Imperial College London

Rumour Source Detection in Social Networks using Partial Observations

Roxana Alexandru and Pier Luigi Dragotti

Communications and Signal Processing Group
Electrical and Electronic Engineering Department
Imperial College London

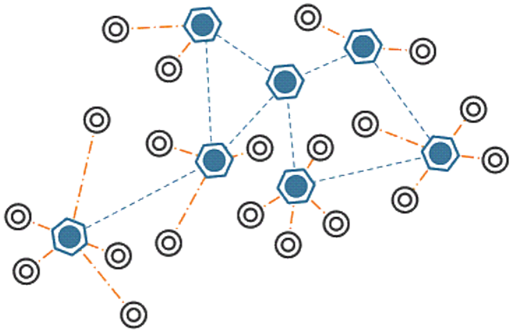
2018 6th IEEE Global Conference on Signal and Information Processing, Anaheim, California, USA

Imperial College London

Content

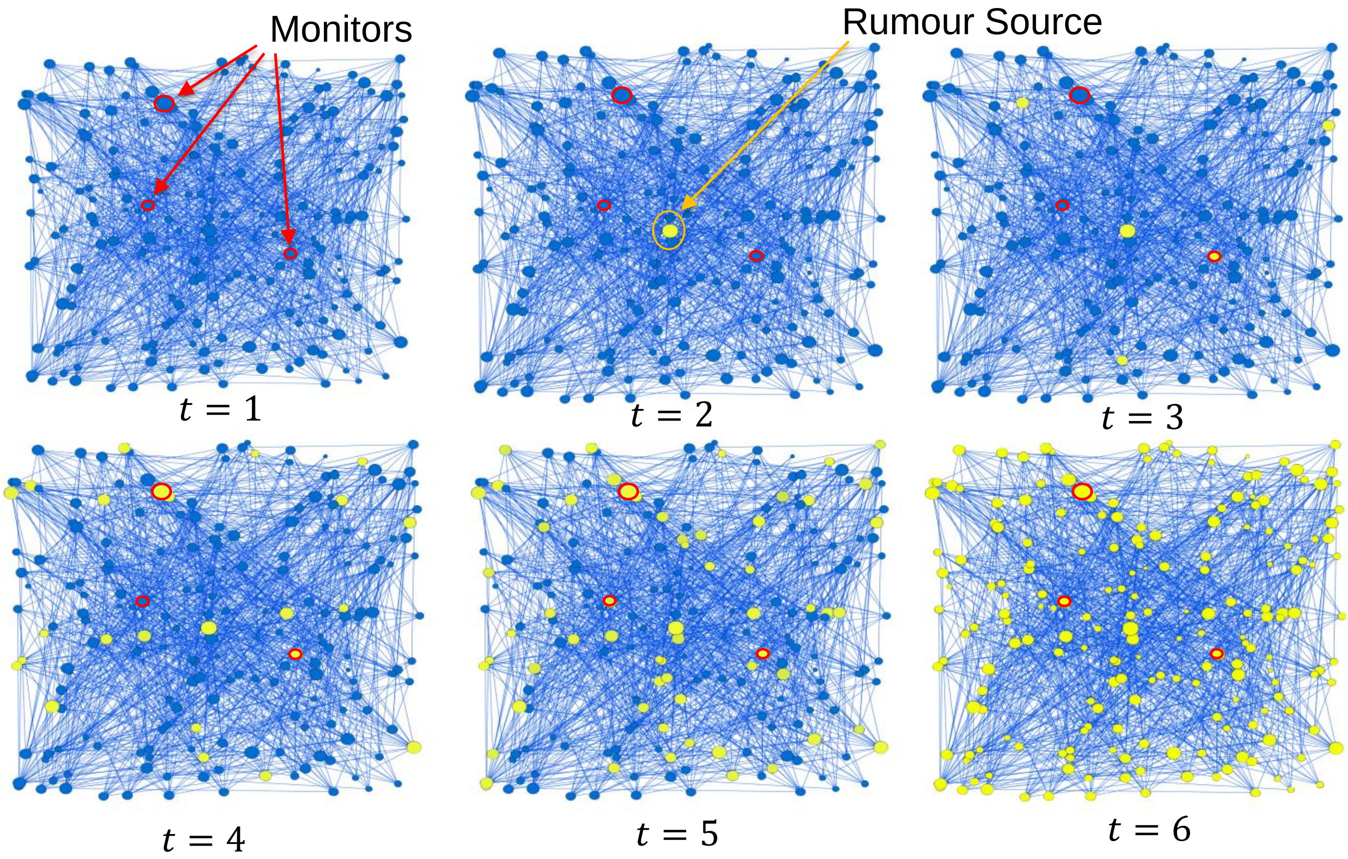
Imperial College London

Motivation



Imperial College London

Problem Statement



Imperial College London

Problem Statement and Assumptions

Network topology

- General graph with small-world property.

Epidemic model

- Discrete-time version of susceptible-infected model.
- Constant transmission rate within the network.

Observation model

- Known graph topology.
- Monitoring of a small fraction of nodes.

Imperial College London

Problem Statement and Assumptions

Source localisation problem

- A source emits R rumours, at $t_0 = 0$.

- We observe some monitors, at discrete times $\mathrm{t} \in \{0, 1, \dots, T\}$.

- The probability of infection of a monitor i at time t is given by:

