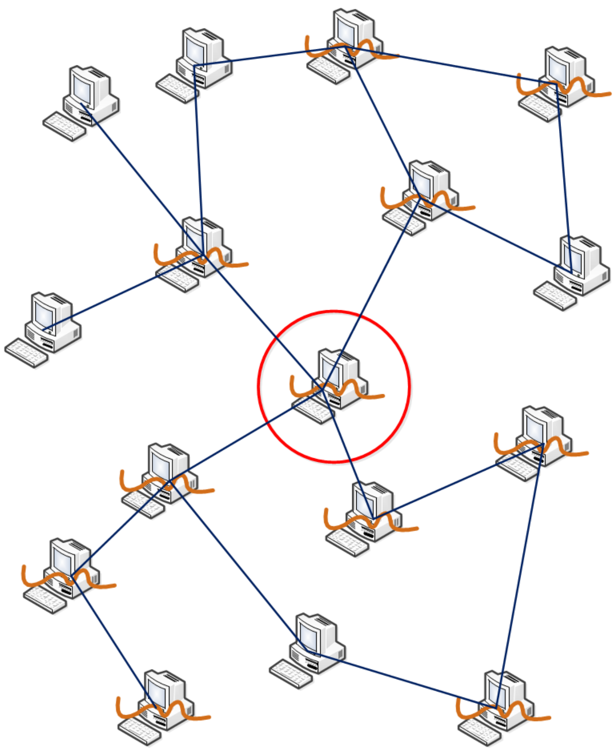
Rumor Source Detection with Multiple Observations: Fundamental Limits and Algorithms

Zhaoxu Wang, Wenxiang Dong, Wenyi Zhang
University of Science and Technology of China
Chee Wei Tan
City University of Hong Kong

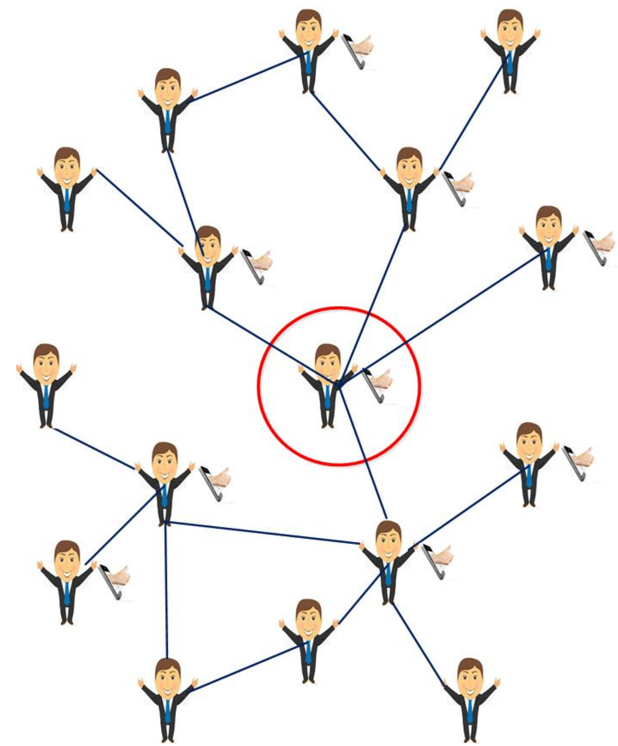
SIGMETRICS, Austin, TX, USA
June 16, 2014

Background

- **How did virus epidemics begin?**



- **Who initiated a rumor in weibo/twitter?**



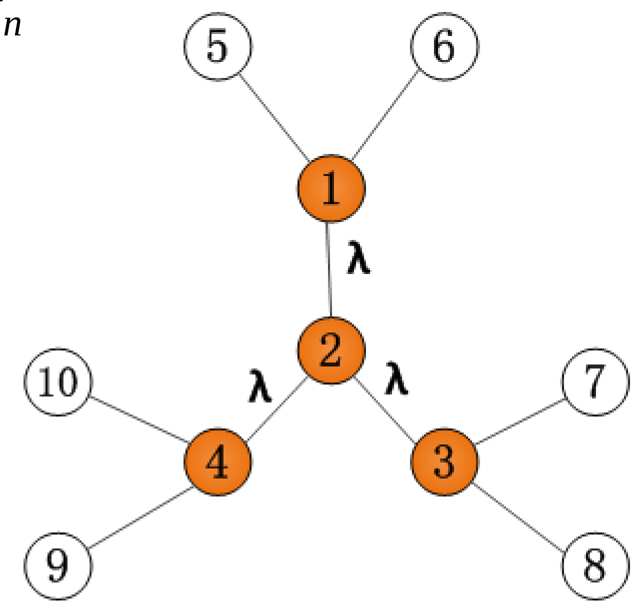
- A rumor, i.e., a “message”, has been passed around in a network.
- At some point we observe those who have possessed the message.
- How and how well can we figure out who initiated this spreading?

Outline

- SI model for rumor spreading
- Rumor center as maximum-likelihood (ML) detector
- Source detection with multiple instances
- Union rumor center
- Performance results
- Experiments
- Conclusions
- Literatures

SI Model for rumor spreading

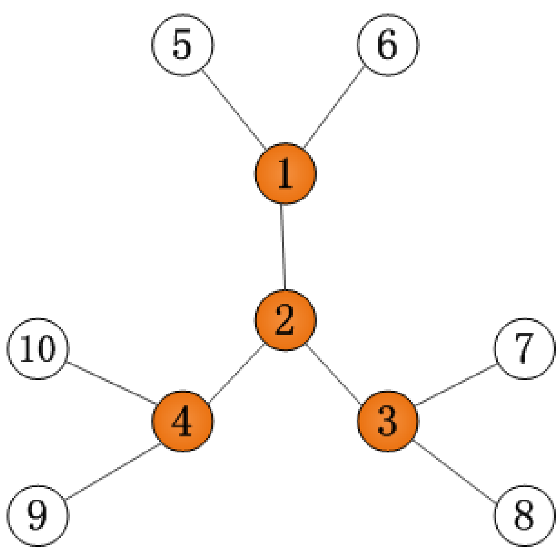
- **SI (susceptible-infectious) model as an undirected graph G =(V,E)**
  - An infected node keeps the rumor forever
- **Exponentially distributed infection time with parameter** λ
- **We observe the network G at some time and find n infected nodes denoted by** G_n



[ShaZamTIT11]

Rumor center
as maximum-likelihood (ML) detector

- **Rumor centrality R(v,G):** The total number of permitted permutations with source node v and infected nodes G.
- **Permitted permutation:** A possible order of infection starting from a postulated source node, obeying causality.
- **Rumor center (RC):** The node with the largest rumor centrality.



$v = 1$

$\{1,2,3,4\},\{1,2,4,3\}$ permitted;

$\{1,3,2,4\}$ not permitted.

