Rumor source inference: an overview and some recent results

Wenyi Zhang
University of Science and Technology of China

February 14, 2014
ITA Workshop, San Diego, CA, USA

Acknowledgement

Collaborators:
- Dr. Chee Wei Tan (City U of Hong Kong)
- Mr. Wenxiang Dong (student, USTC)
- Mr. Zhaoxu Wang (student, USTC)
- Mr. Xin Lou (student, CityU)

Grants:
- SRFDP and RGC ERG Joint Research Scheme
- Chinese Academy of Sciences
- RGC Hong Kong

Background

How did virus epidemics begin?

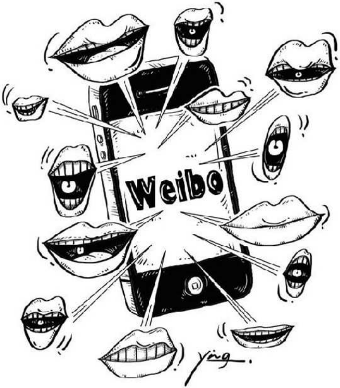
H7N9

H5N1

H3N2

H10N8

Who initiated a rumor in weibo/twitter?



- 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?

A basic model and rumor centrality

Source detection with prior knowledge of suspects

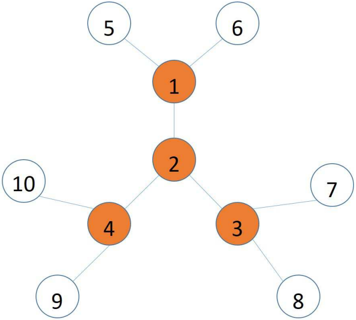
Source detection with multiple instances

Experiments

Beyond and besides the basic model

Wrap-up remarks

A basic model

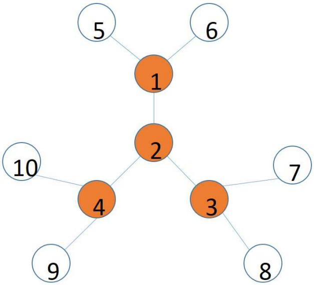


[ShaZamIT11]

- Susceptible-infected (SI) infection process
- Exponentially distributed infection time
- An infinite graph with degree-regular tree topology

Rumor center as maximum-likelihood detector

- *Permitted permutation*: A possible order of infection starting from a postulated source node, obeying causality.
- *Rumor centrality* $R(v, G)$: The total number of permitted permutations with source node v and infected nodes G.
- *Rumor center*: 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 $\propto R(v, G) \Rightarrow \text{ML} = \text{RC}$.

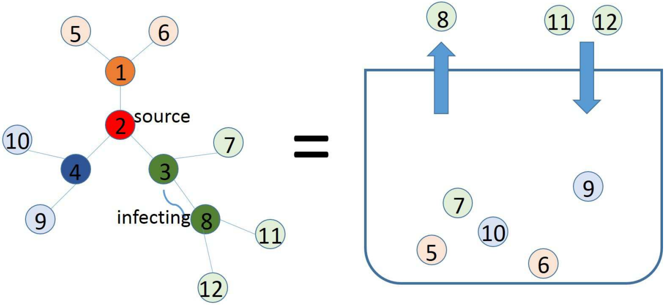
[ShaZamIT11]

Connection with Pólya's urn model

Pólya's urn:

- Initially the urn contains d balls each with one different color.
- Each time a ball is uniformly drawn, and then $(d-1)$ balls of the same color are returned to the urn.
- After n drawings, denote by X_j the number of times that balls of color C_j have been drawn.