$$\tilde{F}_i(t) = \frac{R_i(t)}{R},$$

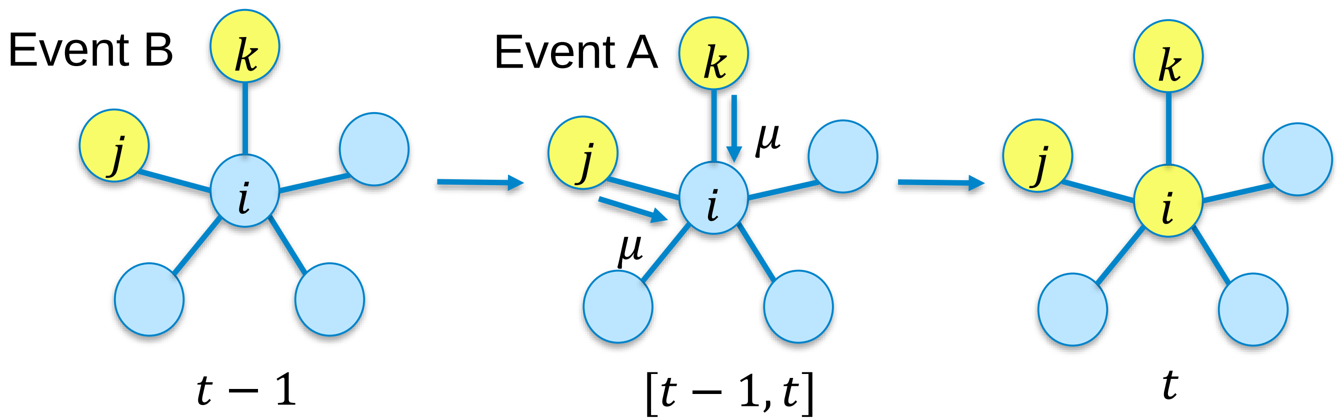
  where $R_i(t)$ is the number of rumours which have reached i by time t.

- We aim to leverage the divergence of the monitor measurements from an analytical probability of infection.

Imperial College London

Approach I to Model Diffusion in a Network

What is the probability a node i gets first infected at time t, $f_i(t)$?



$$f_i(t) = P(A \cap B) = P(A|B)P(B)$$

μ is the constant transmission rate

Derivation in spirit with the methods presented in:

[1] M. Gomez-Rodriguez, D. Balduzzi, B. Schölkopf. *Uncovering the Temporal Dynamics of Diffusion Networks.*

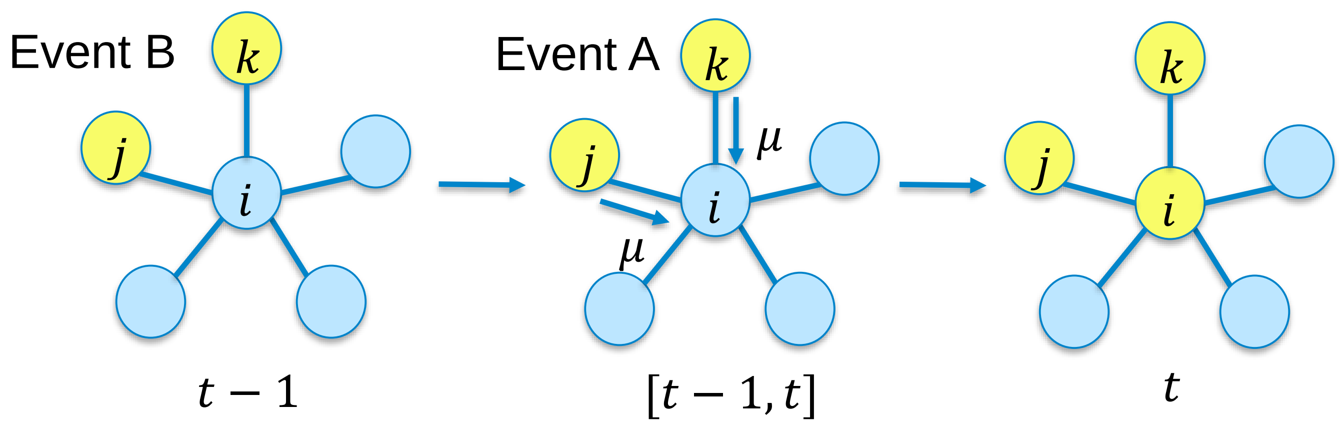
[2] A. Lokhov, M. Mézard, H. Ohta, L. Zdeborová. *Inferring the origin of an epidemic with a dynamic message-passing algorithm.*

[3] N. Ruhi, H. Ahn, B. Hassibi. *Analysis of Exact and Approximated Epidemic Models over Complex Networks.*

Imperial College London

Approach I to Model Diffusion in a Network

What is the probability a node i gets first infected at time t, $f_i(t)$?

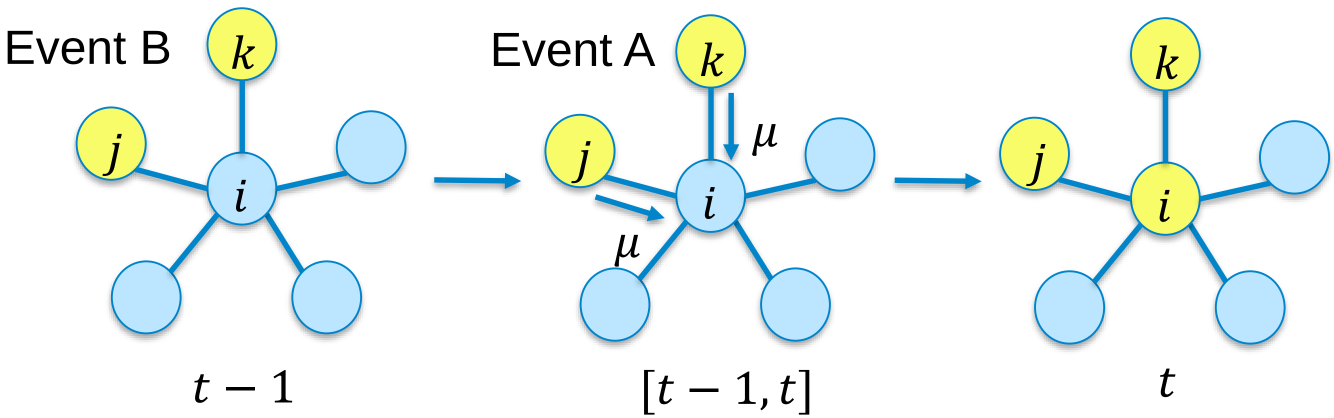


B is the event of node i being in a susceptible state at time $t-1$:

