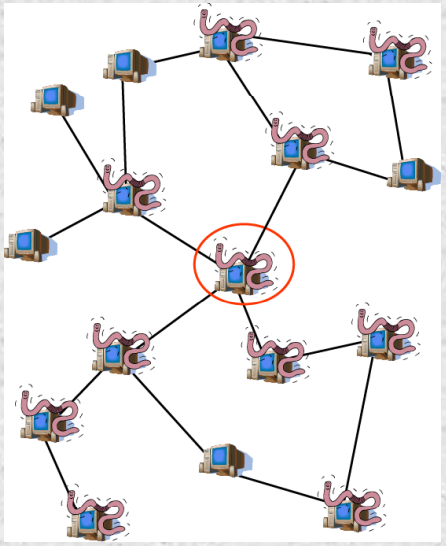
Information Diffusion: Searching for Source

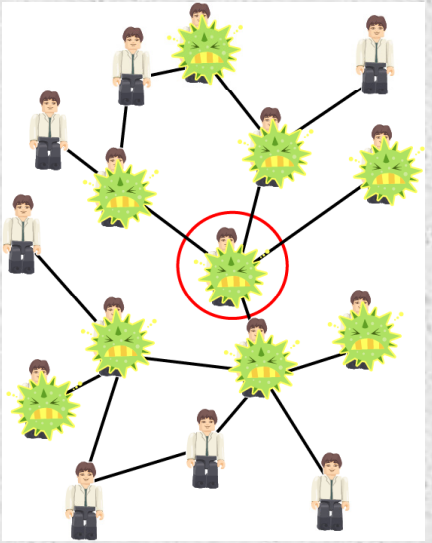
Devavrat Shah Tauhid Zaman

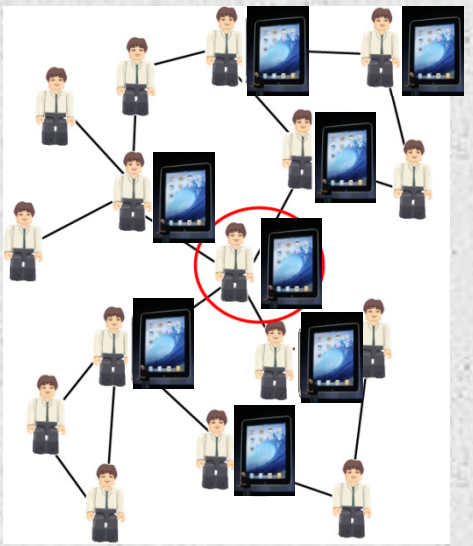
Laboratory for Information & Decision Systems
Massachusetts Institute of Technology

Information diffusion

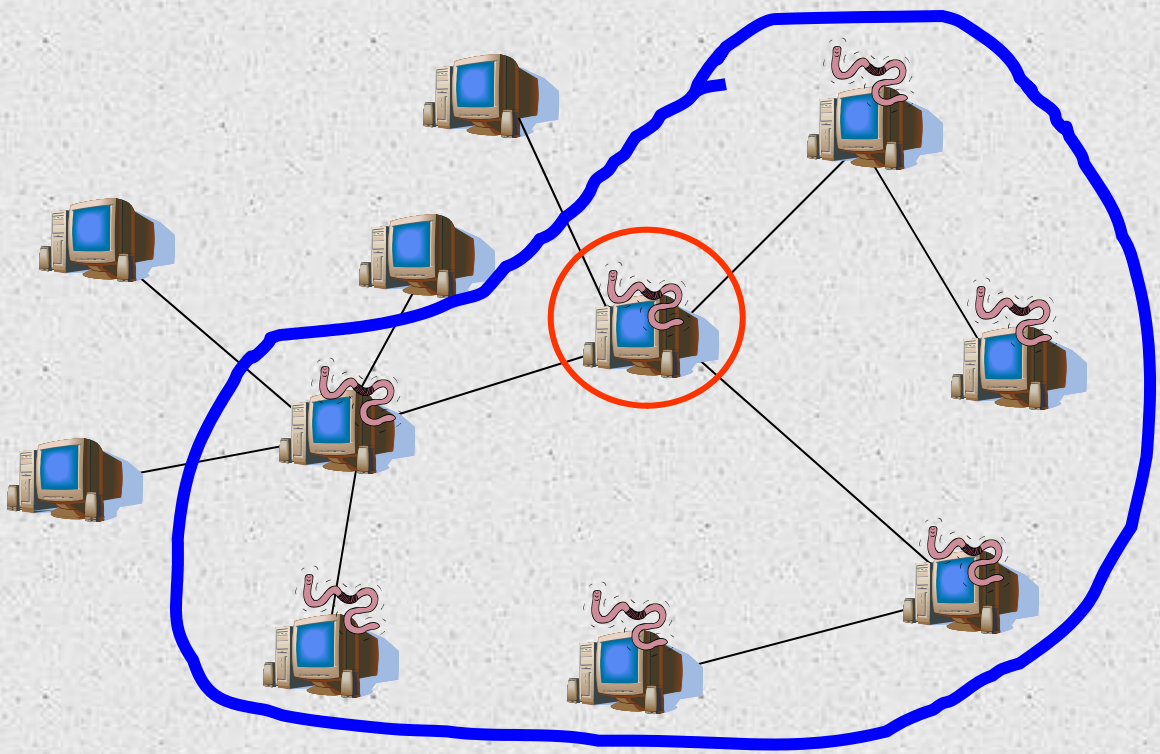
- Unstructured propagation of “information”
  - Virus in the Internet
  - Epidemic of disease
  - Trend in society







Stuxnet: who did it?

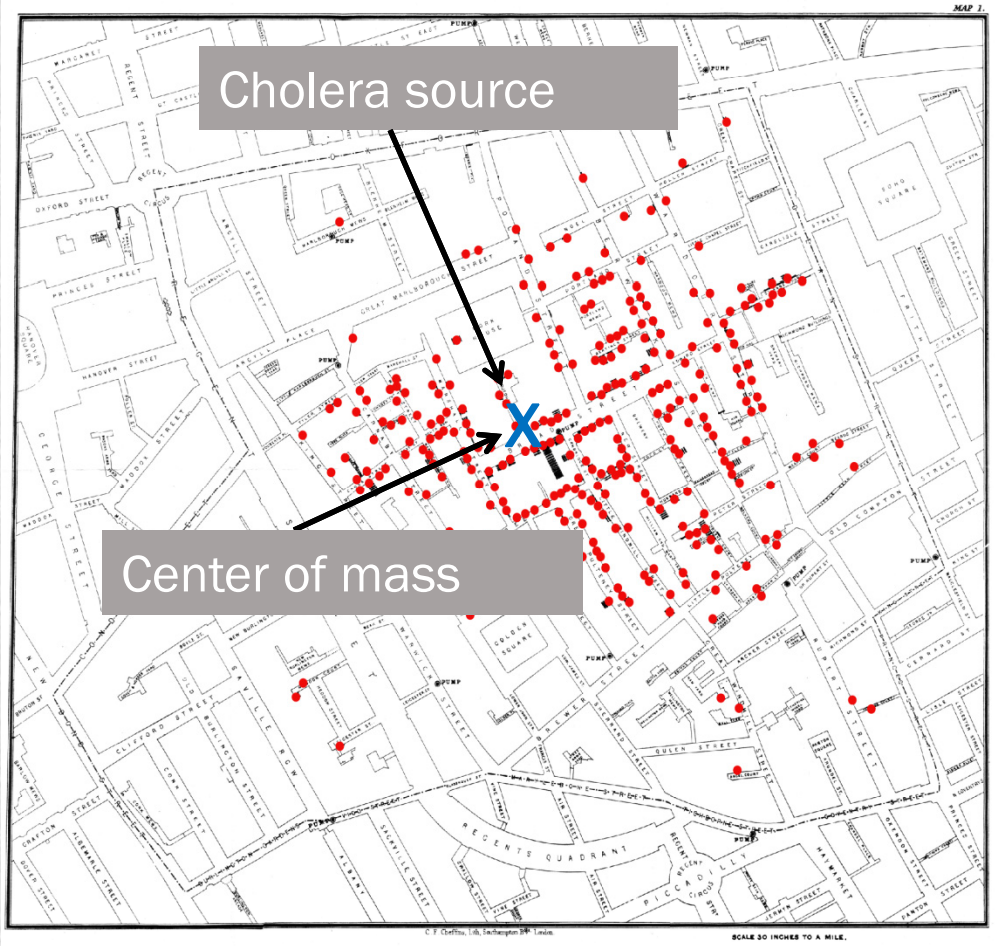


Information diffusion: the culprit?

- Stuxnet
  - *Who* did it ?
  - Or, can *it* remain *hidden* ?

- Basic questions of interest:
  - What aspects of network structure
    - Determine the ability to find the source
  - More generally, when can
    - “Important” agents be found
      - using network structure

1854 London Cholera Epidemic

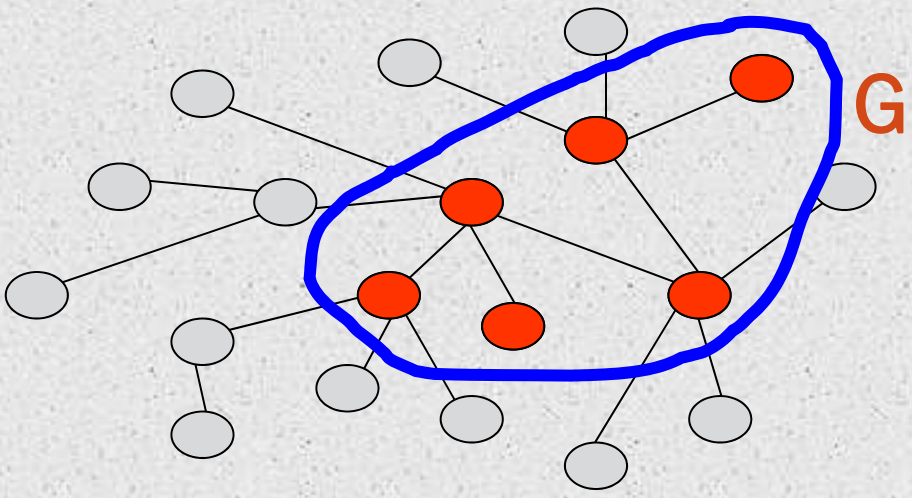


Dr. John Snow

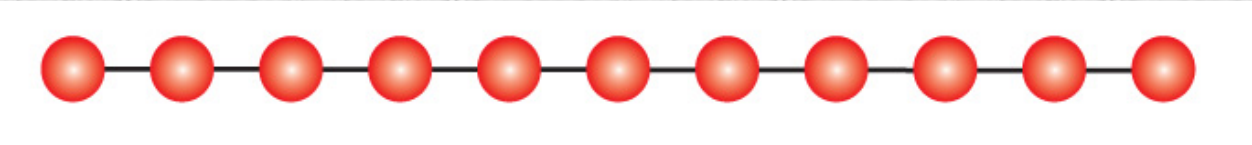
How do we formally extend this center idea to networks?