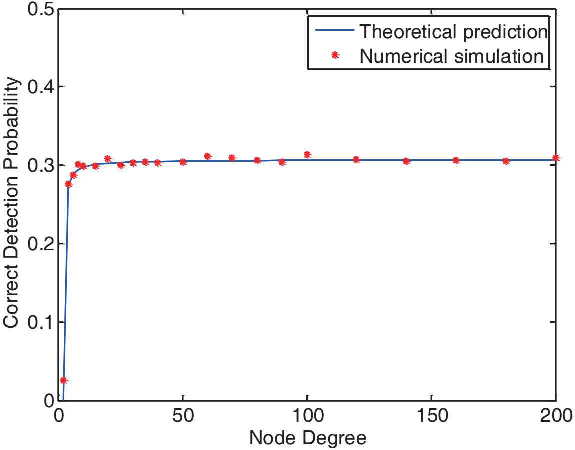
This process exactly describes the growth of the rumor boundary.



Performance results for the basic model

- For node degree $d \geq 3$, "non-trivial" detection: $\lim_{N \to \infty} P_c(\mathrm{G}_N) > 0$.
  - For $d = 2$, detection asymptotically impossible: $\lim_{N \to \infty} P_c(\mathrm{G}_N) = 0$.
- $\lim_{N \to \infty} P_c(\mathrm{G}_N) = d \cdot I_{1/2}\left(\frac{1}{d-2}, \frac{d-1}{d-2}\right) - d + 1$, where $I_x(\alpha, \beta)$ is the incomplete beta function.
  - $\lim_{N \to \infty} P_c(\mathrm{G}_N) \nearrow 0.307$ as $d \to \infty$.

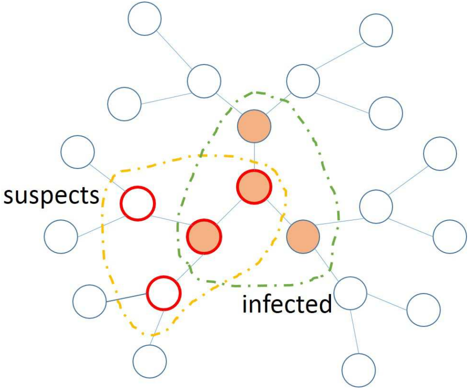
[ShaZamIT11, ShaZamSIGMETRICS12]



Theoretical prediction
Numerical simulation
Correct Detection Probability
Node Degree
0
0.1
0.2
0.3
0.4
0.5
0
50
100
150
200

Source detection with prior knowledge of suspects

- Usually we can not and need not suspect everyone.
- If only those who belong to a suspect set may initiate a spreading, how much can this prior knowledge help?



MAP detector and impact of local structure

- Assume a uniform prior over the suspect set S, $|S| = K$.
- MAP detector $= \arg\max R(v, G_N)$, over $v \in S \cap G_N$.
- Correct detection probability:

$$P_c(\mathrm{G}_N) = \frac{1}{K} \sum_{v \in S} P_c(\mathrm{G}_N | v),$$

  where $P_c(\mathrm{G}_N|v)$ is the correct detection probability conditioned upon that the source is v.
- Key: $P_c(\mathrm{G}_N|v)$ relies on the *local structure* of S,

$$P_c(\mathrm{G}_N|v) = 1 - m\left(1 - I_{1/2}\left(\frac{1}{d-2}, \frac{d-1}{d-2}\right) - \xi(N,d)\right),$$

w/ $\xi(N,d) \to 0$ as $N \to \infty$, and $m = |\mathrm{neighbor}(v) \cap S|$.

Performance results

- Connected S: for any $d \geq 3$,

$$\lim_{N\to\infty} P_c(\mathrm{G}_N) = 1 - \left(2 - \frac{2}{K}\right) \cdot \left(1 - I_{1/2}\left(\frac{1}{d-2}, \frac{d-1}{d-2}\right)\right).$$

- For $d = 2$, detection still asymptotically impossible.
- Better than no prior:
  - For any K, N, and $d \geq 3$: $P_c(\mathrm{G}_N)$ at least 0.5.
  - Asymptotically *reliable* detection (!): $\lim_{d\to\infty} \lim_{N\to\infty} P_c(\mathrm{G}_N) = 1$.
- Independent of the detailed structure of S.

