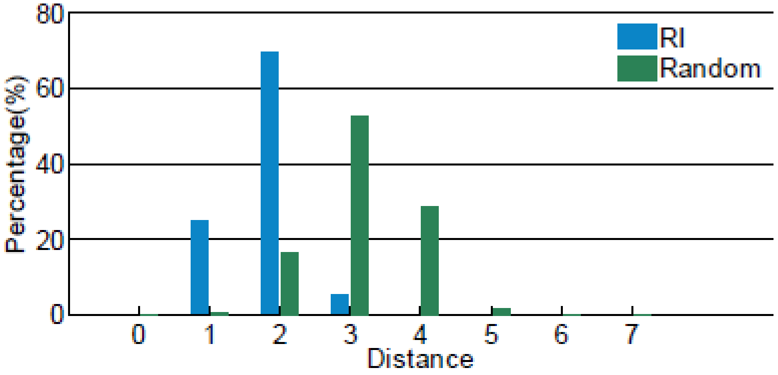
Tree network

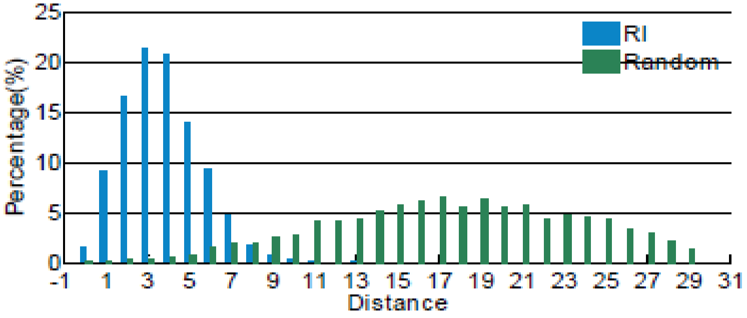
- Binomial random trees
  - The number of children of each node follows a binomial distribution X~B(g’, β). g’=10, β from 0.1 to 0.9
  - RI outperforms the closeness centrality algorithm by 10.16% on average.



Real World Network







- Internet Autonomous system network
  - 10670 nodes and 22002 edges
  - More than 80% are no more than two hops away from the actual sources.
- Wikipedia network
  - 7066 nodes and 100736 links
  - More than 90% are no more than two hops away from the actual sources.
- Power grid network
  - 4941 nodes and 6594 links
  - The peak of the reverse infection algorithm appears at the third hop versus the seventeenth hop under random guessing.

Conclusion

- Develop a **sample path based approach**
- Prove that the sample path based estimator is a node with **minimum infection eccentricity**
- Propose a **reverse infection algorithm**
- Analyze the performance of the RI algorithm and demonstrate the effectiveness.
- Evaluate the performance on **real networks**.

Q & A