$G = \{1,2,3,4\};$

$R(1,G) = 2, R(2,G) = 6,$

$R(3,G) = 2, R(4,G) = 2.$

Rumor center $= 2$

Key: **For the basic model, likelihood** $\propto R(v,G) \Rightarrow \mathrm{ML} = \mathrm{RC}.$

[ShaZamTIT11]

Performance results for the basic model

➢ For general graphs

- For node degree $\delta \geq 3$, "non-trivial" detection:

$$\lim_{n\to\infty} P_c(G_n) > 0$$

- For node degree $\delta = 2$, detection asymptotically impossible:

$$\lim_{n\to\infty} P_c(G_n) = 0$$

➢ For regular trees

- $$\lim_{n\to\infty} P_c(G_n) = \delta \cdot \mathrm{I}_{1/2}\left(\frac{1}{\delta-2}, \frac{\delta-1}{\delta-2}\right) - \delta + 1,$$

  where $\mathrm{I}_x(\alpha, \beta)$ is the incomplete beta function.

- $\lim_{n\to\infty} P_c(G_n) \nearrow 0.307$ as $\delta \to +\infty$.

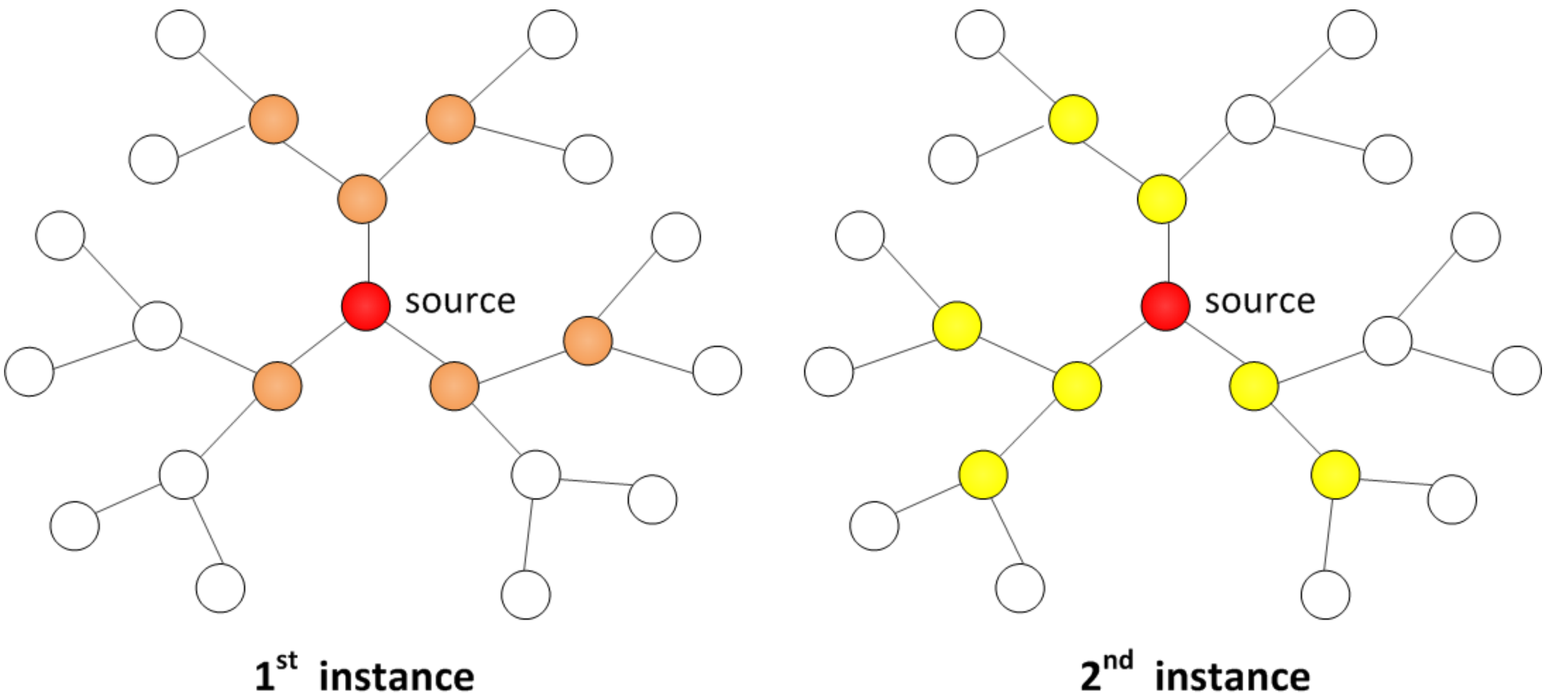
[ShaZamTIT11], [ShaZamSIGMETRICS12]

Source detection with multiple instances

- **A source may initiate multiple instances of spreading, rather than only once.**

  e.g., fraudulent email spams and recurring malcode, which are usually originated from a common culprit.

- **If multiple instances are available, how much can this diversity help?**



Union rumor center

➢ Union rumor center

Assume K independent instances of infected sets G_{n_j} , $j = 1, \ldots, K.$

- Union rumor centrality : $R_K(s^*, G_{n_1}, \ldots, G_{n_K}) = R(s^*, G_{n_1}) \cdots R(s^*, G_{n_K})$
- Union rumor center (URC): The node with the largest union rumor centrality.

➢ ML rumor source estimator

- For regular trees

$$\hat{s} \in \arg\max_{s^* \in \{\cap G_{1\to K}\}} \mathrm{P}_G(G_{n_1}, \ldots, G_{n_K} \mid s^*) = \arg\max_{s^* \in \{\cap G_{1\to K}\}} R_K(s^*, G_{n_1}, \ldots, G_{n_K}),$$

e.g. , **ML detector=URC.**

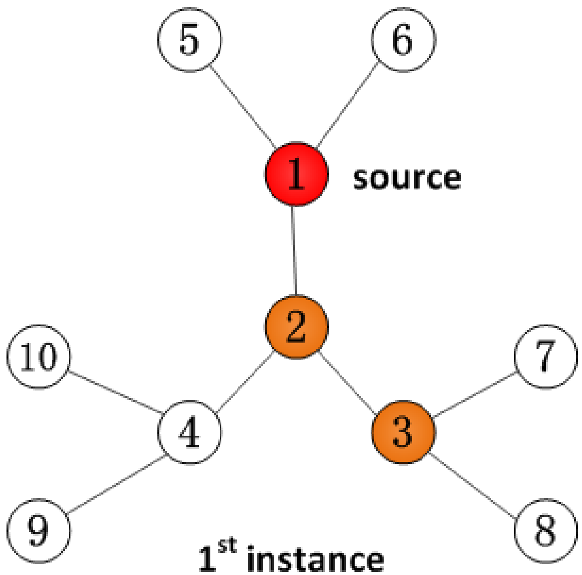
Tip: If s has the largest union rumor centrality among all its neighbors in $G_{n_1} \cap \ldots \cap G_{n_K}$ then s is a **URC**.