$$P(B) = \prod_{\tau=1}^{t-1} (1 - f_i(\tau))$$

Imperial College London

Approach I to Model Diffusion in a Network



$$P(A) = 1 - \prod_{j \in N_i} [1 - \mu \times F(x_j(t-1) = 1)]$$

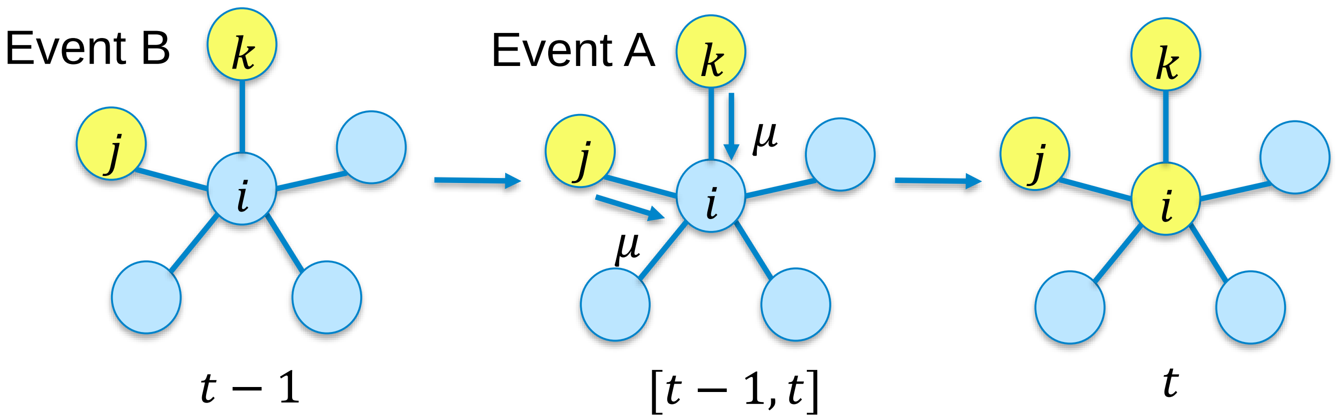
neighbour j infected

neighbour j does not transmit

none of neighbours transmit

Imperial College London

Approach I to Model Diffusion in a Network



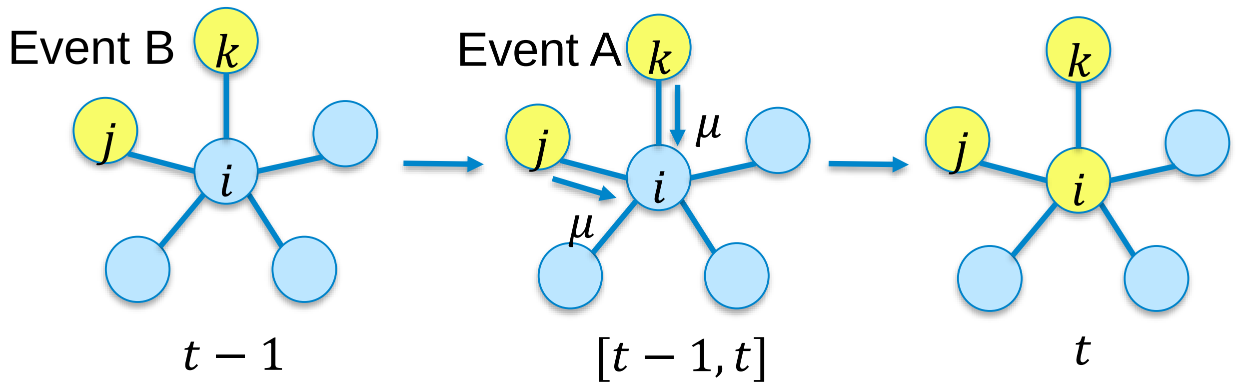
$$P(A) = 1 - \prod_{j \in N_i} [1 - \mu \times F(x_j(t-1) = 1)]$$

The probability i gets the rumour from at least one neighbour, given i was previously in a susceptible state is:

$$P(A|B) = 1 - \prod_{j \in N_i} [1 - \mu \times F(x_j(t-1) = 1 | \boldsymbol{x_i(t-1) = 0})]$$

Imperial College London

Approach I to Model Diffusion in a Network



The probability a node i gets first infected at time t, $f_i(t)$ is:

$$f_i(t) = \underbrace{\left[1 - \prod_{j \in N_i} \left(1 - \mu \times F\big(x_j(t-1) = 1 \,|\, x_i(t-1) = 0\big)\right)\right]}_{P(A|B)} \times \underbrace{\prod_{\tau=1}^{t-1} \left(1 - f_i(\tau)\right)}_{P(B)}$$

Imperial College London

Approach I to Model Diffusion in a Network

- The probability a node i gets first infected at time t is:

$$f_i(t) = \underbrace{[1 - \prod_{j \in N_i} (1 - \mu \times F(x_j(t-1) = 1 | x_i(t-1) = 0))]}_{P(A|B)} \times \underbrace{\prod_{\tau=1}^{t-1} (1 - f_i(\tau))}_{P(B)}$$

- We make the approximation:

$$f_i(t) \approx [1 - \prod_{j \in N_i} (1 - \mu \times F(x_j(t-1) = 1))] \times \prod_{\tau=1}^{t-1} (1 - f_i(\tau))$$

- The approximate probability a node i is infected at time t is:

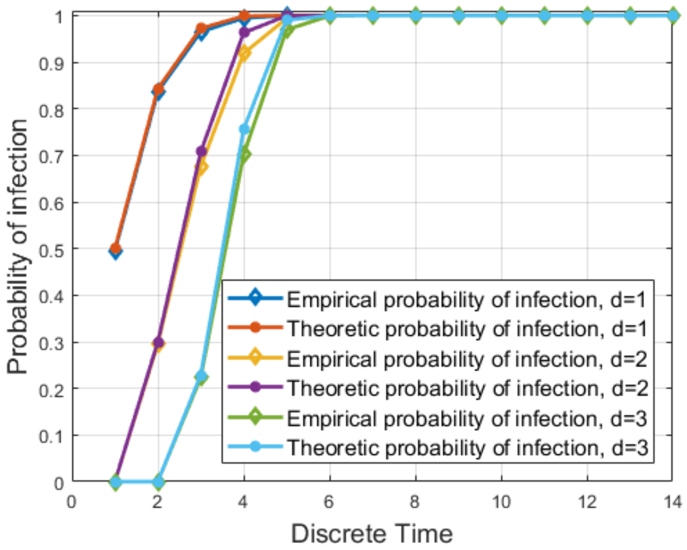
$$F_i(\tau) \approx \sum_{t=1}^{\tau} [1 - \prod_{j \in N_i} (1 - \mu \times F(x_j(t-1) = 1))] \times \prod_{\theta=1}^{t-1} (1 - f_i(\theta))$$

