Examining Key Properties of Diffusion Models for Large-Scale Real-World Networks

Daniel Bernardes, Matthieu Latapy, Fabien Tarissan

Algotel 2012

29 Mai 2012

Diffusion in a network

A **diffusion trace** is composed of:

1. an underlying graph (the network)
2. chronological data of who transmitted information to whom

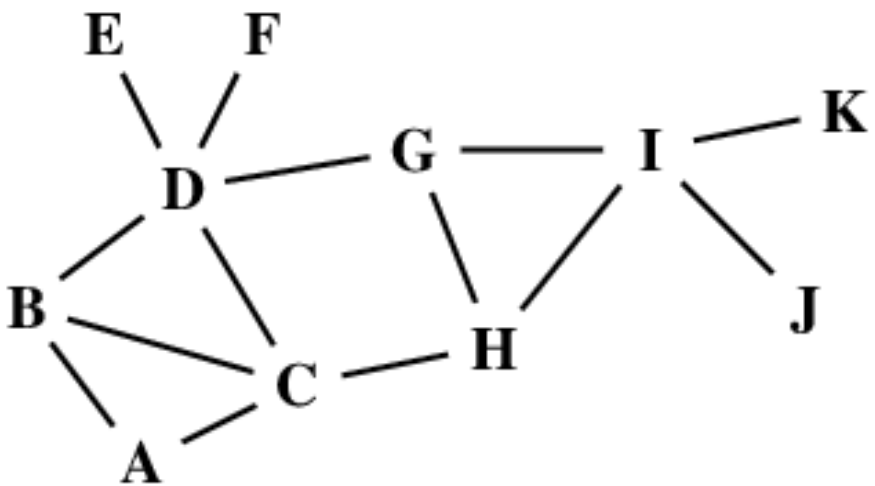
Typical examples:

- virus on a contact or proximity network
- gossip in a social network
- files in a peer-to-peer network
- ...

Diffusion in a network

A **diffusion trace** is composed of:

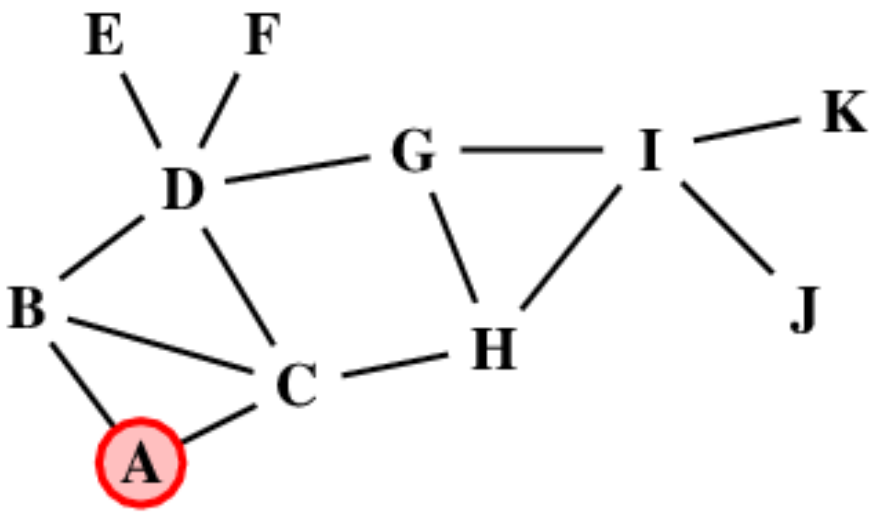
1. an underlying graph (the network)
2. chronological data of who transmitted information to whom



Diffusion in a network

A **diffusion trace** is composed of:

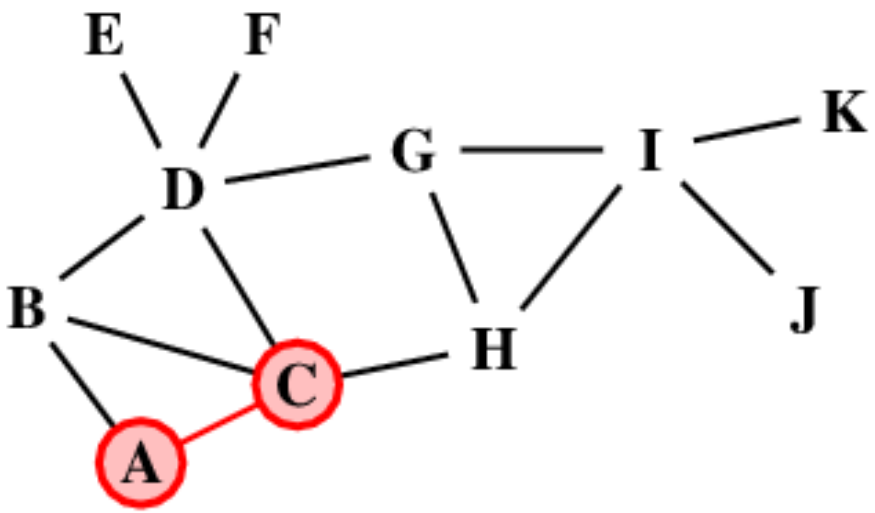
1. an underlying graph (the network)
2. chronological data of who transmitted information to whom