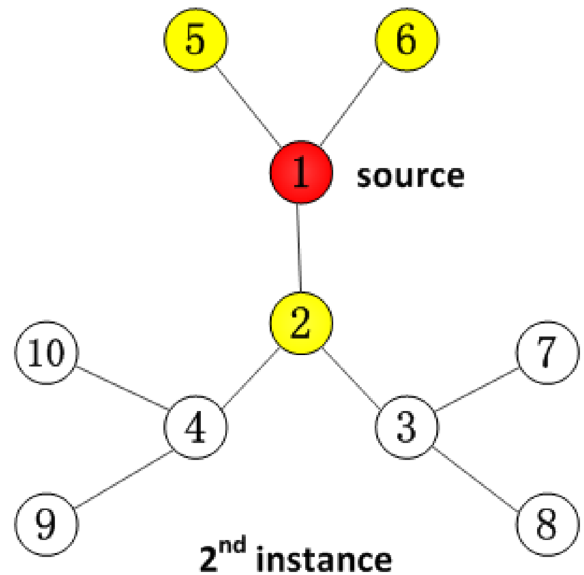
- For general trees

$$\hat{s} \in \arg\max_{s^* \in \{\cap G_{1\to K}\}} \prod_{j=1}^{K} \mathrm{P}(\sigma_s^* \mid s^*, G_{n_j}) \cdot R_K(s^*, G_{n_1}, \ldots, G_{n_K})$$

Union rumor center

- **For example**





$$G_1 = \{1,2,3\};$$
$$R(1,G) = 1,$$
$$R(2,G) = 2,$$
$$R(3,G) = 1.$$

$\Rightarrow$ $\mathrm{RC} = 2 \neq \mathrm{source}$

$$G_2 = \{1,2,5,6\};$$
$$R(1,G) = 6, R(2,G) = 2,$$
$$R(5,G) = 2, R(6,G) = 2.$$

$\Rightarrow$ $\mathrm{RC} = 1 = \mathrm{source}$

Union rumor center

$$G_1 \cap G_2 = \{1,2\};$$
$$R_2(1, G_1, G_2) = 1 \cdot 6 = 6,$$
$$R_2(2, G_1, G_2) = 2 \cdot 2 = 4.$$

$\Rightarrow$ $\mathrm{URC} = 1 = \mathrm{source}$

Performance results

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

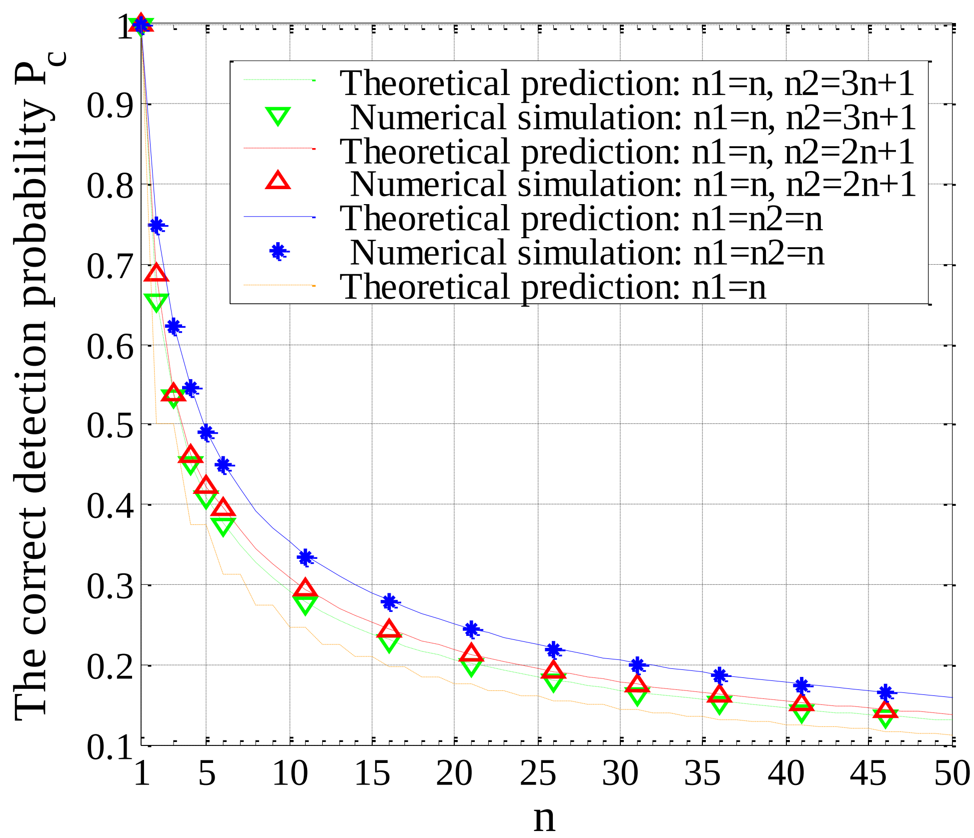
■ (1) $n_1 = n_2 = n \geq 1$

$$p_c = \binom{2n}{n} 2^{-2n+1}, \qquad n \geq 1.$$

■ (2) $n_1 = 2, n_2 > n_1$

$$p_c = \left\{ \sum_{m=0}^{\left\lfloor \frac{n_2}{n_1} \right\rfloor} \binom{n_2 - 1}{m} - \Pi\left(\frac{n_2}{n_1}\right) \binom{n_2 - 1}{\left\lfloor \frac{n_2}{n_1} \right\rfloor} \right\} \cdot 2^{-2n_2+1},$$

where $\Pi\left(\frac{n_2}{n_1}\right) = \begin{cases} 1 & \frac{n_2}{n_1} \in Z \\ 0 & \text{others.} \end{cases}$



$\delta = 2$

■ (3) $n_2 > n_1 \geq 3$

$$p_c = \left\{ \sum_{m=0}^{n_1 - 1} \left[\binom{n_2 - 1}{\left\lfloor m \frac{n_2}{n_1} \right\rfloor} \binom{n_1}{m} + S_{n_2 - 1}(m) \binom{n_1 - 1}{m} \right] - 2 \binom{n_2 - 1}{\left\lfloor \frac{n_2}{n_1} \right\rfloor} - \Pi\left(\frac{n_2}{n_1}\right) \binom{n_2 - 1}{\left\lfloor \frac{n_2}{n_1} \right\rfloor} \right\} \cdot 2^{-(n_1 + n_2)+2}, \quad \text{where } S_{n_2 - 1}(m) = \sum_{i = \left\lfloor m \frac{n_2}{n_1} \right\rfloor + 1}^{\left\lfloor (m+1) \frac{n_2}{n_1} \right\rfloor - 1} \binom{n_2 - 1}{i}.$$

