Identifying Propagation Sources and Modeling Propagation Dynamics in Networks*

Professor Wanlei Zhou
Alfred Deakin Professor
Chair of Information Technology
Deakin University
wanlei@deakin.edu.au
www.deakin.edu.au/~wanlei

* Work is carried out by many colleagues in my group: Prof Wanlei Zhou, Prof Yang Xiang, Dr Shui Yu, Dr Robin Doss, Dr Jun Zhang, Dr Silvio Cesare, Dr Sheng Wen, Dr Yini Wang, Ms Jiaojiao Jiang, Mr Chao Chen, etc.

Outline

- Introduction
- Source Identification Methods
- Comparative Studies
- Our Work in Modelling the Propagation of Worms and Rumours
- Future Work & Discussion

Key Publications

This presentation is based on following publications by my research group:

1. Sheng Wen, Wei Zhou, Jun Zhang, Yang Xiang, Wanlei Zhou, and Weijia Jia, "Modeling Propagation Dynamics of Social Network Worms", ***IEEE Transactions on Parallel and Distributed Systems***, vol. 24, no. 8, pp. 1633-1643, Aug. 2013.
2. Yini Wang, Sheng Wen, Yang Xiang, and Wanlei Zhou, "Modeling the Propagation of Worms in Networks: A Survey", ***IEEE Communications Surveys and Tutorials***, Volume:16, Issue:2, pp 942-960, 2014.
3. Sheng Wen, Wei Zhou, Jun Zhang, Yang Xiang, Wanlei Zhou, Weijia Jia, and Cliff C.Zou "Modeling and Analysis on the Propagation Dynamics of Modern Email Malware", ***IEEE Transactions on Dependable and Secure Computing***, VOL. 11, NO. 4, pp. 361-374, JULY/AUGUST 2014.
4. Sheng Wen, Jiaojiao Jiang, Yang Xiang, Shui Yu, and Wanlei Zhou, "Are the Popular Users Always Important for the Information Dissemination in Online Social Networks?" ***IEEE Network***, pp. 64-67, September/October 2014.
5. Sheng Wen, Jiaojiao Jiang, Yang Xiang, Shui Yu, Wanlei Zhou, Weijia Jia, "To Shut Them Up or to Clarify: Restraining the Spread of Rumours in Online Social Networks", ***IEEE Transactions on Parallel and Distributed Systems***, vol. 25, No. 12, pp. 3306-3316, 2014.
6. Jiaojiao Jiang, Sheng Wen, Shui Yu, Yang Xiang, Wanlei Zhou, and Ekram Hossain, "Identifying Propagation Sources in Networks: State-of-the-Art and Comparative Studies", Accepted by ***IEEE Communications Surveys and Tutorials***, accepted 17/9/2014.
7. Sheng Wen, Mohammad Sayad Haghighi, Chao Chen, Yang Xiang, Wanlei Zhou, and Weijia Jia, "A Sword with Two Edges: Propagation Studies on Both Positive and Negative Information in Online Social Networks", ***IEEE Transactions on Computers***, Vol. 64, No. 3, pp. 640-653, March 2015.
8. Jiaojiao Jiang, Sheng Wen, Shui Yu, Yang Xiang, and Wanlei Zhou, "K-center: An Approach on the Multi-source Identification of Information Diffusion", ***IEEE Transactions on Information Forensics & Security***, Volume:10, Issue: 12, 2015, pp. 2616-2626.

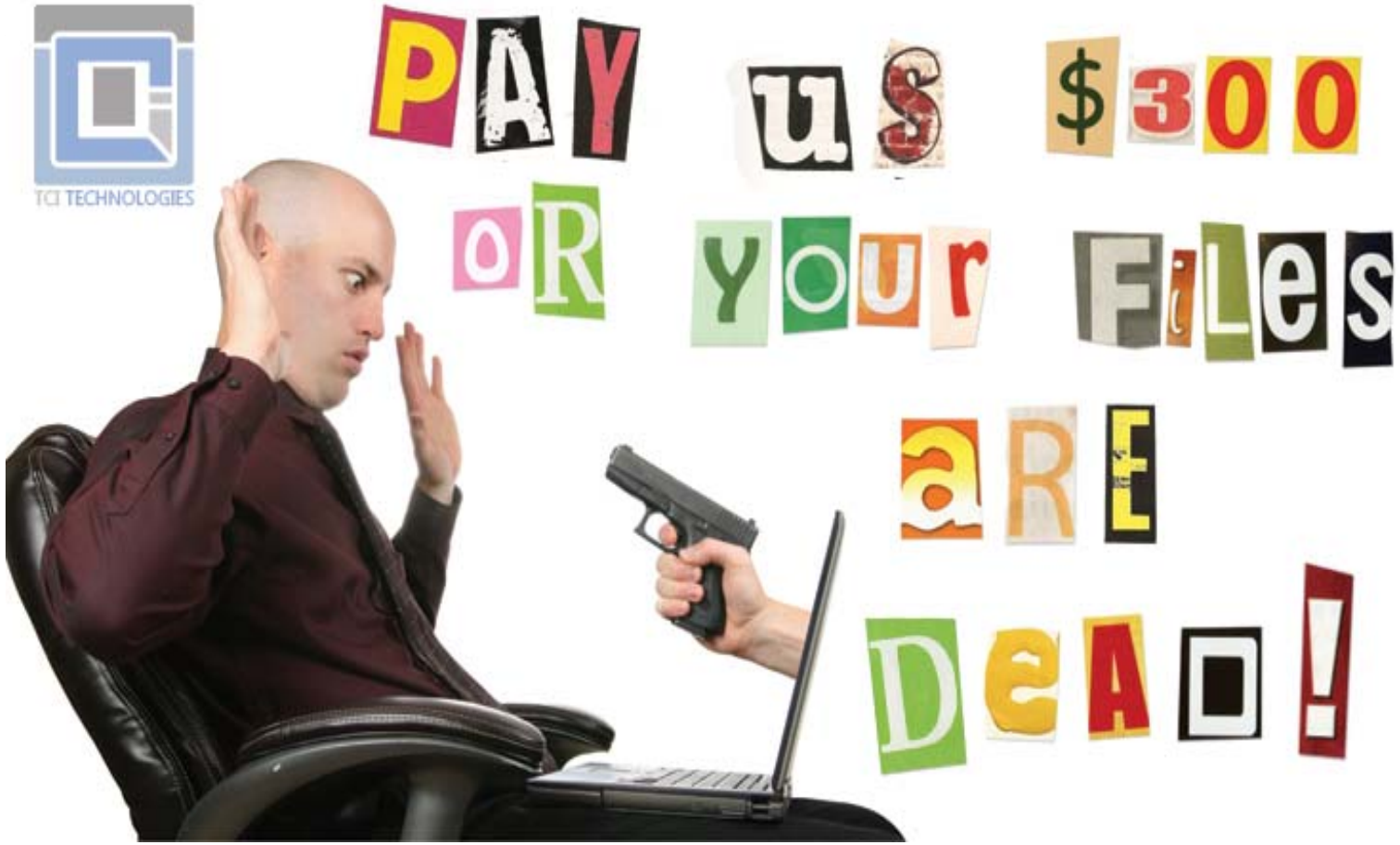
Introduction

- **Background**:
  - We live in a world of ***networks****, e.g., social networks, computer networks, transportation networks, biological networks, etc*.
  - Any complex system can be modeled as a ***network***, where ***nodes*** are the elements of the system and ***edges*** represent the interactions between them.
  - The ***ubiquity*** of networks has made us ***vulnerable*** to various ***network risks***.

Introduction

- A contag on and lead to a
  - For ex rus spread world 10.

Introduction

- 

Introduction

- A ***Social spam*** started by a few individuals can spread quickly through the underlying social network

Introduction

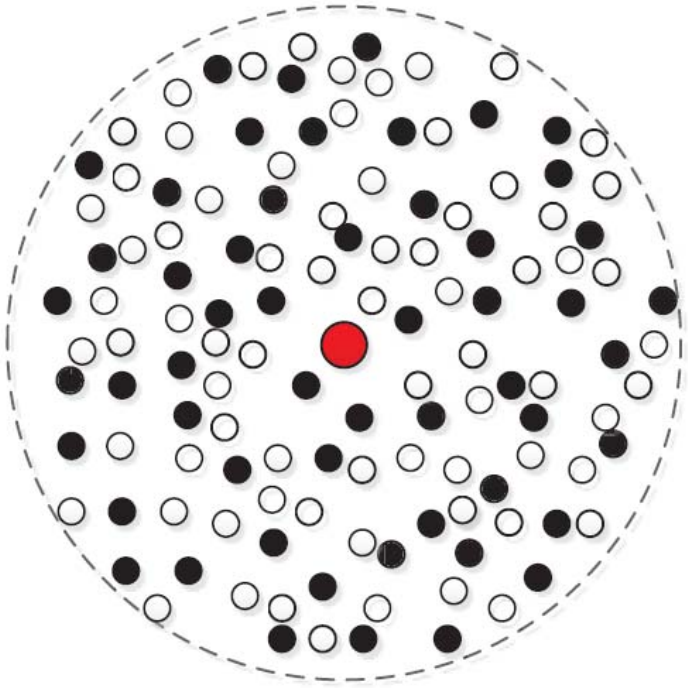
- **Significances of identifying propagation sources and modeling propagation dynamics in networks**:
  - Identifying propagation sources as quickly as possible can help people find the ***causation of risks***, and therefore, ***diminish the damages***.
  - Propagation dynamics modeling can help building better ***defense strategies***.
  - It is important to accurately identify the 'culprit' of the propagation for ***forensic purposes***.
- **Question**:
  - How can we ***identify propagation sources*** in a type of complex networks?

Outline

- Introduction
- Source Identification Methods
- Comparative Studies
- Our Work in Modelling the Propagation of Worms and Rumours
- Future Work & Discussion

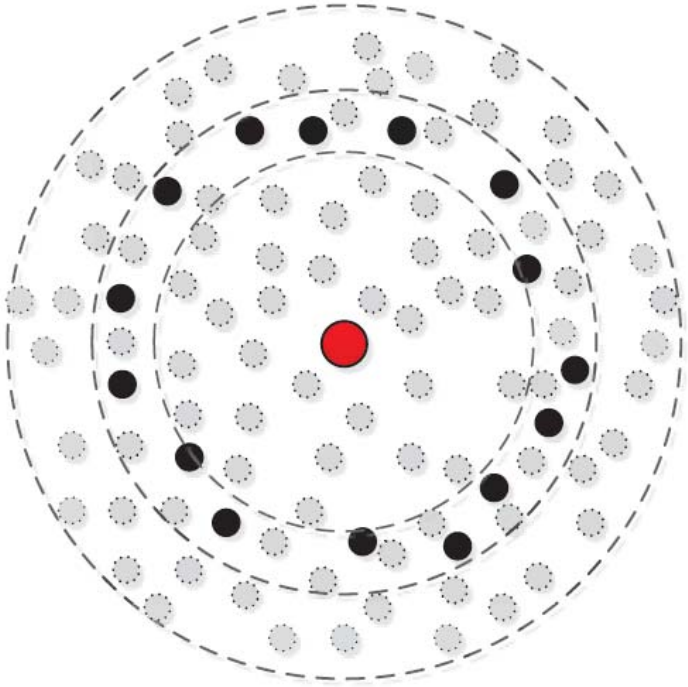
Source Identification Methods

- **Three categories of network observations**:
  - **(i) *Complete Observation***: The exact state of each node is observed in a network.
    - Suitable for ***small-scale*** networks.



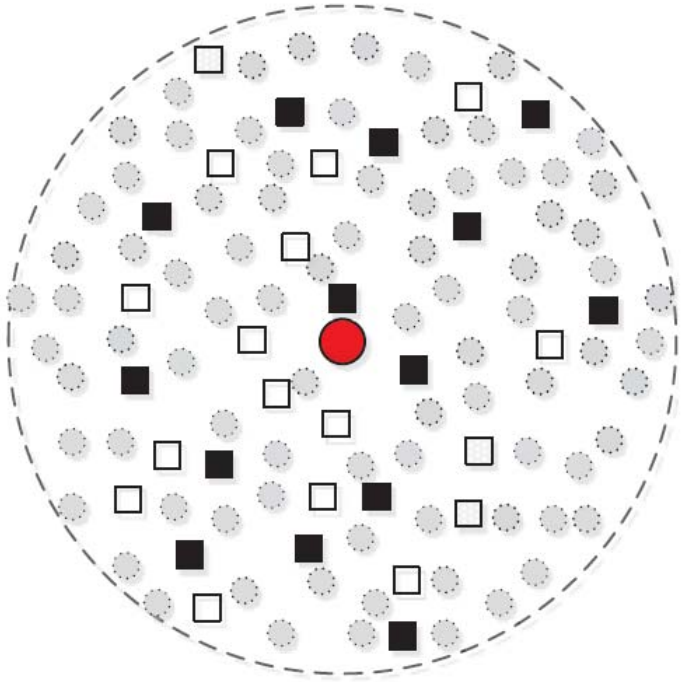
Source Identification Methods

- **(ii) *Snapshot***: A ***portion*** of nodes are observed in a network. It only provides ***partial*** knowledge of a network status.
  - Much more ***usual*** in the real world.



Source Identification Methods

- **(iii) *Sensor Observation***: Sensors record the propagation dynamics over them.
  - Sensors need be ***chosen*** very carefully.

