The Abdus Salam
International Centre for Theoretical Physics

United Nations
Educational, Scientific
and Cultural Organization

International Atomic
Energy Agency

World Year of PHYSICS 2005

SMR.1656 - 5

School and Workshop on
Structure and Function of Complex Networks

16 - 28 May 2005

Traceroute-like exploration of unknown networks: a statistical analysis

Alain BARRAT
Laboratoire de Physique Theorique
Universite de Paris-Sud
Batiment 210
91405 Orsay Cedex
FRANCE

Traceroute-like exploration of unknown networks: a statistical analysis

A. Barrat, LPT, Université Paris-Sud, France

I. Alvarez-Hamelin (LPT, France)
L. Dall'Asta (LPT, France)
A. Vázquez (Notre-Dame University, USA)
A. Vespignani (LPT, France)

COSIN

cond-mat/0406404 to appear in LNCS;
cs.NI/0412007 to appear in TCS;
Phys. Rev E 71 (2005)

DELIS

Plan of the talk

- Context: sampling of complex networks
- Model for `traceroute`–like sampling
- Theoretical approach
- Numerical results
- Conclusions

Main characteristics of complex networks

- Small-world networks
- Heterogeneous networks: broad degree distributions
- Dynamical evolution, self-organisation

...In contrast with usual random graphs

....development of new paradigms (evolving networks, etc...)

Reliability of empirical data ?

Heterogeneity of networks: empirical fact

- social networks: various samplings/networks with similar results
- transportation network: reliable data
- biological networks: incomplete samplings
- Internet: various mapping processes

Statistical analysis of the sampling process

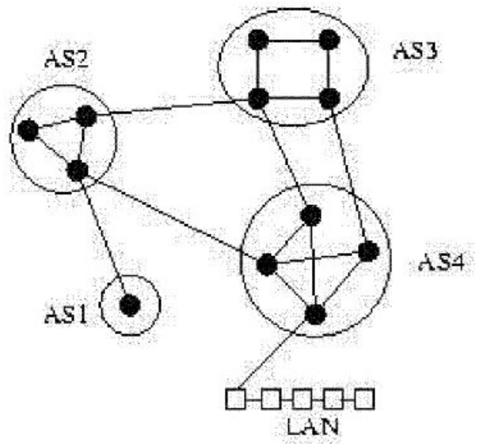
Internet representation

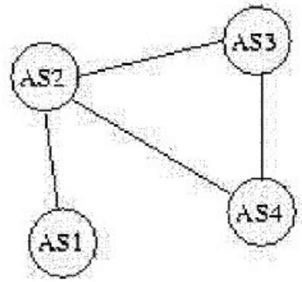
- Multi-probe reconstruction (router-level)
- Use of BGP tables for the Autonomous System level (domains)

Many projects (CAIDA, NLANR, RIPE, IPM, PingER...)