Performance results

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

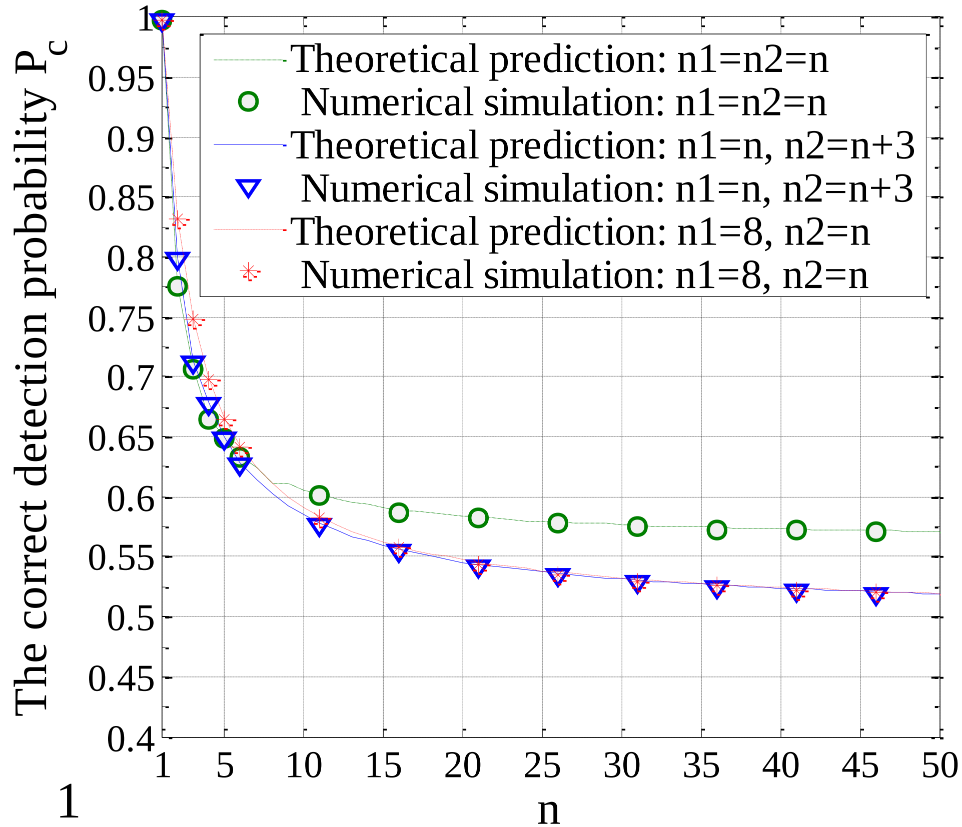
■ (1) $n_1 = n, n_2 = qn\ (q \in Z^+)$

$$\mathrm{P}_c = \frac{qn+q+2}{2(qn+1)}$$

■ (2) $n_1 = n, n_2 = qn+1\ (q \in Z^+)$

$$\mathrm{P}_c = \frac{qn+q+2}{2(qn+1)}$$

■ (3) $n_1 = n, n_2 = qn+t\ (q \in Z^+), t < n$



$\delta = 3$

$$\mathrm{P}_c = \frac{qn+q+2}{2(qn+1)} + \Delta\mathrm{P}_c \quad \text{with} \quad \Delta\mathrm{P}_c < \frac{1}{2(qn+1)}.$$

As $n \to +\infty$, $\lim\limits_{n\to+\infty} \mathrm{P}_c = \frac{1}{2}$

Performance results

➢ Monotonicity

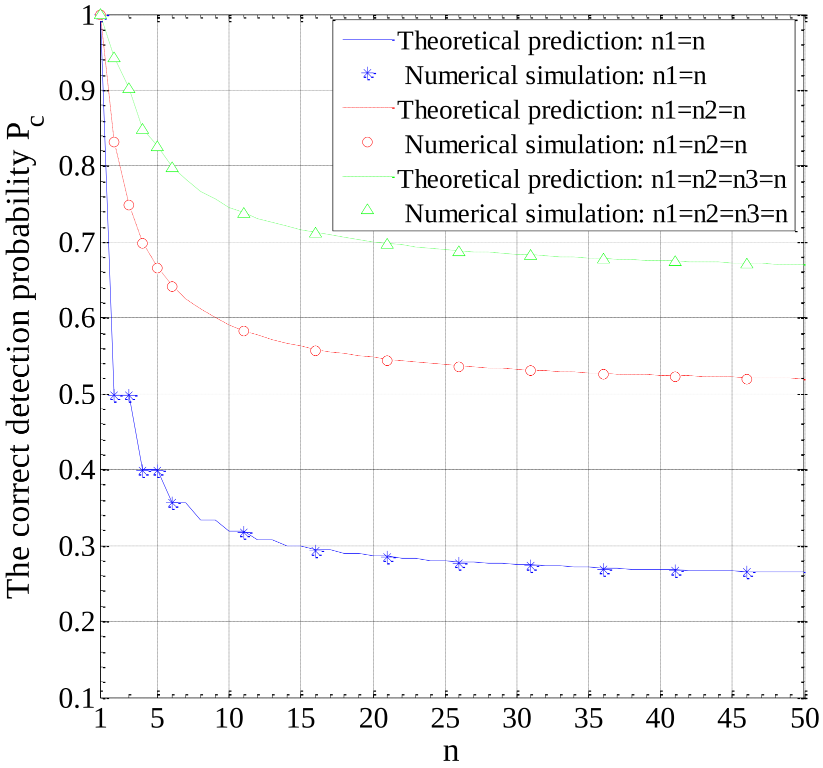
Given $G_{n_1}, G_{n_2}, \ldots, G_{n_K}$

- (1) P_c is increasing with δ
  - **Reliable detection with abundant connectivity**

    For any $K \geq 2$, as $\delta \to +\infty$, $\mathrm{P}_c \to 1$

  - **Reliable detection with abundant diversity**

    For any δ, as $K \to +\infty$, $\mathrm{P}_c \to 1$



$\mathrm{P}_c \ \ vs \ \ n, \delta = 3$

- (2) P_c is non-increasing with n_K, as fixed $n_1, \ldots, n_{K-1}$

Performance results

➢ Asymptotic regime

$$n_1, \ldots, n_K \to +\infty$$

■ (1) Any δ and any K

$$\lim_{n_1,\ldots,n_K \to +\infty} \mathrm{P}_c = \phi_K(\delta) := 1 - \delta\left(1 - \varphi_K\left(\frac{1}{\delta-2}, \frac{\delta-1}{\delta-2}\right)\right),$$