Source estimation

- We know the *infection* graph G
  - We want to find source node
- Basic question:
  - When can we find it?
  - What role does network structure play?



An example



C_t = correct source detection at time *t*

- If G is a line graph, then for *any* algo.

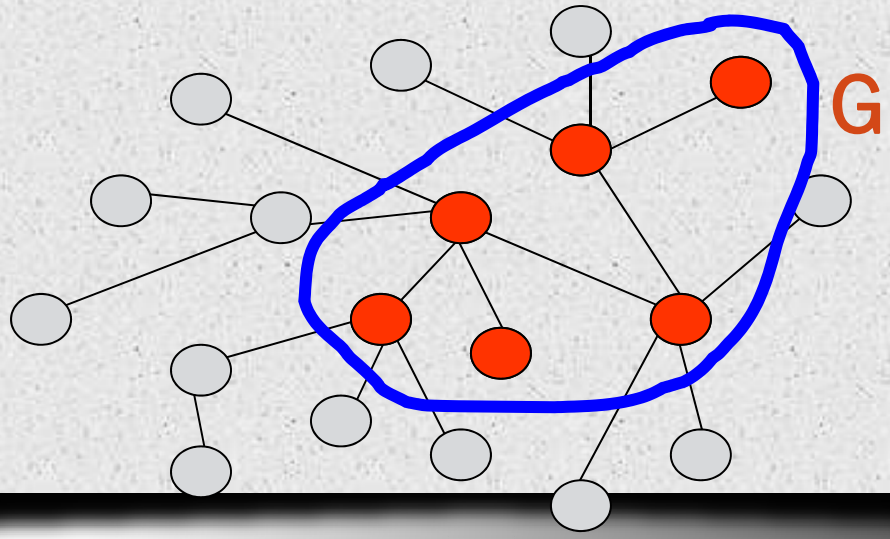
$$P(C_t) = O\left(1 \big/ \sqrt{t} \right) \to 0$$

  - *Needle in a haystack !*
- Is there any hope for other network to allow detection?
  - If so, when ?

Source estimation

- Observation
  - Rumor infected graph G
- Want to find source v* such that

$$v^* = \text{argmax } P(\text{source}=v \mid G)$$

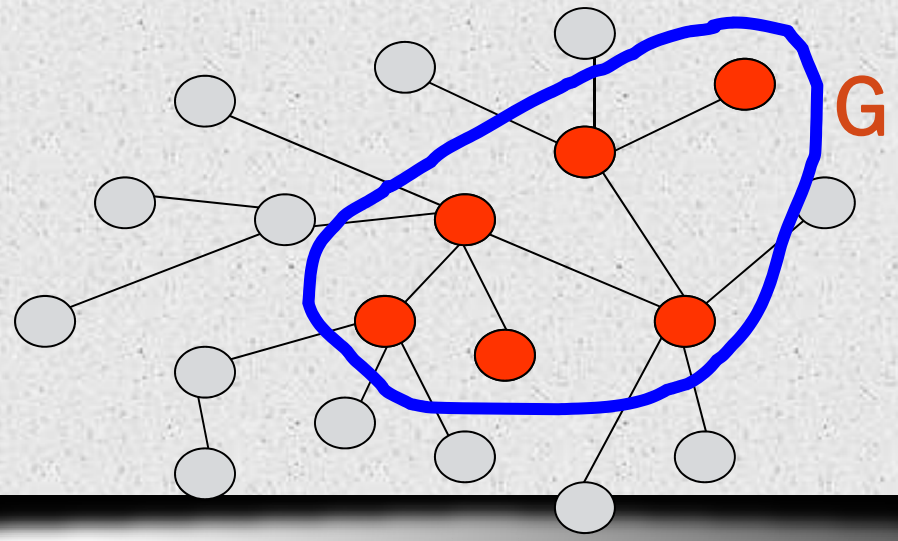


Source estimation

- Observation
  - Rumor infected graph G
- Want to find source v* such that

$$v^* = \text{argmax } P(\text{source}=v|G)$$

$$v^* = \text{argmax } P(G|\text{source}=v)P(\text{source}=v)$$

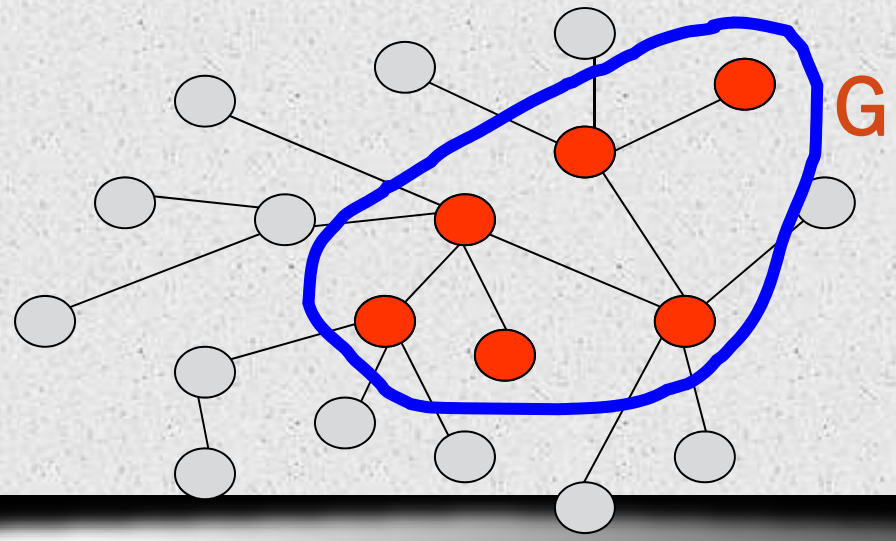


Source estimation

- Observation
  - Rumor infected graph G
- Want to find source v* such that

$$v^* = \text{argmax } P(\text{source}=v \mid G)$$

$$v^* = \text{argmax } P(G \mid \text{source}=v)$$



Source estimation

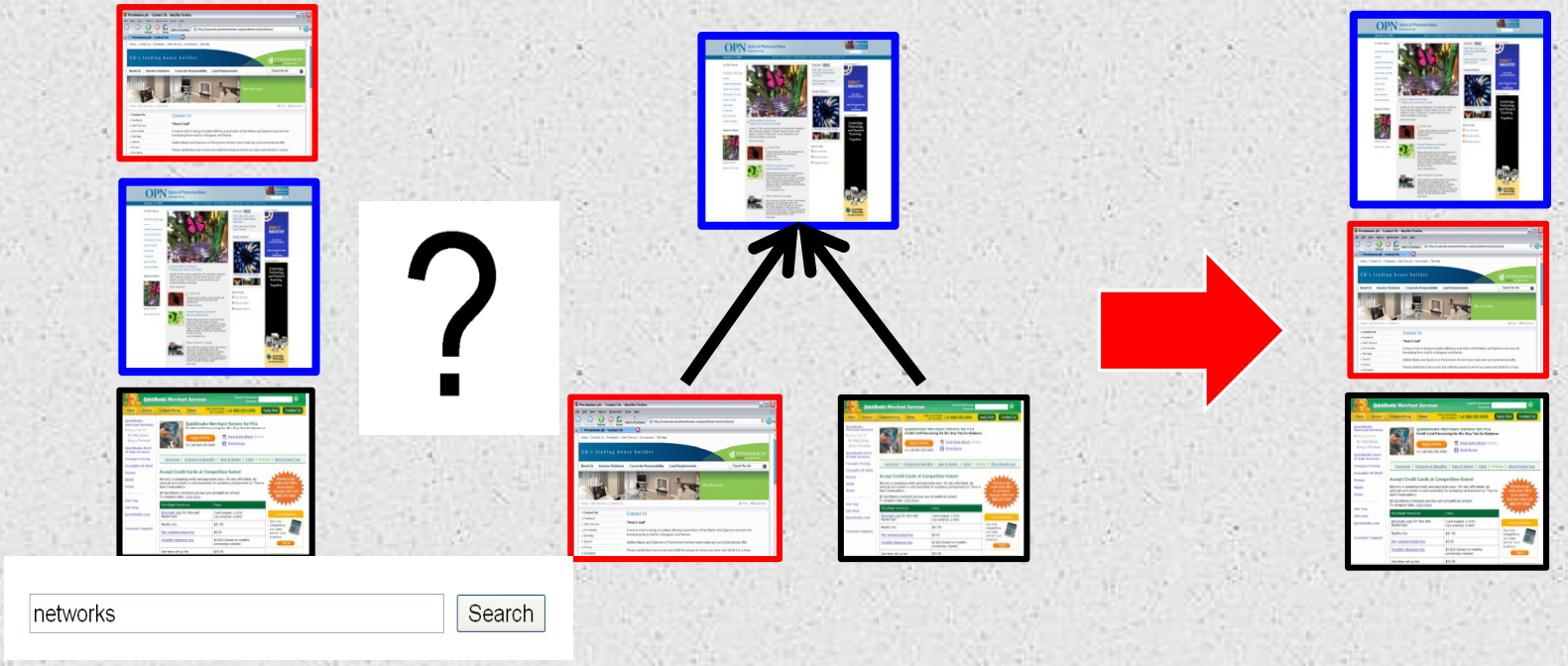
- Goal: find

$$v^* = \text{argmax } P(G|source=v)$$

- Procedure: given *infected* network topology G
  - Find "score" for each node v: P(G|source = v)
  - And choose v with max "score"

- Equivalently, find *network centrality* that captures
  - Importance of v in terms it being the source
  - For example
    - *PageRank* in WWW for web-search

