Detecting the source of spread in complex networks
Part II

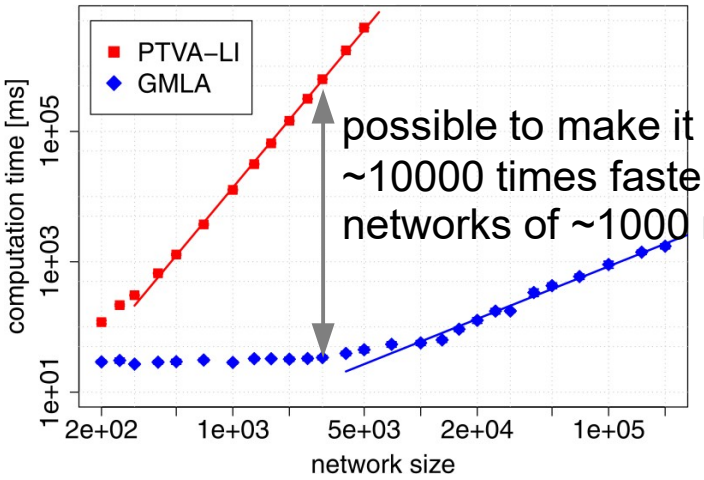
Boleslaw Szymanski and Krzysztof Suchecki

RPI, Troy

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

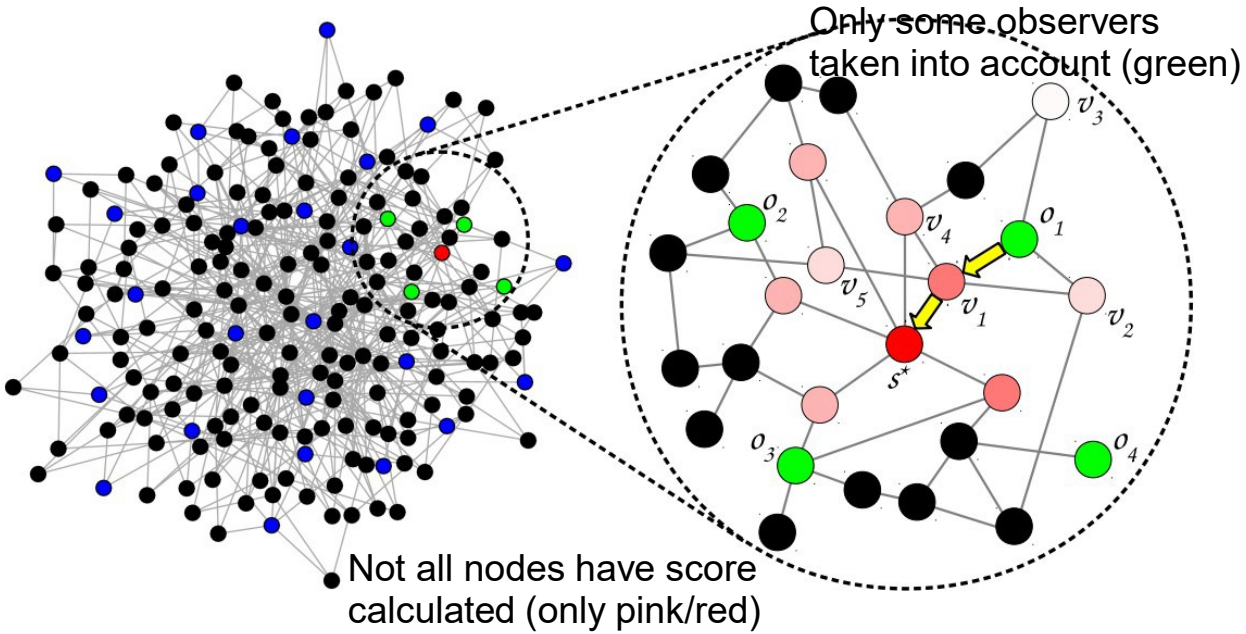
Solution:
- use greedy gradient (limits node to calculate score for)
- use only closest (smallest arrival time) observers to calculate likelihood

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 !