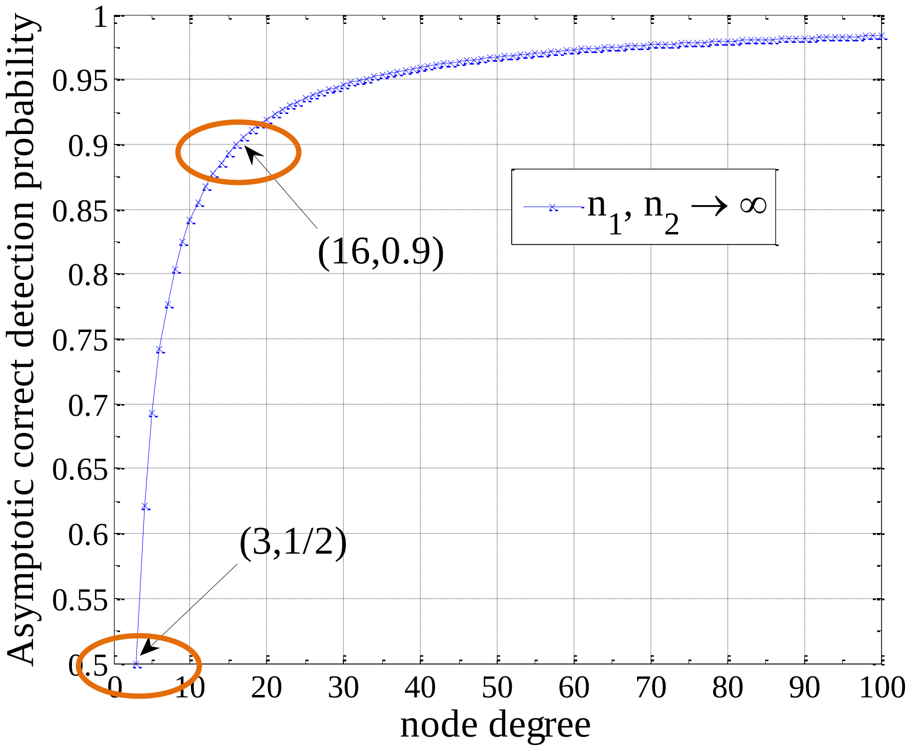
$$\text{where } \varphi_K(\alpha,\beta) = \int \cdots \int \frac{\Gamma(\alpha+\beta)^K}{\Gamma(\alpha)^K \Gamma(\beta)^K} \prod_{j=1}^{K} \left(x_j^{\alpha-1}(1-x_j)^{\beta-1}\right) \mathrm{d}x_1 \ldots \mathrm{d}x_K,$$

$$\alpha = \frac{1}{\delta-2}, \beta = \frac{\delta-1}{\delta-2}.$$

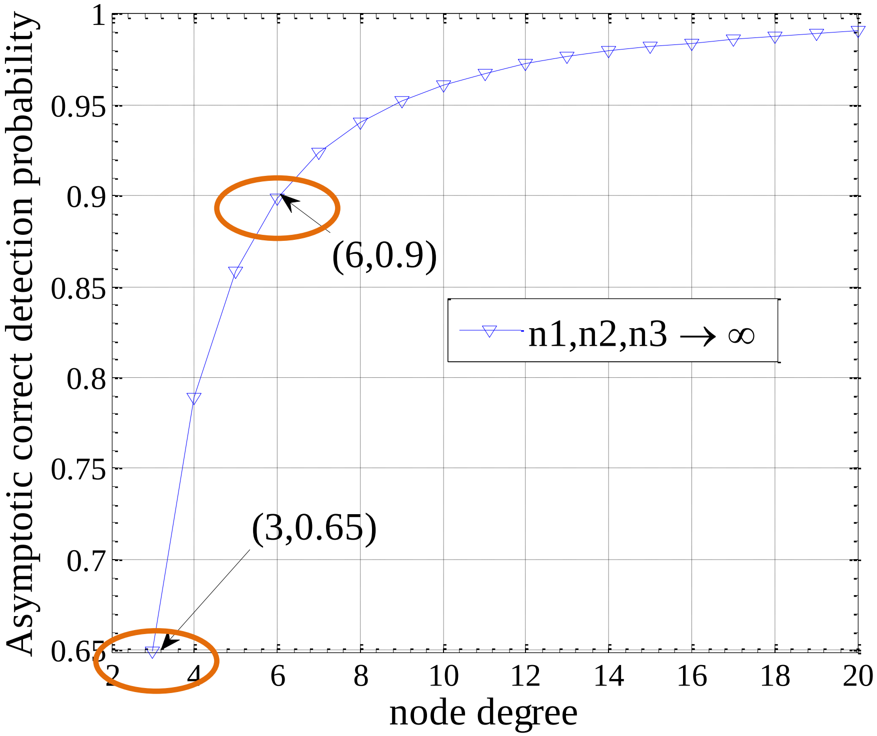
- As $\delta \to +\infty$, $\phi_K(\delta) \to 1$;
- Also, as $K \to +\infty$, $\phi_K(\delta) \to 1$.

Performance results

----Detection performance with multiple instances:



K=2



K=3

K=1 : $\lim_{n\to\infty} P_c(G_n) \nearrow$ **0.307**, $\delta \to +\infty.$

Performance results

➢ Asymptotic regime

■ (2) $\delta = 3$ and any K

$$\phi_K(3) = 1 - 3\cdot 2^{K-2}\int_0^1 \cdots \int_0^1 \frac{\prod_{j=1}^{K-1}\left(x_j(1-x_j)\right)}{\prod_{j=1}^{K-1} x_j + \prod_{j=1}^{K-1}\left(1-x_j\right)} \mathrm{d}x_1 \ldots \mathrm{d}x_{K-1},$$

which is increasing with K and is bounded as follows:

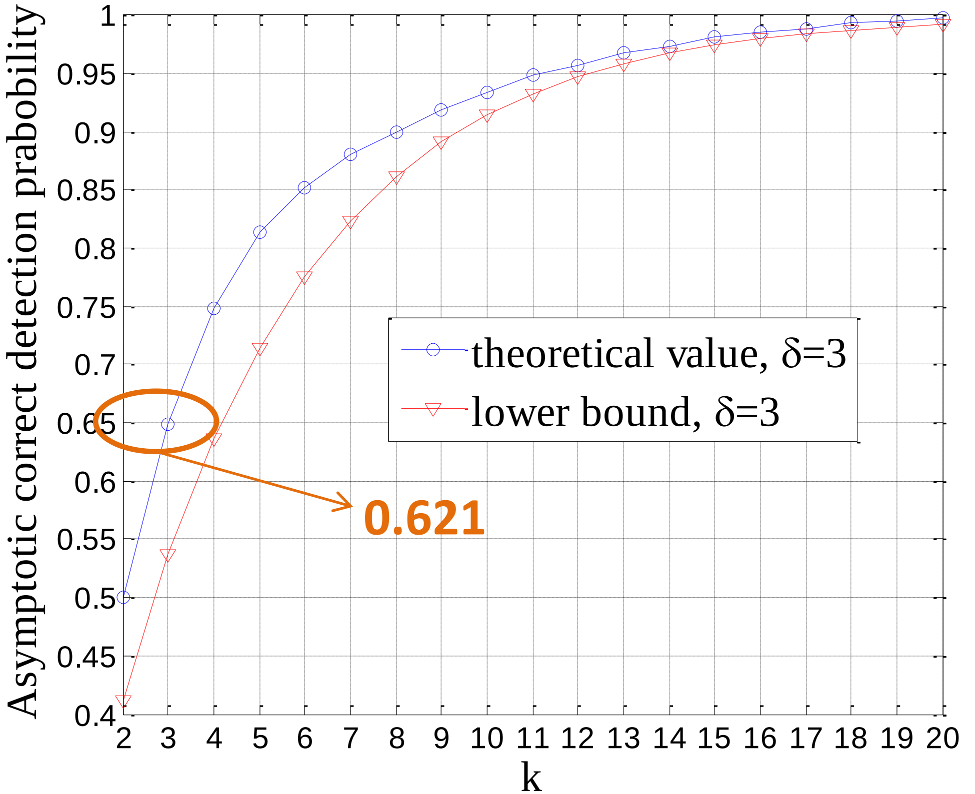
$$1 - \frac{3}{4}\left(\frac{\pi}{4}\right)^{K-1} < \lim_{n_1,\ldots n_K \to +\infty} \mathrm{P}_c < 1,\ K \in Z^+.$$

