Detecting the source of spread in complex networks

Boleslaw Szymanski and Krzysztof Suchecki

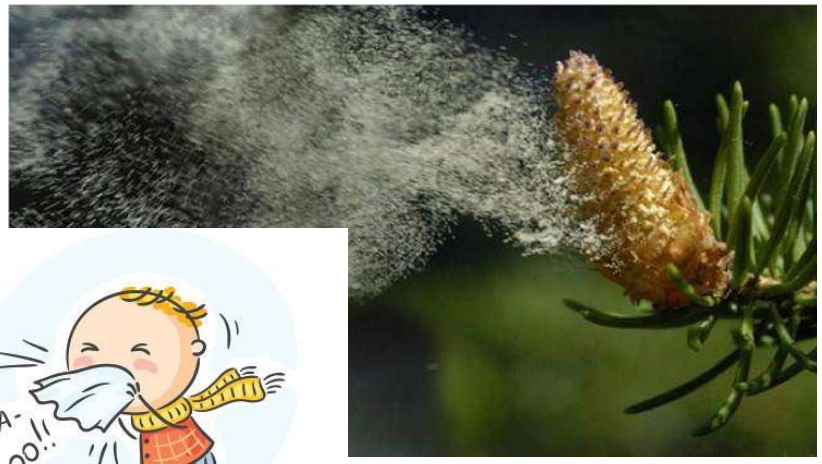
RPI, Troy

Plan

- Spreading processes and sources
- Source search in networks
- Pinto-Thiran-Vetterli algorithm
- Beyond basic methods

Spreading processess and sources

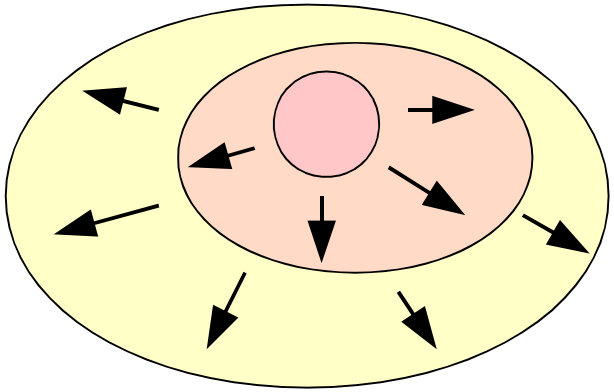
physical substances



infections

waves

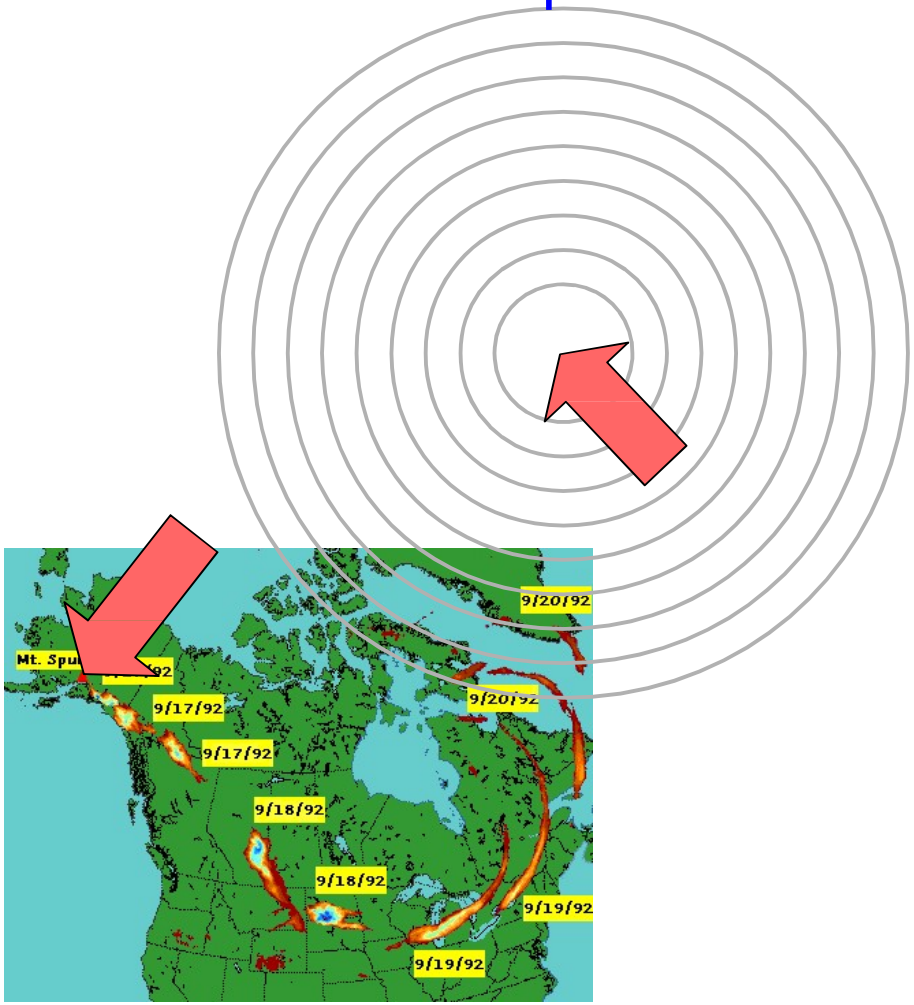
Start small



Become widespread

Spreading processess and sources

Is it possible to identify the source ?



If we have full data, it's obviously easy.

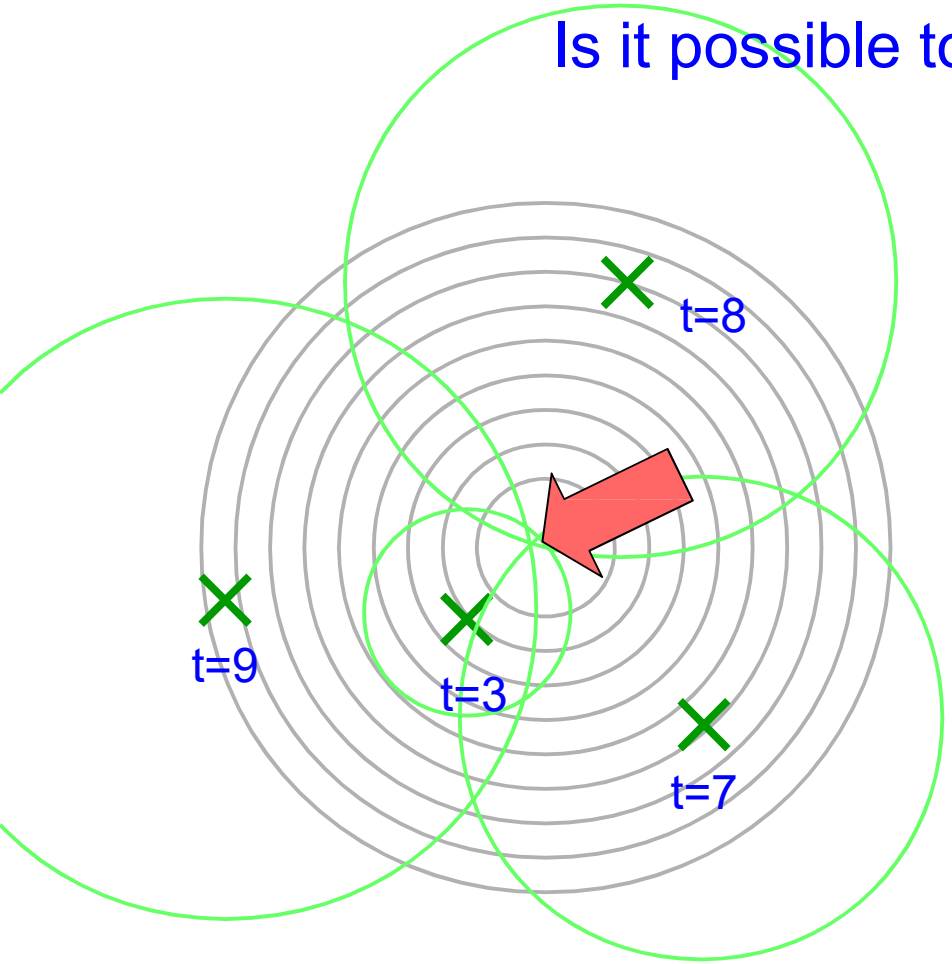
The point at which the wave/cloud/infection/etc. appeared earliest is the source.

Usually we don't have full data, only partial:

- Limited time (only since certain point)
- Limited scope (only know certain points)

Spreading processess and sources

Is it possible to identify the source ?