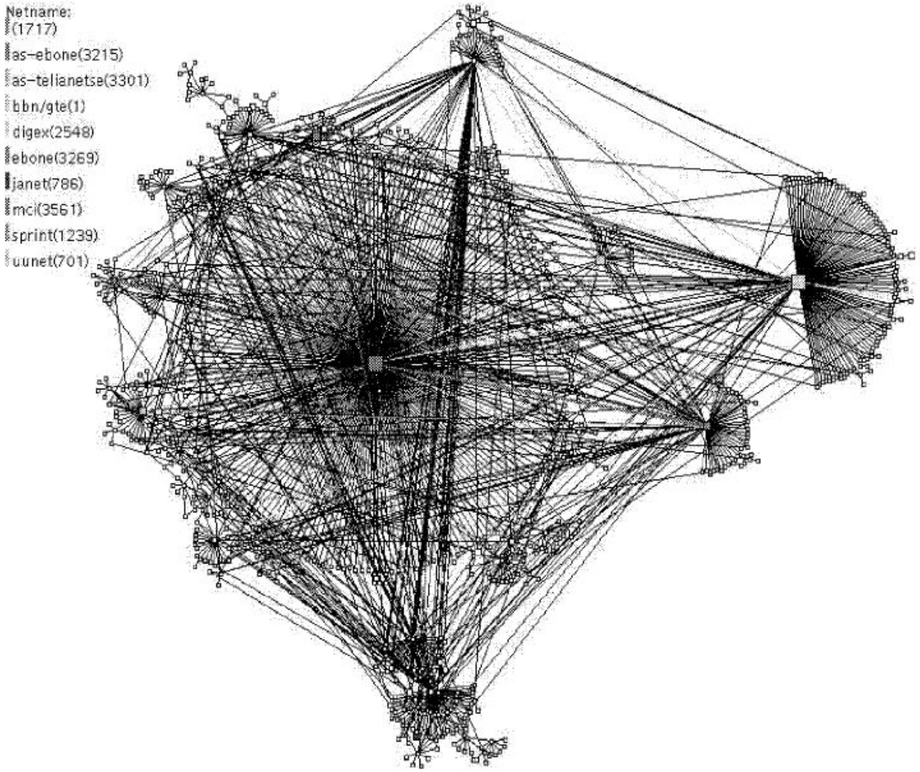
Graph representation

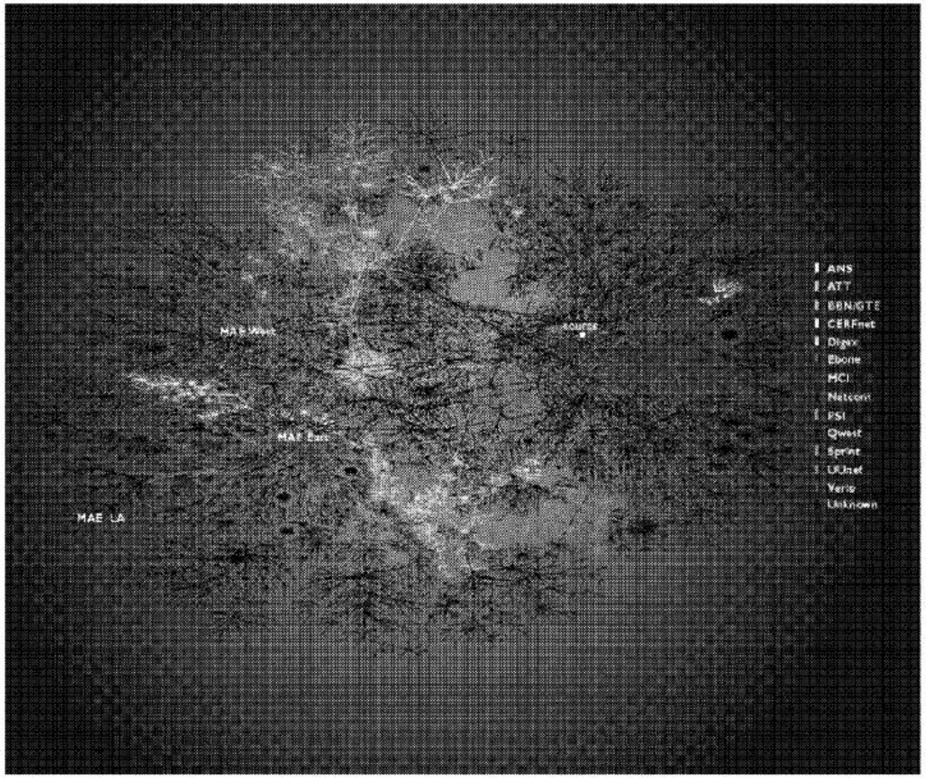
different granularities





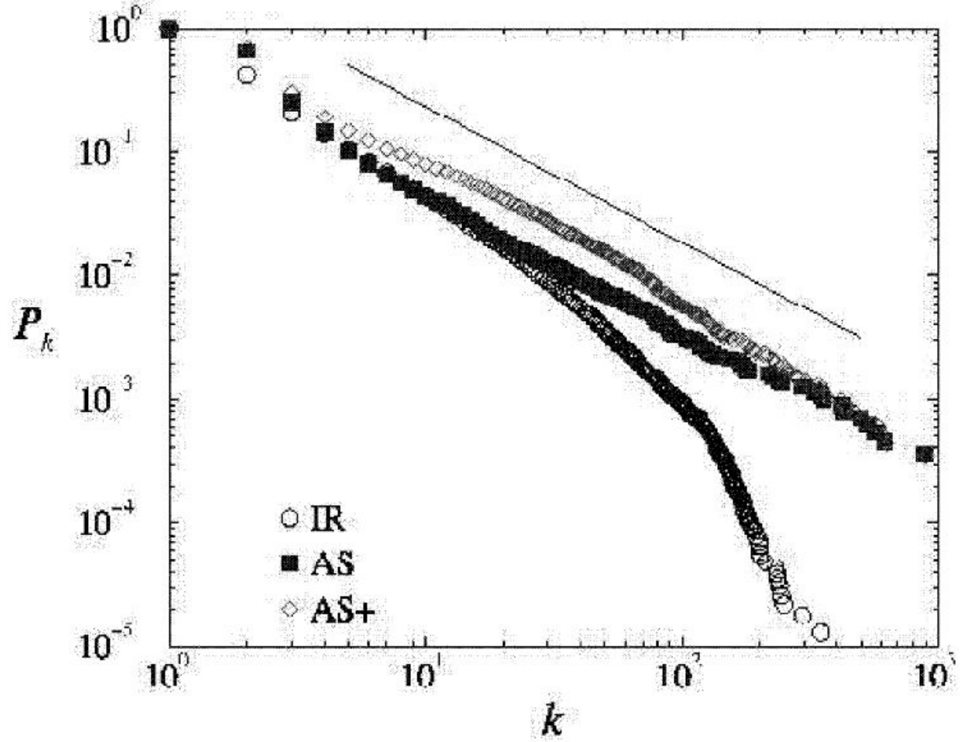
Large-scale visualizations





Topological analysis

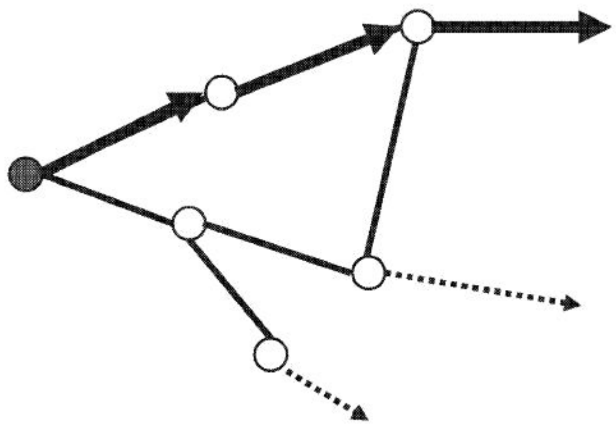
Broad connectivity distributions:
obtained from mapping projects



Result of a sampling:
is this reliable ?

Sampling biases

Internet mapping: traceroute