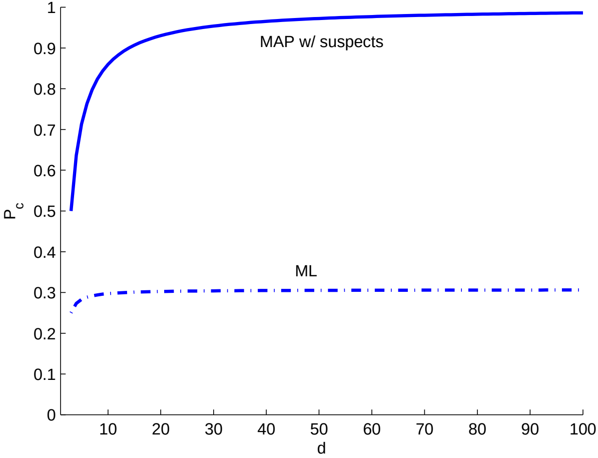
- Connected S leads to the smallest correct detection probability for a given K.

- More: "Rooting out the rumor culprit from suspects," (with W. Dong and C. W. Tan) Preprint; an extended abstract at ISIT 2013.

A closer look

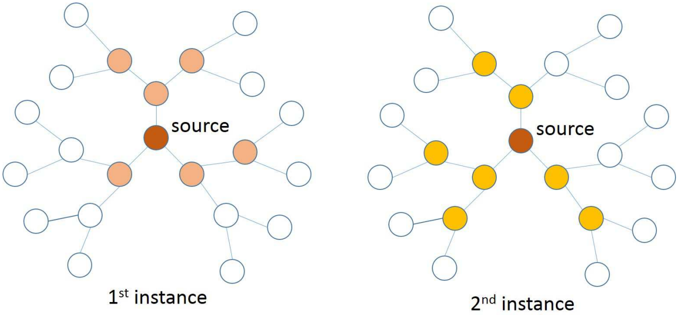
$$P_c(\mathrm{G}_N) = 1 - \left(2 - \frac{2}{K}\right) \cdot \left(1 - I_{1/2}\left(\frac{1}{d-2}, \frac{d-1}{d-2}\right) - \xi(N, d)\right).$$

- K and N can separately grow large, not depending upon each other, $P_c(\mathrm{G}_N) \to 2 \cdot I_{1/2}\left(\frac{1}{d-2}, \frac{d-1}{d-2}\right) - 1$.
- Does *not* degenerate into ML without prior suspects, for which $P_c(\mathrm{G}_N) \to d \cdot I_{1/2}\left(\frac{1}{d-2}, \frac{d-1}{d-2}\right) - d + 1$ (?!)
- The reason lies in the boundary of S, — if the source is located near the boundary, it is easily detected since it has few suspect neighbors. Then the performance is boosted as we average over S.


MAP w/ suspects
ML
P_c
d
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
10
20
30
40
50
60
70
80
90
100

Source detection with multiple instances

- A source may initiate multiple instances of spreading, rather than only once.
- If multiple instances are available, how much can this diversity help?



Union rumor center

- Assume L independent instances of infected sets G_{N_i}, $i = 1, \ldots, L$.
- ML detector $= \arg\max \prod_{i=1}^{L} R(v, G_{N_i})$, over $v \in \bigcap_{i=1,\ldots,L} G_{N_i}$. We call this the *union rumor center*.