Approach I to Model Diffusion in a Network

- The approximate probability a node i is infected at time t is:

$$F_i(\tau) \approx \sum_{t=1}^{\tau} [1 - \prod_{j \in N_i} (1 - \mu \times F(x_j(t-1) = 1))] \times \prod_{\theta=1}^{t-1} (1 - f_i(\theta))$$

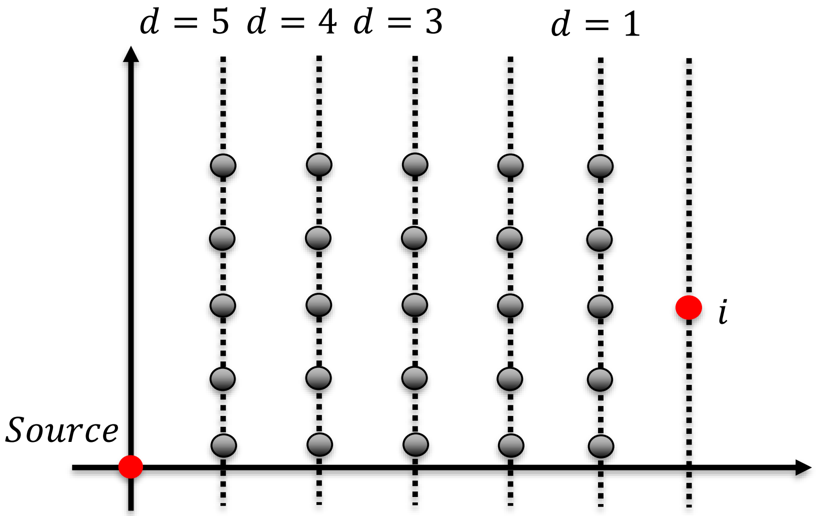
- Spreading of 1000 Rumors, small-world network, 200 Nodes, for distances 1, 2, and 3 from the source:



Imperial College London

Approach II to Model Diffusion in a Network

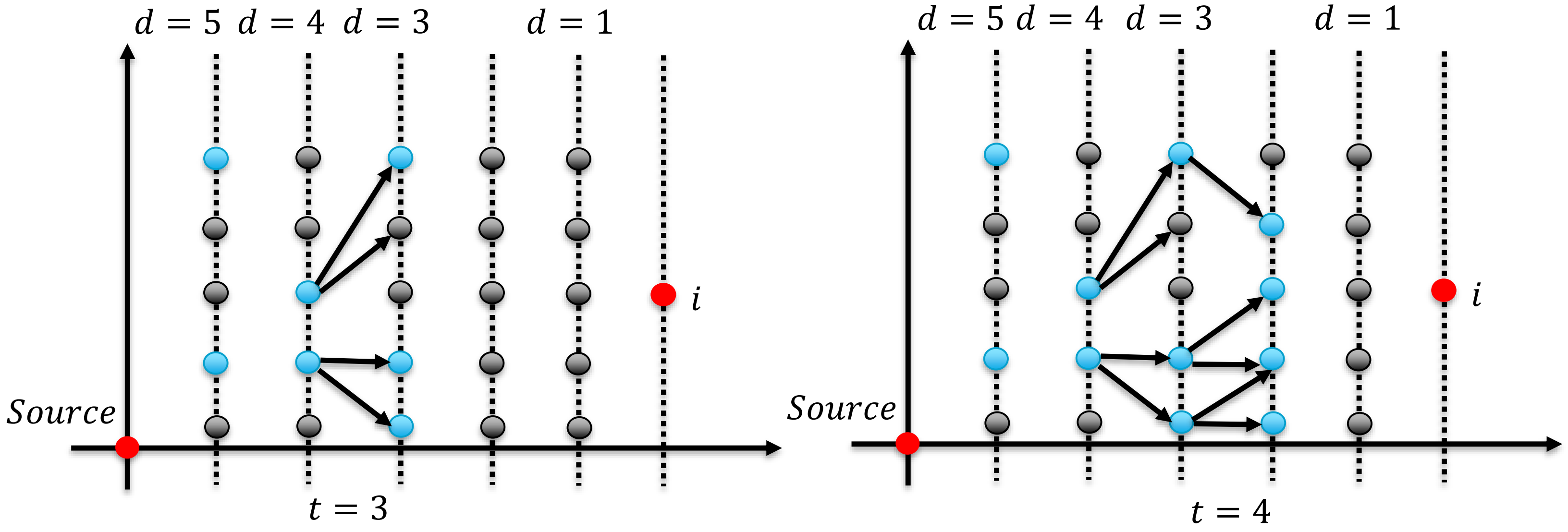
- Probability of infection based on the shortest distance to the source.
- Arrange the nodes according to the shortest distance to the destination.
- What is the probability of first infection of a node i at distance d, at time t?