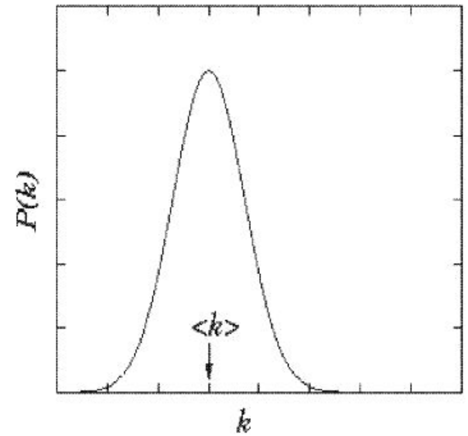
=> spanning tree

Sampling is incomplete
Lateral connectivity is missed (edges are underestimated)
Finite size sample

Sampling biases

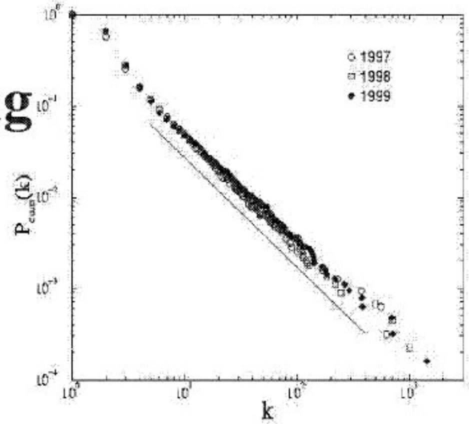
- **Vertices and edges best sampled in the proximity of sources**
- **Bad estimation of some topological properties**

Statistical properties of the sampled graph may sharply differ from the original one



Bad sampling

?