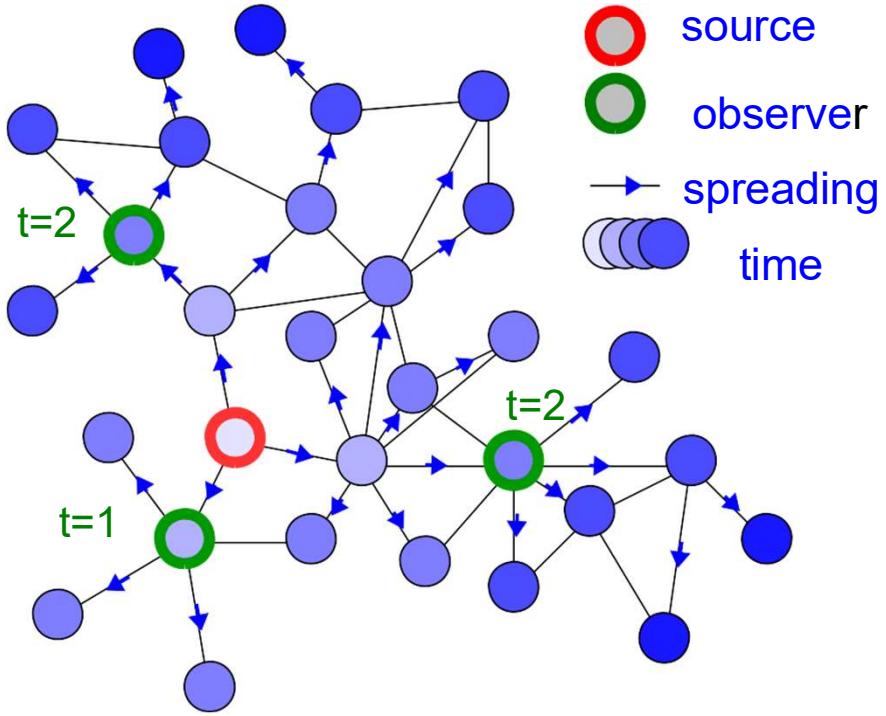
In deterministic spreading (e.g., waves) in space, this is easy.

Given D+1 points with time, or just 2 points with direction, we can tell where the source is.

Problems:

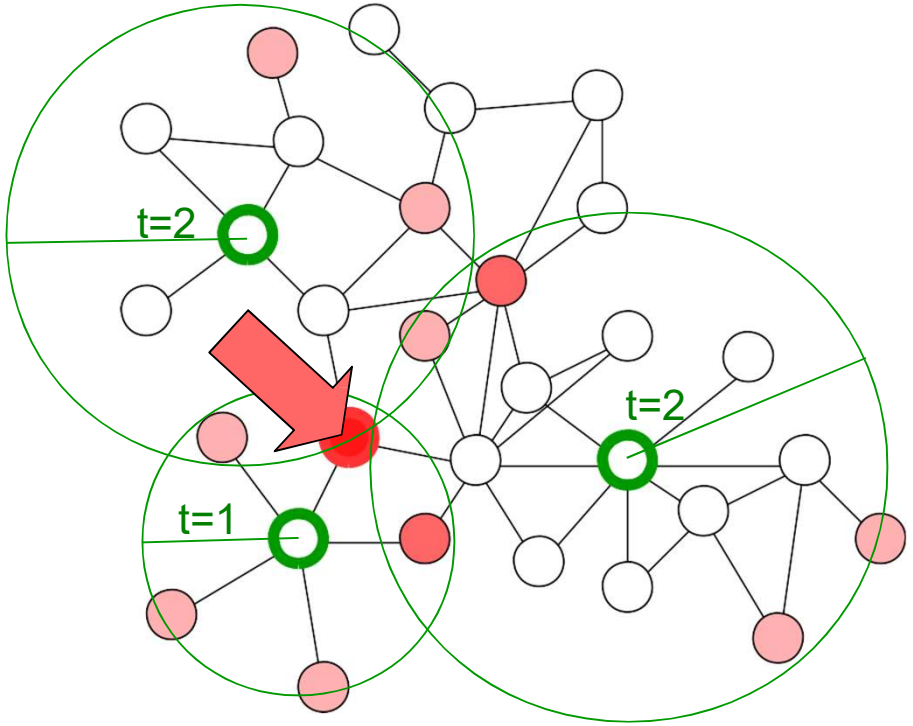
- Stochastic/complex dynamics (epidemics)
- Complex space (spreading in atmosphere)
- Spreading in network (epidemics, information)

Source search in networks

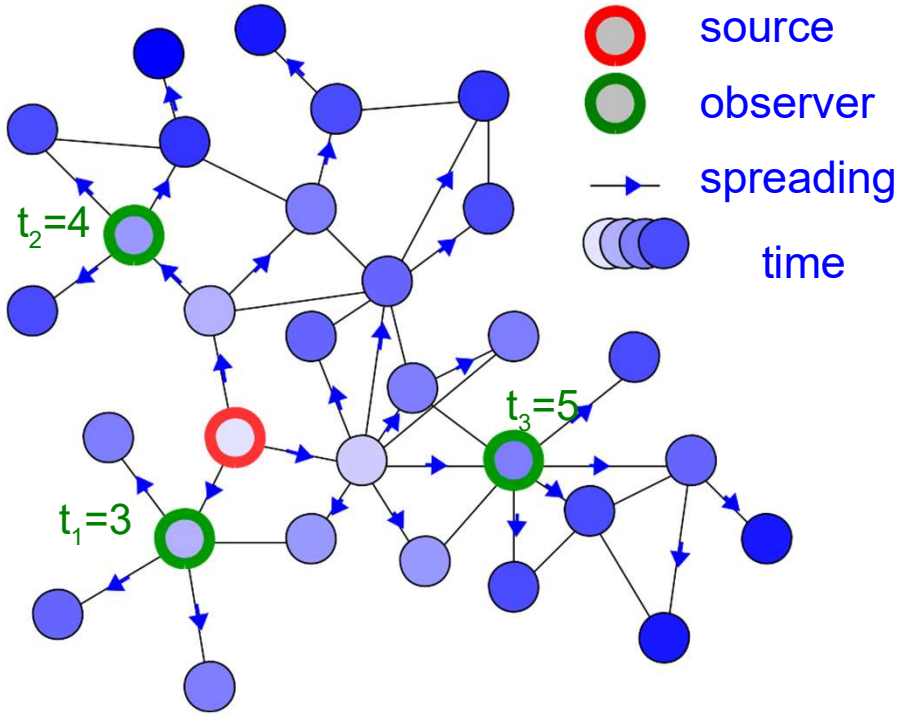


Similar “triangulation” approach could be used in networked environment.

-each observer has a “circle” of radius equal to time of observation
- where all “circles” intersect is the source



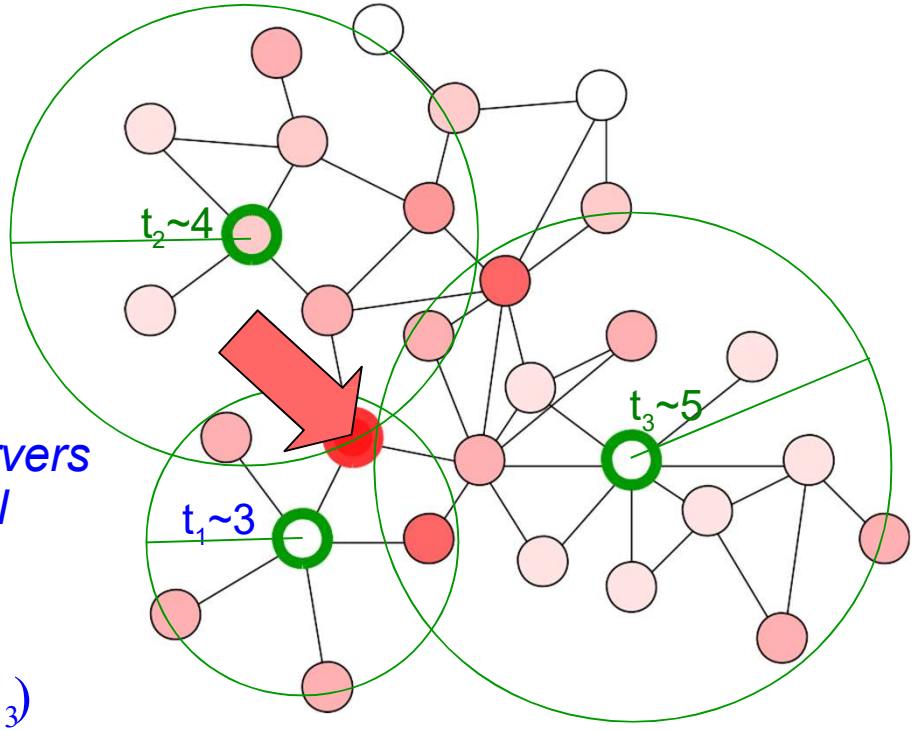
Source search in networks



If the process is stochastic, then the times are random variables and sharp-defined "circles" become blurry distributions.

$$P(s|t_i)$$

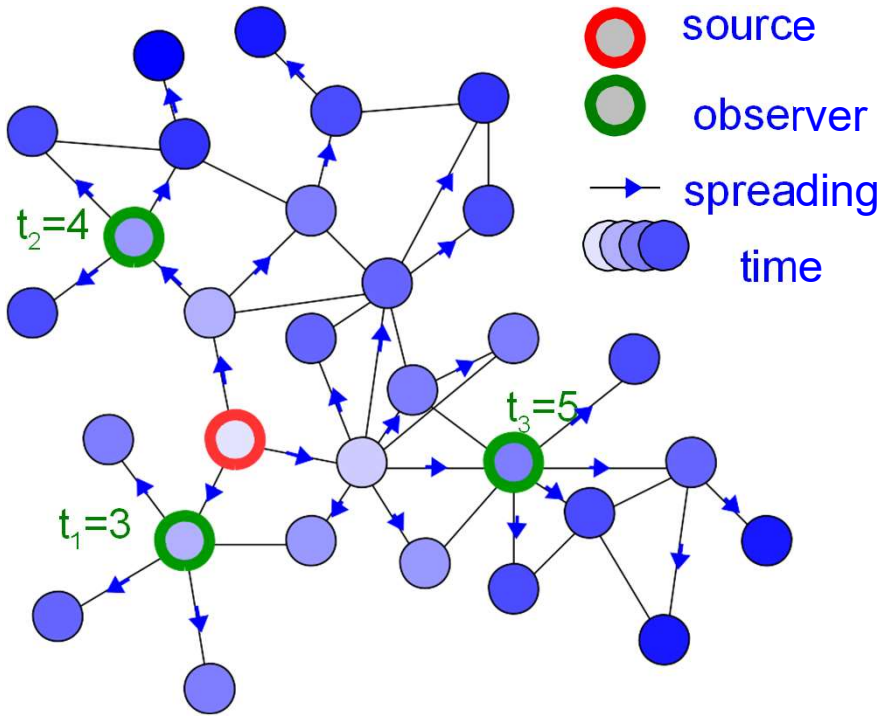
Probability of given node being source conditional on observation time t_i at observer i