Performance results

- For any $d \geq 3$, given L independent instances,

$$\lim_{N_1,\ldots N_L \to \infty} P_c = 1 - d\left(1 - \varphi_L\left(\frac{1}{d-2}, \frac{d-1}{d-2}\right)\right),$$

where $\varphi_L(\alpha,\beta) = \underset{\prod_{j=1}^{L} \frac{x_j}{1-x_j} \leq 1}{\int\cdots\int} \frac{\Gamma(\alpha+\beta)^L}{\Gamma(\alpha)^L \Gamma(\beta)^L} \prod_{j=1}^{L} x_j^{\alpha-1}(1-x_j)^{\beta-1} \mathrm{d}x_1 \cdots \mathrm{d}x_L.$

- For $d = 2$, detection still asymptotically impossible.
- *Reliable* detection with abundant connectivity (!): For any $L \geq 2$, $P_c \to 1$ as $d \to \infty$.
- *Reliable* detection with abundant diversity (!): $P_c \to 1$ as $L \to \infty$.

- Case of $d = 3$:

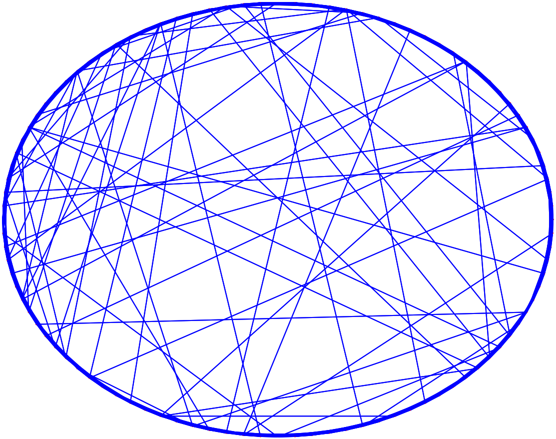
$$1 - \frac{3}{4}\left(\frac{\pi}{4}\right)^{L-1} < \lim_{N_1,\ldots N_L \to \infty} P_c < 1.$$

  Exponential convergence with L.
- Case of $L = 2$:
  - For $d = 3$, $\lim_{N_1,N_2 \to \infty} P_c = 1/2$;
  - for $d = 4$, $\lim_{N_1,N_2 \to \infty} P_c = 16/\pi^2 - 1 \approx 0.621$.

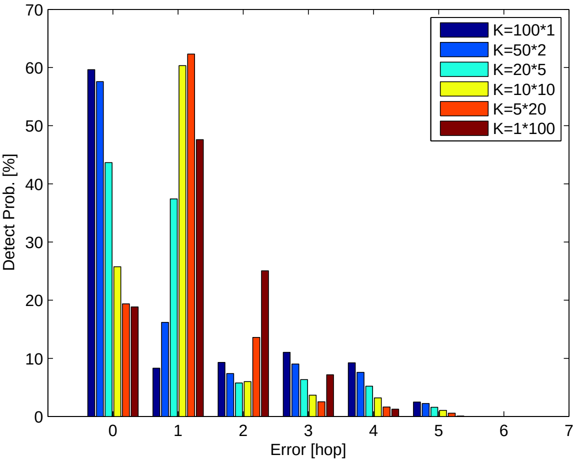
- More: "Rumor source detection with multiple observations: fundamental limits and algorithms," (with Z. Wang, W. Dong, and C. W. Tan) Preprint.

Experiments

Small-world networks (Watts-Strogatz model)



Detection performance with suspects:

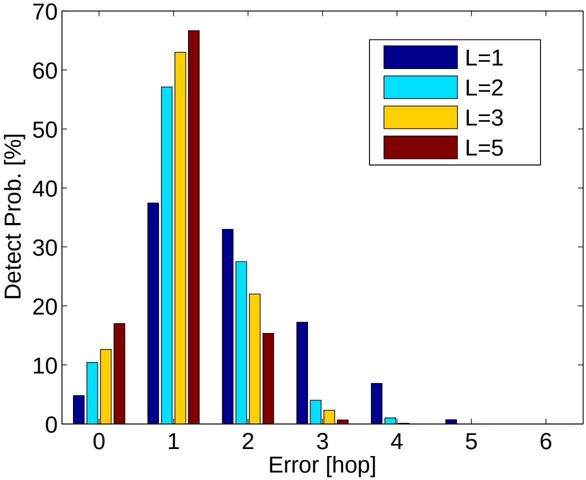


Error $=$ # hops away from the actual source

Network size $=$ 5000, infected set size $=$ 400.