- **Exponential convergence with L.**
- As $K \to +\infty$, $\phi_K(3) \to 1$.

■ (3) $K = 2$

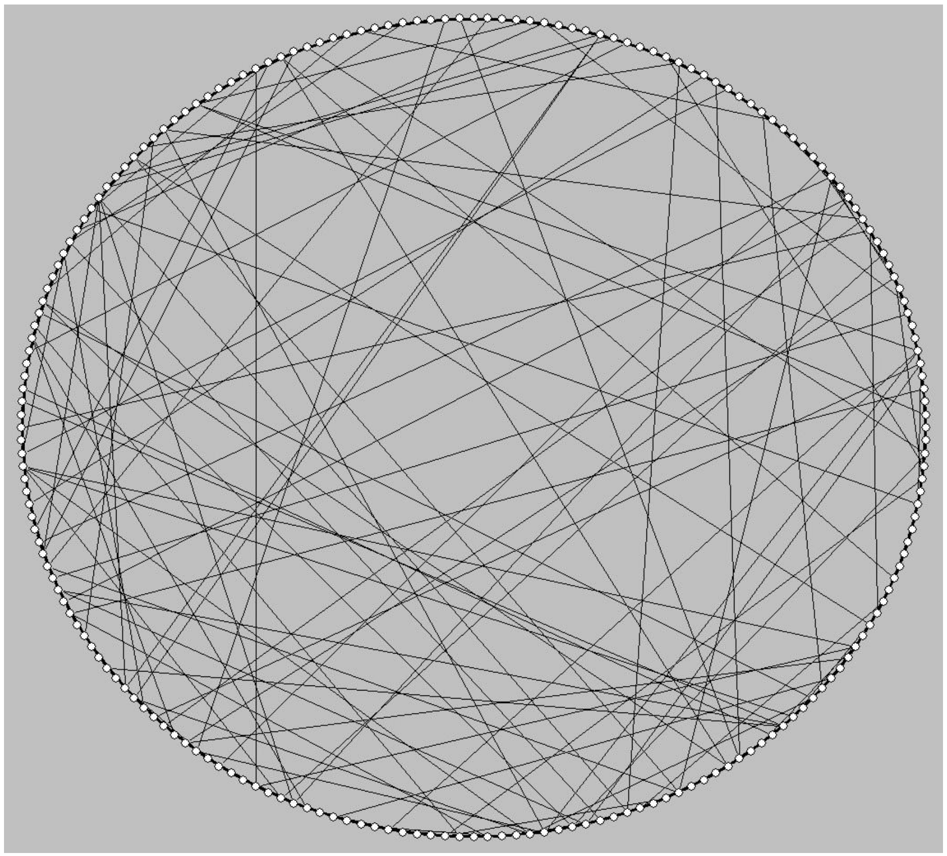
For $\delta = 2$, $\lim_{n_1,n_2 \to +\infty} \mathrm{P}_c = \frac{1}{2}$

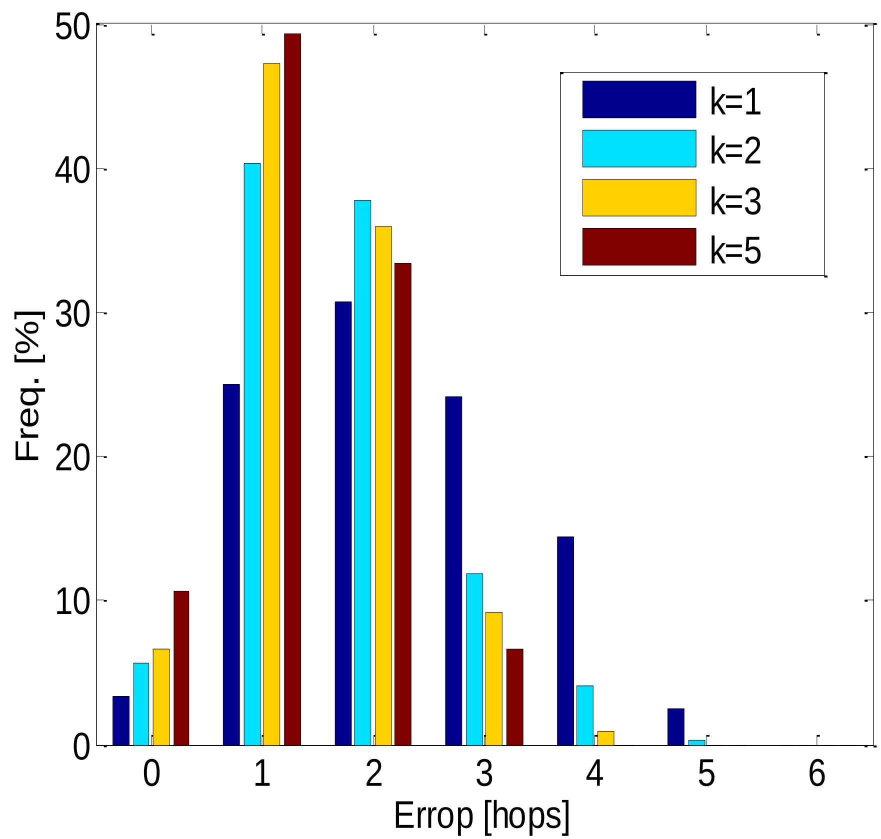
For $\delta = 3$, $\lim_{n_1,n_2 \to +\infty} \mathrm{P}_c = \frac{16}{\pi^2} - 1 \approx 0.621$



Experiments

Small-world networks (Watts-Strogatz model)