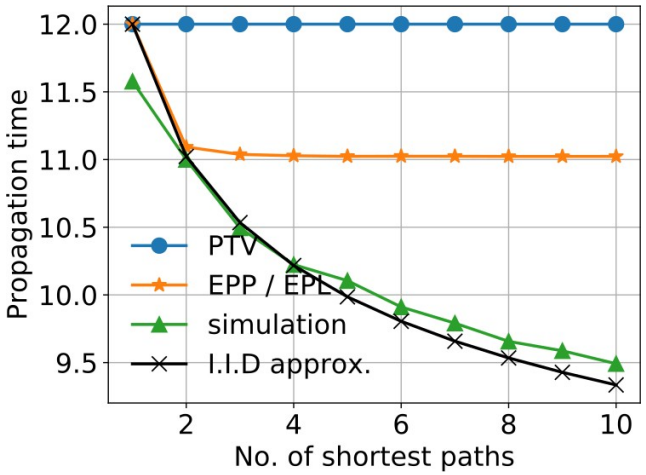
Gradient Maximum Likelihood algorithm

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

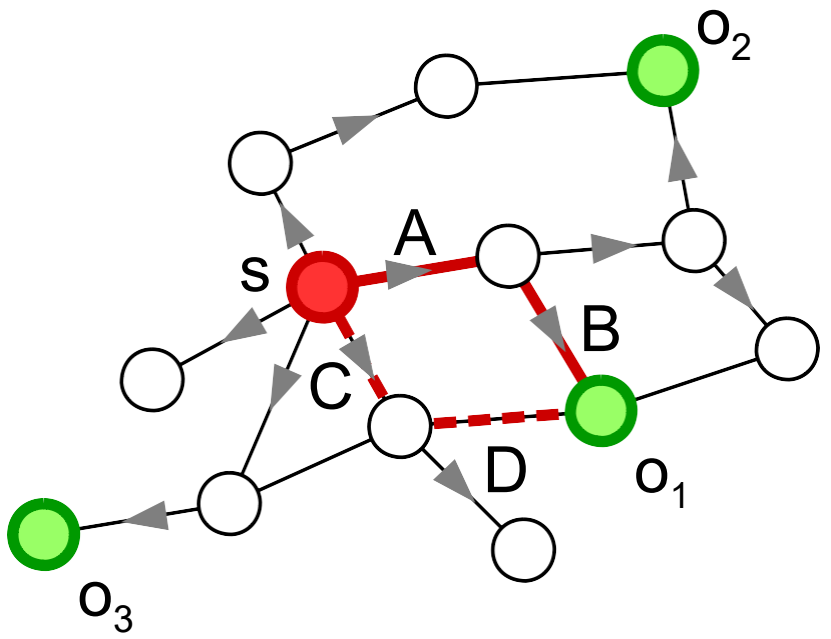
Beyond basic methods

- 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.