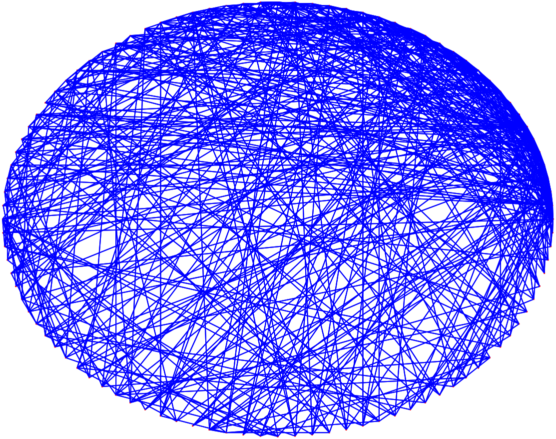
suspects K $=$ # clusters $\times$ # suspects per cluster.

Detection performance with multiple instances:

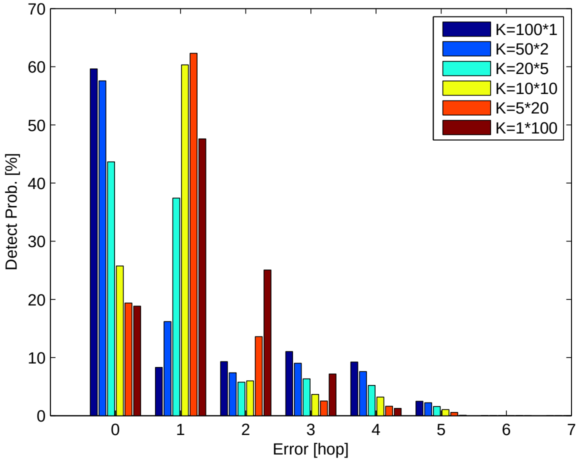


Network size = 5000, infected set size = 400.

Scale-free networks (Barabási-Albert model)



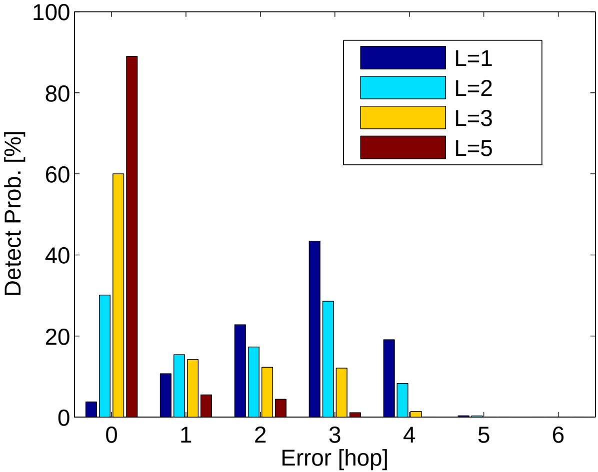
Detection performance with suspects:



Network size $=$ 5000, infected set size $=$ 400.

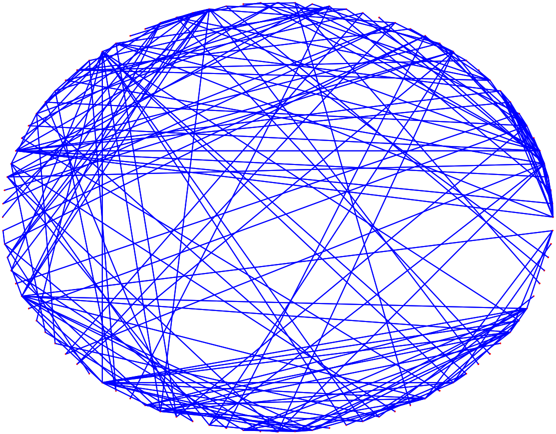
$\#$ suspects $K = \#$ clusters $\times$ $\#$ suspects per cluster.