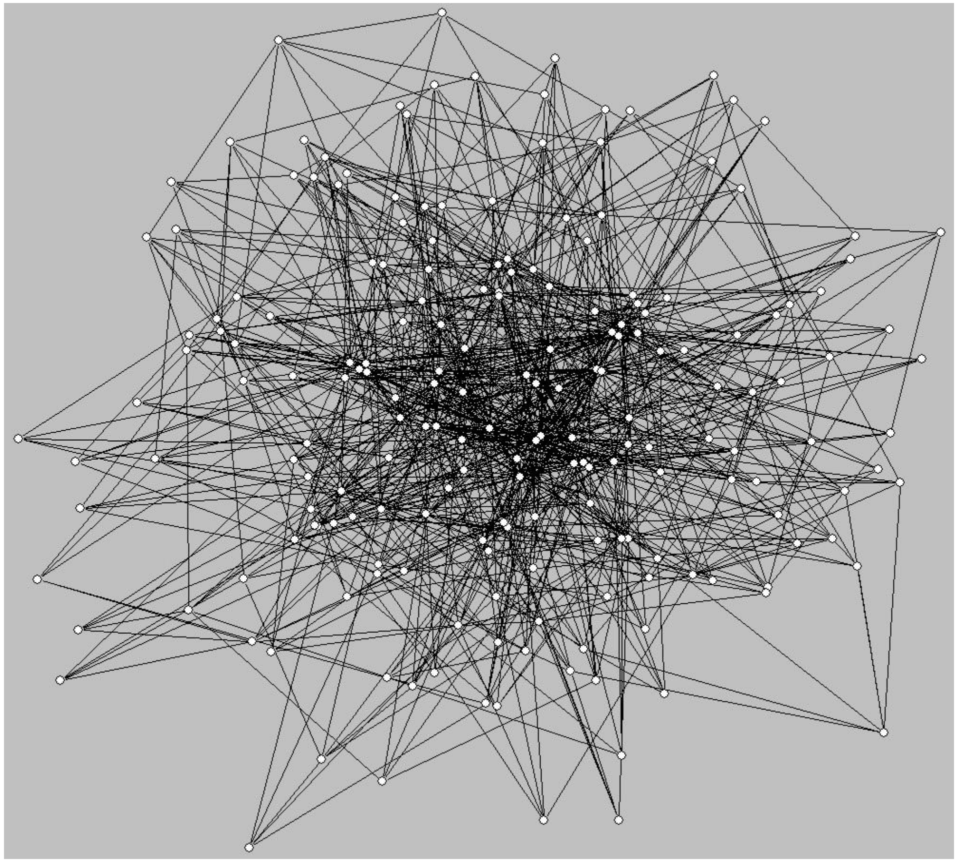


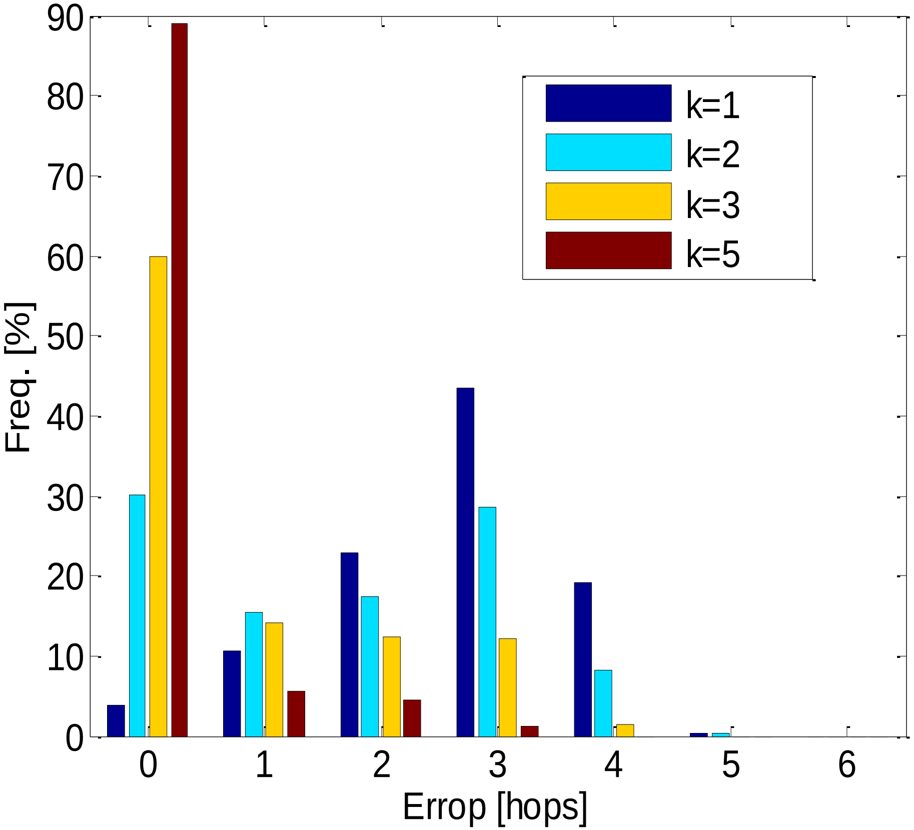


Network size=5000 , infected set size=400.

Experiments

Scale-free networks (Barabasi-Albert model)