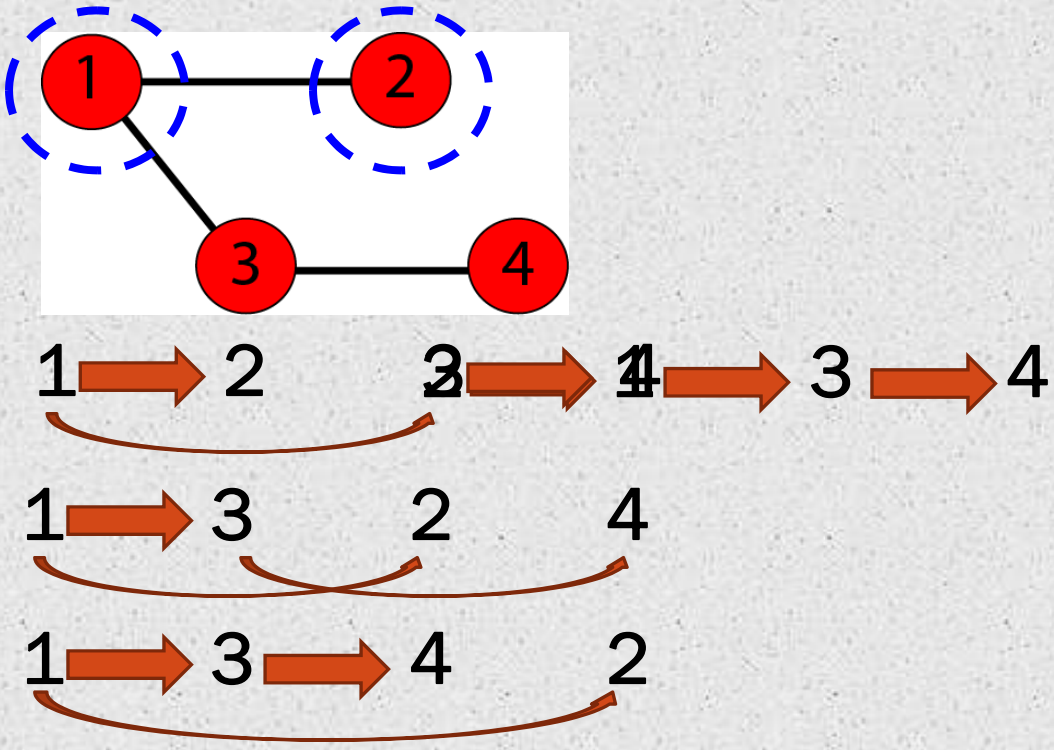
Example: Web search



- Rank pages using network centrality = *PageRank*
  - It works *well* (everyone *knows* that)
  - But where does it all come from (what's the science)

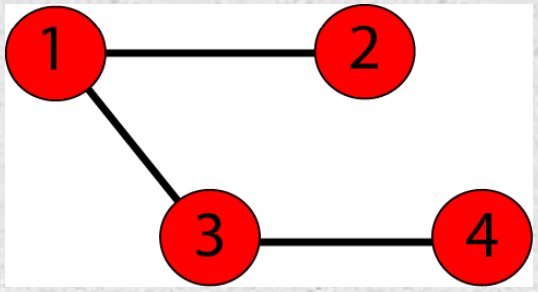
Rumor Spreading Order

- Rumor spreading order not known
- Only spreading constraints are available



More spreading orders = more likely to be source

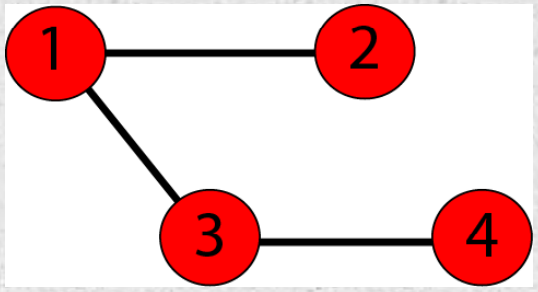
Rumor Spreading Order



More spreading orders = more likely to be source

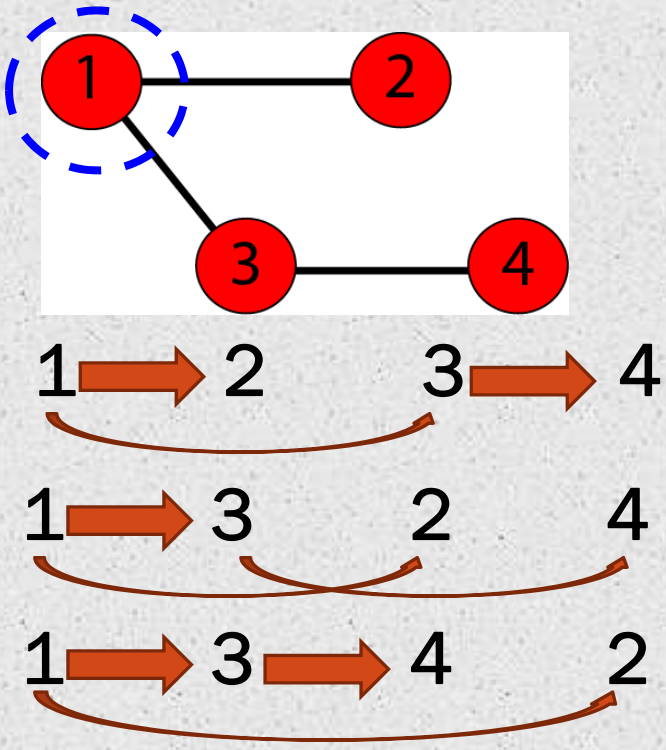
Rumor Spreading Order

- Rumor spreading order not known
- Only spreading constraints are available



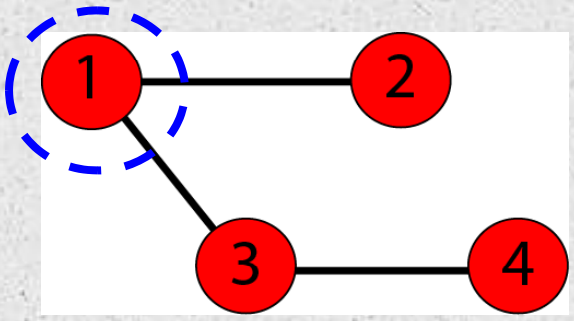
Rumor Spreading Order

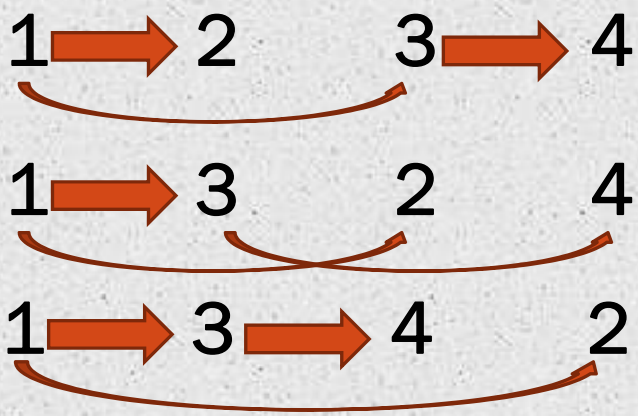
- Rumor spreading order not known
- Only spreading constraints are available



Rumor Spreading Order

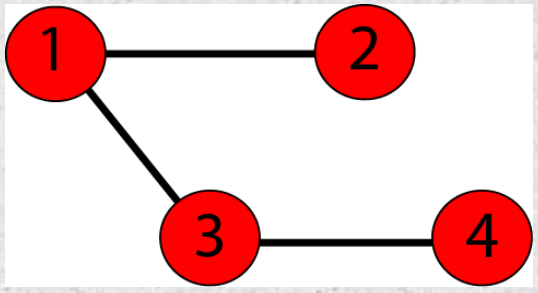
- Rumor spreading order not known
- Only spreading constraints are available

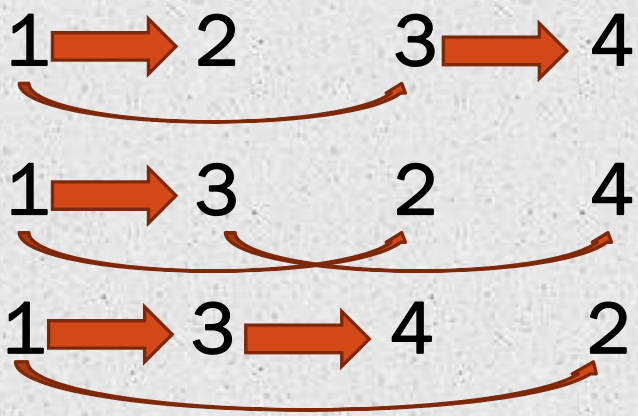




Rumor Spreading Order

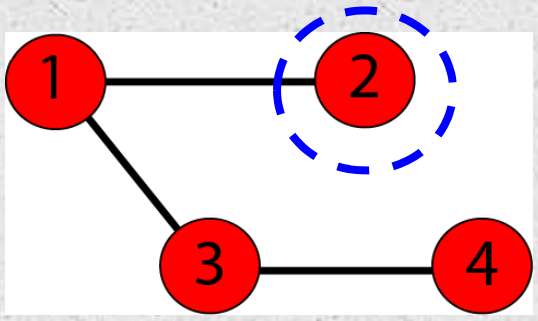
- Rumor spreading order not known
- Only spreading constraints are available

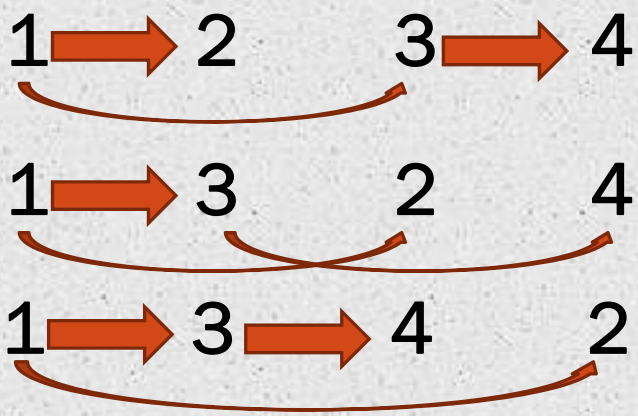




Rumor Spreading Order

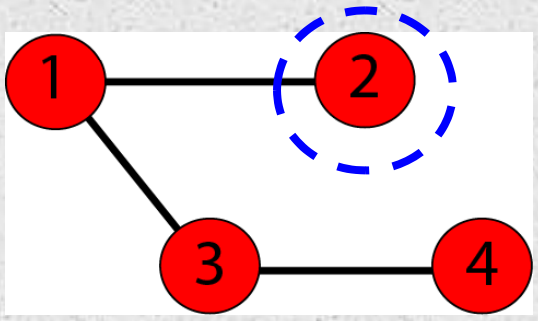
- Rumor spreading order not known
- Only spreading constraints are available





Rumor Spreading Order

- Rumor spreading order not known
- Only spreading constraints are available



1 → 2 3 → 4

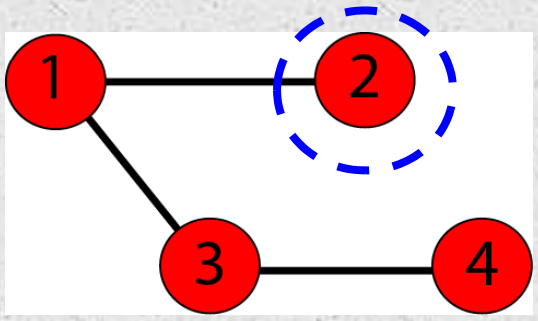
2 → 1 → 3 → 4

1 → 3 2 4

1 → 3 → 4 2

Rumor Spreading Order

- Rumor spreading order not known
- Only spreading constraints are available



1 → 2 3 → 4

2 → 1 → 3 → 4

1 → 3 2 4

1 → 3 → 4 2

More spreading orders = more likely to be source

Rumor centrality (Shah, Zaman '09)