Lakhina et al. 2002

Clauset & Moore 2005

De Los Rios & Petermann 2004

Guillaume & Latapy 2004

Evaluating sampling biases

Real graph G=(V,E)
(known, with given properties)

↓ simulated sampling

Sampled graph G'=(V',E')

↓

Analysis of G', comparison with G

What we have done

I. model for traceroute sampling

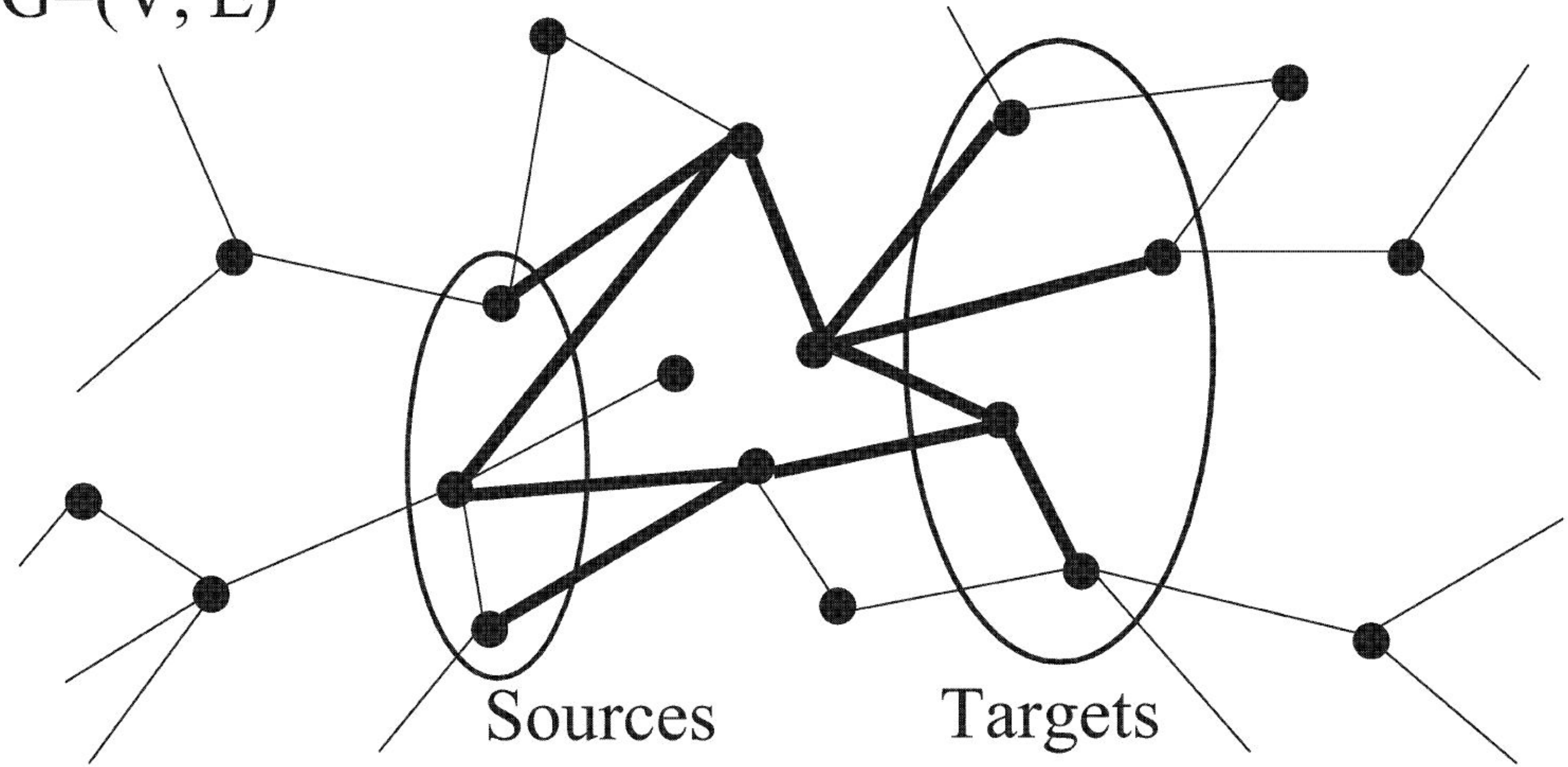
II. analytical analysis with approximations
=> link between topological properties of the sampled network and the sampling biases