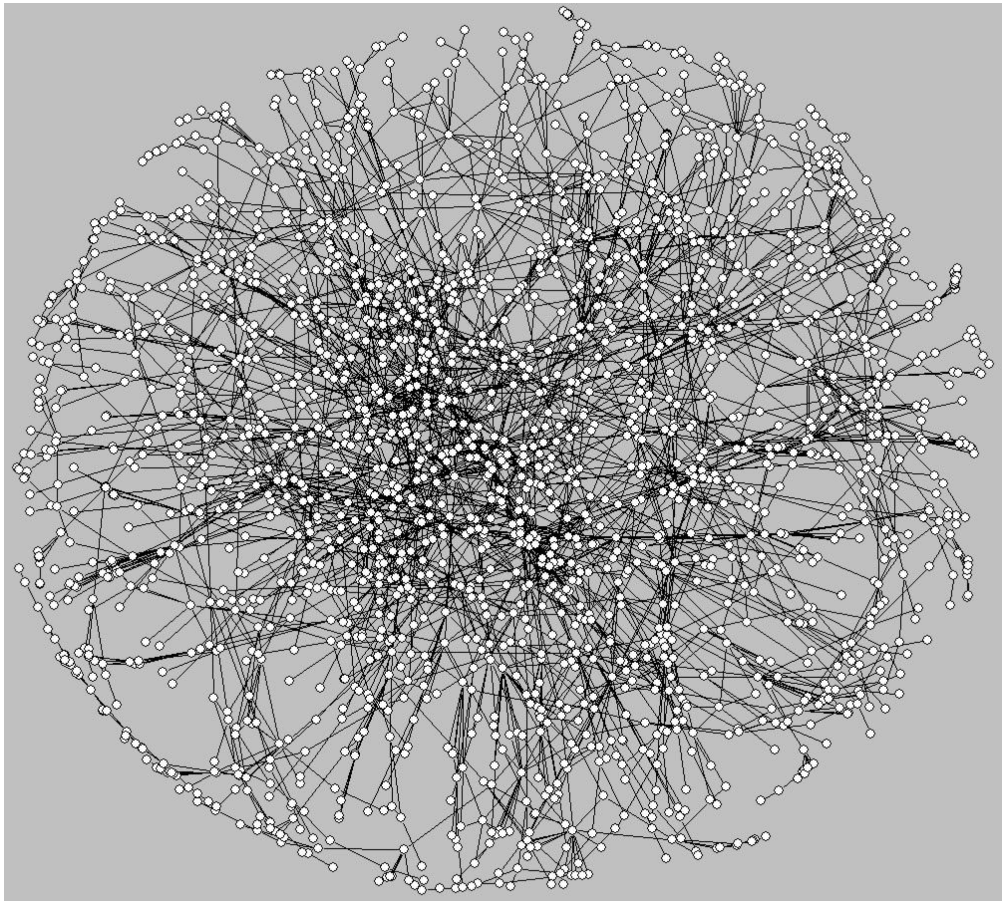


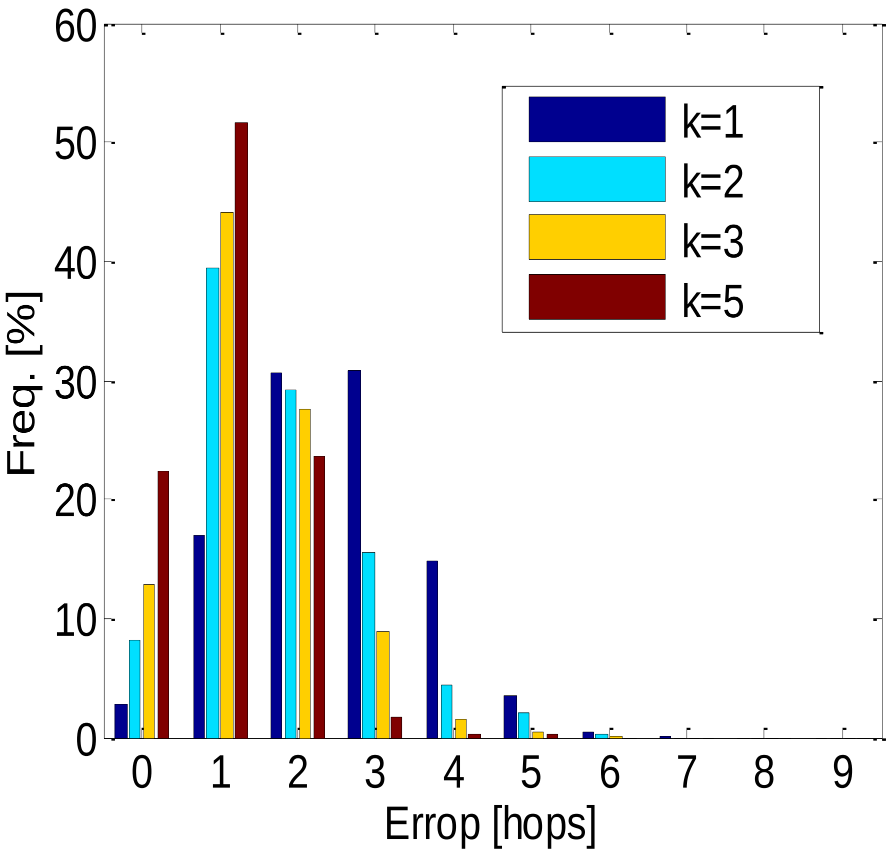


Network size=5000 , infected set size=400.

Experiments

Newman's scientific collaboration network





Source: http://konect.uni-koblenz.de/networks/opsahl-collaboration

Network size=13861 , infected set size=400.

Conclusions

- **Established a unified rumor source inference framework with multiple instances.**
- **Provided explicit analytical results for regular trees.**

— multiple independent observations dramatically enhances detectability.

— even two observations can more than double the detection probability of a single observation.

— P_c increases with δ as well as K, i.e., richer connectivity and diversity both enhance detection.

- **Leveraged the inference framework as an effective heuristic for general graphs.**
- **Still a long march towards a full-grown theory capable of handling all the factors in realistic scenarios.**

Literatures

----Estimation of rumor source

- Regular trees, general trees, general graphs:

breadth-first-search (BFS) heuristic + RC. [ShaZamTIT11]

- General infection time distribution, general random tree:

RC still achieves non-trivial detection (universal detector).

[ShaZamSIGMETRICS12]

- Limited (sparse) observations:

[PinThiVetPRL12], [KarFraISIT13], [LuoTayLenArxiv13]

- Multiple sources:

[LuoTayLenSP13]

- SIS or SIR infection processes:

[LuoTayICASSP13], [ZhuYinITA13]

- Other related models and algorithms:

[PraVreFalICDM12], [LokM'ezOhtZdeArxiv13], [AntLanSteSikSmuArxiv13]

A (partial) bibliography

[ShaZamIT11] D. Shah and T. Zaman, “Rumors in a network: who’s the culprit?” IEEE TIT, 2011.

[ShaZamSIGMETRICS12] D. Shah and T. Zaman, “Rumor centrality: a universal source detector,” ACM SIGMETRICS, 2012.

[PinThiVetPRL12] P. C. Pinto, P. Thiran, and M. Vetterli, “Locating the source of diffusion in large-scale networks,” PRL, 2012.

[KarFraISIT13] N. Karamchandani and M. Franceschetti, “Rumor source detection under probabilistic sampling,” IEEE ISIT 2013.

[LuoTayLenArxiv13] W. Luo, W. P. Tay, and M. Leng, “How to identify an infection source with limited observations,” arXiv, 2013.

[LuoTayLenSP13] W. Luo, W. P. Tay, and M. Leng, “Identifying infection sources and regions in large networks,” IEEE TSP, 2013.

[LuoTayICASSP13] W. Luo and W. P. Tay, “Finding an infection source under SIS model,” IEEE ICASSP 2013.

[ZhuYinITA13] K. Zhu and L. Ying, “Information source detection in the SIR model: a sample path based approach,” ITA 2013.

[PraVreFalICDM12] B. A. Prakash, J. Vreeken, and C. Faloutsos, “Spotting culprits in epidemics: how many and which ones?” IEEE ICDM 2012.

[LokM´ezOhtZdeArxiv13] A. Y. Lokhov, M. M’ezard, H. Ohta, and L. Zdeborav’a, “Inferring the origin of an epidemy with dynamic message-passing algorithm,” arXiv, 2013.

[AntLanSteSikSmuArxiv13] N. Antulov-Fantulin et. al., “Statistical inference framework for source detection of contagion processes on arbitrary network structures,” arXiv, 2013.

The End

Thank you!

Q&A