Imperial College London

Approach II to Model Diffusion in a Network

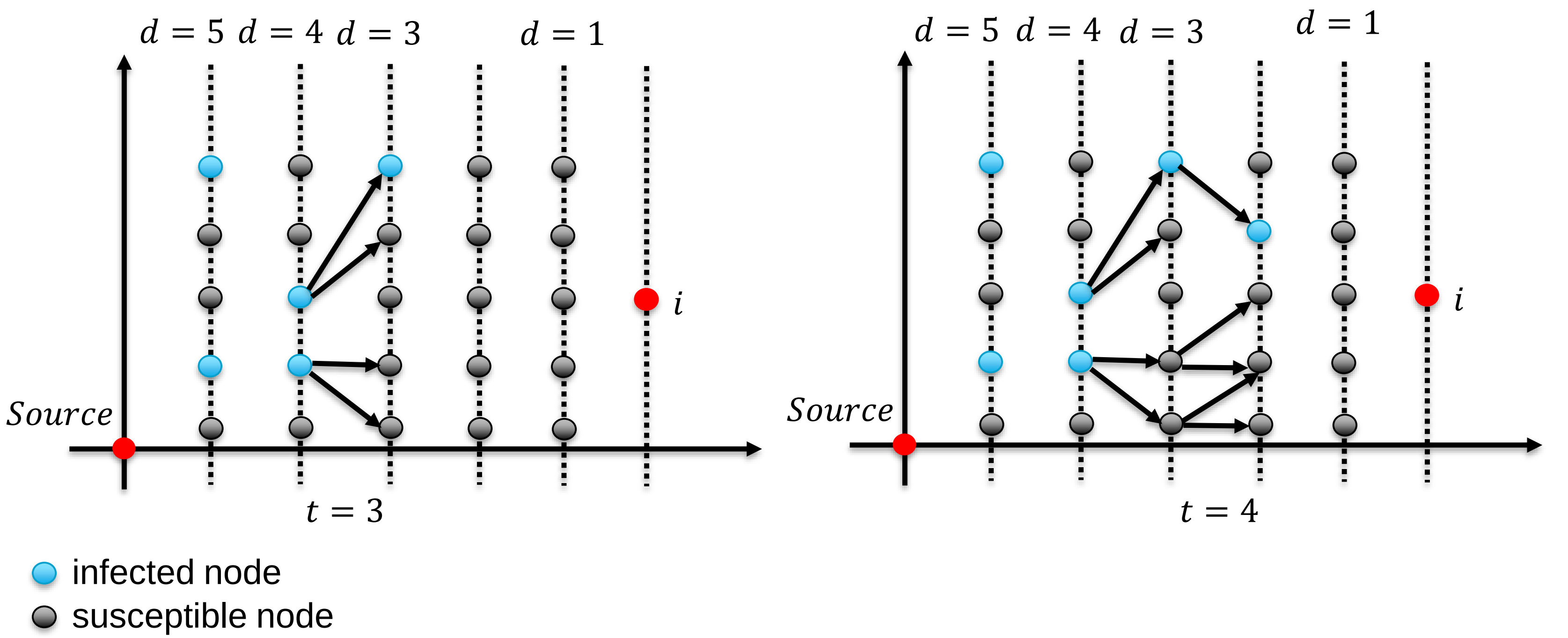
- What is the probability of first infection of a node i at distance d, at time t?
- Success: move closer to node i.



infected node
susceptible node

Approach II to Model Diffusion in a Network

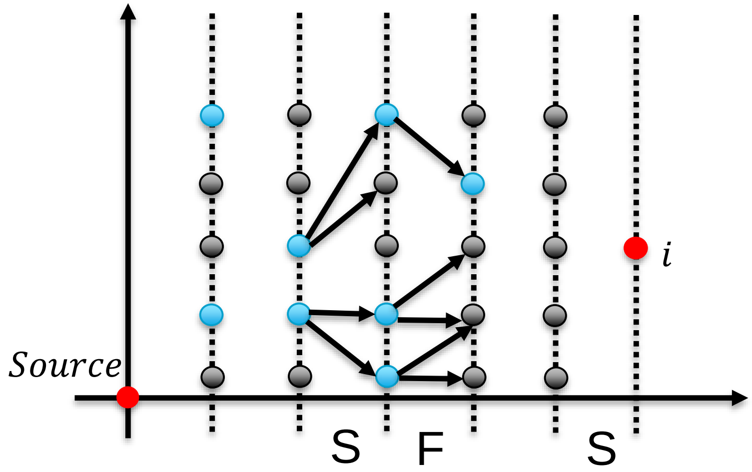
- What is the probability of infection of a node i at distance d, at time t?
- Failure: not spreading the rumour to a sufficient number of nodes closer to the destination.



Imperial College London

Approach II to Model Diffusion in a Network

- Number of paths is the number of ways to choose $d-1$ *success* steps of $t-1$ time steps: $\binom{t-1}{d-1}$.
- Probability of success: p_S.
- Probability of each path: $p_S^d \times (1-p_S)^{t-d}$.
- Approximate $p_S = \alpha_d \mu$, where μ is the constant transmission rate in the graph.



Imperial College London

Approach II to Model Diffusion in a Network

- Number of paths is: $\binom{t-1}{d-1}$.
- Probability of each path: $p_S^d \times (1-p_S)^{t-d}$.
- Set $p_S = \alpha_d \mu$, where μ is the constant transmission rate in the graph.
- The probability of first infection is:

$$f_d(t) = \underbrace{(\alpha_d \mu)^d}_{p_S} \times (1-\alpha_d\mu)^{t-d} \times \underbrace{\binom{t-1}{d-1}}_{\text{\# of paths from source to destination}}$$

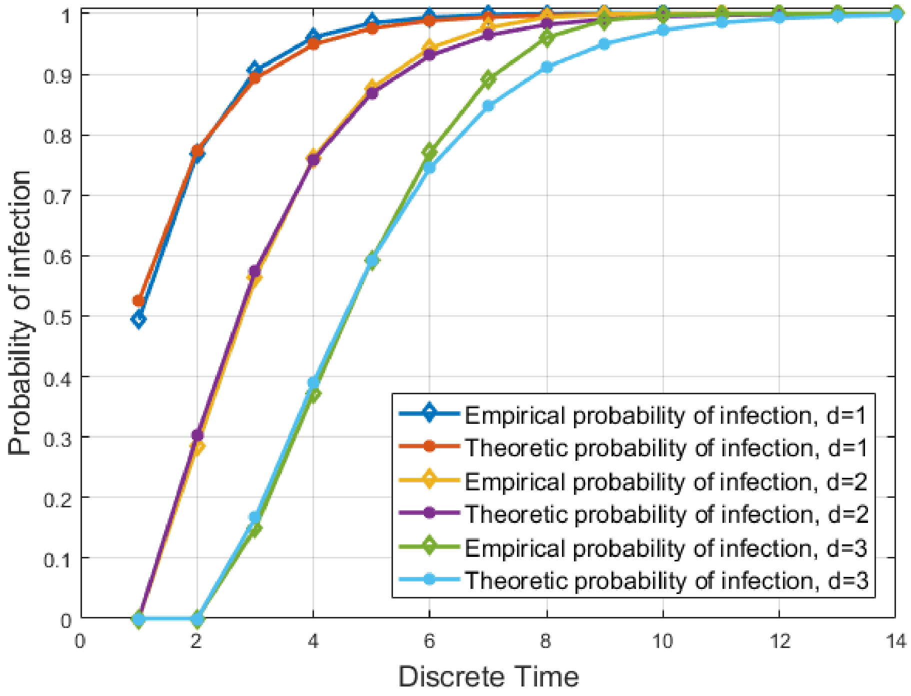
- The probability of infection of a node at distance d from the source at time τ is:

$$F_d(\tau) \approx \sum_{t=d}^{\tau} (\alpha_d\mu)^d \times (1-\alpha_d\mu)^{t-d} \times \binom{t-1}{d-1}$$

Imperial College London

Approach II to Model Diffusion in a Network

- 1000 Rumors, small-world network, 200 Nodes:



Imperial College London

Single Diffusion Source Detection Algorithm

- **Estimate the distances** between each monitor i and the potential source, by computing the dissimilarity between the observed $\tilde{F}_i(t)$ and the theoretical $F_d(t)$.