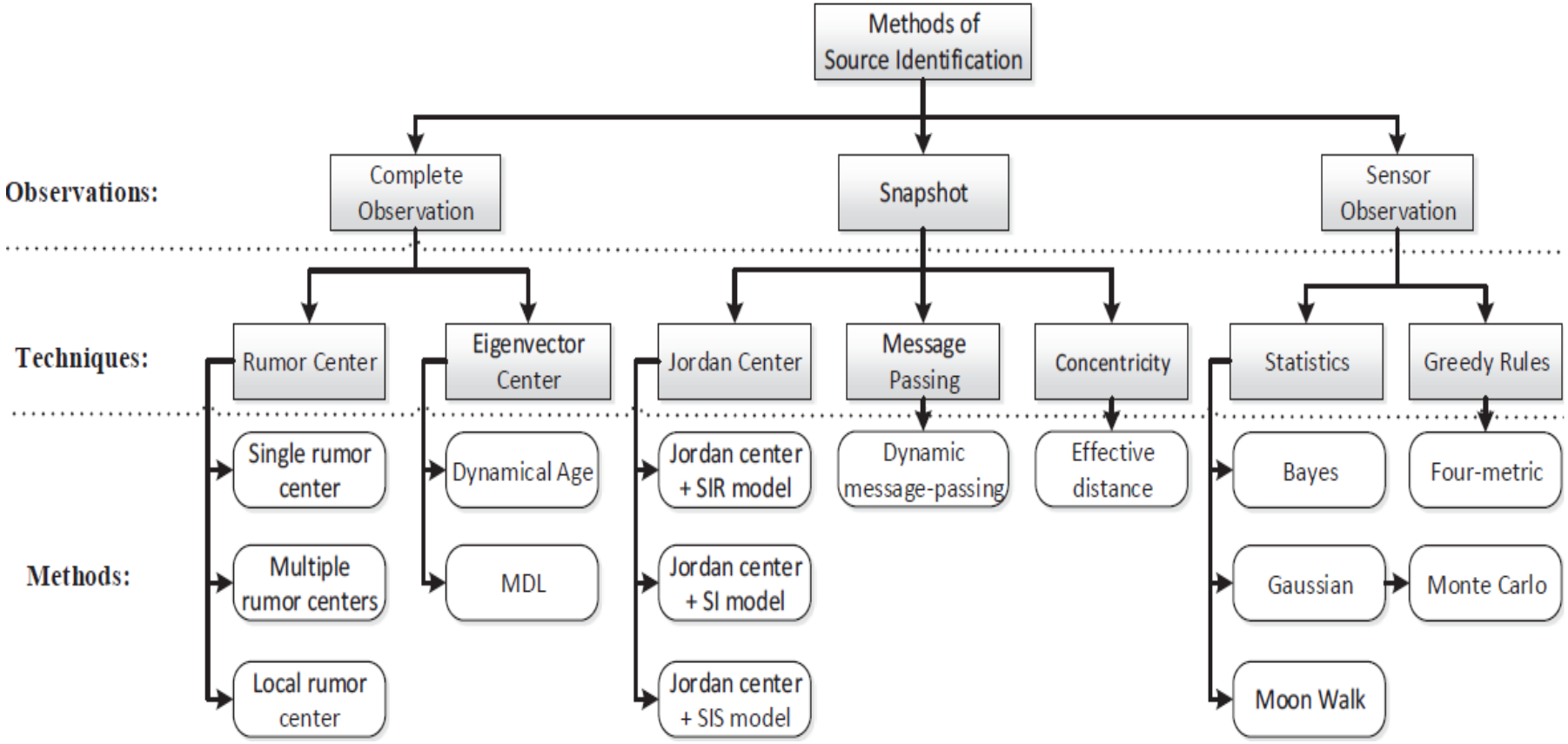


Source Identification Methods

- Epidemic Models
  - *SI* model: In this model, nodes are initially susceptible and can be infected along with the propagation of risks. Once a node is infected, it remains infected forever. This model focuses on the infection process $S \rightarrow I$, regardless of the recovery process.
  - *SIR* model: Recovery processes are considered in this model. Similarly, nodes are initially susceptible and can be infected along with the propagation. Infected nodes can then be recovered, and never become susceptible again. This model deals with the infection and curing process $S \rightarrow I \rightarrow R$.
  - *SIS* model: In this model, infected nodes can become susceptible again after they are cured. This model stands for the infection and recovery process $S \rightarrow I \rightarrow S$.

Source Identification Methods

- Current methods of source identification can be categorized in to



Taxonomy of current source identification methods

Source Identification Methods

- **Rumor-Center Method (*Complete Observation*)**
  - It aims to find a node which is located at the ***center*** of the infection graph.
  - In a tree-like network, the ***rumor center*** is proved to be equivalent to the ***distance center***, i.e., the node have the ***minimum sum of distances*** to all the other infected nodes.

Source Identification Methods

- Single Rumor Center:
  - It assumes that information spreads in tree-like networks and the information propagation follows SI model. Assuming an infected node as the source, its rumor centrality is defined as the number of distinct propagation paths originating from the source. The node with the maximum rumor centrality is called the rumor center.
- Local Rumor Center:
  - Same assumption as before, this method designates a set of nodes as suspicious sources so it reduces the scale of seeking origins. The local rumor center is defined as the node with the highest rumor centrality compared to other suspicious infected nodes. The local rumor center is considered as the source node.
- Multiple Rumor Centers:
  - In addition to the basic assumptions, researchers further assume the maximum number of sources is known for the method of identifying multiple rumor centers. The rumor centrality is extended for a set of nodes, which is defined as the number of distinct propagation paths originating from the set.

Source Identification Methods

- **Eigenvector-based Methods (*Complete Observation*)**
  - Dynamic age: Fioriti et al. claim that the '***oldest***' nodes which are associated to those with ***largest eigenvalues*** of the ***adjacency*** matrix, ***A***, are the propagation sources.
  - The minimum description length (MDL) method: Similarly, Prakash et al. claim that the nodes with the largest score in the ***minimum eigenvector*** of the ***Laplacian*** matrix, ***L = D – A***, refer to the propagation sources.
  - Both methods are considered on generic networks. They assume propagation follows SI model.

Source Identification Methods

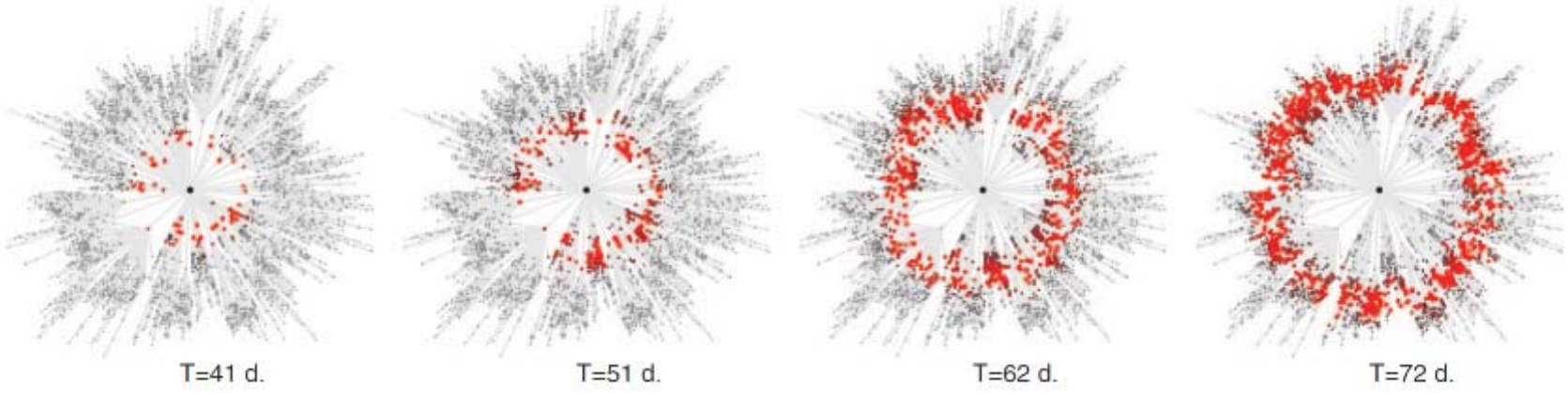
- **Jordan-Center Method (*Snapshot*)**
  - It assumes that information propagates in tree-like networks and the propagation follows SIR, SI or SIS models.
  - It also aims to find a node which located at the ***center*** of the infection graph.
  - It uses a sample path based approach to identify the propagation source. An optimal sample path is the one which most likely leads to the observed snapshot of a network. The source associated with the optimal sample path is proven to be the Jordan center of the infection graph.
  - ***Jordan centers*** have been proved to be the ***graph centers*** that have the ***minimum*** distance to the farthest infected nodes. So Jordan center is considered as a propagation origin.

Source Identification Methods

- **Dynamic Message Passing (*Snapshot*)**
  - The DMP method is based on a set of the dynamic equations and follows SIR model in generic networks. Assuming an arbitrary node as the source node, it first estimates the probabilities of other nodes to be in different states at time t. Then, it multiplies the probabilities of the observed set of nodes being in the observed states. The source node which can obtain the maximum product is considered the propagation origin.
  - The dynamic equations are based on ***maximum likelihood*** techniques and ***discrete*** propagation model.
  - It aims to find a node from which we can obtain the observed snapshot with the ***highest probability***.

Source Identification Methods

- **Effective Distance Based Method (*Snapshot*)**
  - This method assumes that propagation follow SI model in weighted networks. It first proposes the effective distance concept to represent the propagation process. According to the propagation process of wavefronts, the spreading concentricity can only be observed from the perspective of the true source. Then, the node, which has the minimum standard deviation and mean of effective distances to the nodes in the observed wavefront, is considered as the source node.
  - The ***complex spatiotemporal propagation*** can be reduced to simple, homogeneous ***wavefront patterns*** by using effective distance.
  - According to the propagation process of ***wavefronts***, the ***spreading concentricity*** can only be observed from the perspective of the ***source***.



Wavefront propagation patterns

Source Identification Methods

- **Gaussian Method (*Sensor Observation*)**
  - The Gaussian method for single source identification assumes propagation follow SI model in tree-like networks.
  - Every edge has a '***deterministic delay***', and each sensor has an '***observed delay***'.
  - It reduces the scale of seeking origins through a uniquely determined subtree from the direction in which information arrived at the sensors, and is guaranteed to contain the propagation origin. Then it uses the Gaussian technique to seek the source in the subtree: It calculates the 'observed delay' between a sensor and the other sensors and then calculates the 'deterministic delay' for every sensor node relative to a sensor node.
  - The node, which can ***minimize*** the differences between the 'observed delays' and the 'deterministic delays' of sensor nodes, is considered as the propagation source.
- **Monte-Carlo Method**
  - The Gaussian method is extended to the ***Monte-Carlo*** method from ***tree-like*** networks to ***generic networks***.

Outline

- Introduction
- Source Identification Methods
- Comparative Studies
- Our Work in Modelling the Propagation of Worms and Rumours
- Future Work & Discussion

Comparative Studies

- **Crosswise comparison**
  - We conducted experiments on (1) synthetic networks: a regular tree and a small-world network; and (2) two real-world networks.
  - Comparison from ***seven features***.
    - Most of current methods are based on ***tree*** networks.
    - Most of them can only be used to identify ***single*** source.
    - Most of them are very ***computationally complex***.

TABLE I
SUMMARY OF CURRENT SOURCE IDENTIFICATION METHODS.

	Topology	Observation	Model	Number of Sources	Infection Probability	Time Delay	Complexity
Single rumor center	Tree	Complete	SI	Single	HM/HT	Constant	$O(N^2)$
Local rumor center	Tree	Complete	SI	Single	HM	Constant	$O(N^2)$
Multi rumor centers	Tree	Complete	SI	Multiple	HM	Constant	$O(N^k)$
Eigenvector center	Generic	Complete	SI	Multiple	HM	Constant	$O(N^3)$
Jordan center	Tree	Snapshot	SI(R/S)	Single	HM/HT	Constant	$O(N^3)$
DMP	Generic	Snapshot	SIR	Single	HT	Constant	$O(t_0N^2d)$
Effective distance	Generic	Snapshot	SI	Single	HT	Constant	$O(N^3)$
Gaussian	Tree	Sensor	SI	Single	HT	Variable	$O(N^3)$
Monte Carlo	Generic	Sensor	SIR	Single	HT	Variable	$O(N\log N/\varepsilon^2)$
Four-metrics	Generic	Sensor	SI	Single	HT	Variable	$O(N^3)$

Infection Probability: HM stands for homogeneous; HT stands for heterogeneous.

Comparative Studies

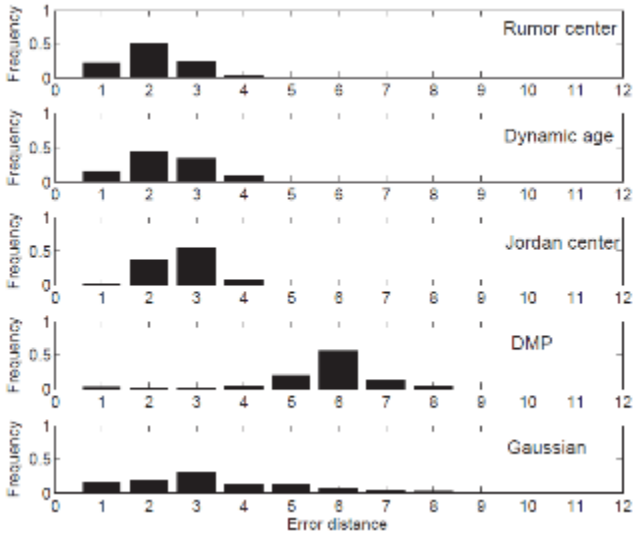
- Crosswise comparison of ***typical methods*** in two ***real networks***
  - The first one is an Enron email network. This network has 143 nodes and 1,246 edges. On average, each node has 8.71 edges. Therefore, the Enron email network is a dense network.
  - The second is a power grid network. This network has 4,941 nodes and 6,594 edges. On average, each node has 1.33 edges. Therefore, the power grid network is a sparse network.
  - Sample topologies of these two real-world networks are shown here

(A) Eron email network

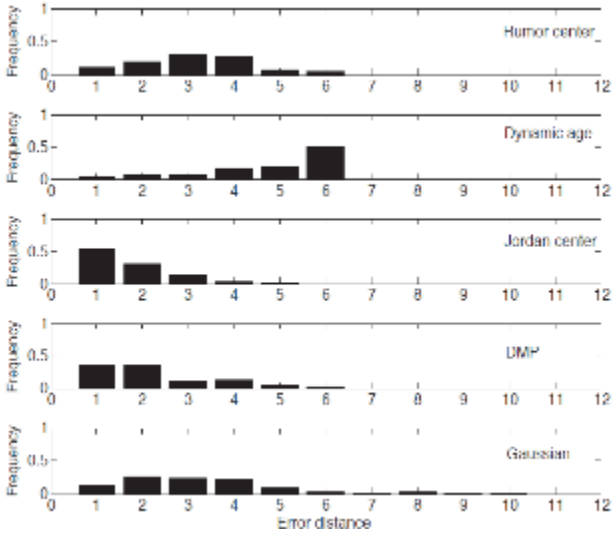
(B) Power grid network

Comparative Studies

- We randomly choose a node as a source to initiate a propagation, and then average the error distance between the estimated sources and the true sources by 100 runs to estimate the accuracy of source identification.
- Evaluated by ***error distance***: the number of ***hops*** between the real sources and the estimated sources.
- The performances ***vary*** in different networks. Therefore, we are inspired to investigate the ***factors*** which may impact source identification methods.