Detection performance with multiple instances:



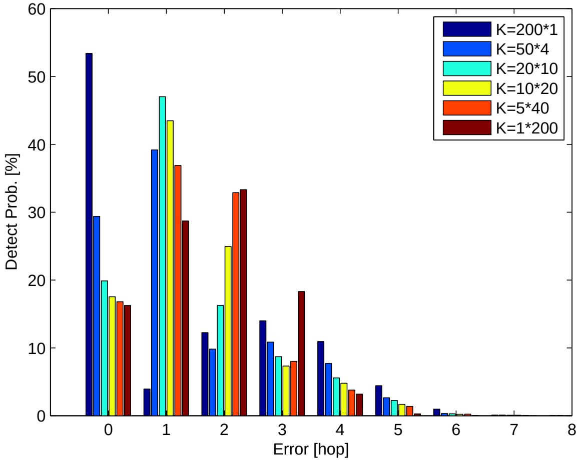
Network size $= 5000$, infected set size $= 400$.

Newman's scientific collaboration network dataset



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

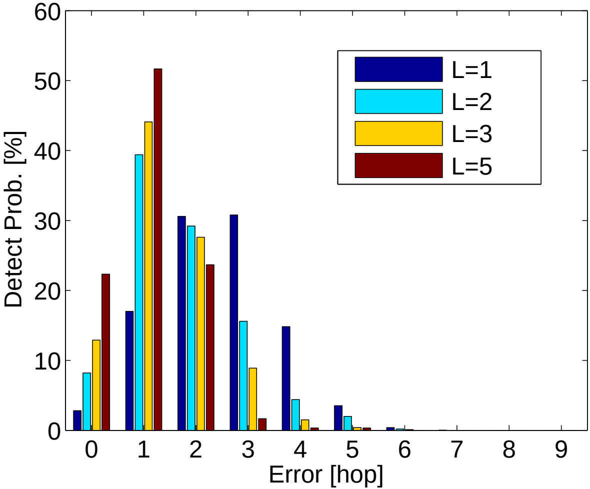
Detection performance with suspects:



Network size $= 13861$, infected set size $= 400$.

$\#$ suspects $K = \#$ clusters $\times$ $\#$ suspects per cluster.

Detection performance with multiple instances:



Network size = 13861, infected set size = 400.

Beyond and besides the basic model

- General tree, general graph:
  breadth-first-search (BFS) heuristic + RC. [ShaZamIT11]
- General infection time distribution, general random tree:
  RC still achieves *non-trivial* detection (*universal* detector).
  [ShaZamSIGMETRICS12]
- Limited (maybe sparse) observations:
  [PinThiVetPRL12], [KarFraISIT13], [LuoTayLenArxiv13]
- Multiple sources:
  [LuoTayLenSP13]
- SIS or SIR infection processes:
  [LuoTayICASSP13], [ZhuYinITA13]
- Other related models and algorithms:
  [PraVreFalICDM12], [LokMézOhtZdeArxiv13], [AntLanSteSikSmuArxiv13]

Wrap-up remarks

- Inference over networks is a fast emerging area merging networking, signal processing, and statistics.
- The basic model of source detection provides an ideal playground for gaining key insights into more realistic scenarios.
- Prior knowledge and diversity are powerful performance boosters.

"Rooting out the rumor culprit from suspects," (with W. Dong and C. W. Tan) Preprint; an extended abstract at ISIT 2013.
"Rumor source detection with multiple observations: fundamental limits and algorithms," (with Z. Wang, W. Dong, and C. W. Tan) Preprint.

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.

[LokMézOhtZdeArxiv13] A. Y. Lokhov, M. Mézard, H. Ohta, and L. Zdeborová, "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.

Still a long march towards a full-grown theory capable of handling the incredible reality...