- Create a set of potential sources using **triangulation**.

- Select the most likely rumour origin, using the approximate model of infection, given a rumour source s:

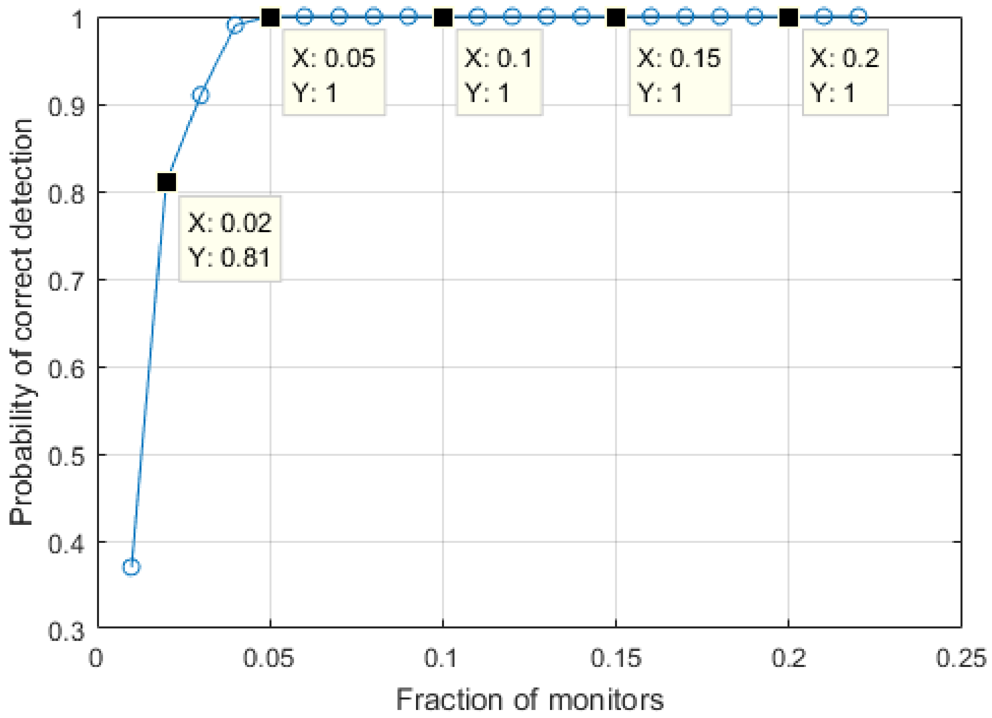
$$F(x_i(\tau) = 1|s) \approx \sum_{t=1}^{\tau}[1 - \prod_{j \in N_i}(1 - \mu \times F(x_j(t-1) = 1))] \times \prod_{\theta=1}^{t-1}(1 - f_i(\theta))$$

- For each potential source s, compute the dissimilarity between empirical $\tilde{F}_i(t)$ and analytical $F(x_i(T) = 1|s)$. The **most likely rumour origin** is the node with the **lowest dissimilarity.**

Imperial College London

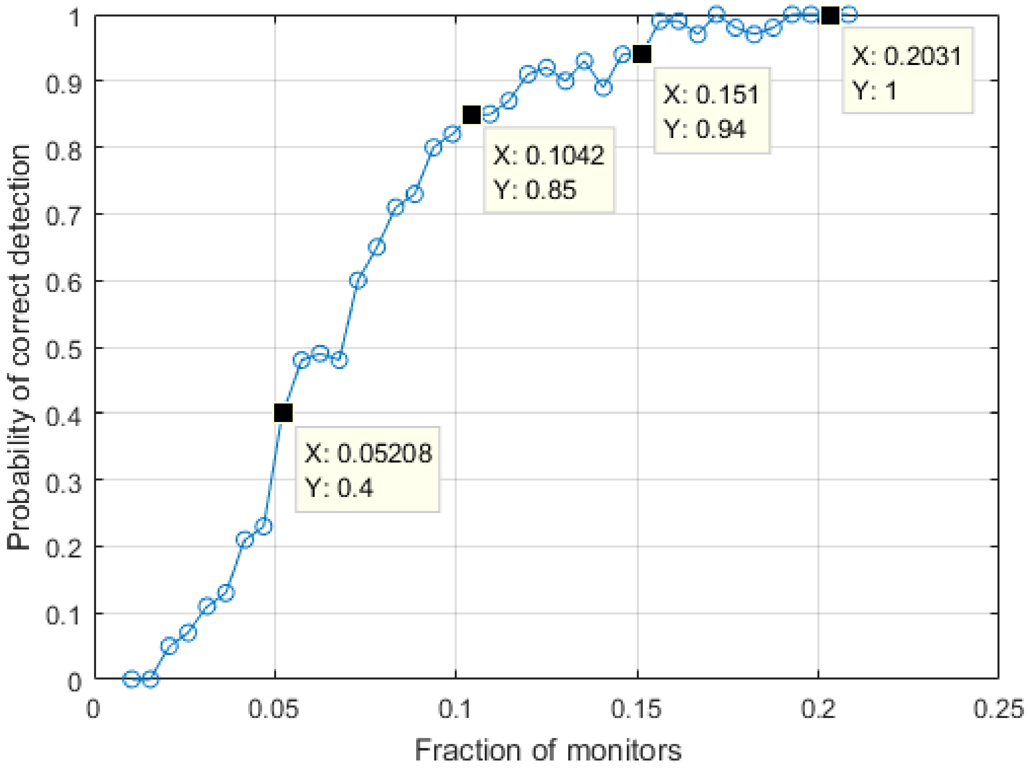
Simulations

- 10 Rumors, small-world network, 1000 Nodes, $\mu = 0.5$, 100 experiments.