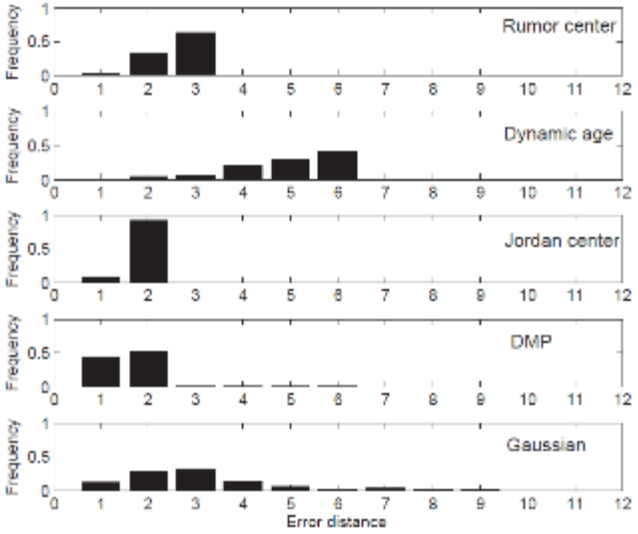
Source identification methods applied on an Enron email network



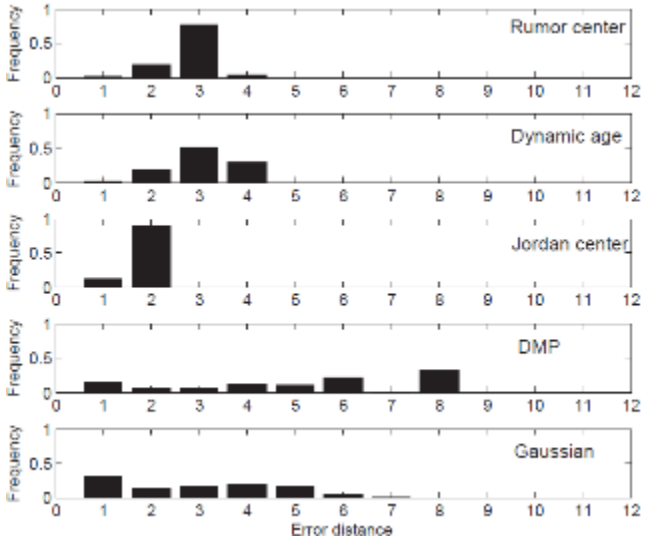
Source identification methods in a power grid network

Comparative Studies

- **Factors that may impact source identification methods**
  - **(i) *Network Topologies***
    - *Regular tree* and *Random tree.*



Source identification methods applied on a 4-regular tree

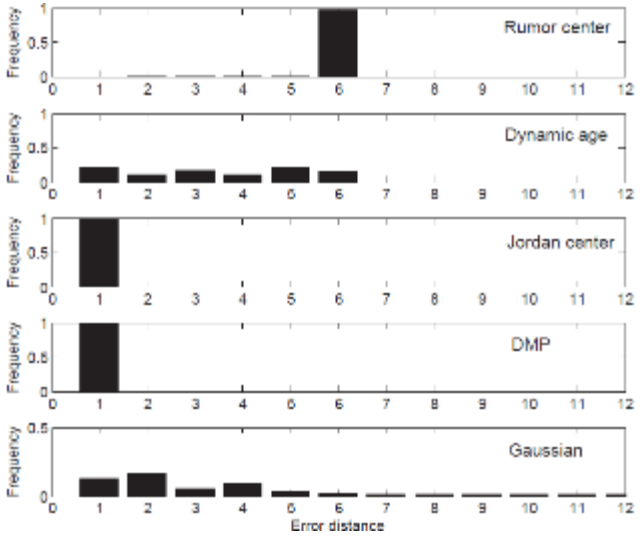


Source identification methods applied on a random tree

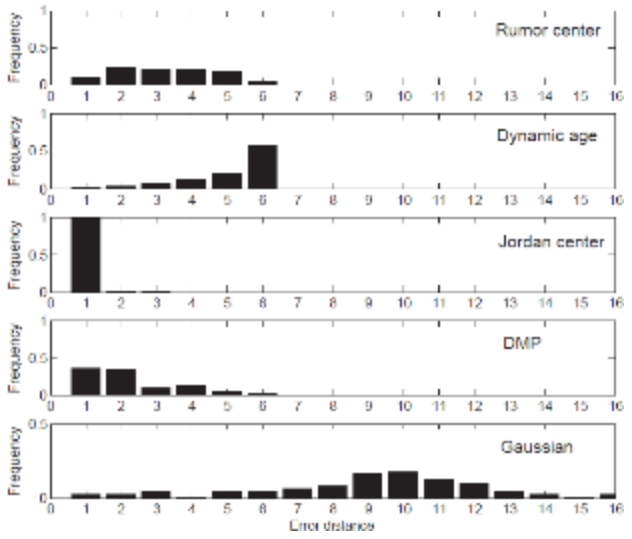
Comparative Studies

- **(i) *Network Topologies (cont'd)***
  - *Regular networks* and *Small-world networks*

Conclusion: Current methods are ***sensitive*** to network topologies.



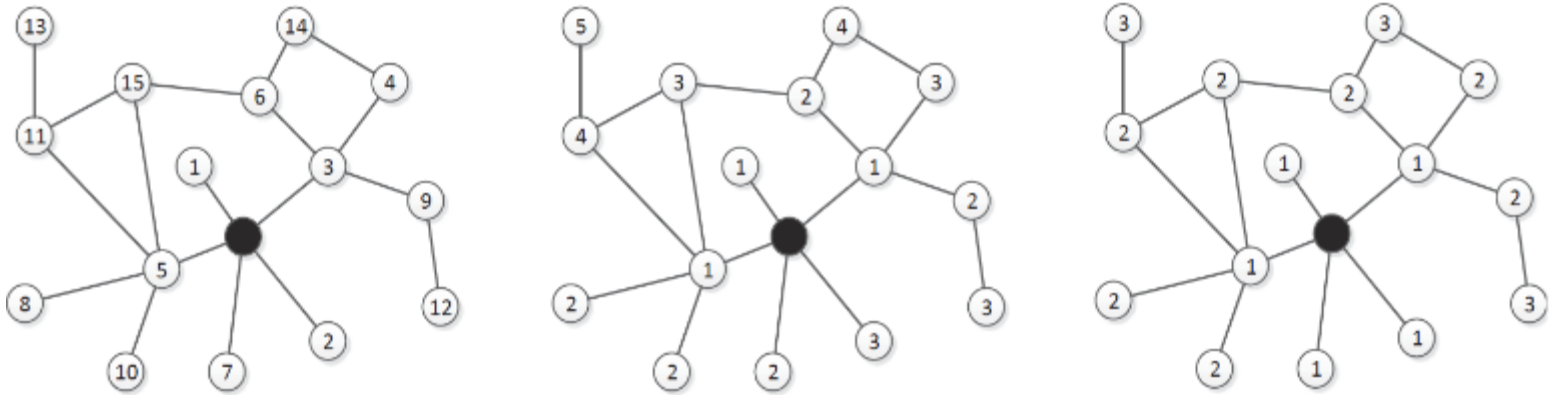
Source identification methods applied on a 4-regular graph



Source identification methods applied on a small-world network

Comparative Studies

- **(ii) *Propagation Schemes***
  - *Random walk: A node can deliver a message randomly to one of its neighbors*
  - *Contact-process (unicast): A node can deliver a message to a group of its neighbors that have expressed interest in receiving the message*
  - *Snowball propagation schemes (broadcast): A node can deliver a message to all of its neighbors*



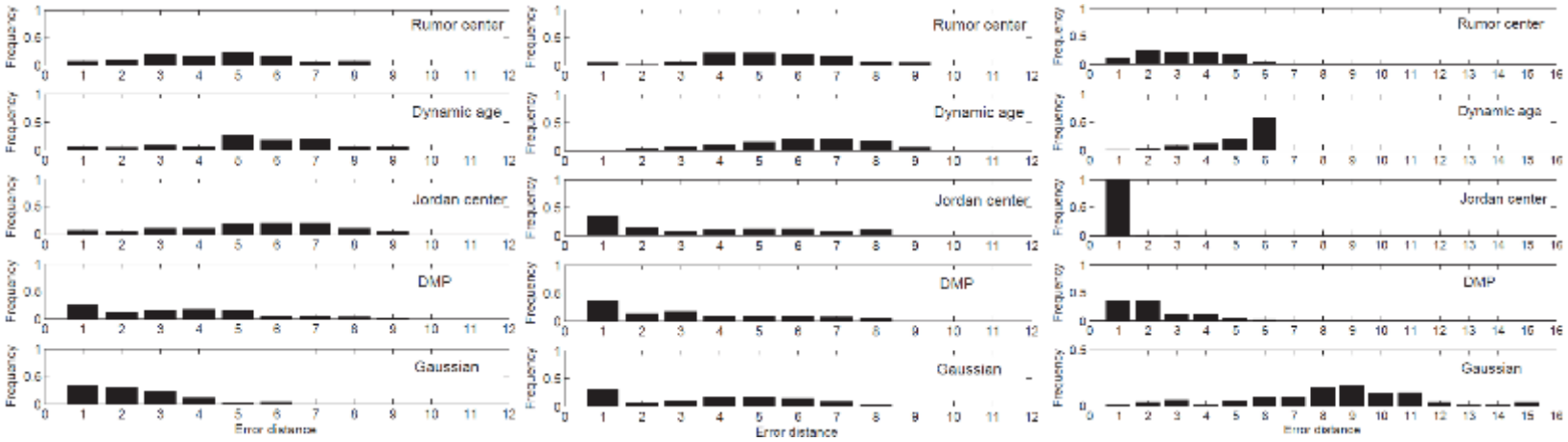
(A) Random walk (B) Contact process (C) Snowball

Illustration of different propagation schemes. The black node stands for the source. The numbers indicate the hierarchical sequence of nodes getting infected

Comparative Studies

- **(ii) *Propagation Schemes (tested on both 4-regular tree and small-world network)***

  Conclusion: Current methods are ***sensitive*** to different propagation schemes.

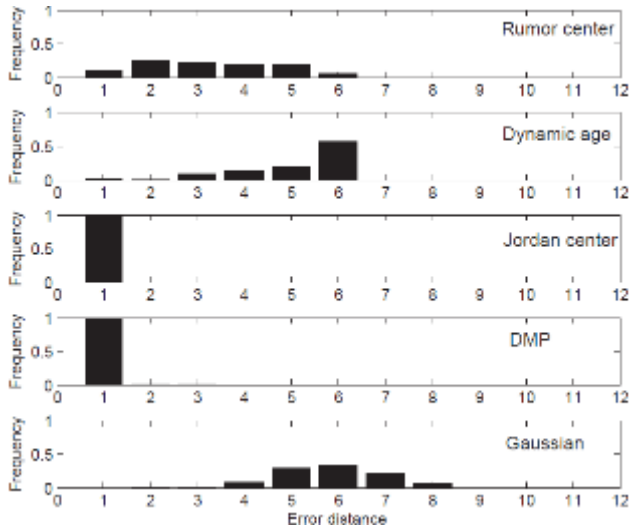


Random-walk Contact-process Snowball

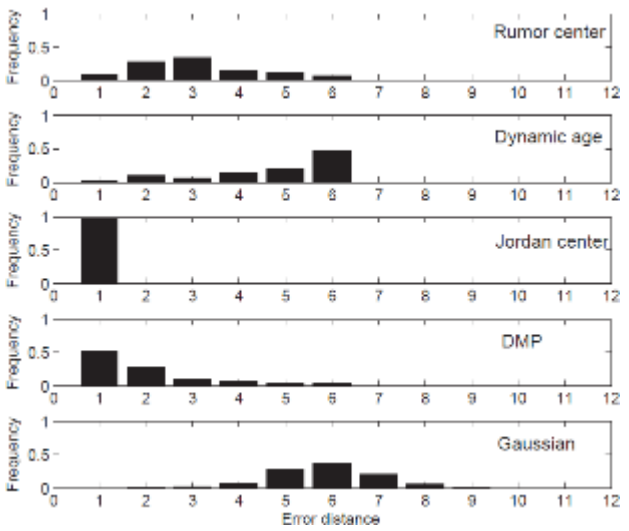
Comparative Studies

- **(iii) *Infection Probabilities (q = 0.5 and q = 0.95)***

Conclusion: Current methods are ***not very sensitive*** to infection probabilities.



q = 0.5



q = 0.95

Outline

- Introduction
- Source Identification Methods
- Comparative Studies
- Our Work in Modelling the Propagation of Worms and Rumours
- Future Work & Discussion

Outline

- Introduction
- Source Identification Methods
- Comparative Studies
- Our Work in Modelling the Propagation of Worms and Rumours
- Future Work & Discussion

Our Work in Modelling the Propagation of Worms and Rumours

- Modelling the Propagation of Scanning Worms
- Modelling the Propagation of Online Social Network Worms
- Modelling the Propagation of Rumours and Truths in Online Social Networks
- Defense Measures in Online Social Networks

Modelling the Propagation of Scanning Worms

- Yini Wang, Sheng Wen, Yang Xiang, and Wanlei Zhou, "Modeling the Propagation of Worms in Networks: A Survey", ***IEEE Communications Surveys and Tutorials***, Volume:16, Issue: 2, 2014, pp 942-960.
- Yini Wang, Sheng Wen, Silvio Cesare, Wanlei Zhou, and Yang Xiang, "Eliminating Errors in Worm Propagation Models", ***IEEE Communication Letters***, VOL. 15, NO. 9, pp. 1022-1024, SEPTEMBER 2011.
- Sheng Wen, Wei Zhou, Yini Wang, Wanlei Zhou, and Yang Xiang, "Locating Defense Positions for Thwarting the Propagation of Topological Worms", ***IEEE Communication Letters***, VOL. 16, NO. 4, pp. 560-563, APRIL 2012.
- Yini Wang, Sheng Wen, Silvio Cesare, Wanlei Zhou and Yang Xiang, "The Microcosmic Model of Worm Propagation", ***The Computer Journal***, Vol. 54 No. 10, pp. 1700-1720, 2011

The Microcosmic Model

Our model has several important components:

Propagation Matrix (PM)
Propagation Source Vector (S)
Vulnerable Distribution Vector (V)
Patching Strategy Vector (Q)
Propagation Ability (PA)

The major contributions are as follows:

> We create a microcosmic landscape on scanning worm propagation and successfully provide useful information for the proposed problems of where, when and how many nodes do we need to patch.