- New centrality: counts number of spreading orders
  - Given (tree) graph G, the centrality of node v is

$$R(\mathrm{v},\mathrm{G}) = \frac{N!}{\prod_{k\in\mathrm{G}} T_k^{\mathrm{v}}}$$

nodes in G

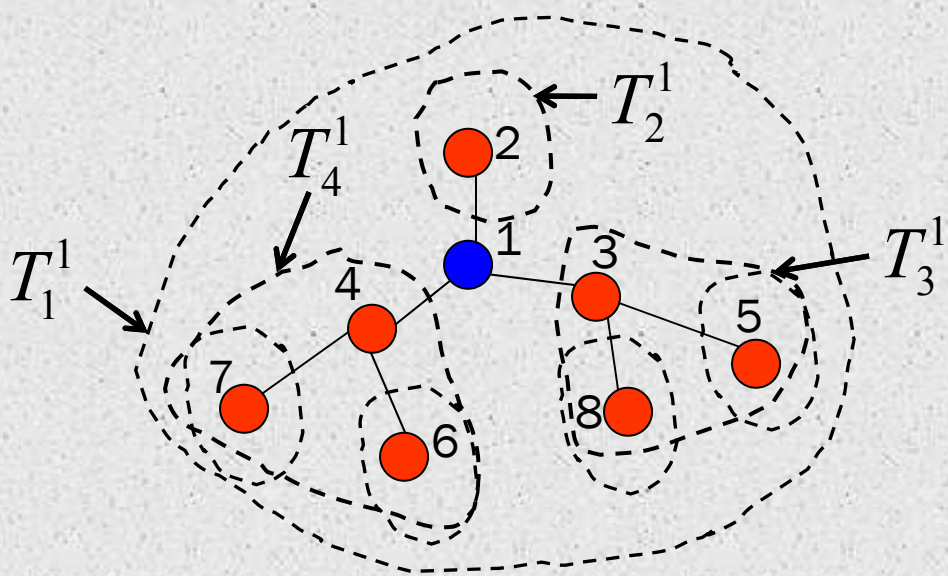
sub-trees of G

  - Linear extensions of a partial order imposed by G
  - Can be computed in linear time (N)
    - In a parallel, distribute manner

- It is the Maximum Likelihood (ML) estimator
  - For regular trees

An example

- Compute $R(1, G)$: N = 8



$$R(1,G) = \frac{8!}{8*1*3*3*1*1*1*1}$$

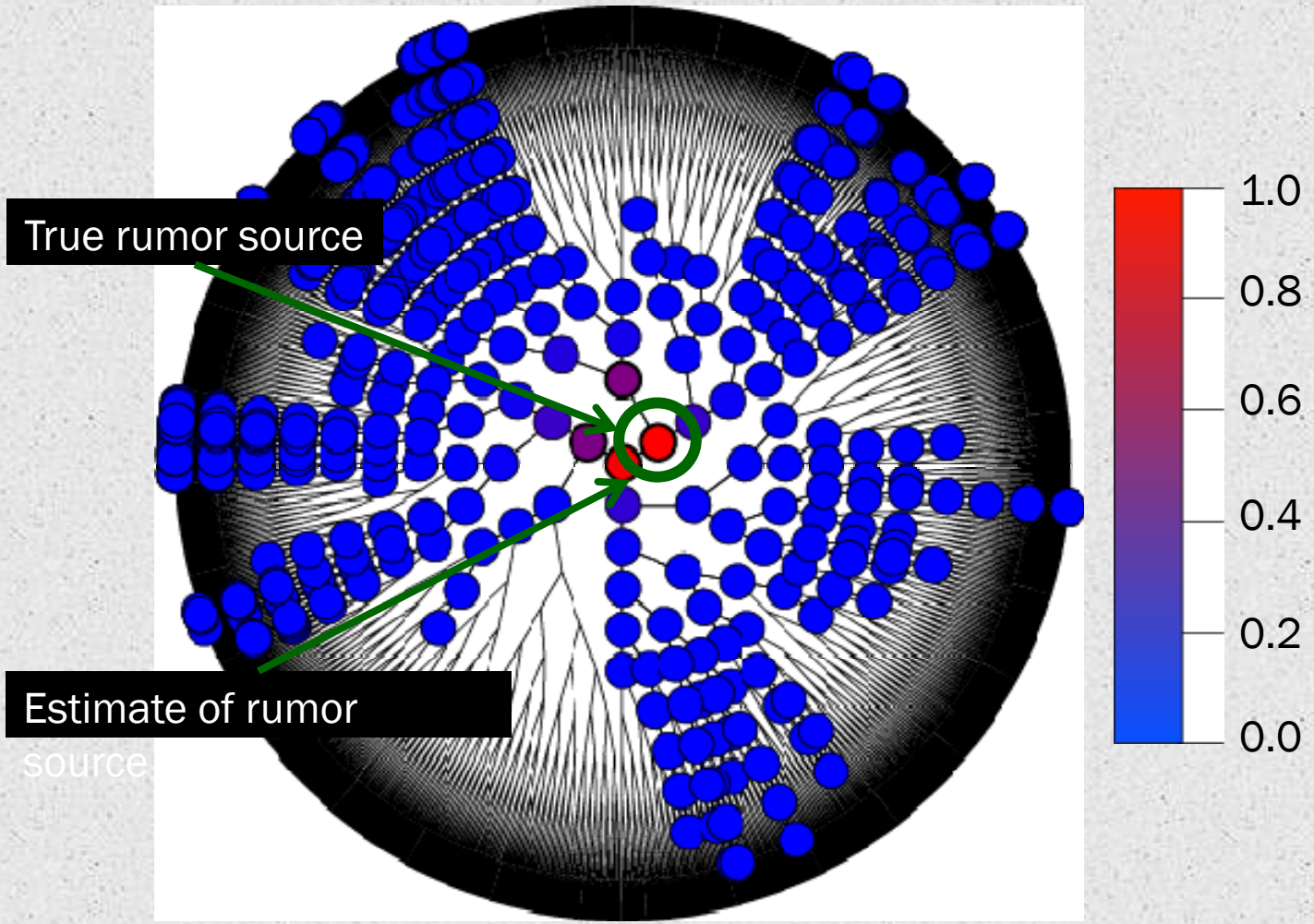
$$R(\mathrm{v},\mathrm{G}) = \frac{N!}{\prod_{k \in \mathrm{G}} T_k^{\mathrm{v}}}$$

Source estimation: rumor centrality

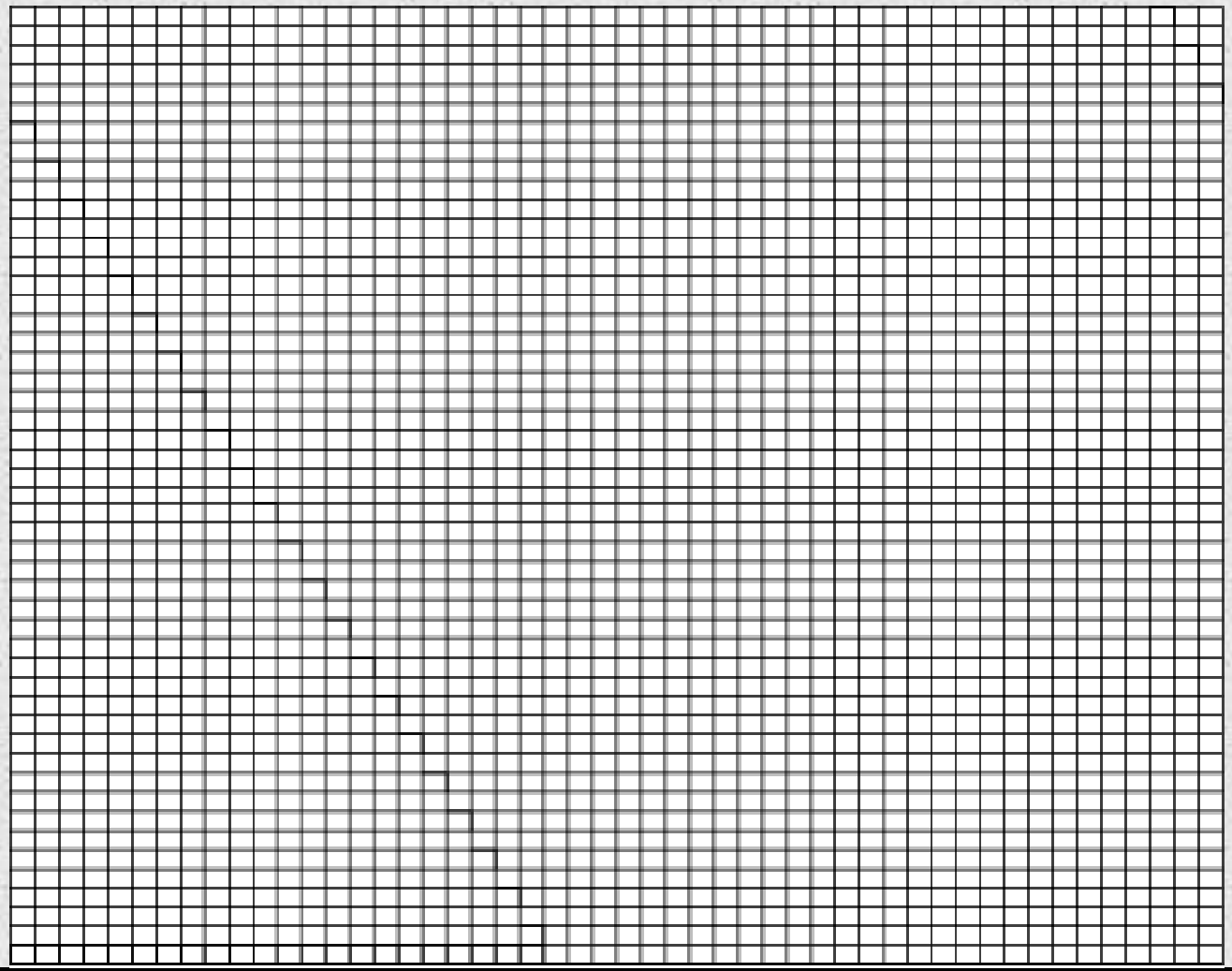
- Rumor centrality is ML
  - We *know* that for line graph it does *not* work
  - Can it work for other graph