Simulations

- 10 Rumors, Facebook network, 192 Nodes, $\mu = 0.5$, 100 experiments.



Imperial College London

Conclusion

- Mathematical models of information propagation, which accurately capture the diffusion process.
- Source detection algorithm, which assumes:
  - Single source, which emits multiple rumours.
  - All rumours start at the same time, which is known.
  - A finite set of monitor nodes is observed at discrete times.
- Future extensions:
  - Source detection with unknown start time.
  - Multiple source detection algorithm.

Imperial College London

Thank you for listening!

Imperial College London

How do we find the optimal parameters α_d in the distance-dependent probabilities?

- The distance-dependent probability of infection for a node at distance d, at time t is:

$$F_d(t) = \sum_{\tau=d}^{t} (\mu \times \alpha_d)^d \times (1 - \mu \times \alpha_d)^{\tau-d} \times \binom{\tau-1}{d-1}$$

- Artificially spread a number of rumours from a random node in the network, and obtain the empirical probabilities $\tilde{F}_i(t)$.
- The optimal parameter α_d minimizes the dissimilarity between $F_d(t)$ and $\tilde{F}_i(t)$ for a particular distance d:

$$\alpha_d^{opt} = argmin_{\alpha_d} \sum_{i \in N_d} \sum_{t=0}^{T} ||F_d(t) - \tilde{F}_i(t)||^2 ,$$

where N_d is the set of nodes at shortest distance d from the source.

How do we estimate the shortest distances between monitor nodes and the source?

- We find the dissimilarity between the distance-dependent analytical probability of infection $F_d(t)$, and the observed infection probability at a node i, using mean-squared error.

- Then, the optimal distance for a monitor i is:

$$d_{i,s} = argmin_d \sum_{t=0}^{T} ||F_d(t) - \tilde{F}_i(t)||^2$$

- We select as potential sources all the nodes at distance $d_{i,s}$ from node i.

Imperial College London

How do we select the most likely rumour origin?

- Select the most likely rumour origin, using the approximate model of infection, given a rumour source s:

$$F(x_i(T) = 1|s) = \sum_{t=1}^{T}[1 - \prod_{j \in N_i} 1 - \mu \times F\left(x_j(t-1) = 1\right)] \times \prod_{\tau=1}^{t-1}(1 - f_i(\tau))$$

- For each potential source s, compute the dissimilarity between the observed infection probabilities of all monitors, and the theoretic model of infection:

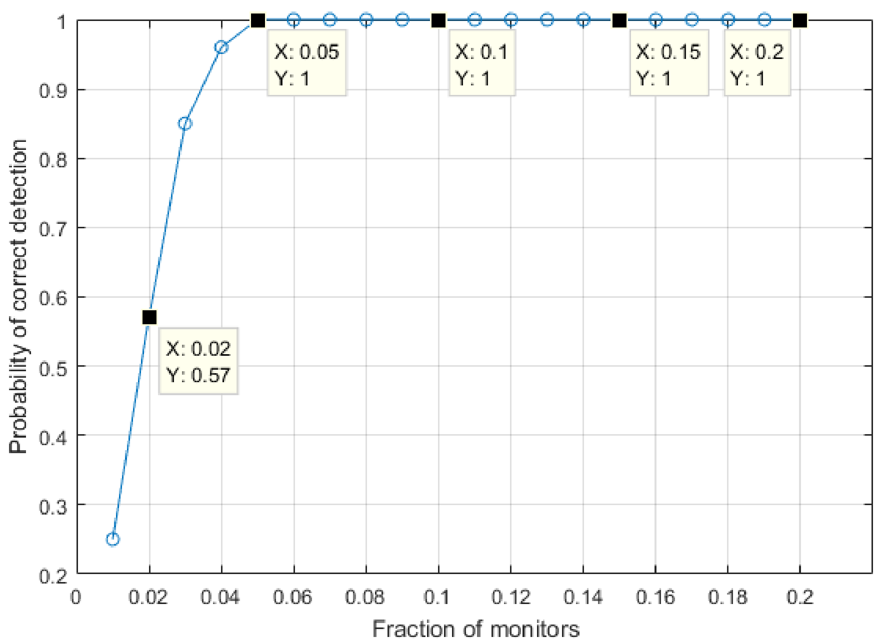
$$\bar{C}(s) = \sum_{i}\sum_{t=0}^{T} ||F(x_i(t) = 1|s) - \tilde{F}_i(t)||^2$$

- The most likely rumour origin is the node with the lowest dissimilarity.

Imperial College London

More simulation results

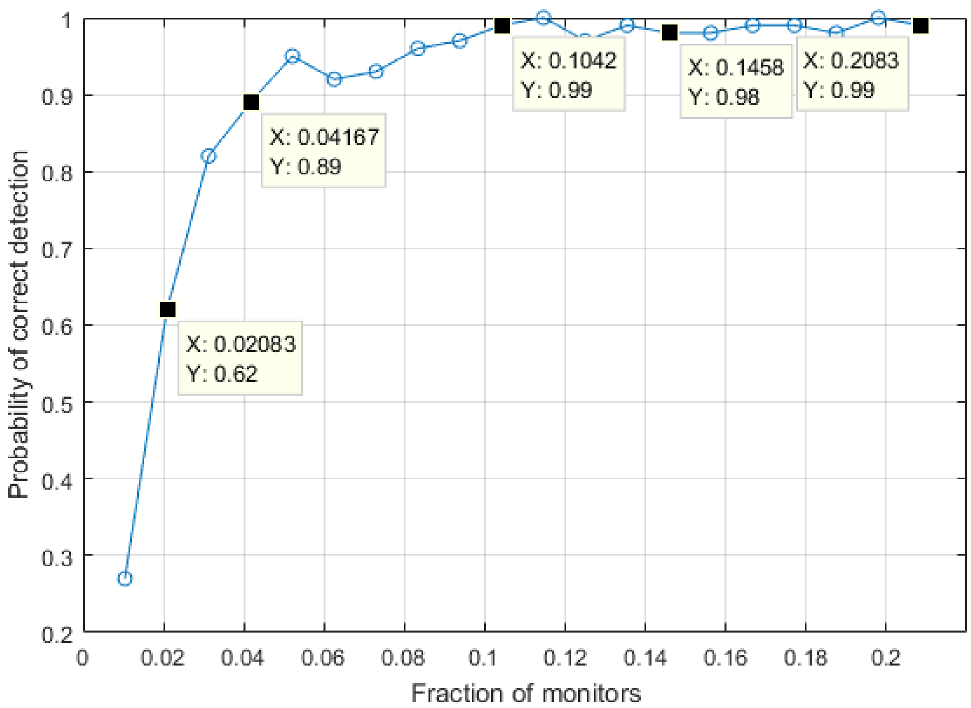
- 10 Rumors, small-world network, 1000 Nodes, *varying* spreading probability