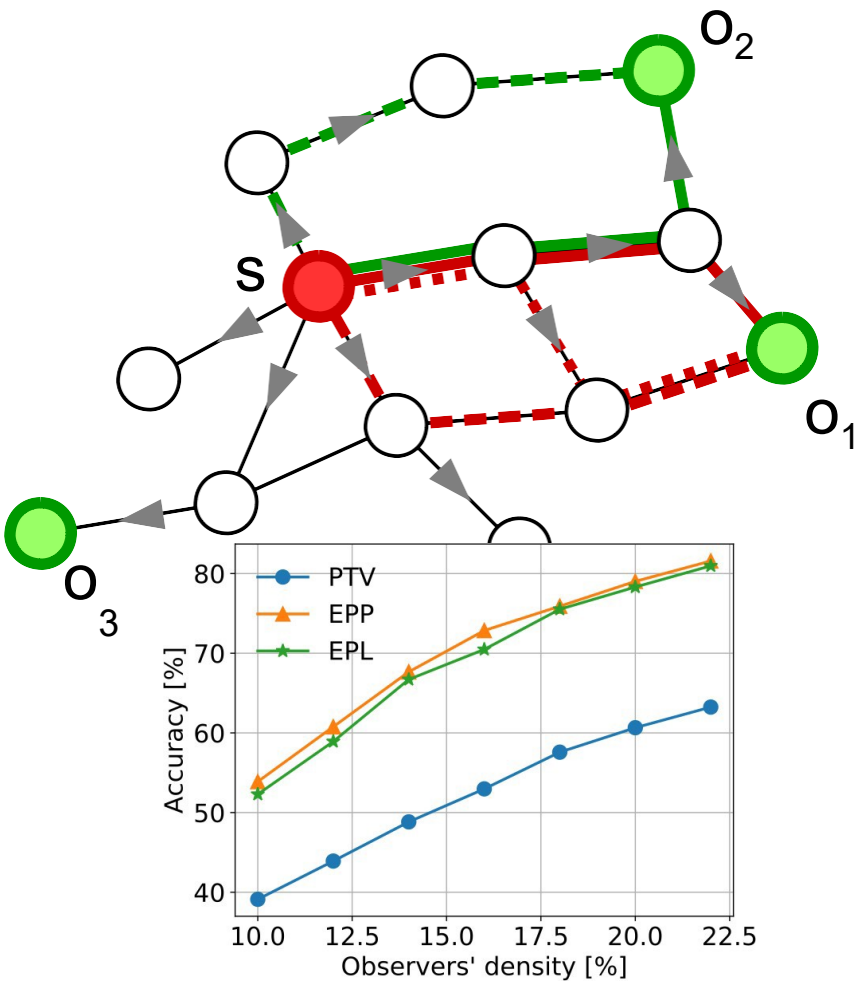
Beyond basic methods

- 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

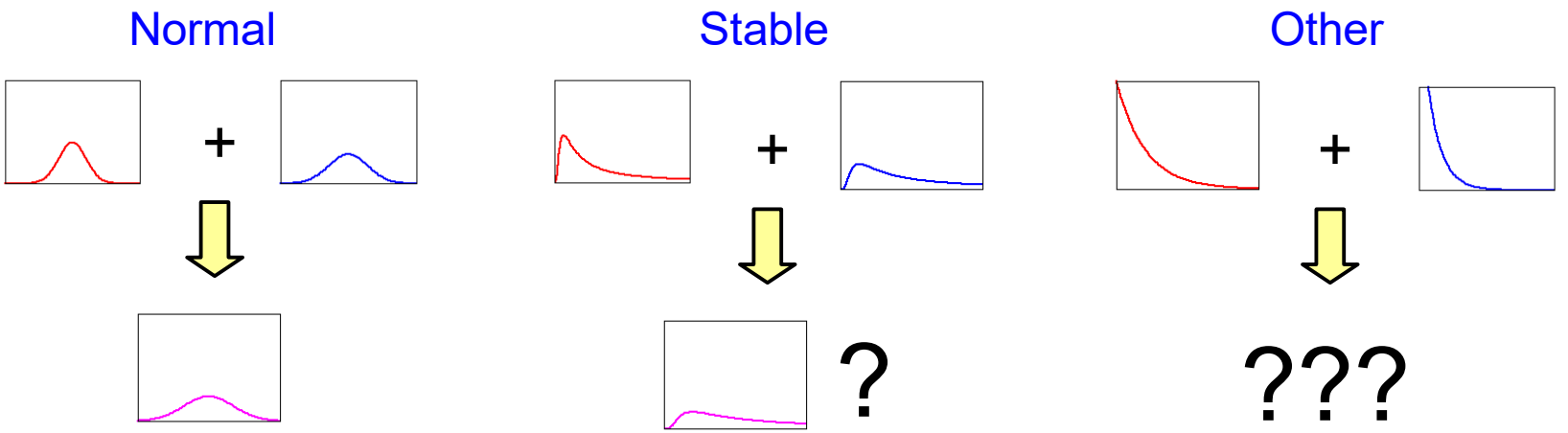
Beyond basic methods

- 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 !