Diffusion in a network

A **diffusion trace** is composed of:

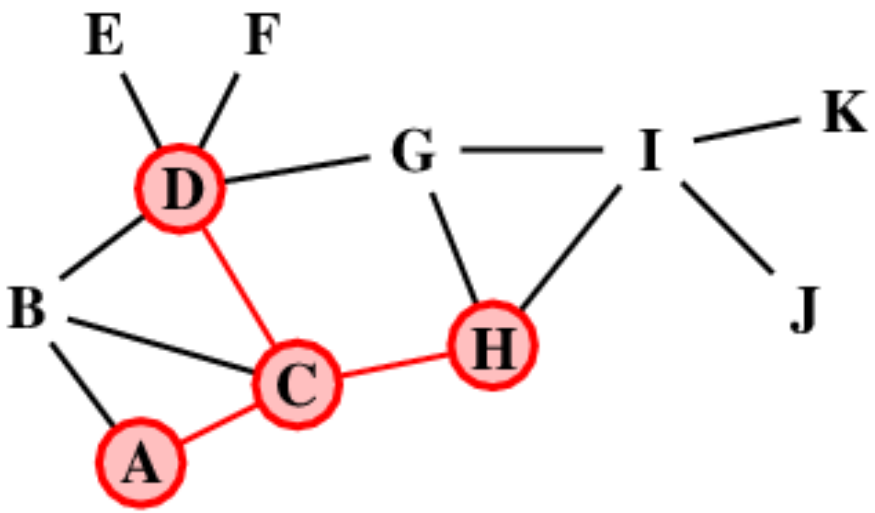
1. an underlying graph (the network)
2. chronological data of who transmitted information to whom



Diffusion in a network

A **diffusion trace** is composed of:

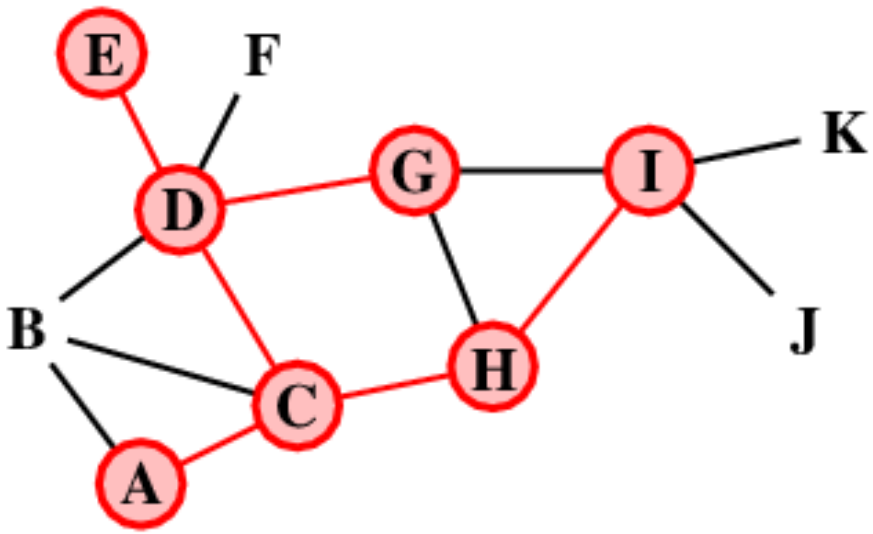
1. an underlying graph (the network)
2. chronological data of who transmitted information to whom



Diffusion in a network

A **diffusion trace** is composed of:

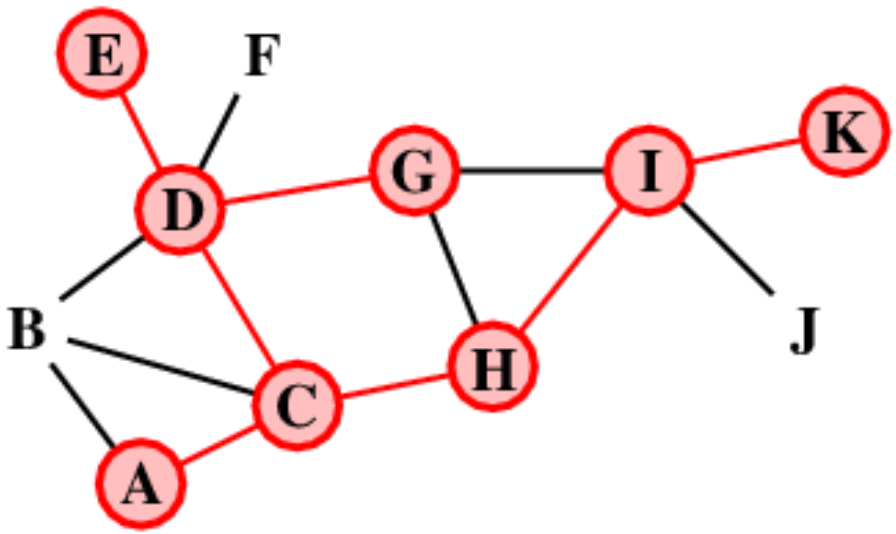
1. an underlying graph (the network)
2. chronological data of who transmitted information to whom



Diffusion in a network

A **diffusion trace** is composed of:

1. an underlying graph (the network)
2. chronological data of who transmitted information to whom



Model

Popular approach in the literature: diffusion as an epidemic

SIR model

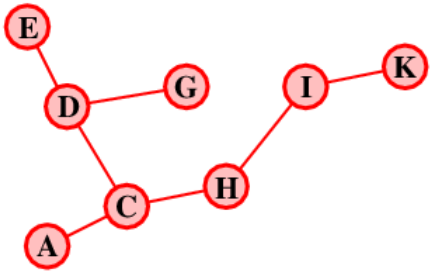
- node states: *susceptible* $\rightarrow$ *infected* $\rightarrow$ *removed*
- infected nodes spread to each of its neighbors with prob. p

Goal: validation using large-scale, real-world diffusion data

Protocol

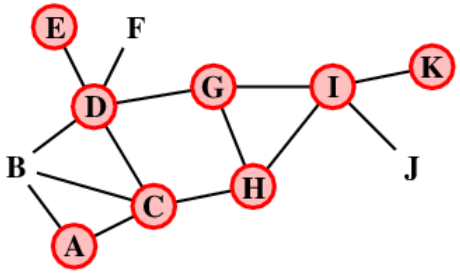
- calculate key properties for the observed diffusion trace
- calibrate SIR model using real data to simulate diffusion
- compare key properties of observed and simulated diffusions

Data challenge: obtaining the complete diffusion trace



Spreading cascade

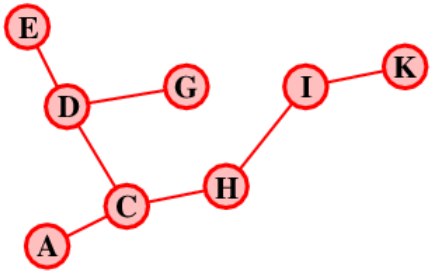
underlying network?



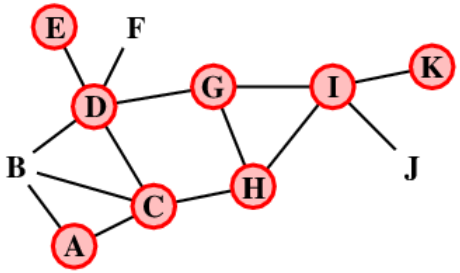
Nodes reached by the diffusion

transmission links?

Data challenge: obtaining the complete diffusion trace



Spreading cascade
underlying network?



Nodes reached by the diffusion
transmission links?