Imperial College London

More simulation results

- 10 Rumors, Facebook network, 192 Nodes, *varying* spreading probability



Probability of infection

- A node has the infection at time t if it got initially infected at any of the times before, $\tau = 1,2, \dots, t$.

- The events of a node getting the initial infection at different times are mutually disjoint.

- Hence, the probability of infection is given by the sum of the likelihoods of first infection at different discrete times:

$$F_i(t) = \sum_{\tau=1}^{t} f_i(\tau)$$

Imperial College London

Probability of being susceptible

- A node is susceptible at time t if it didn't get infected at any of the times before, $\tau = 1,2, \ldots, t$.
- The events of a node not getting the initial infection at different times are mutually disjoint.
- Hence:

$$\bar{F}_i(t) = \prod_{\tau=1}^{T} 1 - f_i(\tau)$$

Imperial College London

Distance-dependent probability of infection

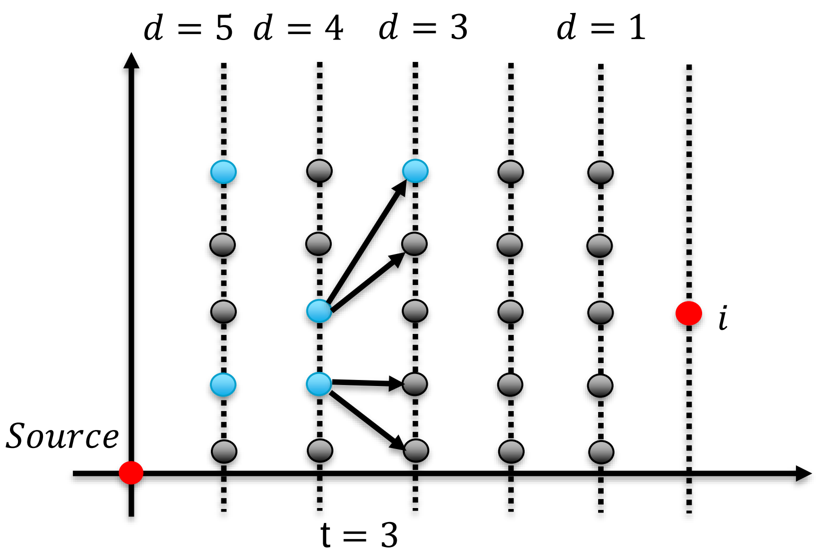
- Number of paths is: $\binom{t-1}{d-1}$.
- Probability of each path: $p_S^d \times (1 - p_S)^{t-d}$.
- A node at distance d gets infected if *any* succession of d success steps, and $t - d$ failure steps happens.
- Different successions of S and F events are mutually disjoint.
- Hence, the probability of first infection is:

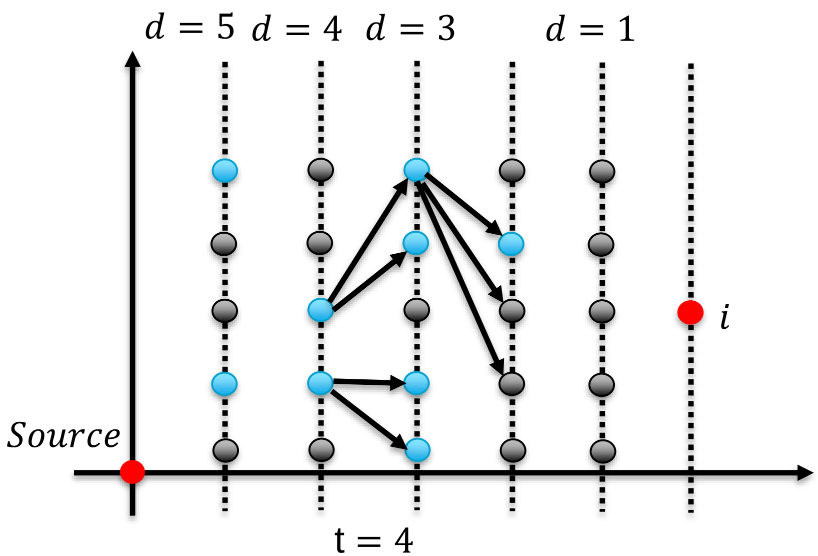
$$f_d(t) = \underbrace{(\alpha_d \mu)}_{p_S}{}^d \times (1 - \alpha_d \mu)^{t-d} \times \underbrace{\binom{t-1}{d-1}}_{\text{\# of paths from source to destination}}$$

Imperial College London

Distance-dependent probability of infection

- A node at distance d gets infected if *any* succession of d success steps, and $t-d$ failure steps happens.
- There can by a success following a failure, at the next time step.





Imperial College London

Comparison to existing methods

- The authors in [1] propose a Monte Carlo method for single source estimation, with unknown infection time. In a **random geometric graph**, the probability of the origin to be within the first **10% ranked nodes** is around **0.5** when **observing 5%** of the network, increasing to **0.9** when **observing the full network**.

- In a **small-world network**, our method achieves correct detection probability of **0.75** when **observing 5%** of the network, and **1** when **observing the full network**. The number of **rumours is 2**, and the rumour **start time is known.**

[1] A. Agaskar and Y. M. Lu. *A fast Monte Carlo algorithm for source localization on graphs.*