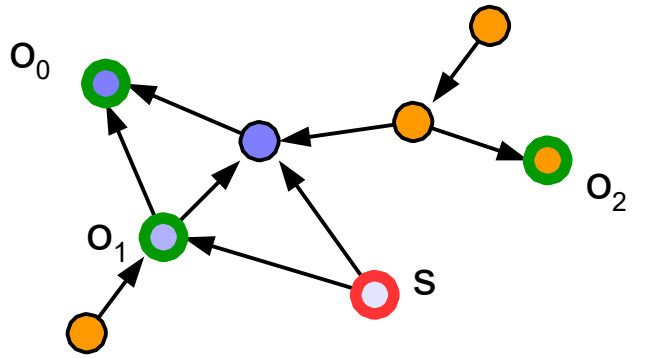


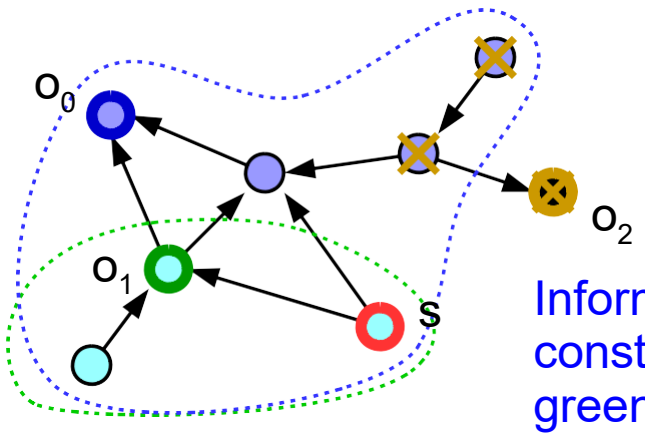
Beyond basic method

- 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





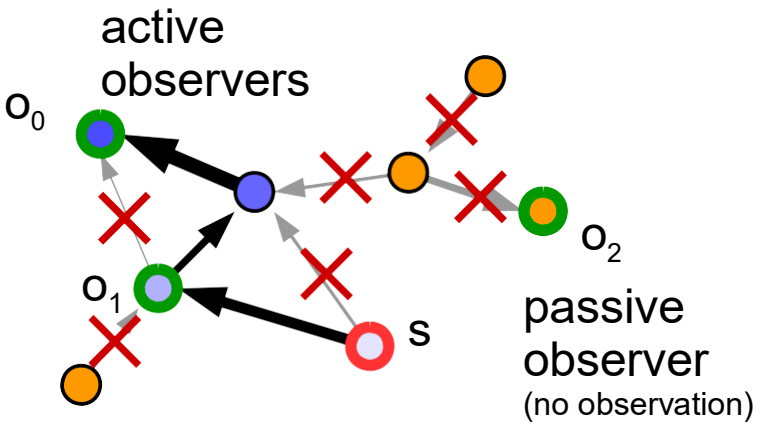
Not every observer will report any time, since parts of network may be unreachable from certain source