III. numerical analysis on various networks with different topologies

Model for `traceroute`

First approximation: union of shortest paths

G=(V, E)

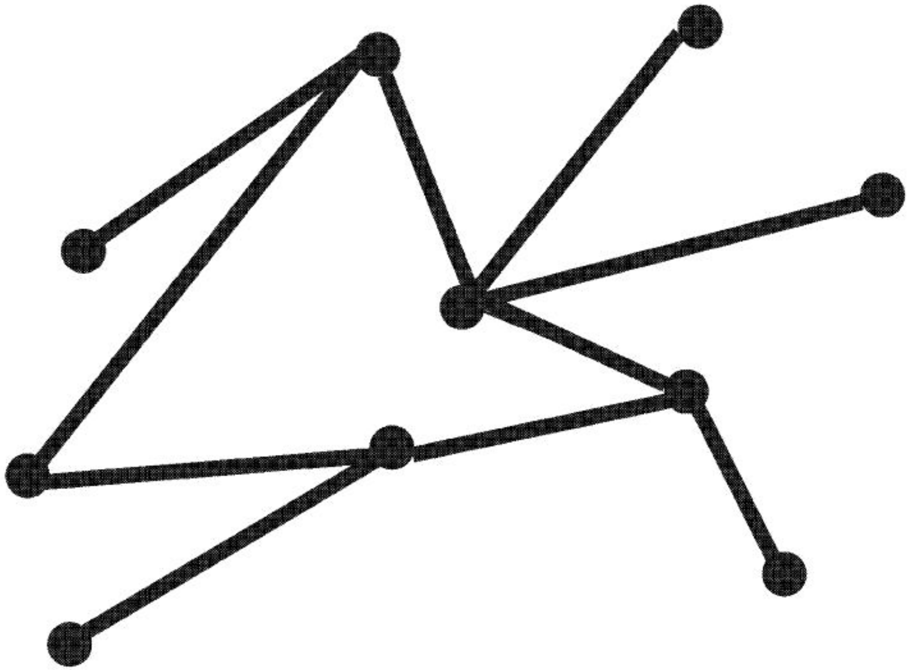


NB: Unique Shortest Path

Model for `traceroute`

First approximation: union of shortest paths

G'=(V', E')



Very simple model, but: allows for some analytical and numerical understanding

More formally…

***G = (V, E)*: sparse undirected graph with a set of**

N_S *sources* $\mathcal{S} = \{ i_1, i_2, \ldots, i_{N_S}\}$

N_T *targets* $\mathcal{T} = \{j_1, j_2, \ldots, j_{N_T}\}$

randomly placed.

The *sampled graph* G'=(V',E') is obtained by considering the union of all the `traceroute`-like paths connecting source-target pairs.

PARAMETERS: $\rho_S = \dfrac{N_S}{N}, \quad \rho_T = \dfrac{N_T}{N}, \quad \varepsilon = \dfrac{N_S N_T}{N}$ (probing effort)

Usually $N_S = O(1)$, $\rho_T = O(1)$

Analysis of the mapping process

For each set $\Omega = \{ \mathcal{S},\mathcal{T} \}$, the indicator function that a given edge *(i, j)* belongs to the sampled graph is

$$\boxed{\pi_{ij} = 1 - \prod_{l \neq m} \left(1 - \sum_{s=1}^{N_S} \delta_{li_s} \sum_{t=1}^{N_T} \delta_{mj_t} \sigma_{ij}^{(l,m)} \right)}$$

with

$$\sigma_{ij}^{(l,m)} = \begin{cases} 1 & \text{if } (i,j) \in path \text{ between } l,m \\ 0 & \text{otherwise,} \end{cases}$$