> Associated with S,V,Q, our model can also help us to evaluate:
  > The mutual effect of initial infectious states and patch strategies.
  > The impact of different distributions of vulnerable hosts.

> We introduce a complex matrix to represent the probabilities and time delay. This extension matches the real case well.

Mathematical Presentation

Propagation Matrix (PM):

We use an *n* by *n* square complex matrix PM with elements c_{xy} to describe a network consisting of n peers.

$$\mathrm{PM} = \begin{bmatrix} c_{11} & \cdots & \cdots \\ \cdots & c_{xy} & \cdots \\ \cdots & \cdots & c_{nn} \end{bmatrix}_{n \times n},$$

$$c_{xy} = p_{xy} + d_{xy} i, \quad c_{xy} = 0 (x = y),$$

$$\mathrm{Re}(c_{xy}) = p_{xy} = p(N_y | N_x) \quad p_{xy} \in [0, 1],$$

$$\mathrm{Im}(c_{xy}) = d_{xy} = t(N_x, N_y) \quad d_{xy} \in (0, 1].$$

The real component of each element *cxy represents the* propagation probability of the worm spreading from node *x* to node *y under the condition that node x is infected.* The imaginary component represents the propagation delay from node x to node y. The calculation rules of complex can reflect the mutual impact between the propagation probability and time delay

> Propagation Probability (p_{xy})

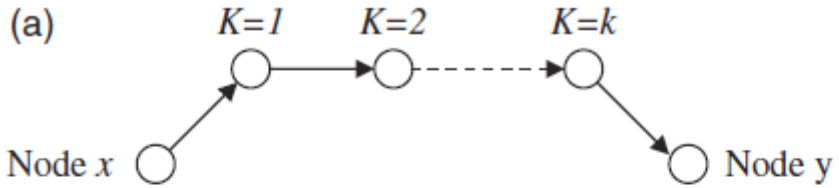
> Propagation Delay (d_{xy})

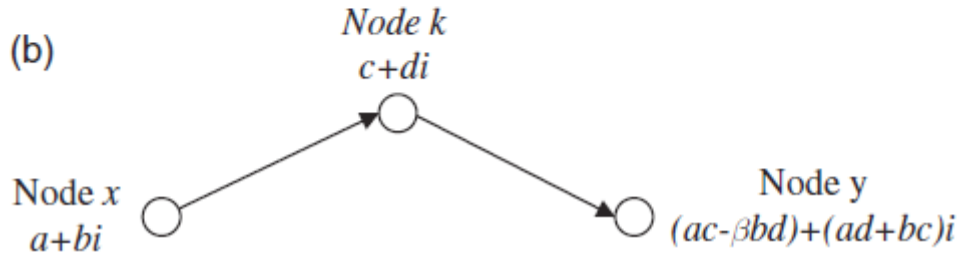
> Maximum time cost of scanning the entire IP address space or the hit-list

Mathematical Presentation

Propagation Function (γ):

Worm propagation from node x (N_x) to node y (N_y) is via and only via k intermediate nodes in a network consisting of n peers.





So that after k steps:

$$\gamma^0(\mathrm{PM}) = \mathrm{PM},$$
$$\gamma^k(\mathrm{PM}) = \underbrace{\mathrm{PM} \times \mathrm{PM} \times \cdots \times \mathrm{PM}}_{k+1}.$$

For each $c_{xy}^{(k)}$:

$$c_{xy}^{(k)} = \sum_{m=1}^{m=n,m\neq x} c_{xm}^{(k-1)} c_{my}$$
$$= \sum_{m=1}^{m=n,m\neq x} ((p_{xm}^{(k-1)} p_{my} - \beta t_{xm}^{(k-1)} t_{my}) + (p_{xm}^{(k-1)} t_{my} + t_{xm}^{(k-1)} p_{my})i),$$
$$k \in [1, n-2], x = 1, \ldots, n, \; y = 1, \ldots, n.$$

> Impact Factor (β): describe the decrease in the propagation probability caused by time delay.

Key Factors: infectious state, vulnerability distribution and patch strategy

Propagation Source Vector (S):

An infectious peer that can propagate worms is represented with a probability of one. The probability of zero means that a peer is healthy and does not have the ability to propagate the worm.

$$S = [s_1, s_2, \ldots, s_x, \ldots, s_n]^{\mathrm{T}}, s_x = 0 \text{ or } 1, x = 1, \ldots, n.$$

Then the iteration procedure can be represented as function γ_s:

$$\gamma_s^0(\mathrm{PM}) = S \,\&_{\mathrm{L}}\, \mathrm{PM},$$
$$\gamma_s^k(\mathrm{PM}) = \gamma_s^{k-1}(\mathrm{PM}) \times \mathrm{PM}$$
$$= (S \,\&_{\mathrm{L}}\, \mathrm{PM}) \times \underbrace{\mathrm{PM} \times \cdots \times \mathrm{PM}}_{k} \; (k \geq 1).$$

Wherein:

$$A \&_{\mathrm{L}} B = \begin{bmatrix} a_1 \\ \cdots \\ a_n \end{bmatrix} \&_{\mathrm{L}} \begin{bmatrix} b_{11} & \cdots & \cdots \\ \cdots & b_{xy} & \cdots \\ \cdots & \cdots & b_{nn} \end{bmatrix}$$
$$= \begin{bmatrix} a_1 \times b_{11} & \cdots & a_1 \times b_{1n} \\ \cdots & a_x \times b_{xy} & \cdots \\ a_n \times b_{n1} & \cdots & a_n \times b_{nn} \end{bmatrix}.$$

During the propagation process, each intermediate node can be infected and become infectious, so we introduce an infected state vector I:

$$ie_x = \sum_y S_y p_{yx}^{(k)} + \frac{\sum_y S_y p_{yx}^{(k)} t_{yx}^{(k)}}{\sum_y S_y p_{yx}^{(k)}} i.$$ ←defined as--

$$I = [ie_1, ie_2, \ldots, ie_x, \ldots, ie_n]^{\mathrm{T}},$$
$$ie_x = 0 \text{ or } 1, \quad x = 1, \ldots, n,$$
$$I_s^{(k)} = \Gamma(S^{\mathrm{T}}, \mathrm{PM}_s^{(k)}) \quad (k \geq 0),$$

Key Factors

Vulnerable Distribution Vector (V):

For an element in V , the value of one represents that a peer is vulnerable. Zero means that the peer is healthy and is not vulnerable.

$$V = [v_1, v_2, \ldots, v_x, \ldots, v_n]^{\mathrm{T}}, \quad v_x = 0 \text{ or } 1, \quad x = 1, \ldots, n.$$

Then the iteration procedure can be represented as function γ_{SV} :

$$\gamma_{sv}^{0}(\mathrm{PM}) = S \;\&_{\mathrm{L}}\; \mathrm{PM} \;\&_{\mathrm{R}}\; V^{\mathrm{T}},$$

$$\gamma_{sv}^{k}(\mathrm{PM}) = \gamma_{sv}^{k-1}(\mathrm{PM}) \times (V \;\&_{\mathrm{L}} \mathrm{PM} \;\&_{\mathrm{R}}\; V^{T}) \quad (k \geq 1).$$

Wherein:

$$B \&_{\mathrm{R}} A = \begin{bmatrix} b_{11} & \cdots & \cdots \\ \cdots & b_{xy} & \cdots \\ \cdots & \cdots & b_{nn} \end{bmatrix} \&_{\mathrm{R}} \begin{bmatrix} a_1 & \cdots & a_n \end{bmatrix}$$

$$= \begin{bmatrix} b_{11} \times a_1 & \cdots & b_{1n} \times a_n \\ \cdots & b_{xy} \times a_y & \cdots \\ b_{n1} \times a_1 & \cdots & b_{nn} \times a_n \end{bmatrix}.$$

The PM and infected probability vector I can be represented by :

$$\mathrm{PM}_{sv}^{(k)} = \gamma_{sv}^{k}(\mathrm{PM}) \quad (k \geq 0),$$

$$I_{sv}^{(k)} = \Gamma(S^{\mathrm{T}}, \mathrm{PM}_{sv}^{(k)}) \quad (k \geq 0).$$

Key Factors

Patch strategy vector (*Q*):

For an element in *Q* , the value of one represents that a peer has been patched and becomes a healthy node. Zero means that a peer is still vulnerable.

$$Q = [q_1, q_2, \ldots, q_x, \ldots, q_n]^{\mathrm{T}}, q_x = 0 \text{ or } 1, \quad x = 1, \ldots, n.$$

Then the iteration procedure can be represented as function γ_{SVQ} :

$$Q' = V \underline{\&} Q,$$

$$\gamma_{svq}^{0}(\mathrm{PM}) = S \,\&_{\mathrm{L}}\, \mathrm{PM} \,\&_{\mathrm{R}}\, Q'^{\mathrm{T}},$$

$$\gamma_{svq}^{k}(\mathrm{PM}) = \gamma_{svq}^{k-1}(\mathrm{PM}) \times (Q' \,\&_{\mathrm{L}}\, \mathrm{PM} \,\&_{\mathrm{R}}\, Q'^{\mathrm{T}}) \quad (k \geq 1).$$

The PM and infected probability vector *I* can be represented by :

$$\mathrm{PM}_{svq}^{(k)} = \gamma_{svq}^{k}(\mathrm{PM}) \quad (k \geq 0),$$

$$I_{svq}^{(k)} = \Gamma(S^{\mathrm{T}}, \mathrm{PM}_{svq}^{(k)}) \quad (k \geq 0).$$

The Contributions

We tested our model using a typical local preference worm: Code Red II and implemented a series of experiments to evaluate the effects of each major component in our microcosmic model.

Through the microcosmic model we are able to understand the detailed propagation process of scanning worms at the node-to-node level so as to enable us to develop effective countermeasures.

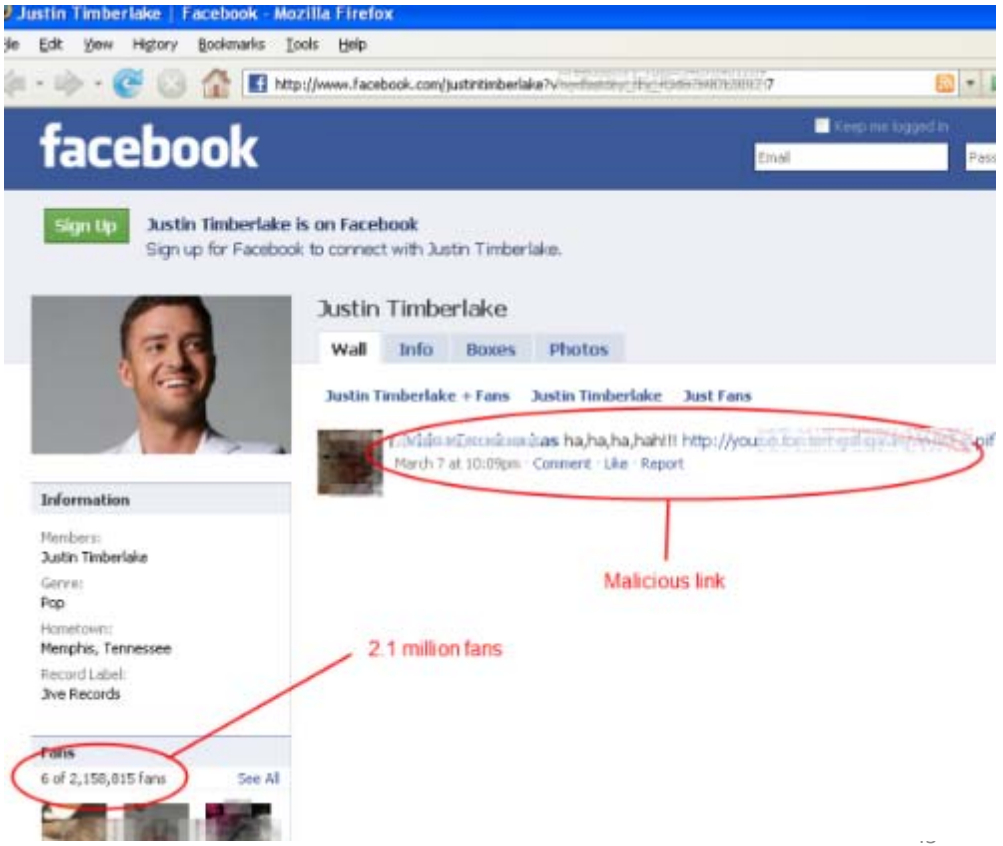
Based on the results drawn from the experiments, for high-risk vulnerabilities, it is critical that networks reduce the number of vulnerable nodes to below 80%. We believe our microcosmic model can benefit the security industry by allowing them to save significant money in the deployment of their security patching schemes.

Modelling the Propagation of Online Social Network Worms

- Sheng Wen, Wei Zhou, Jun Zhang, Yang Xiang, Wanlei Zhou, and Weijia Jia, "Modeling Propagation Dynamics of Social Network Worms", ***IEEE Transactions on Parallel and Distributed Systems***, vol. 24, no. 8, pp. 1633-1643, Aug. 2013.
- Silvio Cesare, Yang Xiang, and Wanlei Zhou, "Malwise - An Effective and Efficient Classification System for Packed and Polymorphic Malware", ***IEEE Transactions on Computers***, Vol. 62, Issue 6. pp.1193-1206, June 2013
- Sheng Wen, Wei Zhou, Jun Zhang, Yang Xiang, Wanlei Zhou, Weijia Jia, and Cliff C.Zou "Modeling and Analysis on the Propagation Dynamics of Modern Email Malware", ***IEEE Transactions on Dependable and Secure Computing***, VOL. 11, NO. 4, JULY/AUGUST 2014, pp. 361-374.