  - Let us consider regular tree with degree 3
    - Line is degree 2 : each node has two neighbors

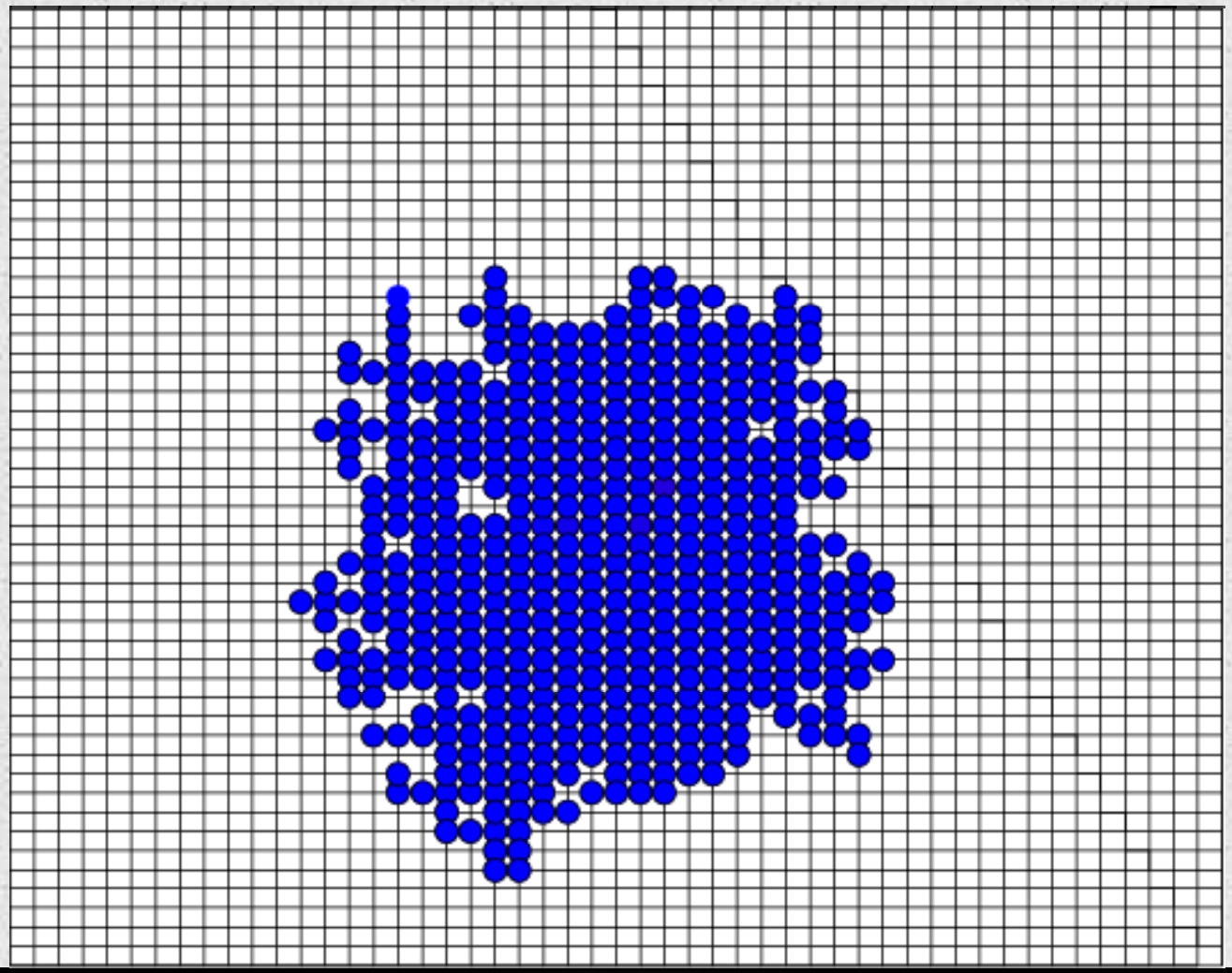
Precision for regular tree: deg=3



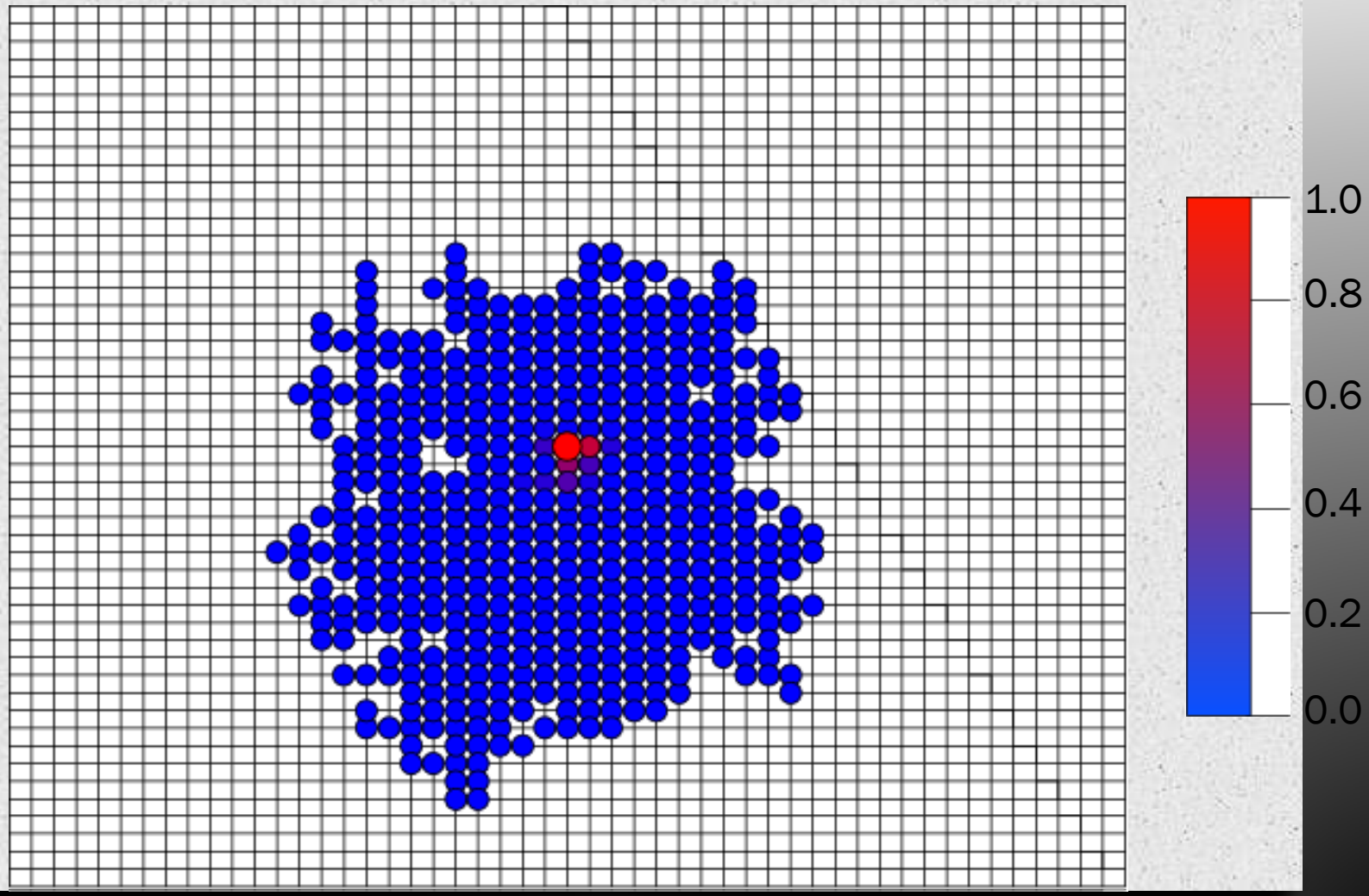
Precision of Rumor Centrality: General network



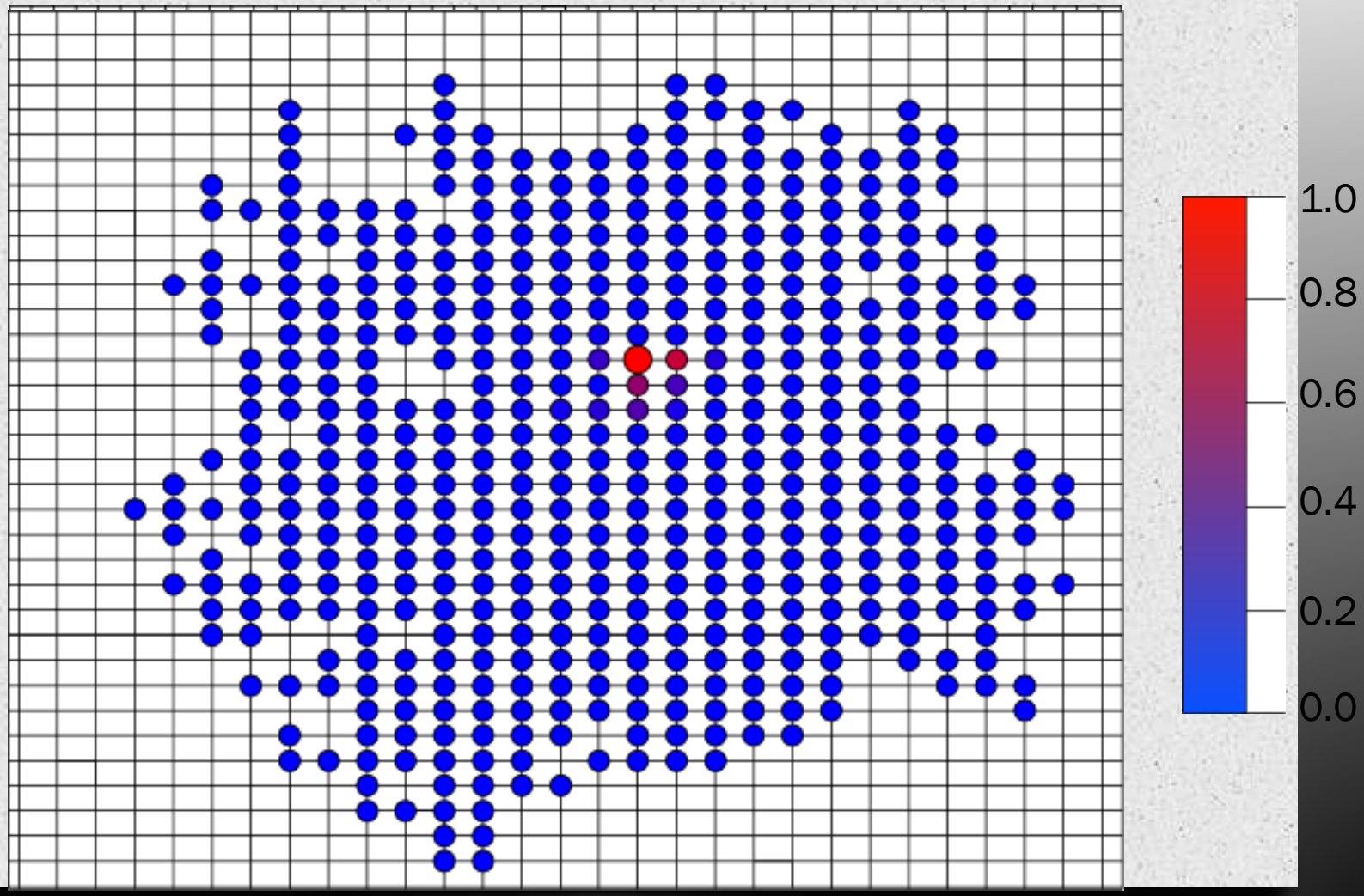
Precision of Rumor Centrality: General network