Note: on the right, the sum of probabilities from different observers are added up – this is not overall probability for given node to be source

$$P(s|t_1)+P(s|t_2)+P(s|t_3)\neq P(s|t_1,t_2,t_3)$$

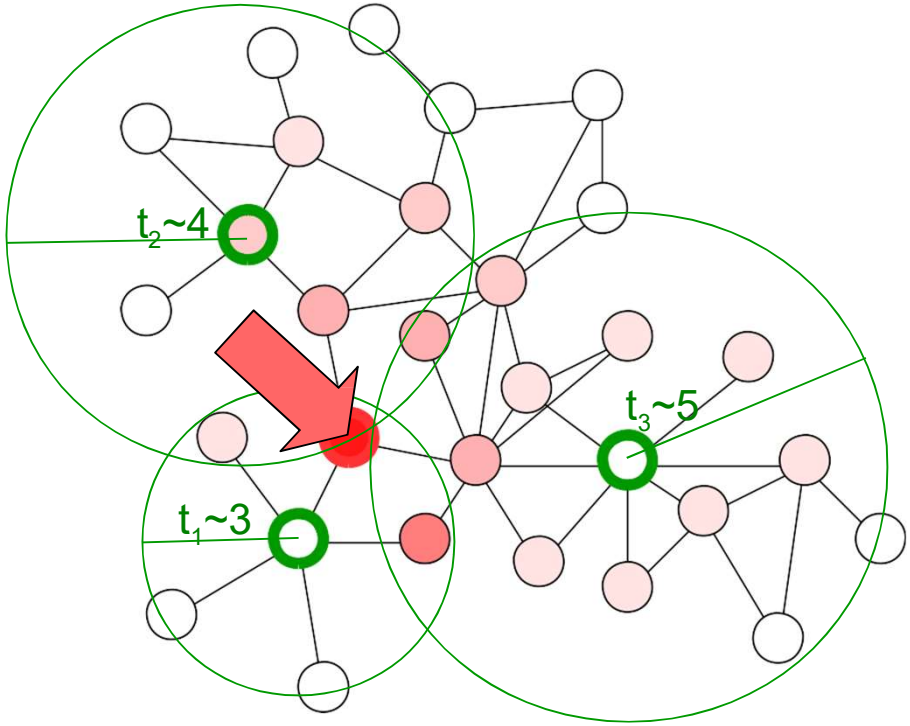
Source search in networks



If we look at all observers together, could we determine the overall probability ?

$$P(s|t_1, t_2, t_3) \equiv P(s|t)$$

If we have this, we could determine the most likely source.



Source search in networks

Bayes' Theorem:

$$P(s|t) = P(t|s)\frac{P(s)}{P(t)}$$

With this, we can calculate P(s|t) if we know P(t|s) – distribution of observed times if given node would be source

P(s) – usually we know nothing about which node could be real source, so we assume uniform 1/N distribution over all nodes

P(t) – we can calculate as

$$P(t) = \sum_s P(t,s) = \sum_s P(s)P(t|s)$$

Which we will need only for single value of t (the one that was observed)

In other words:

If we can calculate distribution of times given a source, we can calculate distribution of probability of being source given observation times.

To calculate P(t|s) we need to know something about the spreading process.

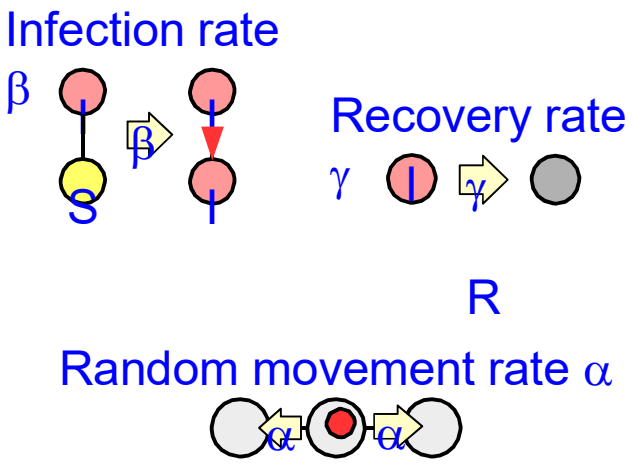
Source search in networks

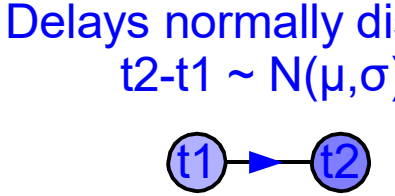
The better model we have for spreading, the more accurately we can calculate P(t|s), and thus make more accurate calculation of P(s|t) and find the source.

- Susceptible-Infected(-Recovered) model, created to describe spread of infectious diseases, is one of most commonly used to describe complex behavior, by reducing it to randomness.

- Diffusion/random walks, could be used to describe spread that conserves some “mass”

- Assume normally distributed delays on edges
  this is not really accurate model for anything, but unlike others, is possible to precisely calculate P(t|s) analytically
  - could be used to approximate other models

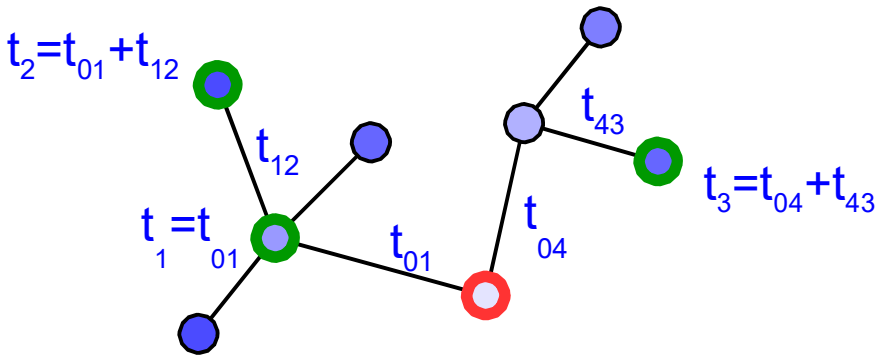




Source search in networks

Assume:

- normal delays on links $t_{ij} \sim N(\mu,\sigma)$
- tree topology ← *unfortunately necessary for analytical solution*



assuming IID delays

Mean:

$\mu_1=\mu_{01}$ $=\mu$

$\mu_2=\mu_{01}+\mu_{12}$ $=2\mu$

$\mu_3=\mu_{04}+\mu_{43}$ $=2\mu$

Variance:

$\sigma^2_1=\sigma^2_{01}$ $=\sigma^2$

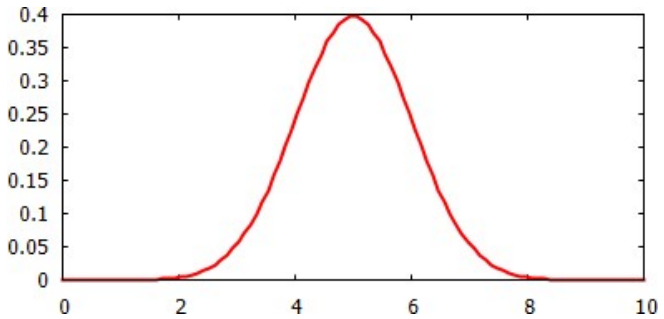
$\sigma^2_2=\sigma^2_{01}+\sigma^2_{12}$ $=2\sigma^2$

$\sigma^2_3=\sigma^2_{04}+\sigma^2_{43}$ $=2\sigma^2$

Sum of normally distributed variables t_{ij} =

= normally distributed variables t_i

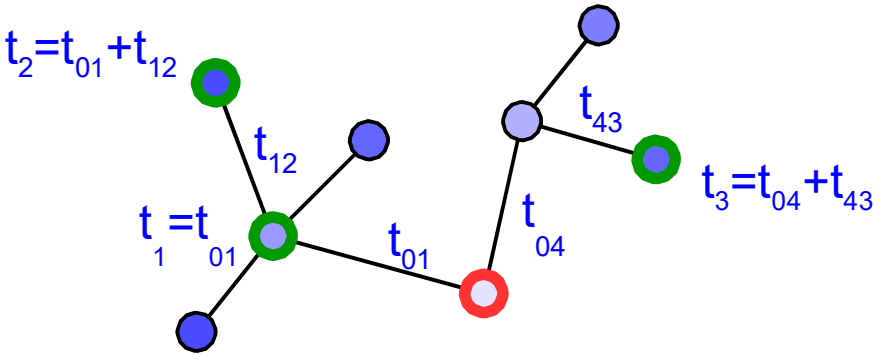
$$P(t_i)=\frac{1}{\sqrt{2\pi\sigma_i^2}}\exp\left(\frac{(t_i-\mu_i)^2}{2\sigma_i^2}\right)$$



Source search in networks

Assume:

- normal delays on links $t_{ij} \sim N(\mu,\sigma)$
- tree topology ← *unfortunately necessary for analytical solution*



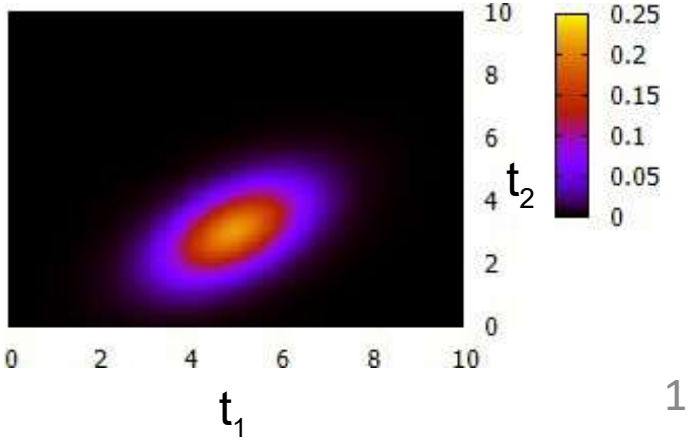
Mean:
μ=?