What cause Epidemics in OSN?

- People share news, interests and ideas in OSNs,
- **but** these platforms also spread email malware, rumors and malicious links.



How Viruses Spread in OSNs?

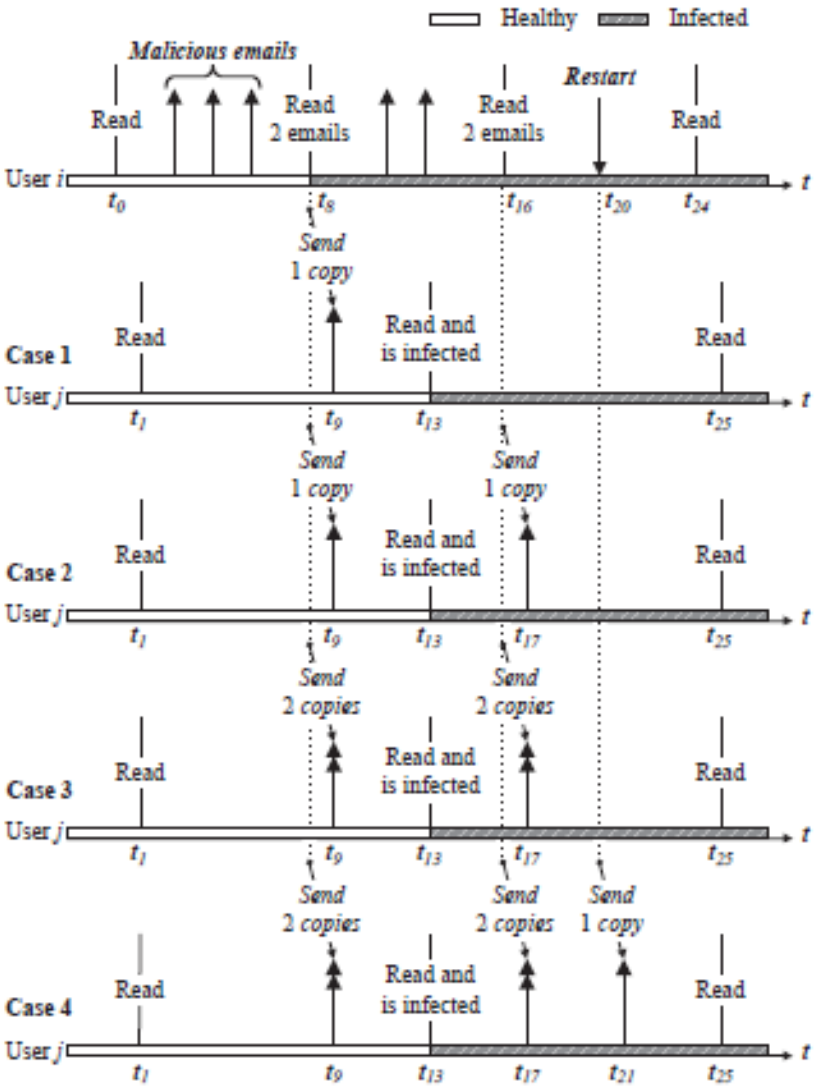


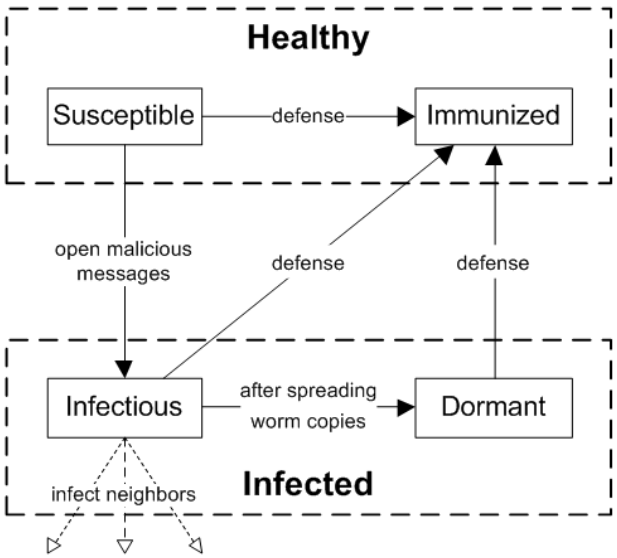
Fig. 1: Recipient user j's behavior for different types of malware emails. User i reads two of three malware emails at t_8 and another two malware emails at t_{16}, and then restarts at t_{20}. Case 1: nonreinfection; Case 2: reinfection in the work [3]; Case 3: reinfection of modern email malware; Case 4: both self-start mechanism in modern email malware. We assume a user will visit the malicious hyperlink or attachment if the user reads emails in this figure.

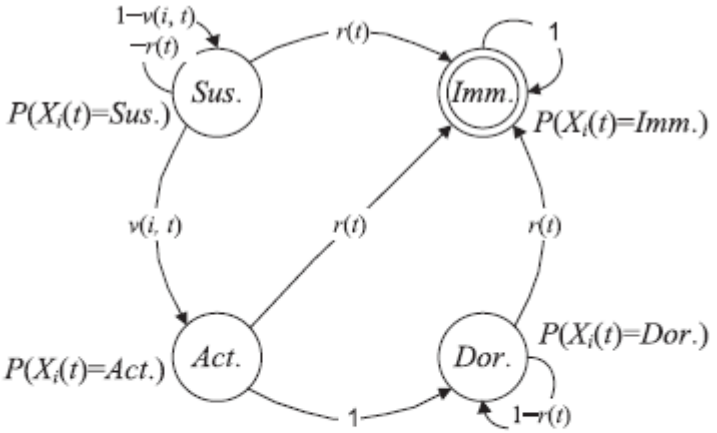
Important Features of OSN Worm

- **Feature 1**: The propagation of social network worms is triggered by user clicking or reading malicious hyperlinks or messages. *Users have different patterns to visit these malicious things*. So in the previous k-hops modeling where the worms spread from the origins to their neighbors, and then to the neighbors' neighbors in a hop-by-hop pattern. This may cause a problem.
  Solution: different patterns of users in the modelling (temporal dynamics).
- **Feature 2**: Social network worms, such as email worms, search new targets in the contact lists of compromised computers. This means *social network worms can only spread to topological neighbors in social networks*.
  Solution: consider topological neighbors in the modelling (spatial dependence).
- **Feature 3**: Users can be infected and send out worms copies again even though they have been infected before. Furthermore, *infected users can send out worm copies when compromised computers restart or certain events are triggered*.
  Solution: Implementing reinfection and self-start.

The susceptible-infectious-immunized (SII) model

State Transition Graph





State transition graph of a node in the topology.

$$X_i(t) = \begin{cases} \boldsymbol{Hea.}, \text{healthy} \begin{cases} \boldsymbol{Sus.}, \text{susceptible}, \\ \boldsymbol{Imm.}, \text{immunizied}, \end{cases} \\ \boldsymbol{Inf.}, \text{infected} \begin{cases} \boldsymbol{Act.}, \text{active}, \\ \boldsymbol{Dor.}, \text{ dormant}. \end{cases} \end{cases}$$

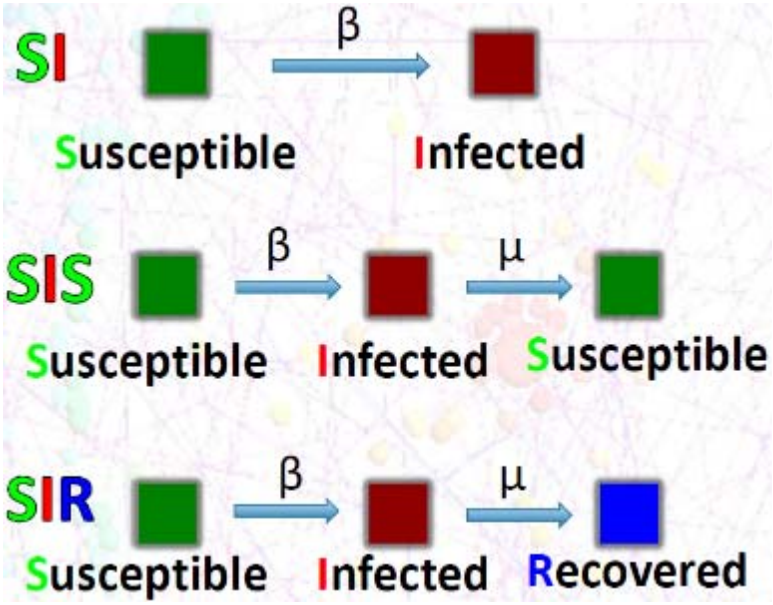
46

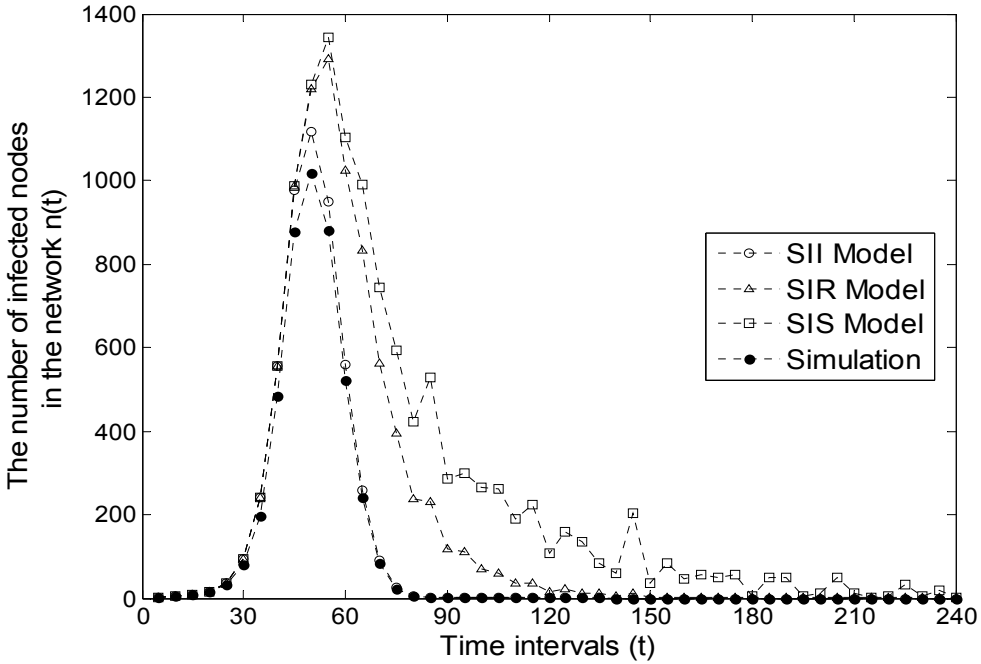
Comparison of SIS, SIR, and SII models

- The difference among these models is caused by different considerations on the state transition of nodes.
- SIS models assume infected nodes become susceptible again after recovery.
- If infected nodes cannot become susceptible again once they are cured, the models are called SIR models.
- Considering the propagation of modern email malware, after users clean their infected computers or become more vigilant against a type of malware, they are unlikely to be infected any more. Therefore, SIS models are not appropriate to model the propagation of modern email malware.
- SIR models may suit for modern email malware, but the real case is that a susceptible user can be immunized directly without being infected at first.
- Thus, the state transition of our SII model is similar to SIR model except nodes at the susceptible state can directly transit to the immunized state..

Results Exhibition of the Modelling via Simulation

The traditional epidemic models have three categories: SI, SIS, SIR.





Comparison Result: Our modelling result is more close to the simulations. Our susceptible-infectious-immunized (SII) model can precisely present the propagation / dynamics of social network worms.

Modelling the Propagation of Rumours and Truths in Online Social Networks

- Sheng Wen, Jiaojiao Jiang, Yang Xiang, Shui Yu, and Wanlei Zhou, "Are the Popular Users Always Important for the Information Dissemination in Online Social Networks?" ***IEEE Network***, pp. 64-67, September/October 2014.
- Sheng Wen, Jiaojiao Jiang, Yang Xiang, Shui Yu, Wanlei Zhou, Weijia Jia, “To Shut Them Up or to Clarify: Restraining the Spread of Rumours in Online Social Networks”, ***IEEE Transactions on Parallel and Distributed Systems***, vol. 25, No. 12, pp. 3306-3316, 2014.
- Sheng Wen, Mohammad Sayad Haghighi, Chao Chen, Yang Xiang, Wanlei Zhou, and Weijia Jia, "A Sword with Two Edges: Propagation Studies on Both Positive and Negative Information in Online Social Networks", ***IEEE Transactions on Computers***, Vol. 64, No. 3, pp. 640-653, March 2015.
- Jiaojiao Jiang, Sheng Wen, Shui Yu, Yang Xiang, and Wanlei Zhou, "K-center: An Approach on the Multi-source Identification of Information Diffusion", ***IEEE Transactions on Information Forensics & Security***, Volume:10, Issue: 12, pp. 2616-2626.
- Jiaojiao Jiang, Sheng Wen, Shui Yu, Yang Xiang, Wanlei Zhou, and Ekram Hossain, "Identifying Propagation Sources in Networks: State-of-the-Art and Comparative Studies", Accepted by ***IEEE Communications Surveys and Tutorials***, accepted 17/9/2014.

How Rumours and Truths Spread in OSNs

Case 1 (23/04/13): Market chaos after fake Obama explosion tweet

- NEGATIVE INFORMATION

Hackers took control of the Associated Press Twitter account overnight and sent a false tweet about two explosions at the White House that briefly sent US financial markets reeling.

- POSITIVE INFORMATION

The White House quickly issued a statement saying the report of explosions was not correct.

Negative Results

Reuters data showed the tweet briefly wiped out $US136.5 billion of the S&P 500 index's value before markets recovered.

Technical Perspective (optimistic choices)

The probabilities of people believing positive information (a) and negative information (b), we have $0<a<1$, $b=0$. We let $a+b<1$ since they can contradict both kinds of information like "there was an explosion in White House but Obama was not injured".