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

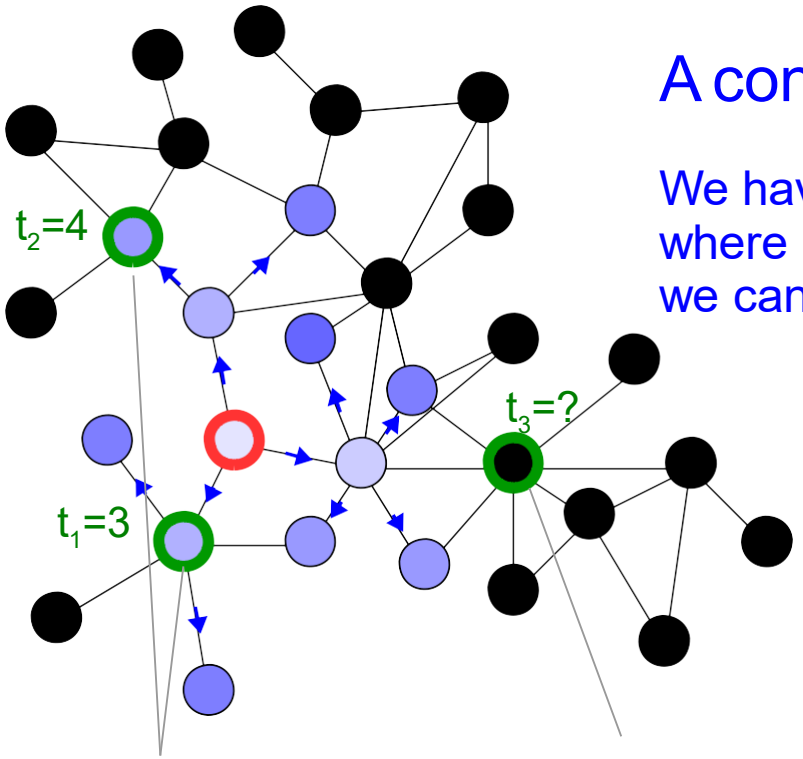
$$\vec{\mu} = \mu \begin{bmatrix} P_{s1} | - | P_{s0} | \\ P_{s2} | - | P_{s0} | \end{bmatrix} \rightarrow \left[\mu(P_{si}) - \mu(P_{s0}) \right]$$

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 ?

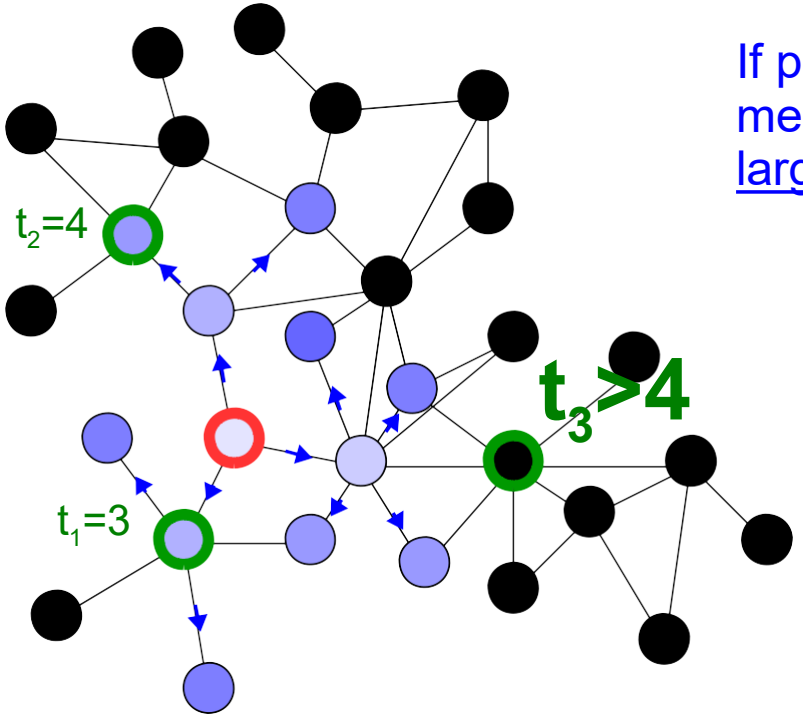
What about the 3rd place, where it did not reach yet ?