Covariance:
Σ=?

Take all times – multivariate normal distribution

$$P(\vec{t}\,) = \frac{1}{\exp\left(-\frac{1}{2}(\vec{t}-\vec{\mu})^T \Sigma^{-1}(\vec{t}-\vec{\mu})\right)}$$

Note: times may be correlated !

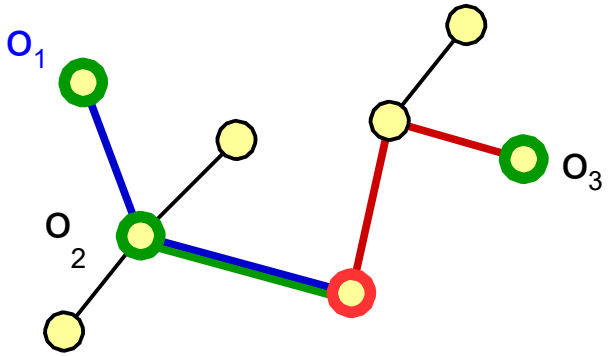


Source search in networks

Mean:

$$\vec{\mu} = \begin{bmatrix} \mu|P_{s1}| \\ \mu|P_{s2}| \\ \mu|P_{s3}| \end{bmatrix} = \mu \begin{bmatrix} 2 \\ 1 \\ 2 \end{bmatrix}$$

Mean is just length of path P_{si} from source to observer times mean delay on link



Covariance:

Covariance of random random variables made of sum of random variables is just the part that repeats in both – path overlap

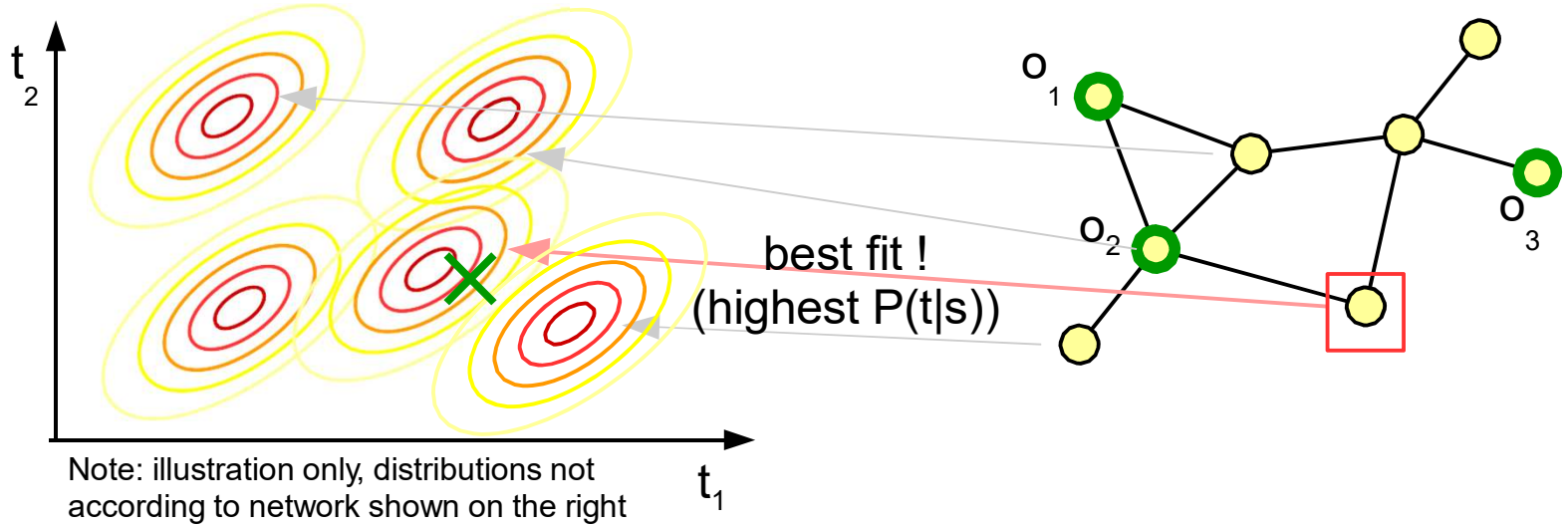
$$\begin{bmatrix} |P_{s1}| & |P_{s1} \cap P_{s2}| & |P_{s1} \cap P_{s3}| \\ |\overline{P_{s3}} \cap \overline{P_{s1}}| & |\overline{P_{s3}} \cap P_{s2}| & |\overline{P_{s3}}| \end{bmatrix} \sigma^2 = \sigma^2 \begin{bmatrix} 2 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 2 \end{bmatrix}$$

Note: P_{ij} here is path between observers i and j, not probability

Source search in networks

We know how to calculate P(t|s) as multivariate normal distribution under few assumptions.

We can get what is probability P(t|s) for the observed time and calculate P(t)



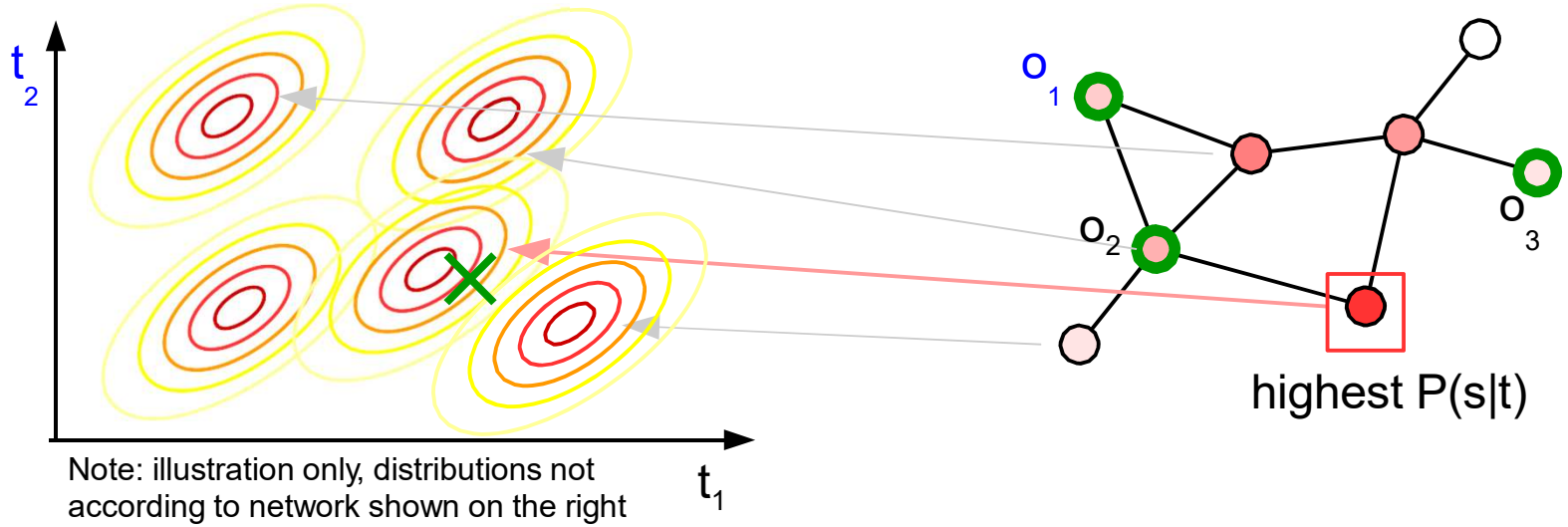
Given $P(s|t) = P(t|s)\frac{P(s)}{P(t)}$ and P(s) (a priori), P(t) (from P(t|s) and P(s))

We know that node s with highest P(s|t) is the one where P(t|s) is highest (what distribution fits the real data best)

Source search in networks

We know how to calculate P(t|s) as multivariate normal distribution under few assumptions.