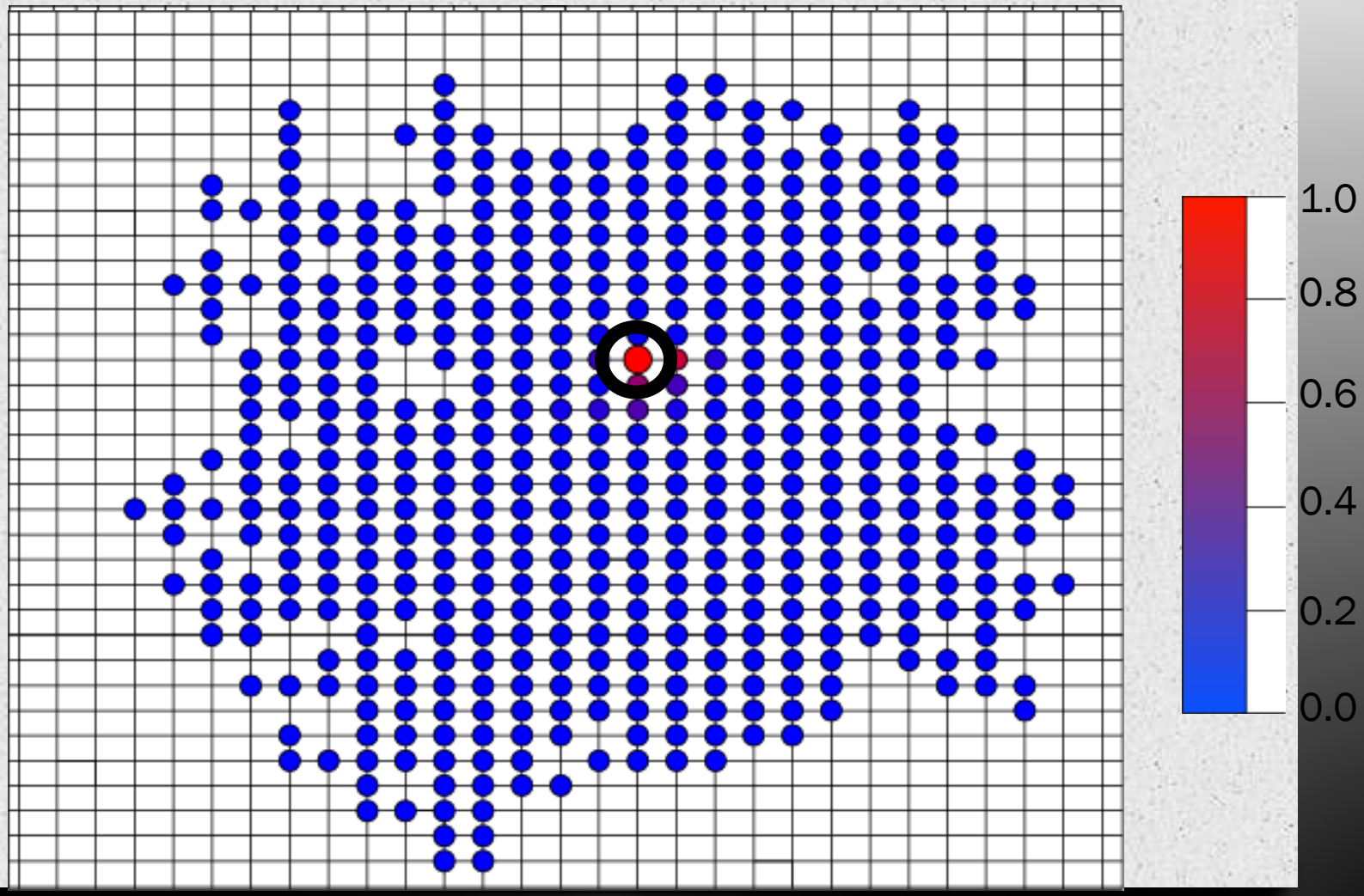
Precision of Rumor Centrality: General network



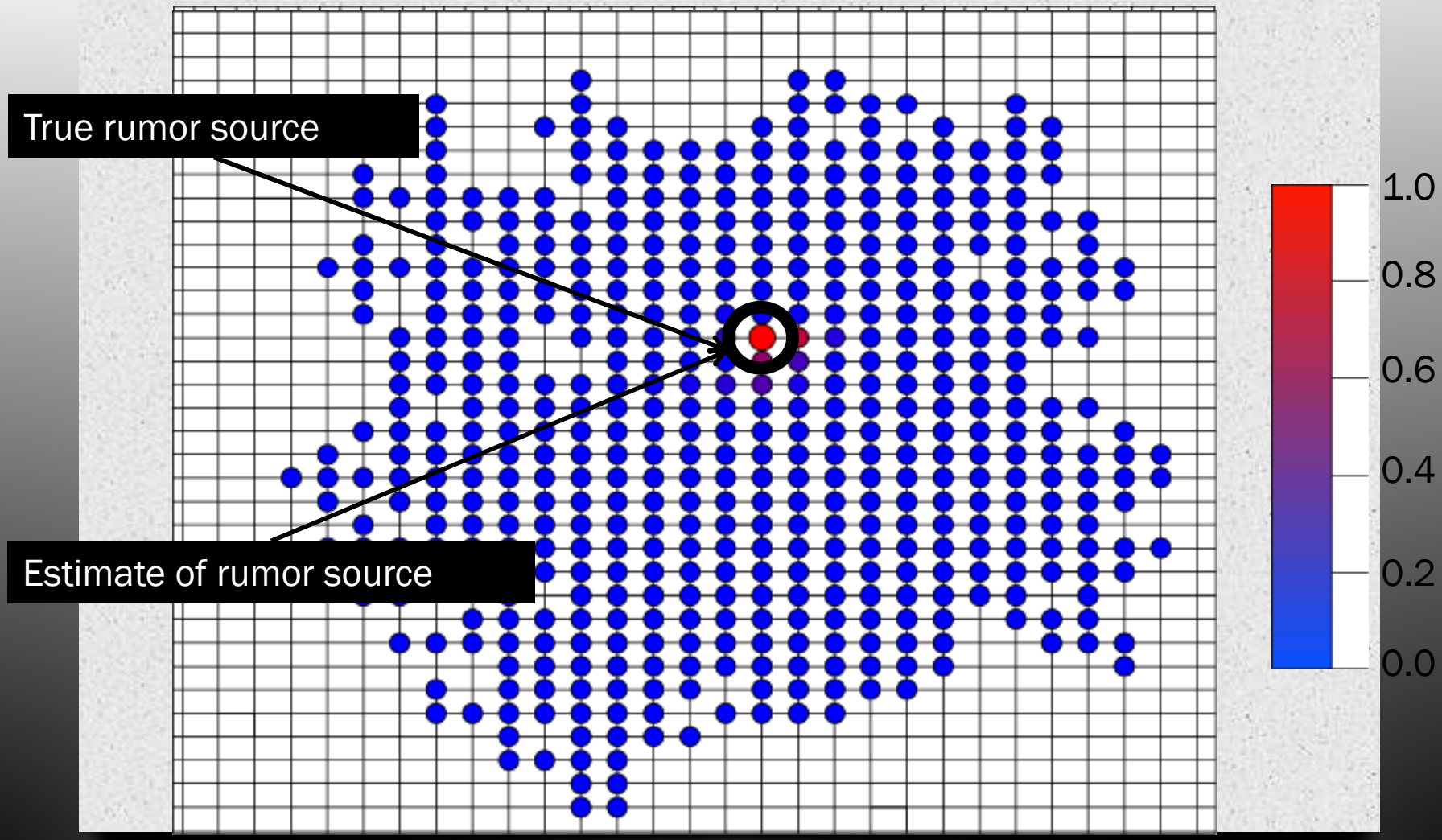
Precision of Rumor Centrality: General network



Precision of Rumor Centrality: General network



Precision of Rumor Centrality: General network

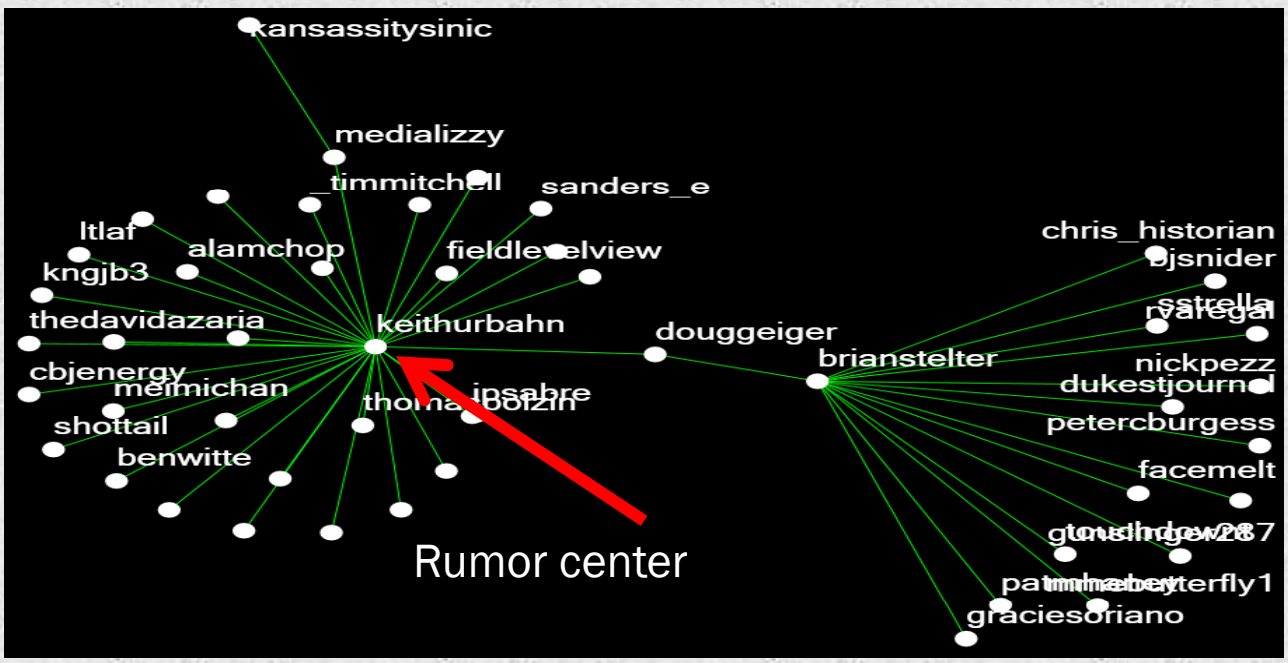


More structure = better source detection

Twitter: identifying source of rumor

- Who leaked “Bin-Laden” death news before official announcement:
  - Keith Urbahn: the first to tweet about the death of Osama bin Laden

keithurbahn Keith Urbahn
So I'm told by a reputable person they have killed Osama Bin Laden. Hot damn.
1 May