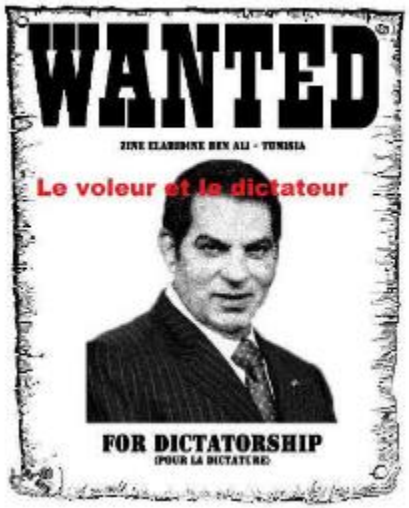
> Ops! @ap get owned by Syrian Electronic Army! #SEA #Syria #ByeByeObama twitter.com/Official_SEA6/…
> — SyrianElectronicArmy (@Official_SEA6) April 23, 2013

How Rumours and Truths Spread in OSNs

Case 2 (11/01/11): Coup in Tunisia - the fall of president Ben Ali

Negative Information

A rumor from tweeters went round that the army has seized power and ousted the Tunisia president.



Positive Information

The coup story was later suggested to be untrue by Egyptian Chronicles etc.

Unconfirmed News :Coup D'état in Tunisia " Updated"

~~I have just got confirmation from Tunisia that President Ben Ali of Tunisia has been overthrown by the Tunisian army~~
~~More and more to come~~
~~General Ammar is officially ruling the country~~
~~Nothing yet on TV channels at all or any other mainstream media , you read here first ;)~~
Update#1
~~The terrestrial TV channel in Tunisia has announced General Ammar as the temporary president for the country. Ammar promised to have an election with in a year.~~

Technical Perspective (preferable choices or Minority being subordinate to majority)

The probabilities of people believing positive information (a) and negative information (b), we have $0<b<a<1$, $a+b<1$. On the contrary, if people prefer negative information, we have $0<a<b<1$, $a+b<1$. People can contradict both positive and negative information.

Negative Results

Anonymous 1/12/2011 06:22:00 AM

its not confirmed yet but we are praying!!

Reply

How Rumours and Truths Spread in OSNs

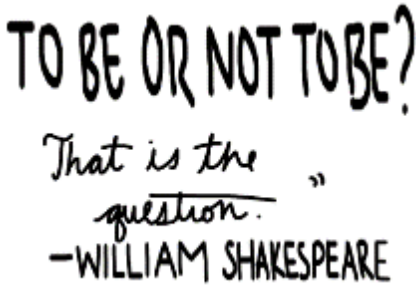
Case 3 (19/06/13): R.I.P. Jackie Chan Dead

- NEGATIVE INFORMATION

The action star Jackie Chan was reported to be dead in Facebook sending thousands of his devout fans into shock.

- POSITIVE INFORMATION

Jackie Chan posted to Facebook a photo of himself with a newspaper.

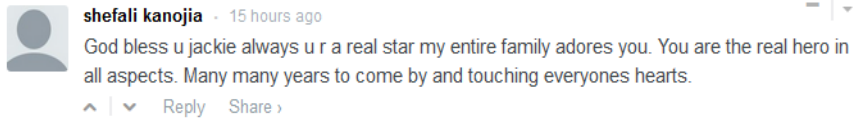


facebook

Technical Perspective (alternative choices)

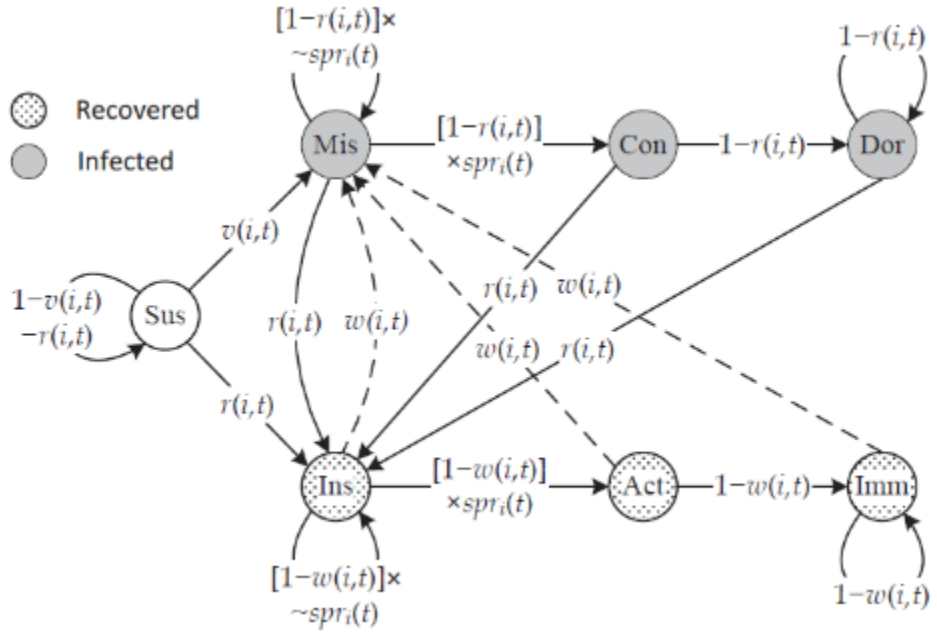
People making alternative choices is people answering “yes-or-no” questions. People must take one side. They cannot say Jackie Chan is neither dead nor alive. If people believe Jackie Chan has died with probability (a), there must be a probability (b) that people believe he is still alive and we have 0<a, b<1, a+b=1. People cannot refute both of the information.

Negative Results



52

Modelling the Spread of Rumours and Truths



$$X_i(t) = \begin{cases} \textit{Sus.}, & \textit{susceptible} \\ \textit{Rec.}, & \textit{recovered} \begin{cases} \textit{Ins.}, & \textit{insider} \\ \textit{Act.}, & \textit{active} \\ \textit{Imm.}, & \textit{immunized} \end{cases} \\ \textit{Inf.}, & \textit{infected} \begin{cases} \textit{Mis.}, & \textit{misled} \\ \textit{Con.}, & \textit{contagious} \\ \textit{Dor.}, & \textit{dormant} \end{cases} \end{cases}$$

$$\begin{pmatrix} \eta_{11} & \cdots & \eta_{1m} \\ \vdots & \eta_{ij} & \vdots \\ \eta_{m1} & \cdots & \eta_{mm} \end{pmatrix} \quad \eta_{ij} \in [0,1]$$

Susceptible-X-X

Susceptible-X-Recovered

1. Modelling nodes, topology and social events
2. SXX and SXR
3. Modelling propagation dynamics

$$P(X_i(t) = Sus.) = \left[1 - v(i,t) - r(i,t)\right] \cdot P(X_i(t-1) = Sus.)$$

$$P(X_i(t) = Rec.) = 1 - P(X_i(t) = Sus.) - P(X_i(t) = Inf.)$$

$$P(X_i(t) = Inf.) = v(i,t) \cdot P(X_i(t-1) = Sus.) + [1 - r(t)] \cdot P(X_i(t-1) = Inf.) + w(i,t) \cdot P(X_i(t-1) = Rec.)$$

$$\begin{cases} S(t) = m - I(t) - R(t) \\ I(t) = \sum_{i=1}^{m} P(X_i(t) = Inf.) \\ R(t) = \sum_{i=1}^{m} P(X_i(t) = Rec.) \end{cases}$$

Correctness Validation

Basic Properties of the Network Topologies

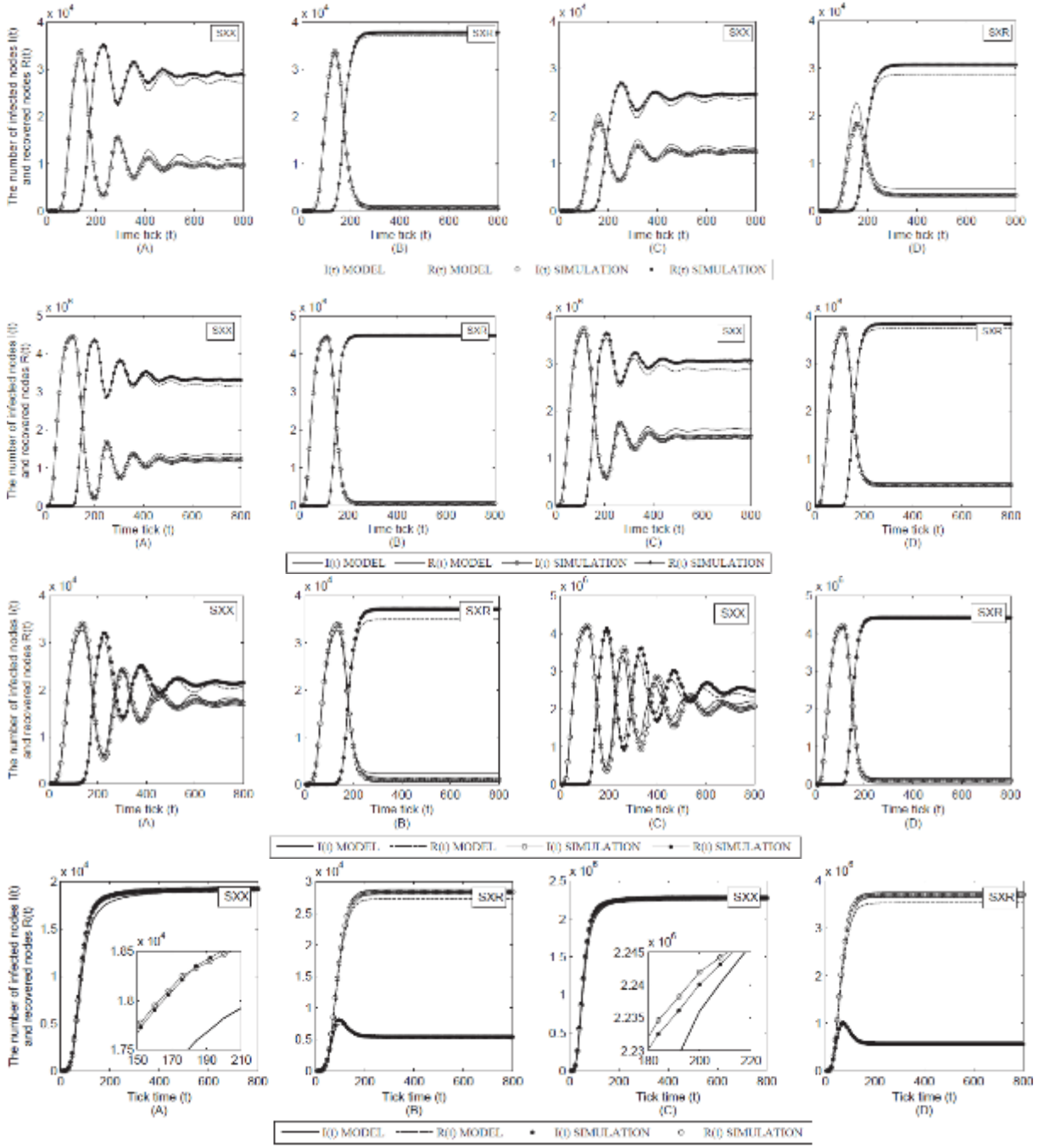
	Facebook	Google Plus
Number of nodes	45814	264004
Number of links	4693129	47130325
Average degree	5.76	10.04
Max outdegree	199	5739
Max indegree	157	3063

Fig1. Optimistic choices (FB)

Fig2. Optimistic choices (G+)

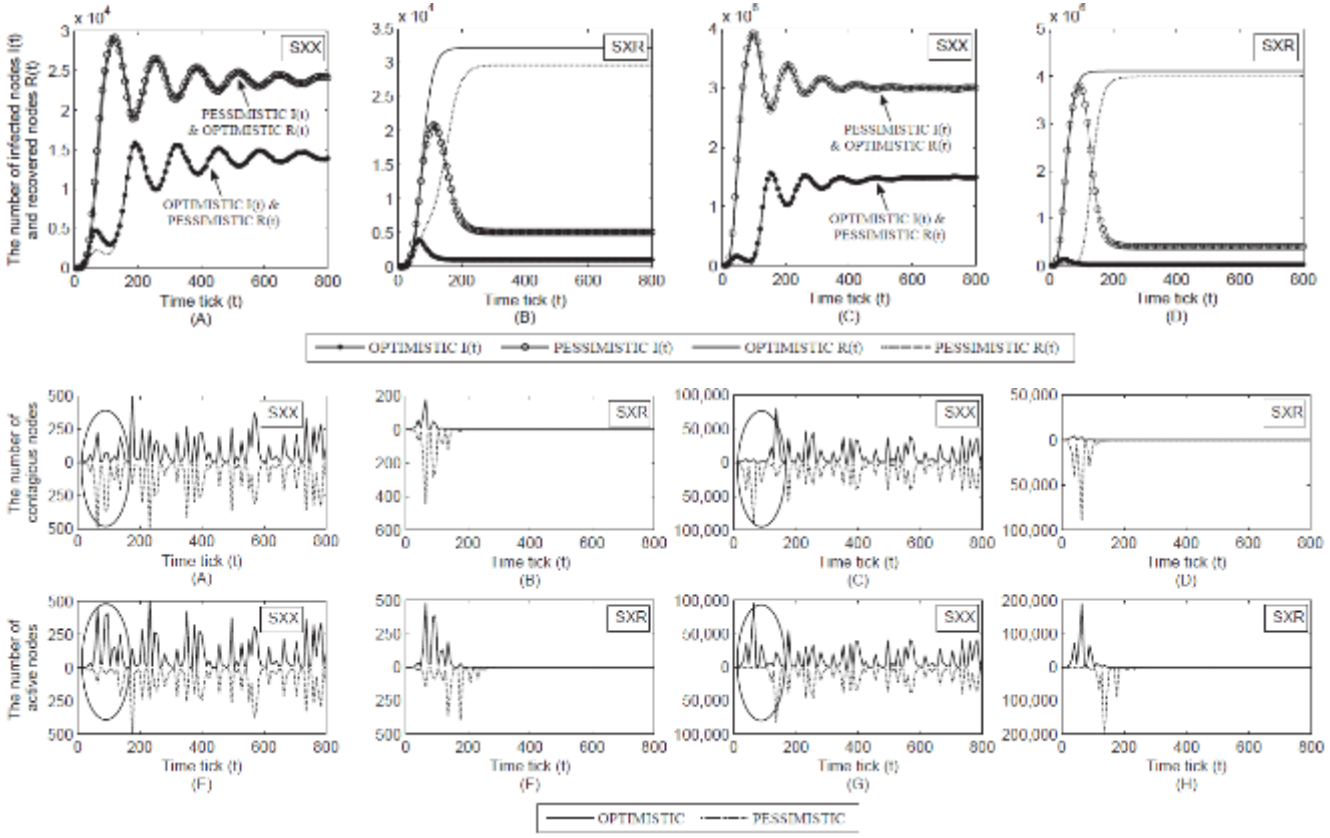
Fig3. Preferable choices (FB & G+)