We can get what is probability P(t|s) for the observed time and calculate P(t)



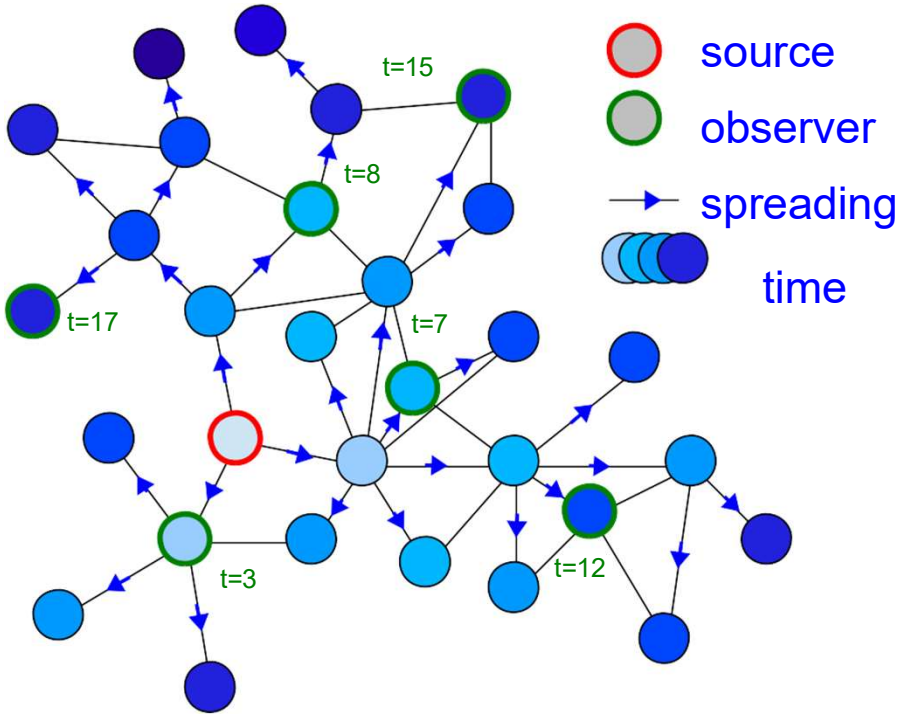
We can also calculate P(s|t) and thus calculate how likely it is for each node to be source.
(distribution of P(s|t) on nodes)

Pinto-Thiran-Vetterli algorithm

Known:
- Network topology
- Times when spreading arrived at observers
- Mean time it takes to infect along a single link
- Variance of that time

Want to know
- True source of the spread

Assumes
- Network is a tree (or approximates as such)
- Normally distributed delays on links



Not known
- When spread started (not necessarily at t=0)

P.C. Pinto, P. Thiran, M. Vetterli, “Locating the source of diffusion in large-scale networks”, Physical Review Letters 109, 068702 (2012)

Pinto-Thiran-Vetterli algorithm

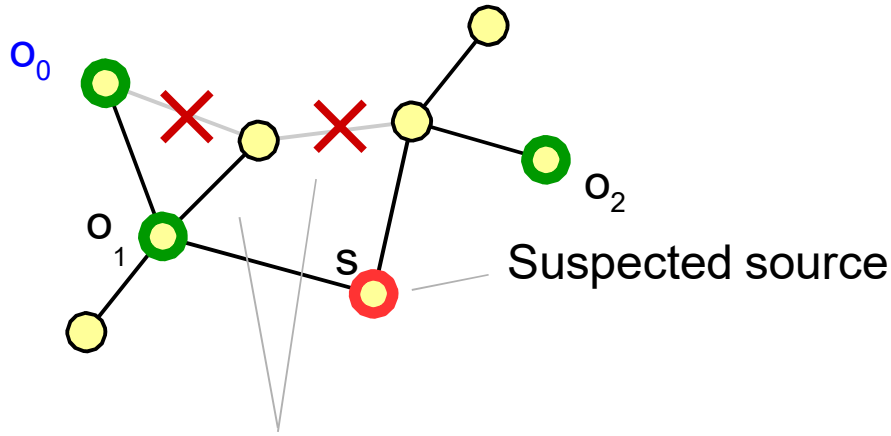
Issue: network is not a tree

Solution: make a tree out of it !

Since spreading process uses *fastest* path, it *usually* means the shortest topologically.

Use Breadth-First Search to make a tree (BFS tree) rooted at suspected source.

Note: each suspected source may have different BFS tree, unless original network is actually a tree.



Which link to take ?

Shortest paths are not unique, so we have to take one of the trees. Different trees may give different results.

Pinto-Thiran-Vetterli algorithm

Issue: we don't know the “zero” time (when spread started)

Solution: look at relative times only – use one observer as reference (e.g. observer 1 becomes 0 (reference), 2→1, 3→2)

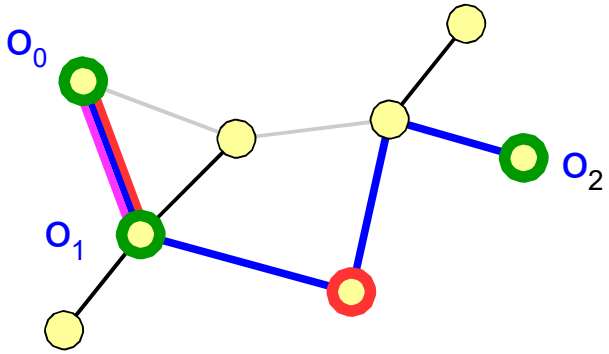
Mean: use time relative to reference

$$\vec{\mu} = \mu \begin{bmatrix} |P_{s1}| - |P_{s0}| \\ |P_{s2}| - |P_{s0}| \end{bmatrix} = \mu \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

Covariance: use paths anchored at reference, not suspected source

$$\begin{bmatrix} |P_{02} \cap P_{01}| & |P_{02}| \\ |P_{02}| & |P_{02}| \end{bmatrix} = \sigma^2 \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$$

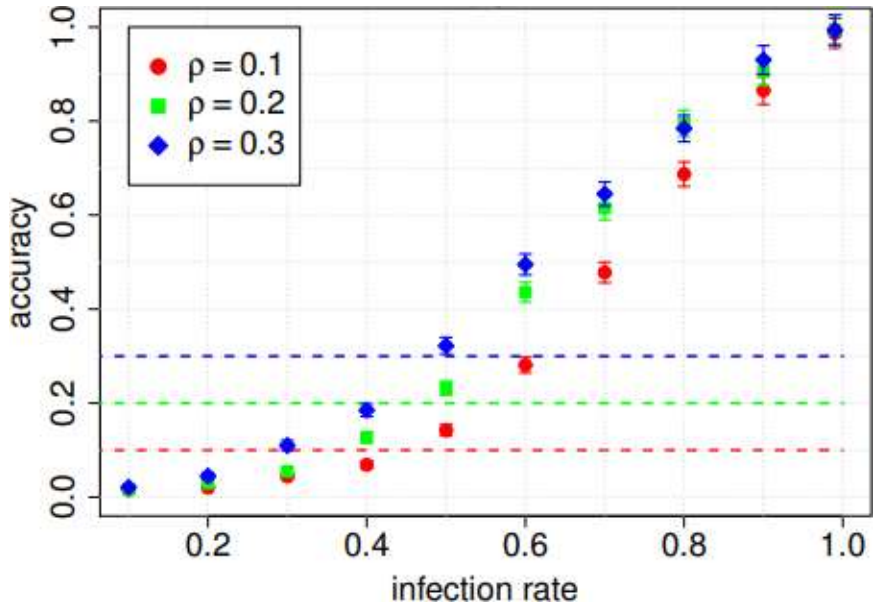
reference observer also introduces randomness, which is added or substracted from relative results (depend on situation)



Note: since the correlations are correct for tree only, for non-trees it's only approximation. Using closest observer (with smallest time) as reference minimizes this error for non-tree networks.

Pinto-Thiran-Vetterli algorithm

Performance of PTV algorithm:



Only really works when infection rate is high → so called propagation ratio μ / σ .

High propagation ratio – process is more deterministic. Low propagation ratio – process is more stochastic.

Can't expect to find a needle in a haystack with few measurement points, but still performs reasonably well if the process isn't too random.

Note: broken horizontal lines show accuracy of naive method that says that observer with lowest time is actual source, accuracy is equal density of observers then

Beyond basic methods

What can be we improve ?

- Make it faster (because it's slow $O(N^3)$ or worse)
- Don't approximate with a tree
- Use other distribution than normal
- Adapt for directed, weighted network
- Early estimation of source using yet silent observers

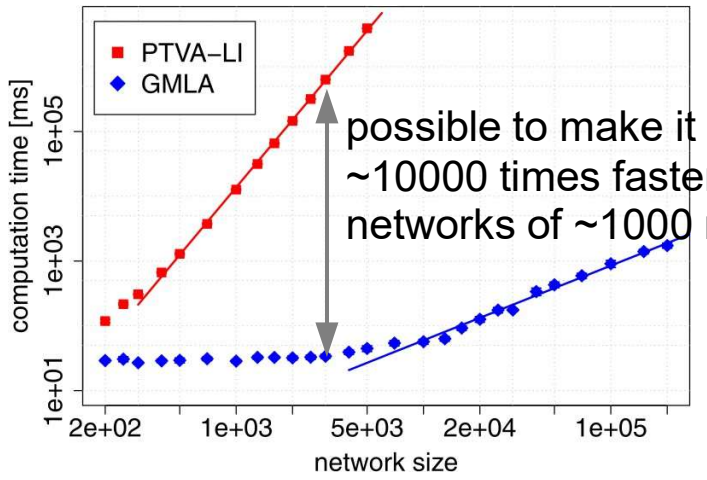
Note:

red – not attempted or done, hard to solve
yellow – only approximation done
green – done
black – under investigation

Beyond basic methods

- Make it faster (because it's slow $O(N^3)$ or worse)

In particular $O(N \cdot (N^2+K^3))$, where N is network size and K is number of observers. If K~N, then it is as bad as $O(N^4)$!