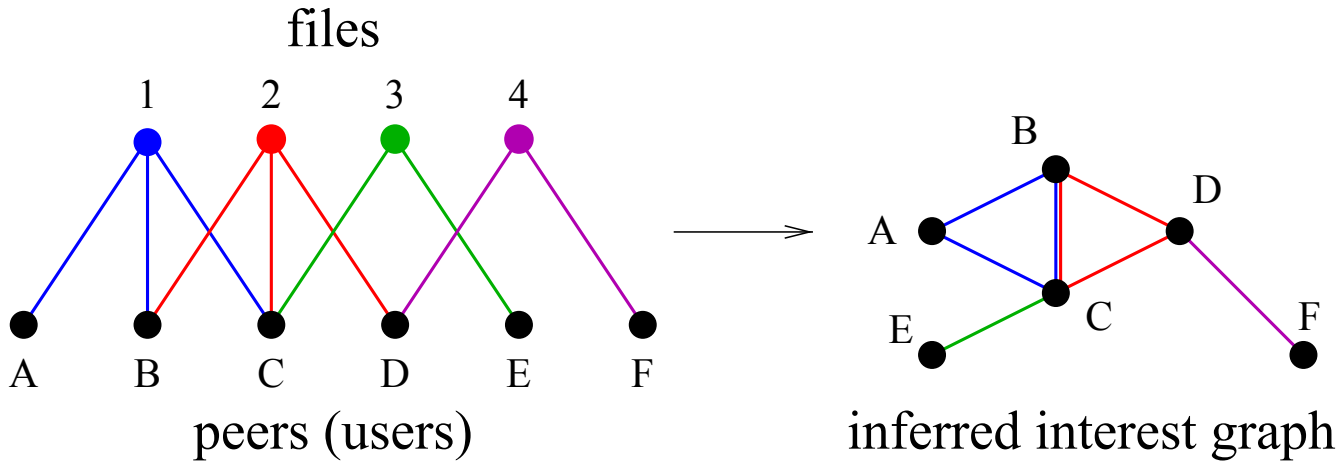
Our dataset: trace of file queries to an eDonkey server

satisfied query: (*timestamp*, *providers id*, *peer id*, *file id*)

6h with 2 million peers, 800k files and 23 million queries

Interest graph

peers are connected if they have requested/provided the same file



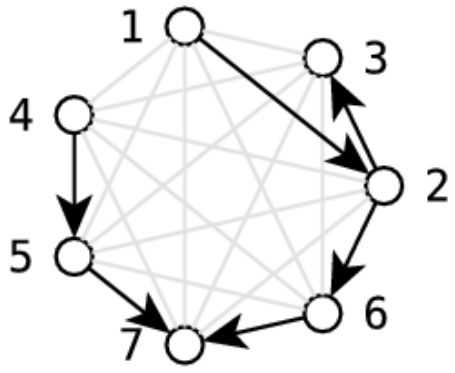
the diffusion takes place in the interest graph by construction