Influence tracking engine

TRUMOR

Query: syria

Start Date: 06-01-2011 Stop Date: 07-05-2011

Trumor Search

Trumor found 1012 users tweeting about **syria** from **Wednesday, June 01, 2011** to **Tuesday, July 05, 2011**. Took 48.421 seconds

_syriatweet: Trumor score = 243.66

1. **July 1, 2011, 7:09 p.m. Tweet:** RT @Syria2daycom: #Syria 2day | تجدد المظاهرات في مدن سورية .. وأنباء عن سقوط ضحايا http://tinyurl.com/3hmv2dp
2. **July 1, 2011, 3:06 a.m. Tweet:** RT @SyrianFreePress: #SPF #Syria مظاهرات كفرنبل احتجاجا على دخول الجيش لجبل الزاوية 29 6 2011 http://ow.ly/5umv6
3. **June 26, 2011, 5:06 a.m. Tweet:** RT @nalsowaidi: 1.Join the demonstration in London to condemn Iran's crimes against Humanity in Syria, Bahrain, Ahwaz

_syrianfreepress: Trumor score = 143.27

1. **July 1, 2011, 3:06 a.m. Tweet:** #spf #syria مظاهرات كفرنبل احتجاجا على دخول الجيش لجبل الزاوية 29 6 2011 http://ow.ly/5umv6
2. **June 26, 2011, 5:48 a.m. Tweet:** #sfp #syria دوما اضراب عام في المدينة الخميس 23-6-2011 http://ow.ly/5qf45
3. **June 24, 2011, 11:02 a.m. Tweet:** #sfp #syria أحرار المعضمية يردون على خطاب السفاح با بشار و با كذاب تضرب أنت و هالخطاب 20 6 2011 http://ow.ly/5pgfz

_antigovsyrians: Trumor score = 135.69

1. **June 25, 2011, 2:07 p.m. Tweet:** @halagorani thank you for going into #syria please speak to anti regime protestors to expose #assadcrimes #sy11

Trumor vs. Twitter

TRUMOR

Query: S&P

Start Date: 08-06-2011 Stop Date: 08-07-2011

Trumor Search

Trumor found 485 users tweeting about **S&P** from **08-06-2011** to **08-07-2011**. Took 0.703 seconds

BreakingNews: Trumor score = 19.0

Aug. 6, 2011, 12:22 a.m. Tweet: Credit rating agency S&P downgrades US debt from AAA, first debt downgrade in US history - AP

Aug. 6, 2011, 12:28 a.m. Tweet: More: S&P downgrade of US credit is 1 notch to AA-plus from AAA http://reut.rs/rbl8sC

Aug. 6, 2011, 1:01 a.m. Tweet: More from S&P on debt downgrade: Effectiveness, stability, predictability of US policymaking weakened http://on.wsj.com/oSB3jQ

thinkprogress: Trumor score = 18.0

Aug. 6, 2011, 12:25 a.m. Tweet: This is the first downgrade in U.S. history. S&P reportedly pins blame on GOP for intransigence on taxes. http://thkpr.gs/ogQJgI

Aug. 6, 2011, 12:27 a.m. Tweet: Reuters reports that S&P signaled "another downgrade is possible in the next 12 to 18 months." http://thkpr.gs/necATf

Aug. 6, 2011, 12:40 a.m. Tweet: S+P: "The effectiveness, stability, and predictability of American policymaking and political institutions have weakened." #teaparty

Aug. 6, 2011, 12:41 a.m. Tweet: S+P cites "the prolonged controversy over raising the statutory debt ceiling" as a reason for downgrade. #thanksgop

twitter S&P

Results for S&P

Tweets · Top Refine results »

20 new tweets

tinyrevolution Jonathan Schwarz
Possible ways to get our AAA back: (1) raise taxes; (2) cut spending; or (3) just start paying **S&P** for ratings the way Wall Street does
6 Aug
Retweeted 100+ times

NYTimeskrugman Paul Krugman
S&P Precedents http://nyti.ms/qSMJ3o
11 minutes ago

DougKass Douglas Kass
one of the big stories this week will be the lost credibility of **s&p** - because of their political spin and math miscalcullations -
11 minutes ago

EricBoehlert Eric Boehlert
prob solved....RT @thinkprogress: S+P wanted $4T in deficit reduction over 10 yrs. Allowing Bush tax cuts 2 expire would save $5.4T.
22 minutes ago

OTOOLEFAN Don Millard
RT @thinkprogress: S+P wanted $4T in deficit reduction over 10 yrs. Allowing Bush tax cuts 2 expire would save $5.4T. http://thkpr.gs/nrtDp7
29 minutes ago

NewYorker The New Yorker
John Cassidy: Why S. & P. Got the U.S. Government Downgrade Right: nyr.kr/oitOBk
37 minutes ago

mediaite Mediaite
Chris Wallace Challenges Paul Ryan On Whether 'Failing To Compromise' Led To **S&P** Downgrade (VIDEO) mediaite.com/a/gozpj JN
2 hours ago

Rumor centrality: theory

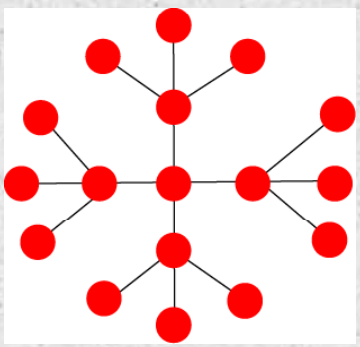
Rumor centrality: d-regular trees

- Number of nodes distance *r* from source
  - Grows as $(d-1)^r$: ***exponential*** growth
  - Leads to *high* variation in "rumor graph"

- **Theorem** [S., Zaman]: For a regular tree with **d>2**, if the rumor spreads for time *t*, then

$$\lim_t P(C_t) = \alpha_d > 0.$$

where $\alpha_d = d\, I_{1/2}\left(\frac{1}{d-2}, 1+\frac{1}{d-2}\right) - (d-1)$

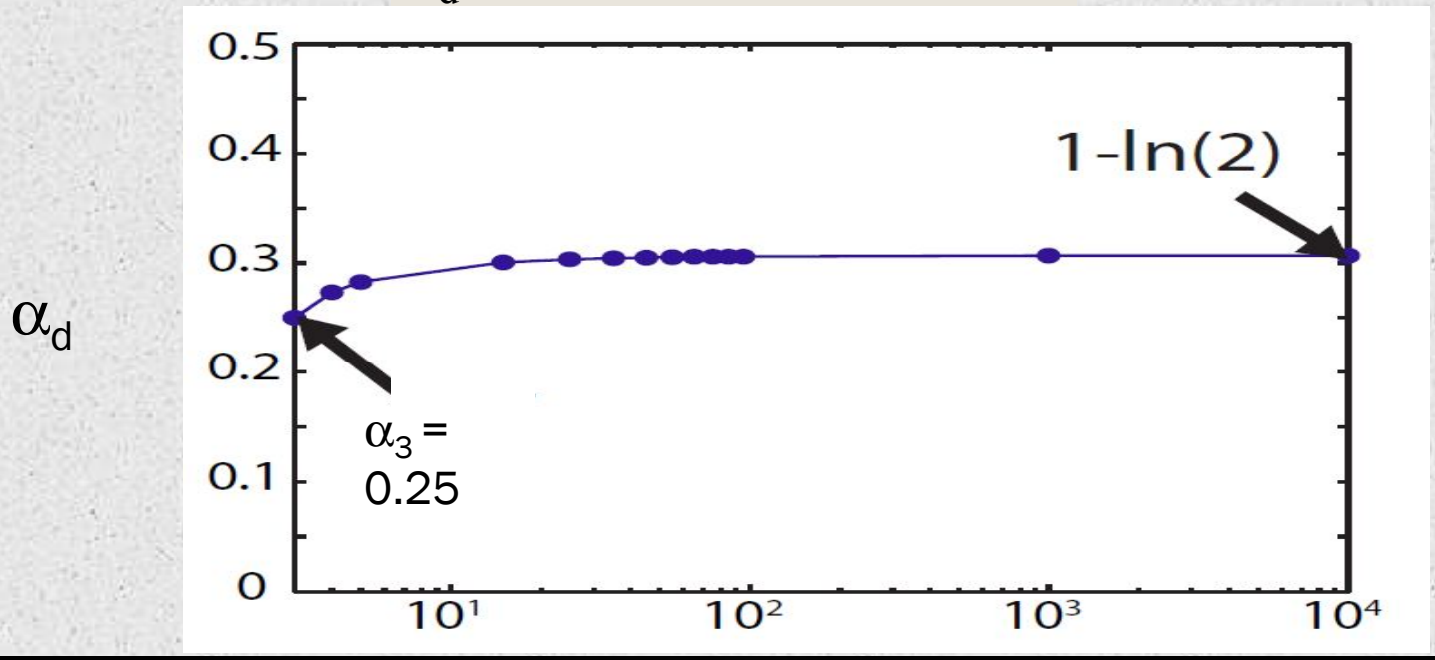


with $I_x(a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_0^x t^{a-1}(1-t)^{b-1}\,dt$