Fig4. Alternative choices (FB &G+)



Results Exhibition of the Modelling

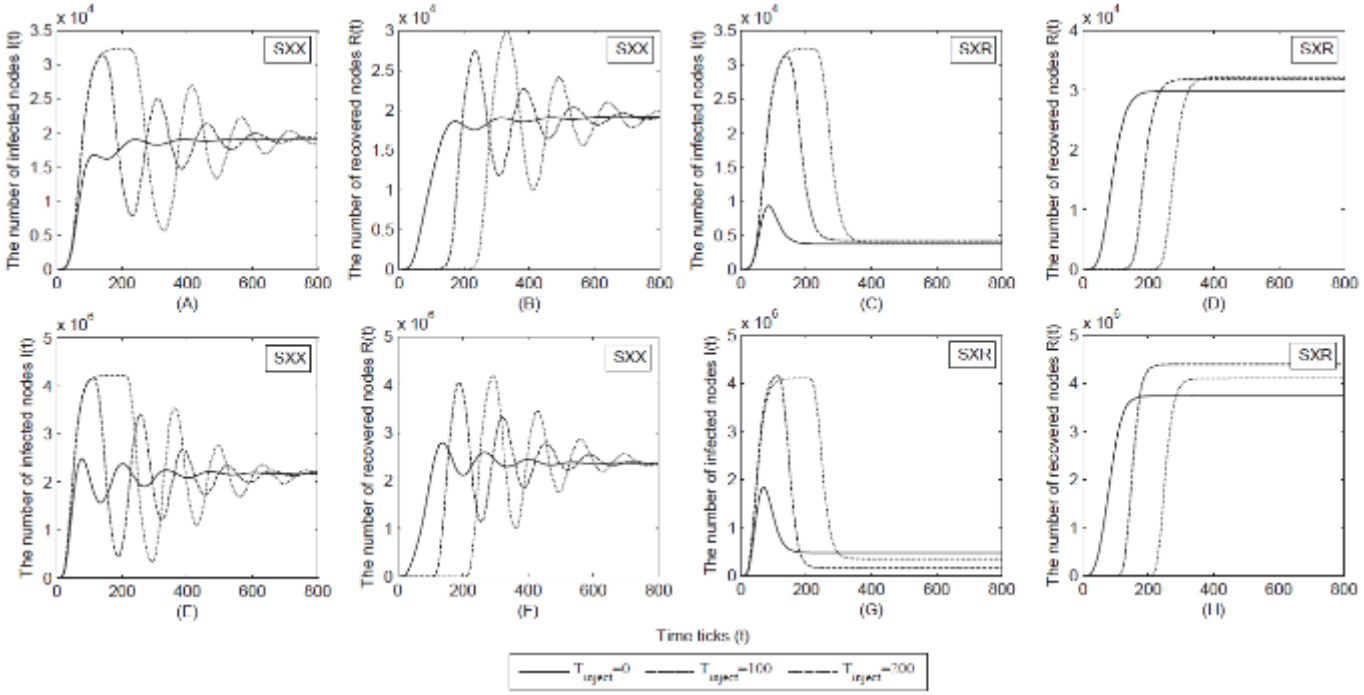
- Optimistic and Pessimistic



We find that the propagation is mainly decided by the early spreading dynamics if people make optimistic or pessimistic choices on their receiving.

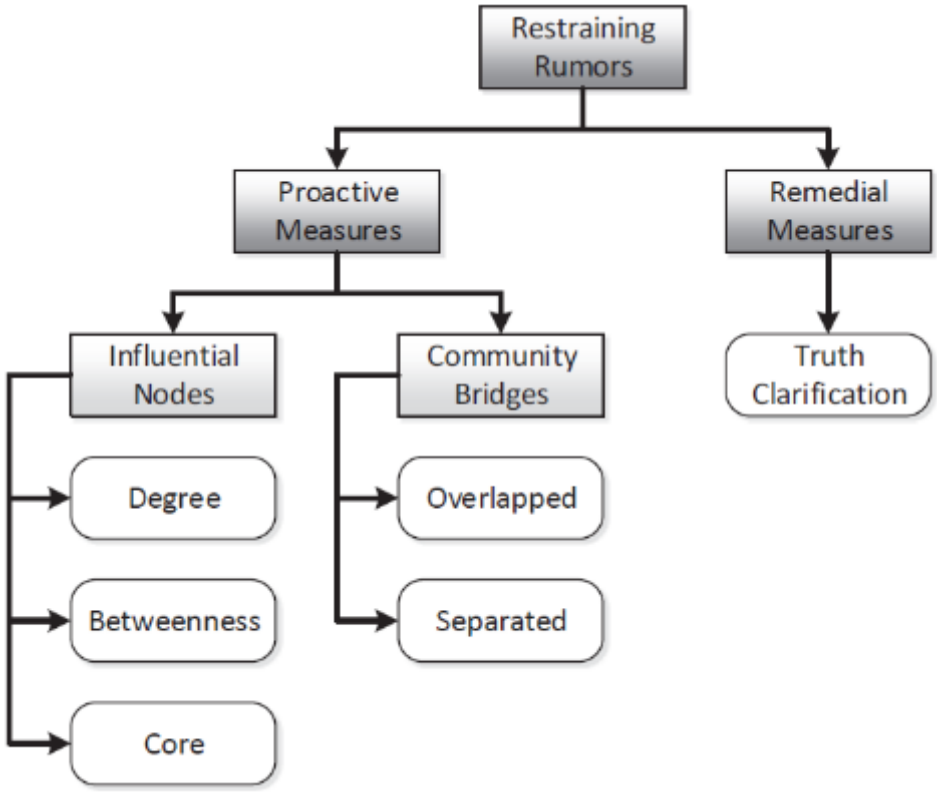
Results Exhibition of the Modelling

- Impact of the start time of positive information



Firstly, the propagation under different settings will finally become steady even though there are oscillations. The final results will be approximately equal to a constant. Secondly, we can observe that the spread of negative information reaches the largest scale at the early time stage.

Defense Measures in Online Social Networks



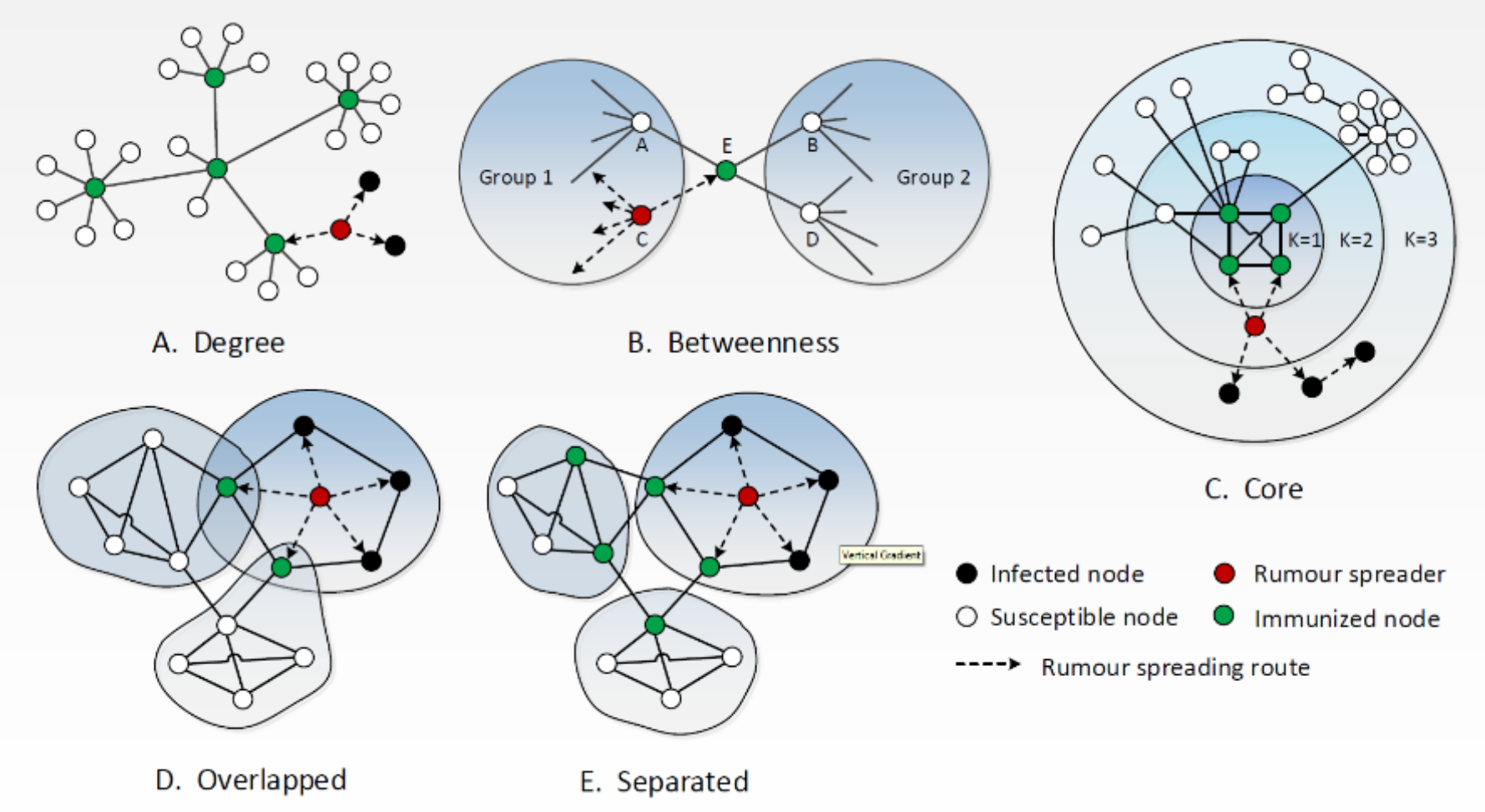
Terminology of Immunization Techniques

Take the methods of restraining rumors as an example.

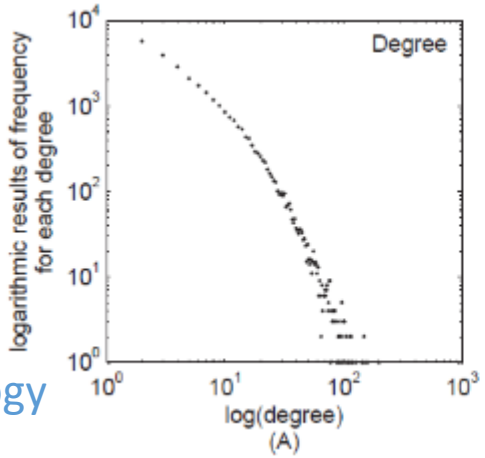
- Degree: "Error and attack tolerance of complex networks," *Nature*, Jul. 2000.
- Betweenness: "Controllability of complex networks," *Nature*, 2011.
- Core: "Identification of influential spreaders in complex networks," *Nature*, Aug 2010.
- Overlapped: "Uncovering the overlapping community structure of complex networks in nature and society," *Nature*, 2005.
- Separated: "A social network based patching scheme for worm containment in cellular networks," in *INFOCOM, Proceedings IEEE*, 2009

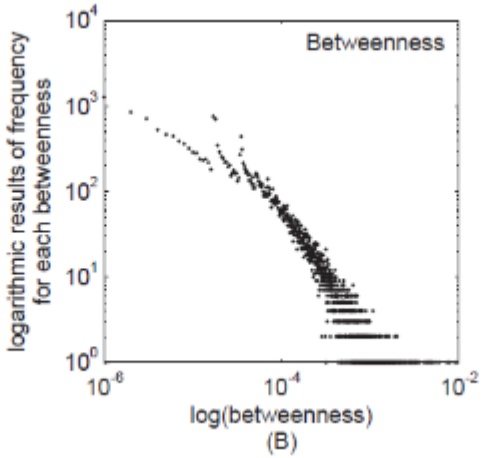
57

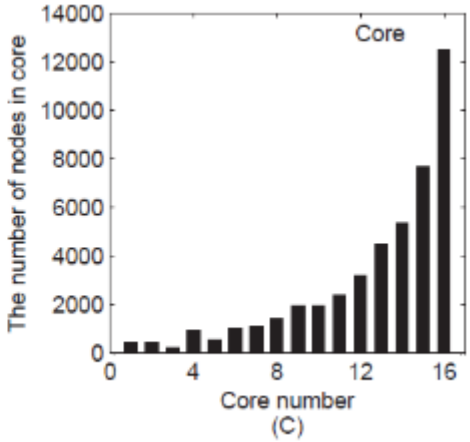
Illustration of Various Proactive Measures



Important Users in OSNs

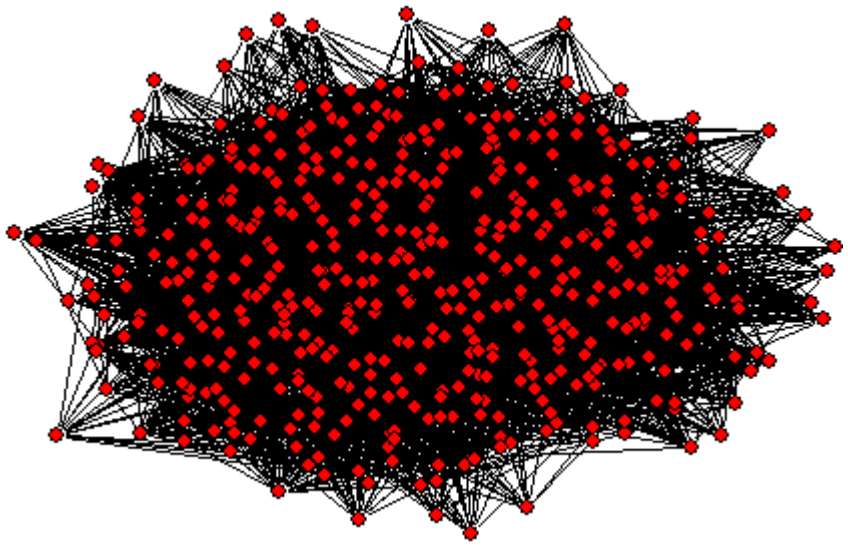


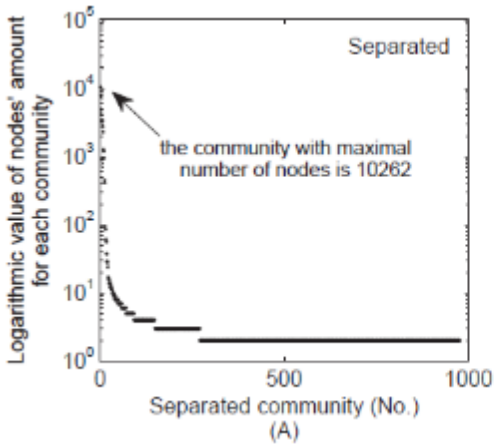


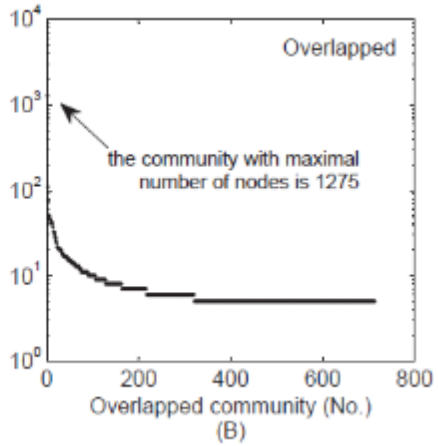


Facebook Topology

Just an example.