Spreading cascades

Trace log example

t0	1	2	F
t1	2	3	F
t2	4	5	F
t3	2	6	F
t4	6	7	F
t4	5	7	F
t5	7	3	F

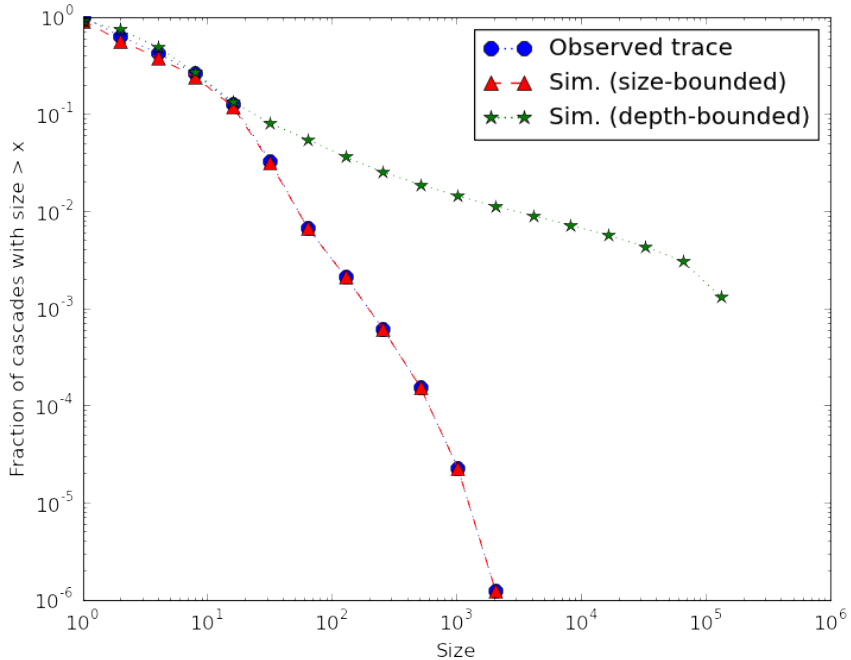
Spreading cascade

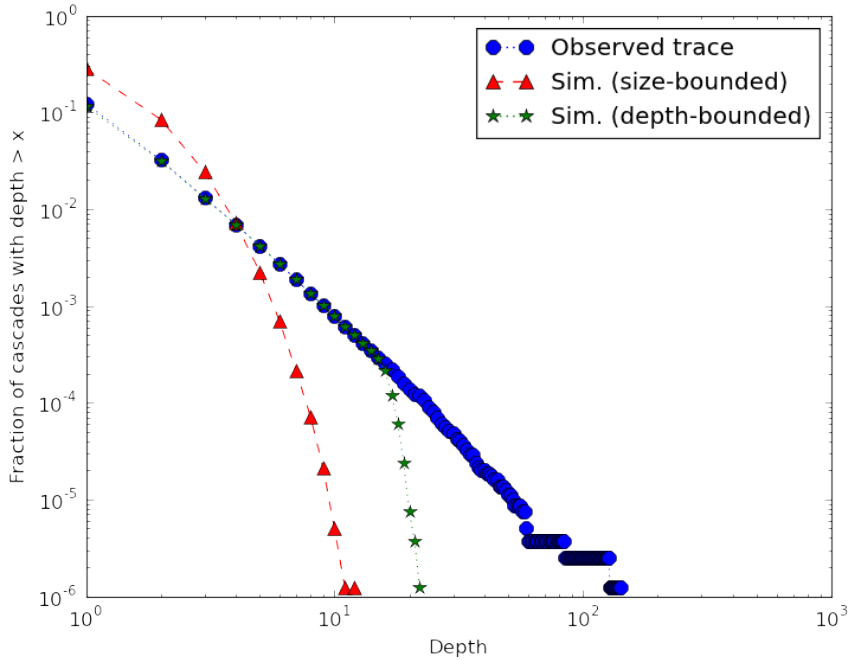


Key spreading cascade properties

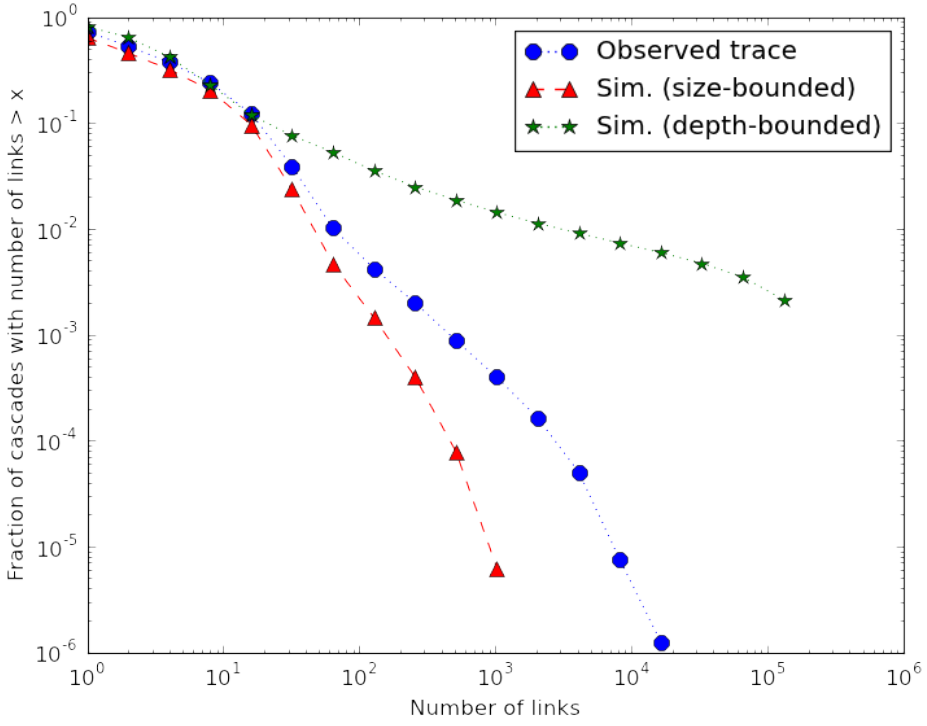
- size (number of nodes)
- depth (length of the longest path)
- number of links

Real vs Simulated: size and depth





Real vs Simulated: num. of links



Results

Real cascades are denser and more *elongated* than simulated ones

Conclusion and perspectives

Conclusion

The classical SIR model fails to reproduce key cascade properties in this context

Major perspectives

- improved epidemic models: heterogeneous SIR
- enhanced underlying network: weighted interest graph
- alternative diffusion models: adoption/threshold models

Questions welcomed!

complexnetworks.fr