Rumor centrality: d-regular trees

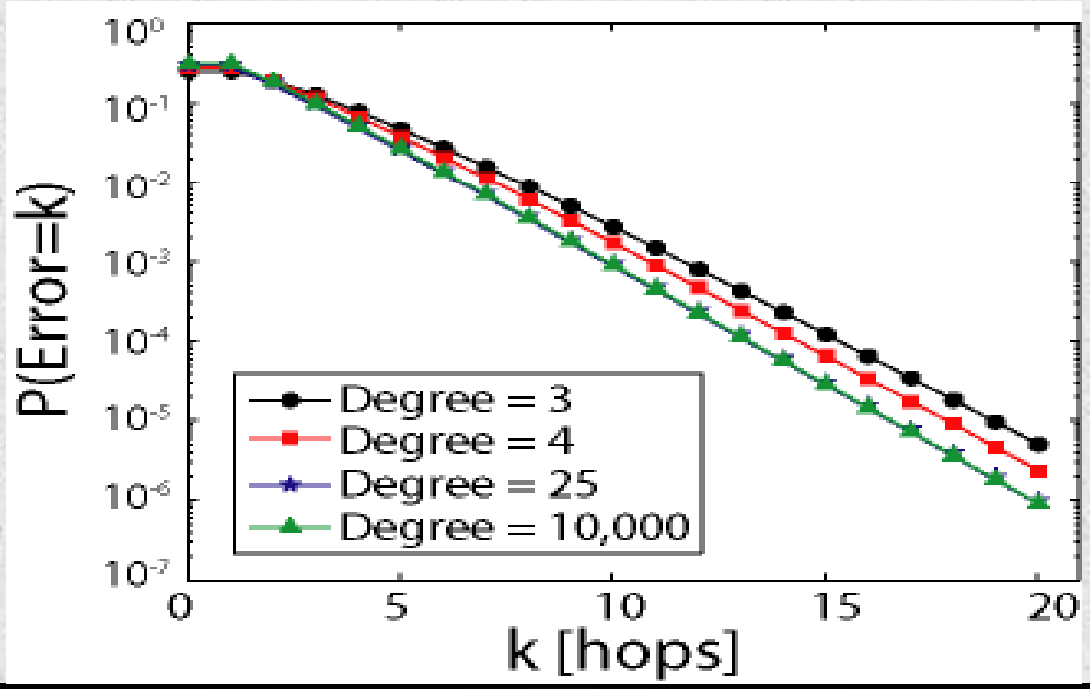
- Some implications:
  - For d = 3: $\alpha_3 = 0.25$
  - As d increases, α_d increases:

$$\lim_{d} \alpha_d = 1 - \ln 2 \approx 0.301$$



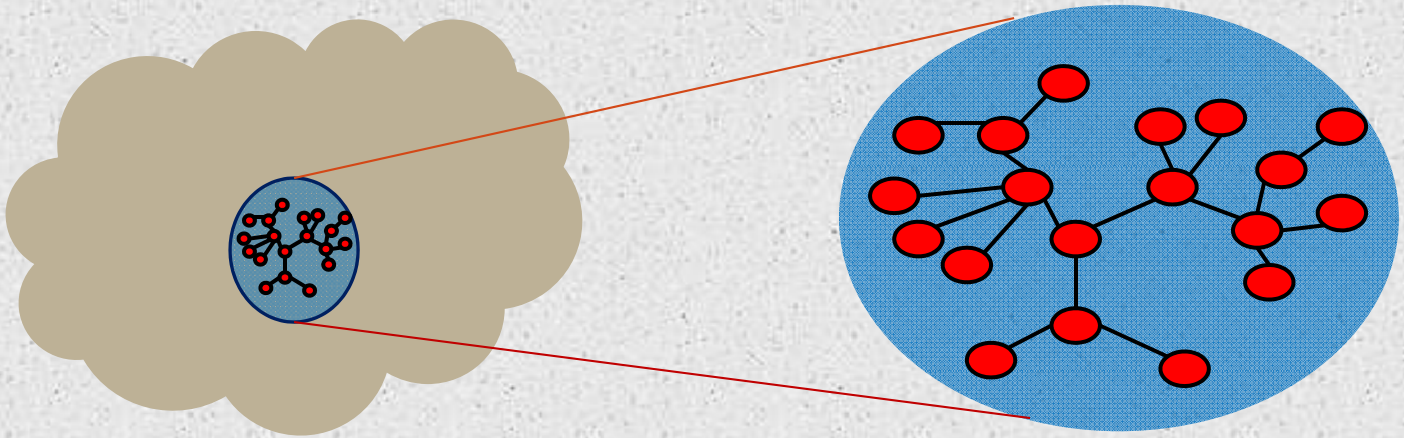
Rumor centrality: d-regular trees

- Some implications:
  - Probability of rumor center being k^{th} infected node
    - Can be calculated exactly for any d regular tree
    - It decreases exponentially in k



Random regular graphs

- Sparse random graphs
  - Locally tree-like
  - For example, random d regular graph
    - Locally like a d regular tree
    - Therefore, result for regular tree extend *verbatim*
  - Therefore, it is reasonable to focus on
    - Probability of source detection for generic tree



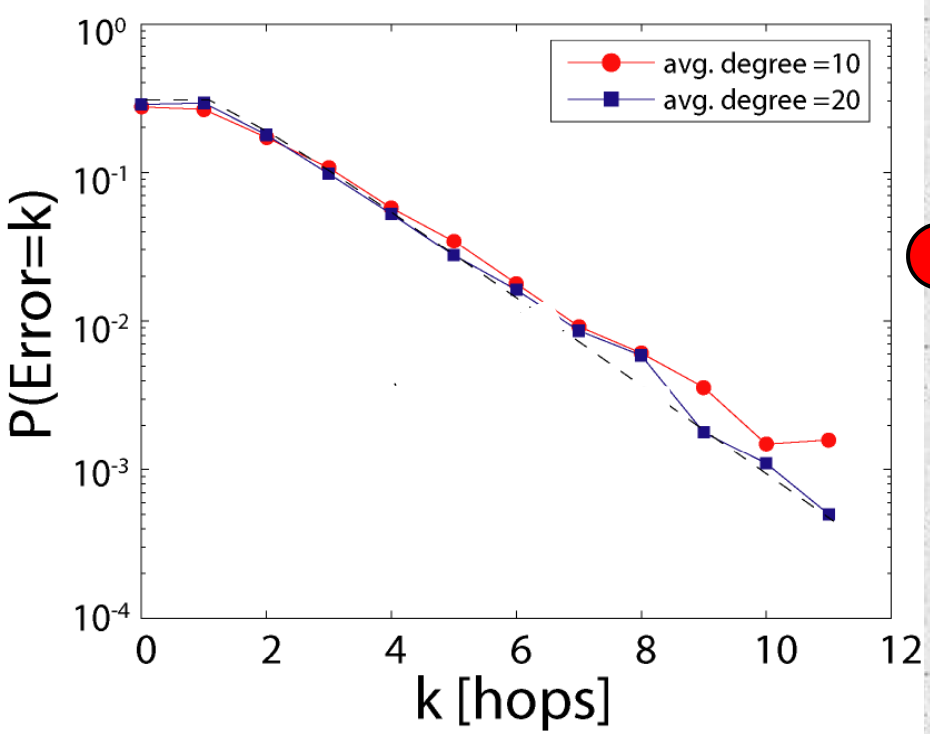
Generic random graph

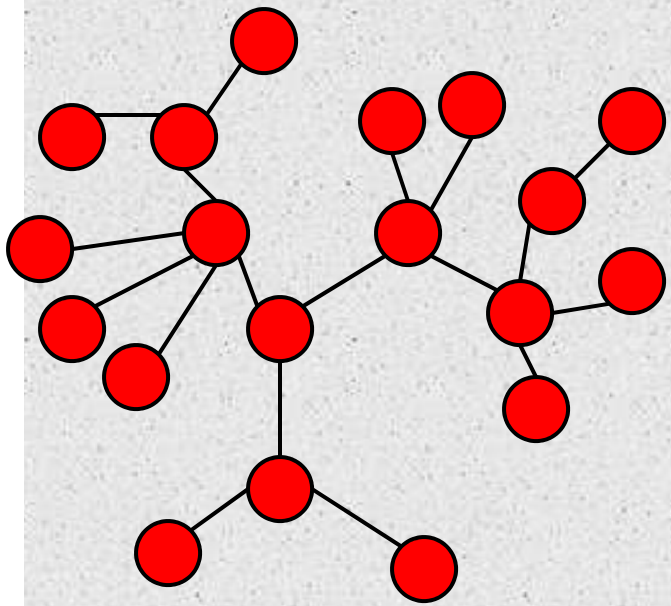
- Root node
  - Number of children, say n_0, with distribution D_0
  - These nodes have children distributed as per
    - Distribution D_1 in i.i.d. manner (say random variable n_1)
    - Just like Galton-Watson process

- **Theorem** [Shah, Zaman '11]**:** The probability of asymptotic detection is strictly positive (non-trivial) if
  - $P(n_0 > 2) > 0$
  - $E[n_1^2] < \infty$

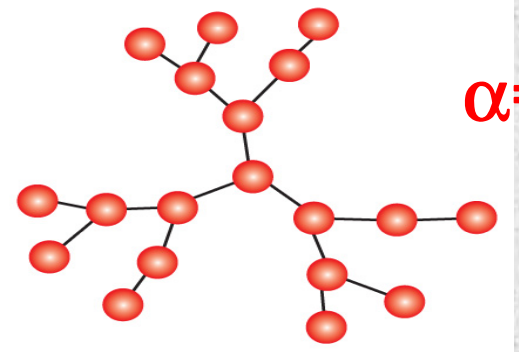
Erdos-Renyi (ER) graphs

- ER Graph
  - Each edge exists independently with probability $p=c/n$
  - Locally tree-like with Poisson(c) degree distribution





Geometric trees



α=1

Polynomial growth:

nodes at distance *r* from any node grows like r^{α} for some α > 0: line graph if α =0

Theorem [Shah, Zaman]: For such trees (root deg > 2)

$$\lim_{t} \mathrm{P}(C_t) = 1$$

Rumor centrality

- When can source be detected
  - No detection for line (or line like) graph
  - But good detection for graphs growing
    - Faster than line
  - Essentially
    - Graph expansion matters
    - Physics jargon:
      - phase transition in expansion parameter

- Rumor centrality
  - A principled approach towards *network centrality*

Remarks

- Information diffusion is fundamental to
  - Natural and Engineered networks
  - Two ubiquitous problems are
    - Spreading rate, computation
      - Gossip algorithms (monograph)
    - Source detection

- Principle : network expansion is useful for
  - Propagating information quicker, and
  - Finding source of the diffusion/rumor

Remarks

- Computational network science
  - Fast computation, meaningful answers
  - An enabling concept:
    - network centrality through lens of statistical inference

  - Rumor centrality for Twitter search
    - Trumor: influence tracking engine
    - *Recent Press:* Tech Review

  - Recommendation, Decision making
    - Celect: decision making engine
    - *Recent Press:* New York Times