Feasible to calculate for networks of even millions of nodes (will not take 1000 years)

Cause:
- calculating likelihood score for each node
-using potentially large number of observers, requiring large tree and matrix operations

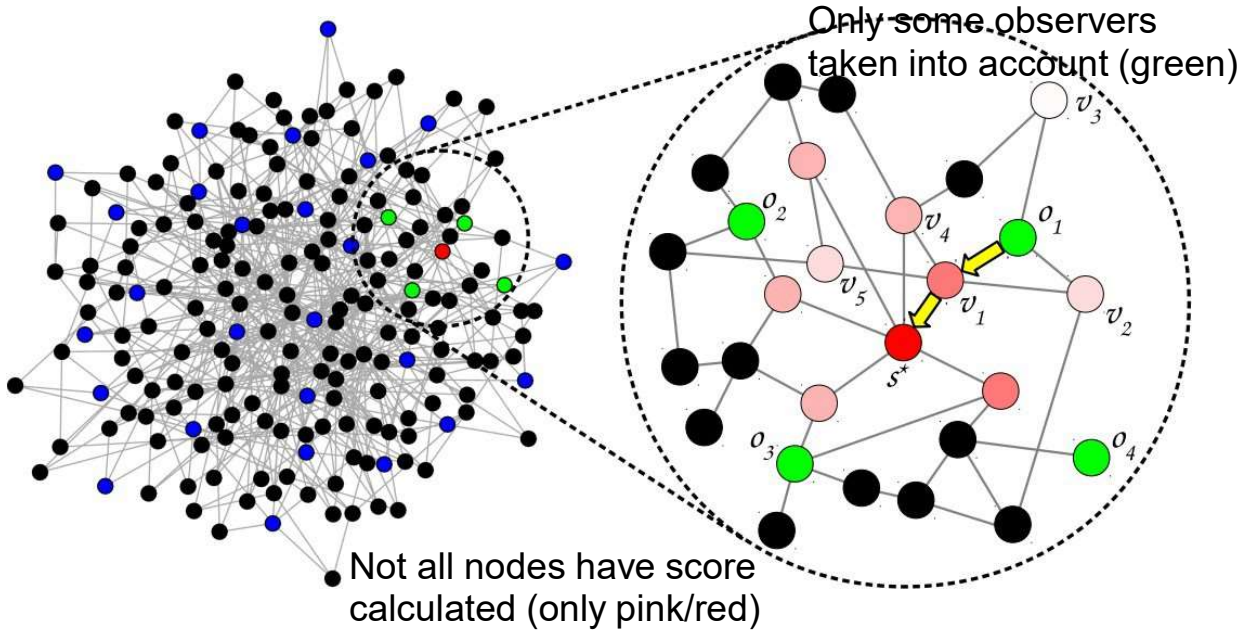
Solution:
- use greedy gradient (limits node to calculate score for)
- use only closest (smallest arrival time) observers to calculate likelihood

Beyond basic methods

- Make it faster (because it's slow $O(N^3)$ or worse)

Solution:
-use greedy gradient (limits node to calculate score for)
- use only closest (smallest arrival time) observers to calculate likelihood

Note: accuracy does not decrease in most situations, sometimes even increases !



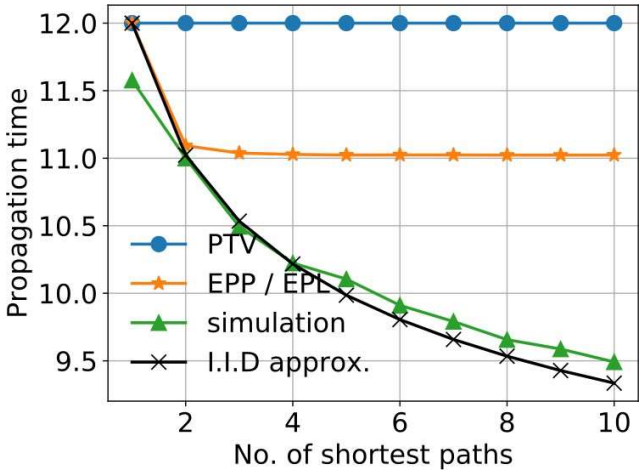
Gradient Maximum Likelihood algorithm

R. Paluch, X. Lu, K. Suchecki, B.K. Szymański, J.A. Hołyst, "Fast and accurate detection of spread source in large complex networks", Scientific Reports 8, 2508 (2018), doi: 10.1038/s41598-018-20546-3

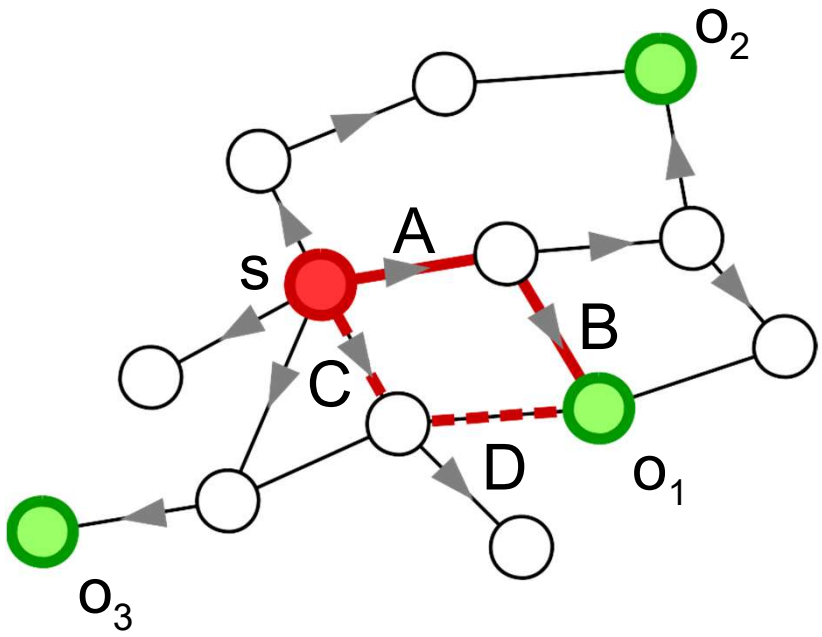
Beyond basic methods

- Don't approximate with a tree

Multiple paths change mean time, even if they have same length



Multiple paths can be taken into account when calculating expected mean times μ. Issue: correlations between them (which change mean of minimum)



$t_2=\min(A+B, C+D)$

$\langle t_2\rangle=\langle\min(A+B, C+D)\rangle\neq\min(\langle A+B\rangle, \langle C+D\rangle)=2\mu$

Mean of the minimum of two IID random variables is smaller than mean of that variable.

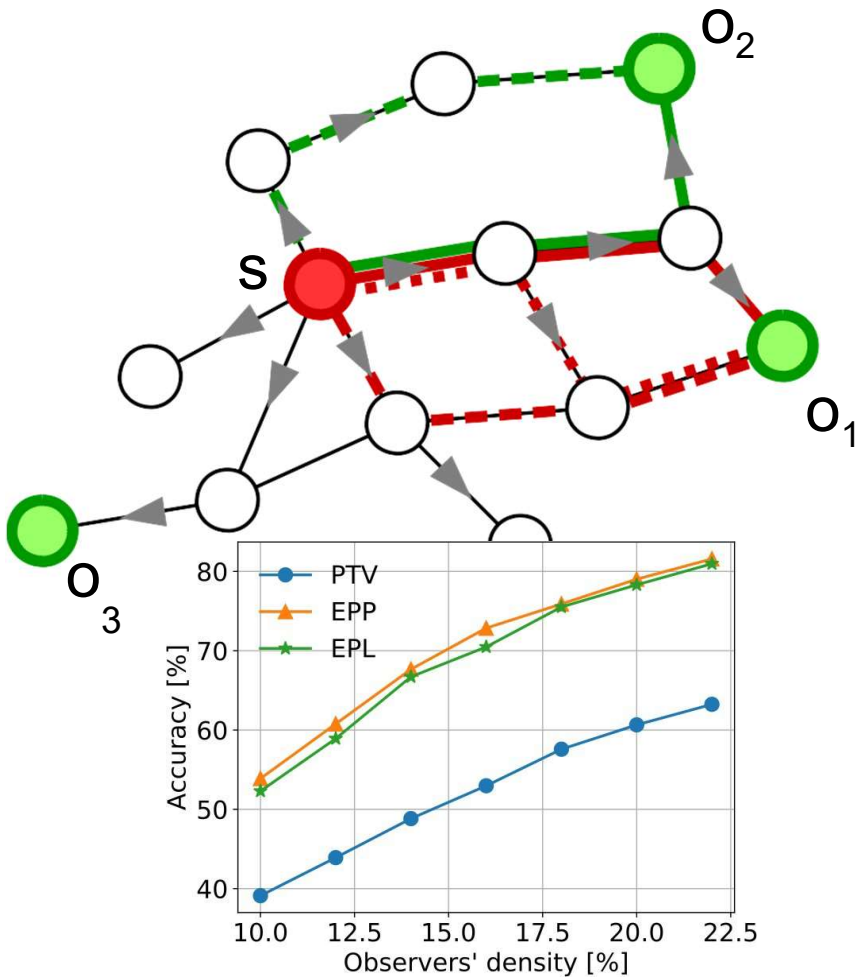
Beyond basic methods

- Don't approximate with a tree

No exact analytical solutions – only approximations possible.