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

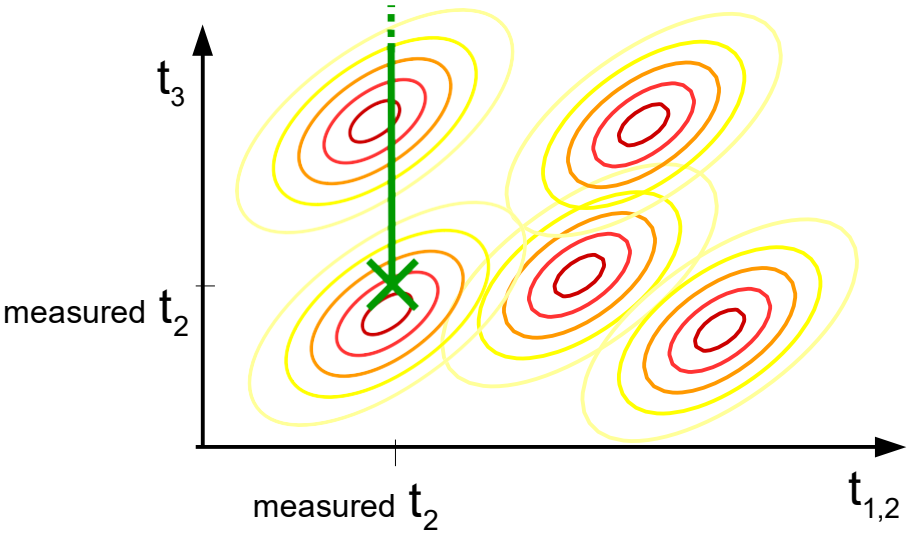
Yes, we can.

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.

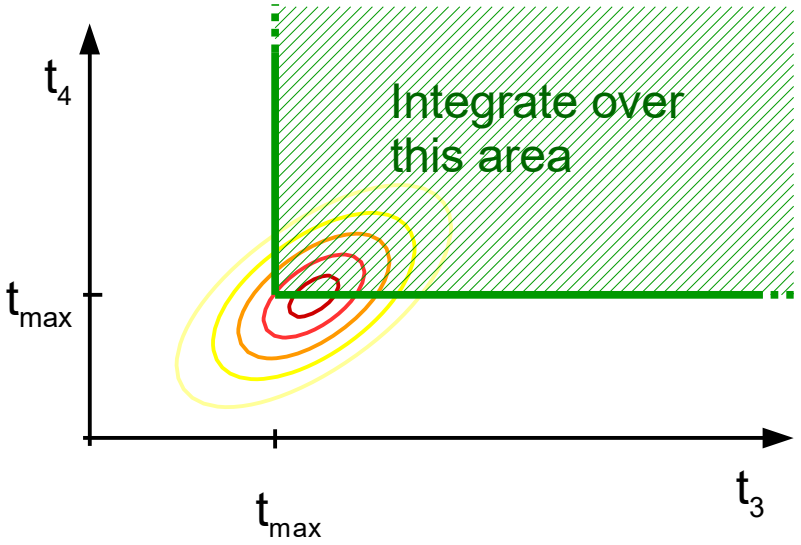


Effectively, measurement is not a point, but a part of space of arrival times (here, a line because we have 1 passive observer, but for more observers it's more dimensional)

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

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

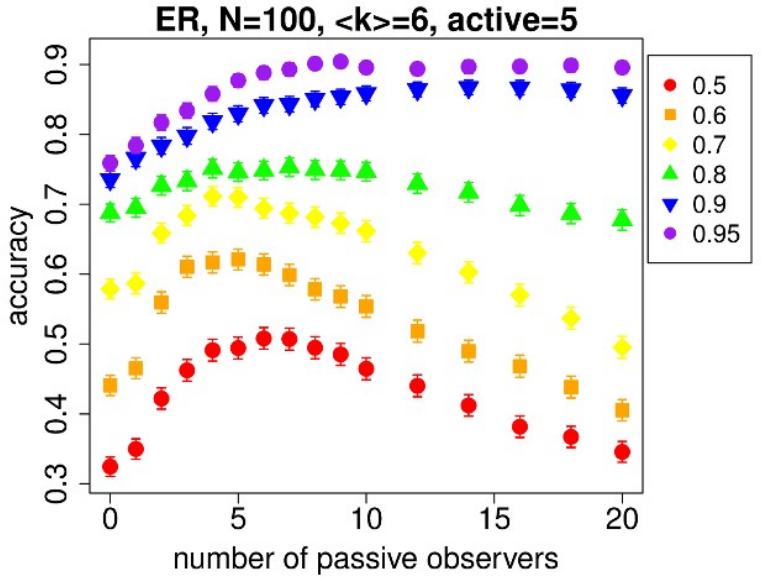
- Numerical solutions

  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*