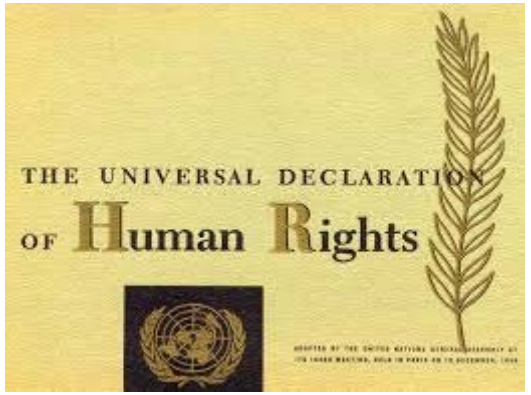




Why is the Truth Spread Necessary?

Case 1: The place of birth of Obama

Case 2: Free press and human right



Case 3: Uncontrollable deceptive reporter

Fake photo report

Actually accident in Hangzhou

The question is whether truth propaganda can outperform traditional controlling techniques?

Maximal Infected Users and Final Infected Users

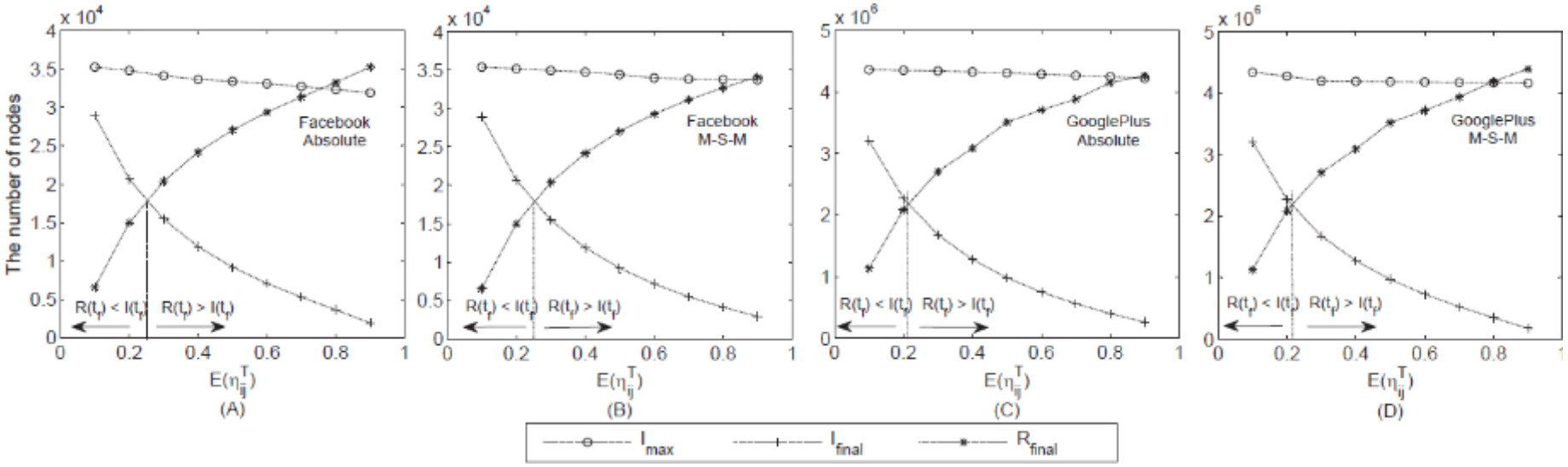


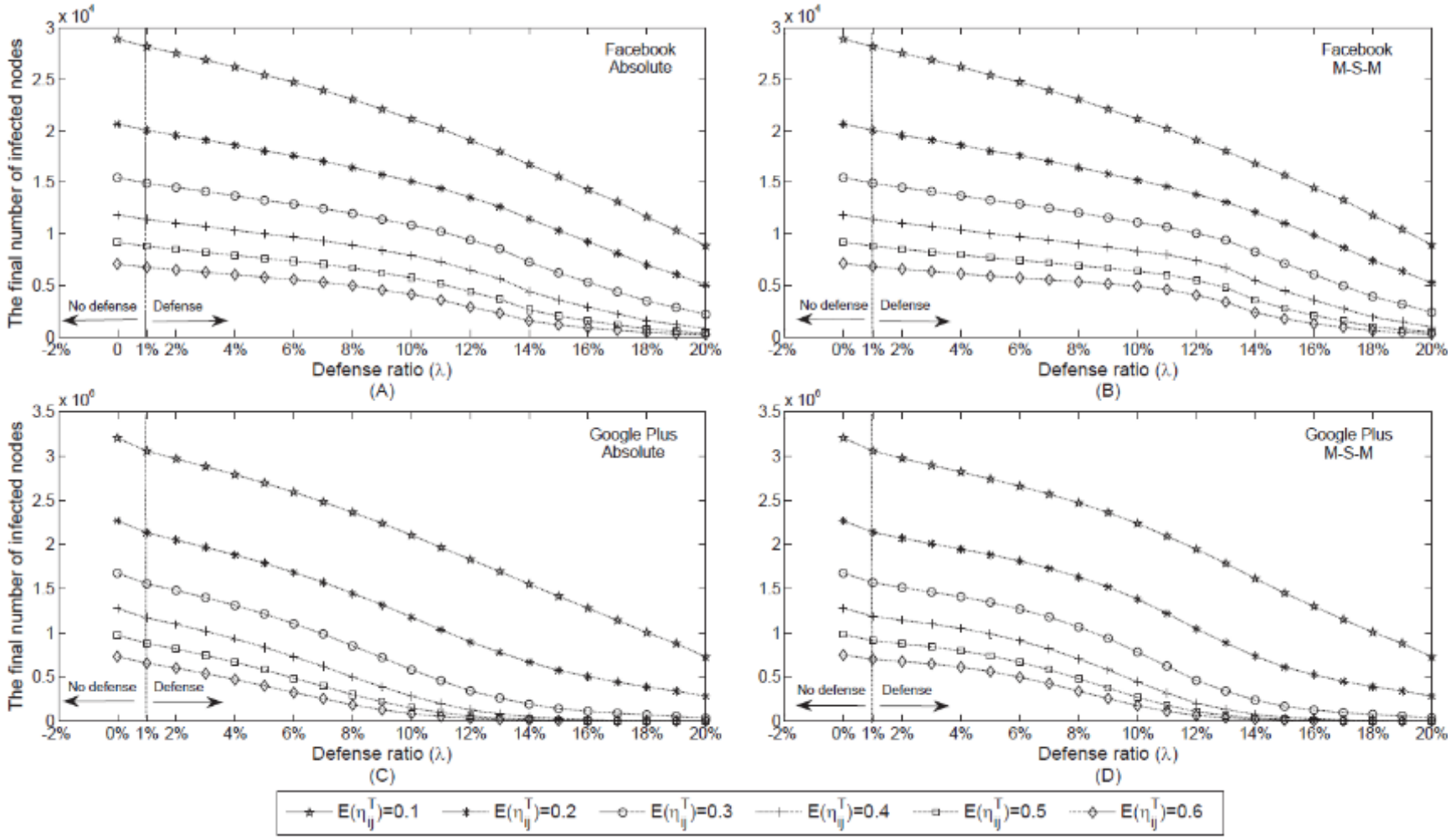
Fig. 17. The maximum number of infected users (I_{max}), the final number of infected users (I_{final}) and the final number of recovered users (R_{final}). Settings: $t_{inject} = 3$, $E(\eta_{ij}^R) = 0.75$, $E(\eta_{ij}^T) \in [0.1, 0.9]$.

Maximal infected users: stay almost constant when using truth clarification.

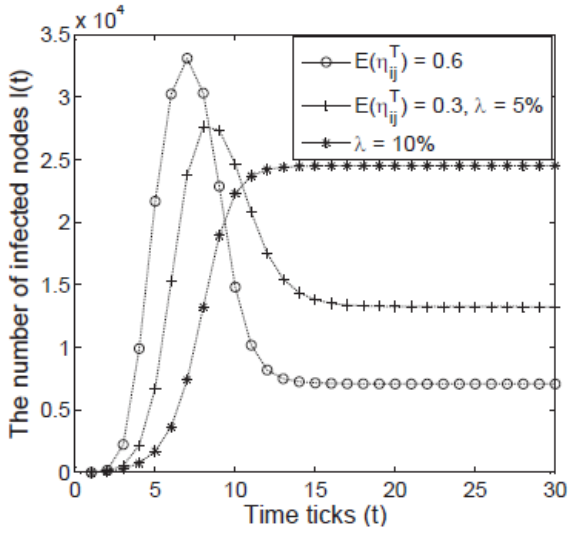
Final infected users: decrease gradually when people are more likely to believe truths.

The Equivalence between Blocking Rumors and Spreading Truth

Fig. 19. The final number of infected nodes (I_{final}) when we set a series of different defense ratios (λ) and truth spreading probabilities $E(\eta_{ij}^{truth})$. Setting: $t_{inject} = 3$, $E(\eta_{ij}^{R}) = 0.75$.



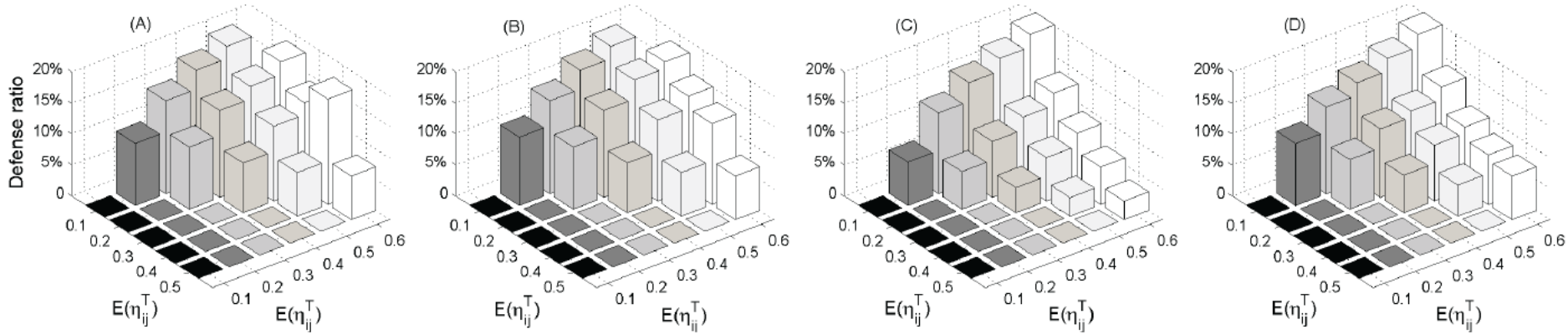
Put Eggs in Different Baskets



A Case Study: Both the maximal infected users and the final infected users decrease.

The exact equivalence: Two methods working together.

Fig. 20. The numeric equivalence between the degree measure and the remedial measure when we set a series of different defense ratios (λ) and truth spreading probabilities $E(\eta_{ij}^T)$. Setting: $t_{inject} = 3$, $E(\eta_{ij}^R) = 0.75$.



Outline

- Introduction
- Source Identification Methods
- Comparative Studies
- Our Work in Modelling the Propagation of Worms and Rumours
- Future Work & Discussion

Future Work & Discussion

- **Research problems**:
  - Analyze the ***factors*** that may affect the process of identifying propagation sources.
    - Factors like types of network topologies, propagation schemes, propagation models, etc.
  - Identifying Propagation Sources in ***Mobile*** Networks.
  - Identifying ***Multiple*** Propagation Sources.
  - Identifying Propagation Sources in ***interconnected*** Networks.

Acknowledgement

The research is partially supported by the following ARC grants:

- ARC Discovery Project DP0773264: 2007-2009, Wanlei Zhou and Yang Xiang, "Development of methods to address internet crime".
- ARC Linkage Project LP100100208, 2010-2012, Wanlei Zhou and Yang Xiang, "An active approach to detect and defend against peer-to-peer botnets".
- ARC Discovery Project DP1095498, 2010-2012, Yang Xiang, Wanlei Zhou, and Yong Xiang, "Tracing real Internet attackers through information correlation".
- ARC Linkage Project LP120200266, 2012-2015, Yang Xiang, Wanlei Zhou, Vijay Varadharajan, and Jonathan Oliver, "Developing an active defence system to identify malicious domains and websites".
- ARC Discovery Project DP140103649, 2014-2016, Wanlei Zhou and Yang Xiang, "Modelling and defence against malware propagation".

Thank you!