Mean:
> Exact value for least correlated single pair.
> As if paths are uncorrelated

Covariance:
> Equiprobable Paths (EPP) – assume it's equal to mean of covariances of all path pairs in the two sets.
> Equiprobable Links (EPL) – assume it's equal to overlap between sets of links of both path sets.



Ł.G. Gajewski, K. Suchecki, J.A. Hołyst, Multiple propagation paths enhance locating the source of diffusion in complex networks, Physica A 519, 34-41 (2019), doi: 10.1016/j.physa.2018.12.012

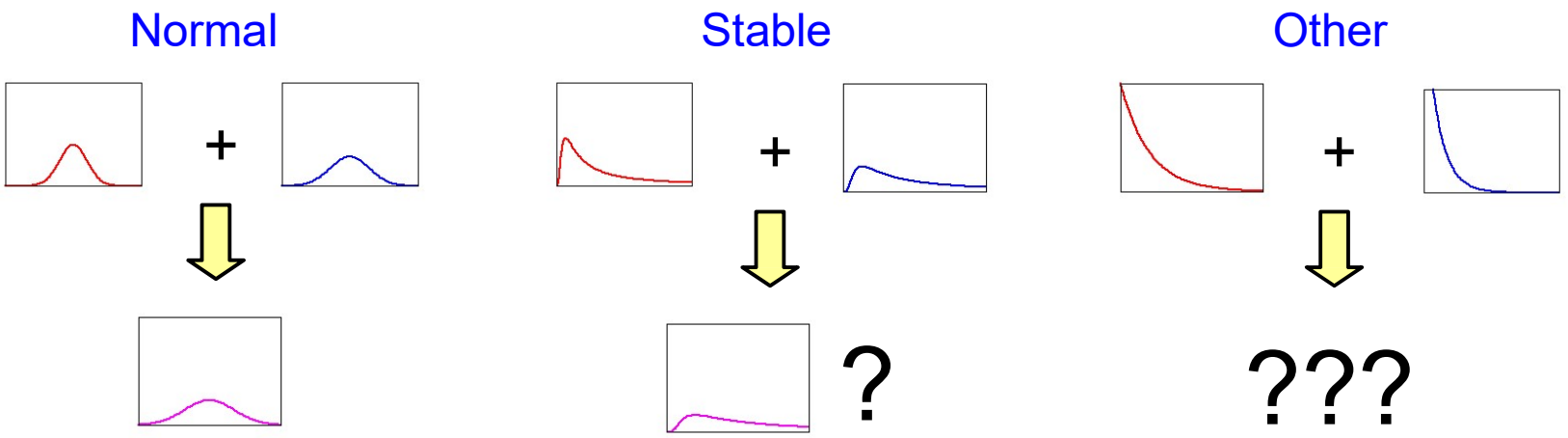
Beyond basic methods

- Use other distribution than normal

Idea is obvious, but solution is hard:

- If sum of 2 variables is from different distribution than each, number of variables can affect the shape of distribution, not only parameters
- assuming stable distribution (sum comes form same distribution) mean of sum will be sum of means, but how do other parameters of distribution change ?

  Extra issue: analytical stable distributions have infinite (Levy) or undefined (Cauchy) mean !

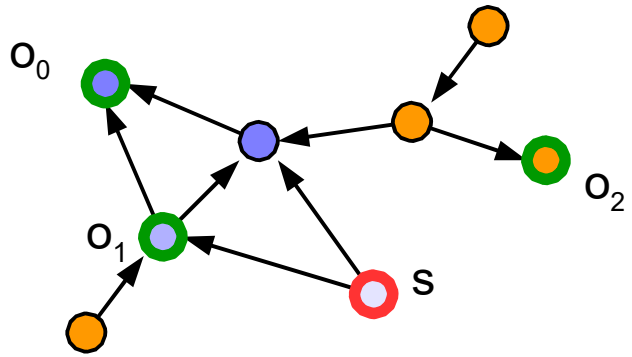


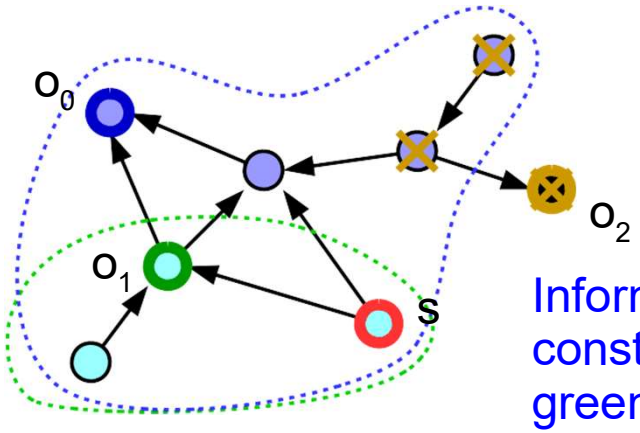
Beyond basic method

- Adapt for directed, weighted network

If the probability of infection depends on the link ? → weighed networks

If the link is one-sided (e.g. only reader of infecfed e-mail can catch computer virus) → directed networks





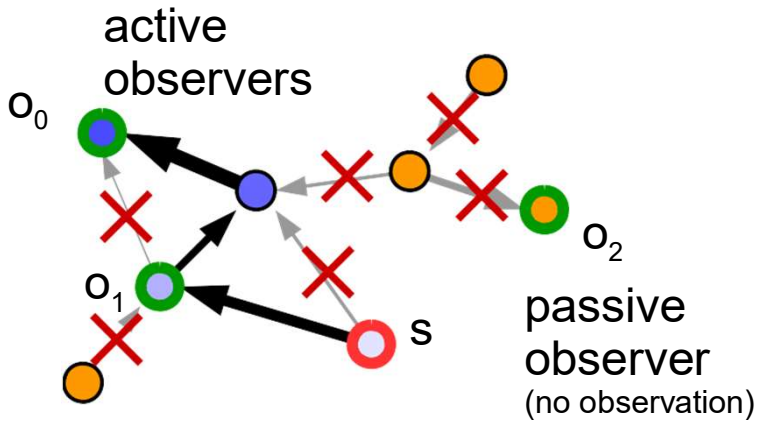
Not every observer will report any time, since parts of network may be unreachable from certain source

Information where spread arrived at all gives constrains on where the source can be (blue, green observers) or can't be (yellow observer), before we even consider time distribution

Beyond basic method

- Adapt for directed, weighted network

Weights on links mean BFS will be according to shortest mean time, not topological distance.
They also span only part of network reachable from given node in directed networks.
Only active observers are taken.



Mean: path lengths become sums of delays on paths

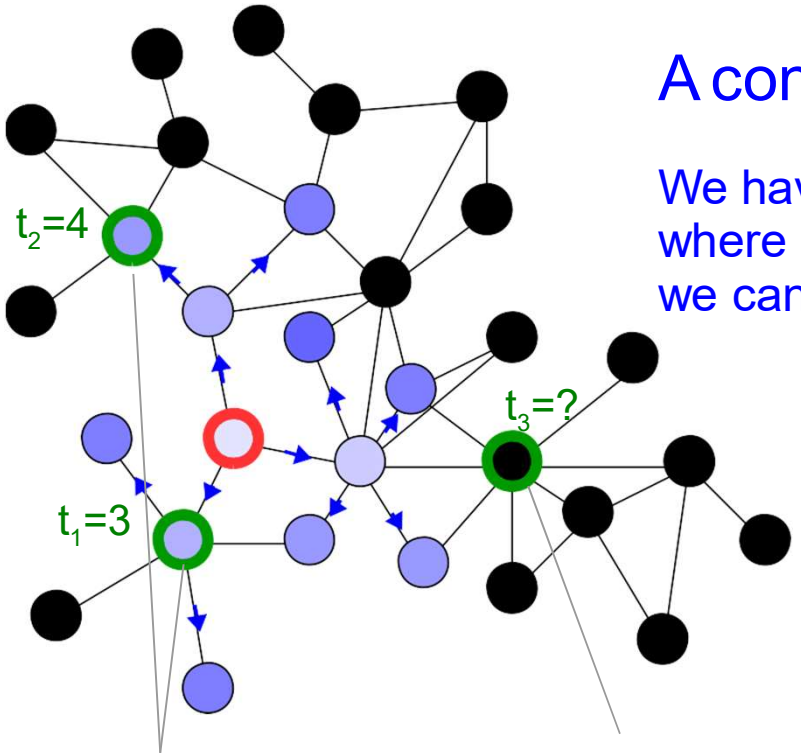
$$\vec{\mu} = \mu \begin{bmatrix} P_{s1} | - | P_{s0} | \\ P_{s2} | - | P_{s0} | \end{bmatrix} \rightarrow \left[\mu(P_{si}) - \mu(P_{s0}) \right]$$

Variance: can't use paths to/from reference because they are always from source towards observer – use source→observer paths instead; variance depends on path

$$\begin{bmatrix} |P_{02} \cap P_{01}| & |P_{02}| \end{bmatrix} \rightarrow \left[\sigma^2(P_{si} \cap P_{sj} / P_{s0}) + \sigma^2(P_{s0} / (P_{si} \cup P_{sj})) \right]$$

Beyond basic method

- Early estimation of source using yet silent observers



2 active observers

passive observer
(no time measurement yet)

A contagion started spreding out !

We have this situation now, we know 2 places where it already reached. Is this all information we can use to detect the source ?

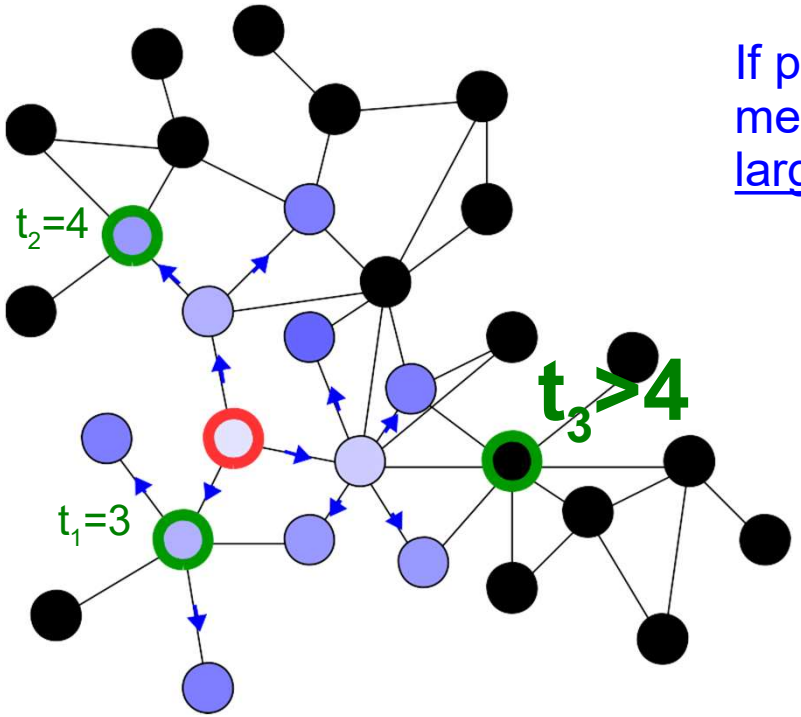
What about the 3rd place, where it did not reach yet ?

Can we use that information to increase the chances of succesfully finding the source early on ?

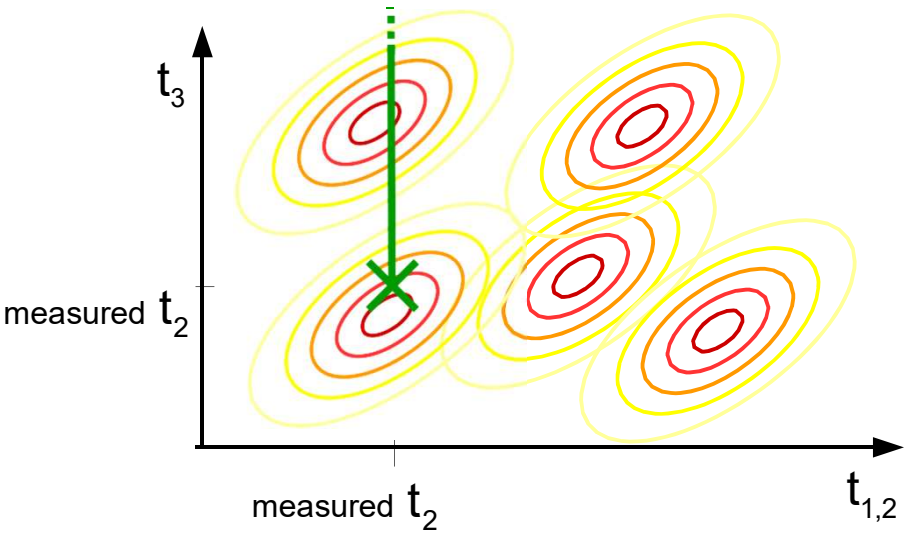
Yes, we can.

Beyond basic method

- Early estimation of source using yet silent observers



If passive observers are not infected yet, it means that time to reach that observer is larger than largest observed time.

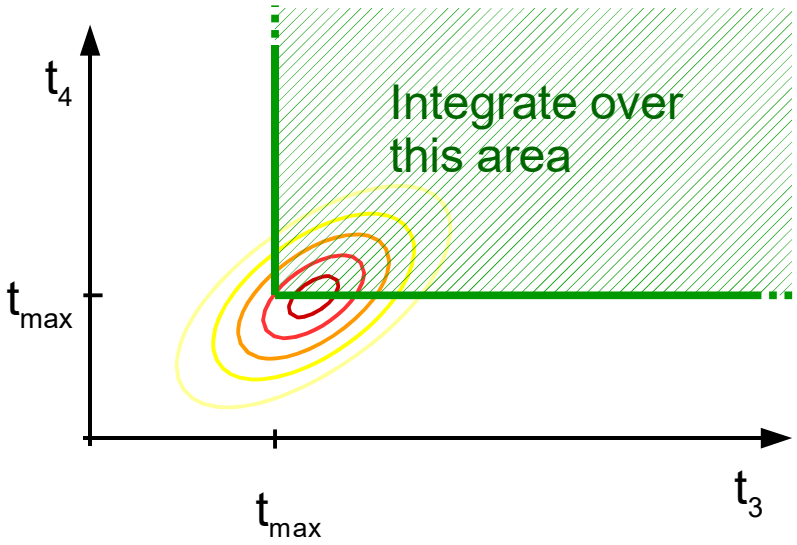


Effectively, measurement is not a point, but a part of space of arrival times (here, a line because we have 1 passive observer, but for more observers it's more dimensional)

Need to integrate over passive times > max time

Beyond basic method

- Early estimation of source using yet silent observers



Integrating over an arbitrary cut of correlated multivariate normal distribution (gaussian orthant problem) is a hard problem – closed form analytical solutions exist only for up to 3 dimensions

Possible approximations:

- Independent passive observers

$$P(t^*|s) = P(t_a|s) \prod_{i\,\text{passive}} P(t_i > t_{max})$$

- Mutually independent passive observers

$$P(t^*|s) = P(t_a|s) \prod_{i\,\text{passive}} P(t_i > t_{max}|t_p)$$

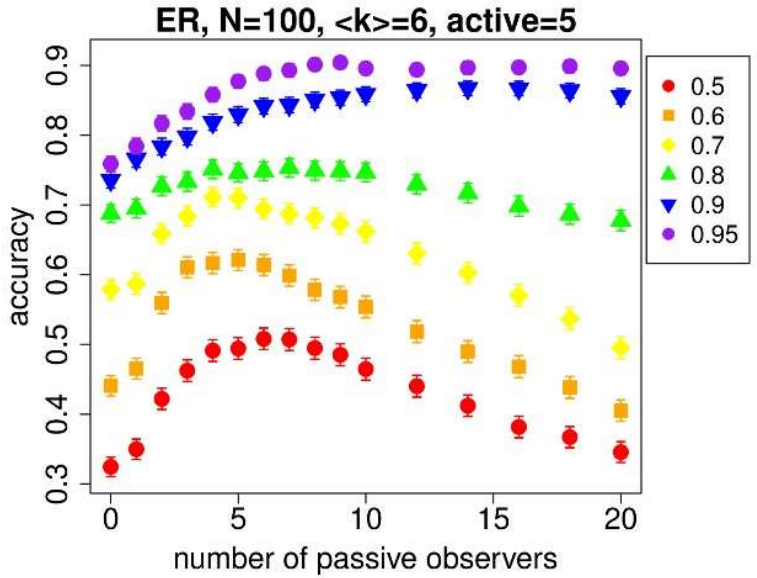
- Numerical solutions

Can be too expensive computationally

Beyond basic method

- Early estimation of source using yet silent observers

Does it actually work ?



Results for independent passive observers approximation show that it does, but only if we take not too many of them.

Why can taking too many decrease the accuracy ?

- since we assume independent, they don't take correlations into account and if they outnumber real observers, they shift “best” towards the “uncorrelated best”

- mutually independent passive observers approximation should solve that (at least partially)

Beyond basic method

Other issues or extensions:

- Using different spread model, where spreading is not certain (for example full SIR with recovery)
- Where to put observers in a network if we want to maximize accuracy ?
- Inverse: how to design spreading method to hide the source ?
- Other methods of finding source than maximum likelihood

Thank you

P.C. Pinto, P. Thiran, M. Vetterli, “Locating the source of diffusion in large-scale networks”, Physical Review Letters 109, 068702 (2012), doi: 10.1103/PhysRevLett.109.068702

R. Paluch, X. Lu, K. Suchecki, **B.K. Szymański**, J.A. Hołyst, “Fast and accurate detection of spread source in large complex networks”, Scientific Reports 8, 2508 (2018), doi: 10.1038/s41598-018-20546-3

Ł.G. Gajewski, K. Suchecki, J.A. Hołyst, “Multiple propagation paths enhance locating the source of diffusion in complex networks”, Physica A 519, 34-41 (2019), doi: 10.1016/j.physa.2018.12.012

R. Paluch, **B.K. Szymański**, J.A. Hołyst, “Efficient observers for source localization in complex networks: the state-of-the-art and comparative study”, Future Generation of Computer Systems, 112(11):1070-1092 June 22, 2020.

Y. Lytkin, R. Paluch, Ł. Gajewski, K. Suchecki, K. Bochenina, B.K. Szymanski, J.A. Hołyst, “How much information is in silence”, *in preparation*