Mean-field statistical analysis

Averaging over all the possible realizations of the set $\Omega = \{\, \mathcal{S}, \mathcal{T} \,\}$,

$$\langle \pi_{ij} \rangle = 1 - \left\langle \prod_{l \neq m} \left[1 - \sum_{s=1}^{N_S} \delta_{l i_s} \sum_{t=1}^{N_T} \delta_{m j_t} \sigma_{ij}^{(l,m)} \right] \right\rangle \approx 1 - \prod_{l \neq m} (1 - \rho_S \rho_T \langle \sigma_{ij}^{(l,m)} \rangle)$$

WE HAVE NEGLECTED CORRELATIONS !!

$$<\pi_{ij}> \simeq 1 - \exp(- \rho_S \, \rho_T \, b_{ij})$$

- *usually* $\rho_S \, \rho_T << 1$
- $b_{ij} = \sum_{s \neq i \neq j \neq t \in V} \left[\frac{\sigma_{ij}^{(s,t)}}{\sigma^{(s,t)}} \right]_{USP} = \sum_{l \neq m \neq i \neq j \in V} \langle \sigma_{ij}^{(l,m)} \rangle$

betweenness

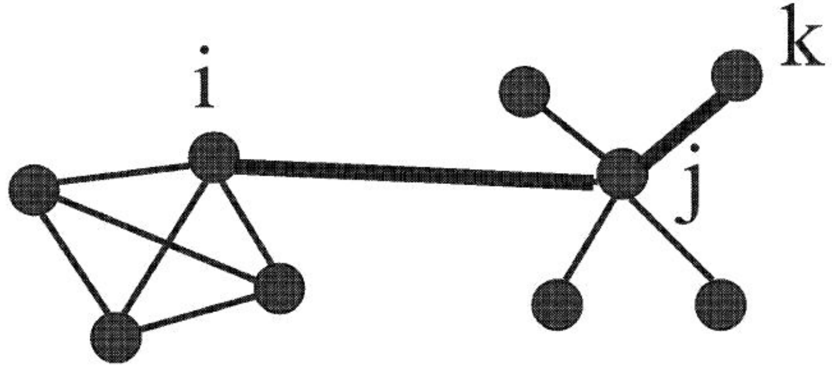
Betweenness centrality b

for each pair of nodes (l,m) in the graph, there are

ν^{lm} shortest paths between l and m

ν_{ij}^{lm} shortest paths going through ij

b_{ij} is the sum of ν_{ij}^{lm} / ν^{lm} over all pairs (l,m)



ij: large betweenness

jk: small betweenness

Similar concept:
node betweenness b

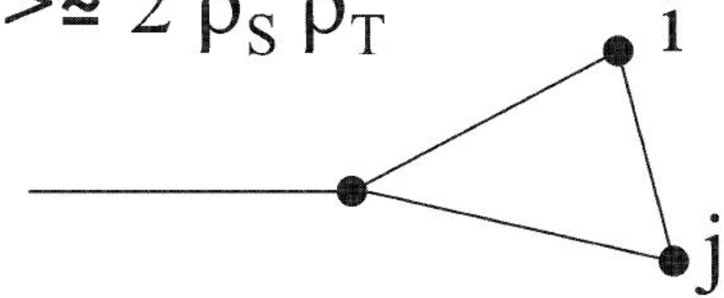
Also: flow of information if each individual sends a message to all other individuals

Consequences of the analysis

$$<\pi_{ij}> \simeq 1 - \exp(-\rho_S \rho_T b_{ij})$$

- Smallest betweenness

  $b_{ij}=2 \Rightarrow <\pi_{ij}> \simeq 2 \rho_S \rho_T$



  (i.e. i and j have to be source and target to discover i-j)

- Largest betweenness

  $b_{ij}=O(N^2) \Rightarrow <\pi_{ij}> \simeq 1$

Results for the vertices

$$\langle \pi_{ij} \rangle \simeq 1 - \exp(- \varepsilon\, b_{ij}/N)$$

$\langle \pi_i \rangle \simeq 1 - (1-\rho_T) \exp(- \varepsilon\, b_i/N)$ (discovery probability)

$N^*_k/N_k \simeq 1 - \exp(- \varepsilon\, b(k)/N)$ (discovery frequency)

$\langle k^* \rangle/k \simeq \varepsilon\, (1 + b(k)/N)/k$ (discovered connectivity)

Summary:

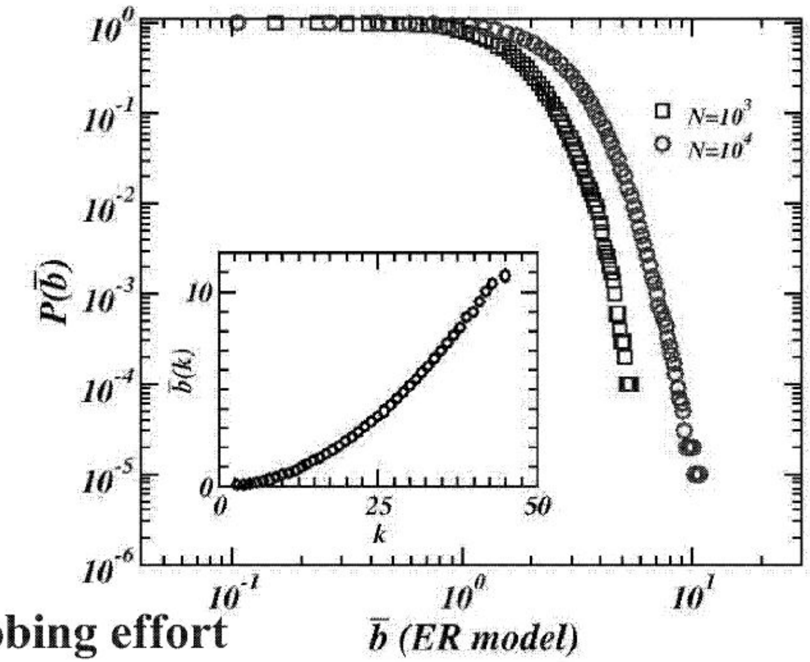
- **discovery probability strongly related with the *centrality* ;**
- vertex discovery favored by a ***finite density*** of targets ;
- accuracy increased by increasing ***probing effort*** ε.

Numerical simulations

1. Homogeneous graphs:

(ex: ER random graphs)

- peaked distributions of k and b
- narrow range of betweenness



=> Good sampling expected only for high probing effort

$$\boxed{\varepsilon >> \max\left[\overline{b}^{-1}, \overline{b_e}^{-1}\right]}$$

Numerical simulations

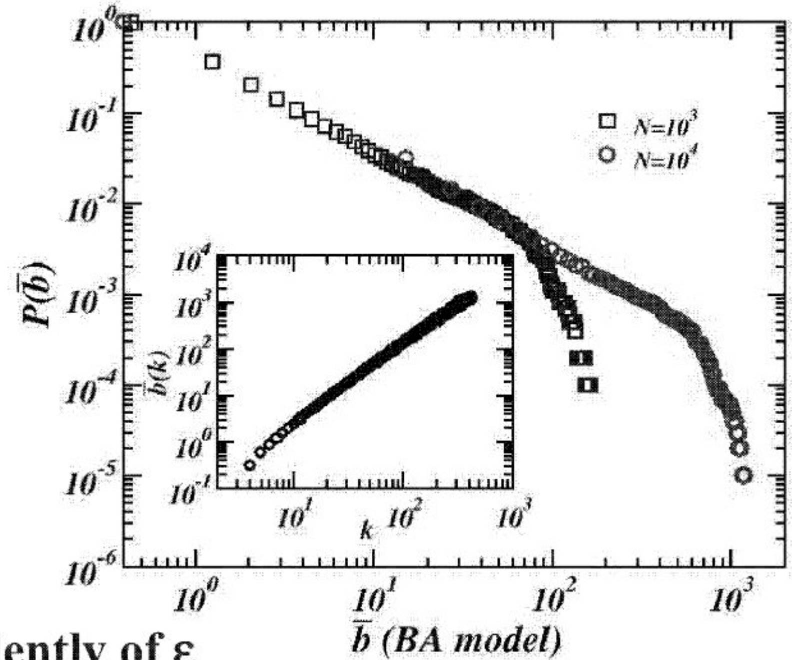
1. ***Homogeneous graphs***
2. ***Heavy-tailed graphs***

(ex: Scale-free BA model)

- broad distributions of k and b

P(k) ~ k^{-3}

- large range of available values

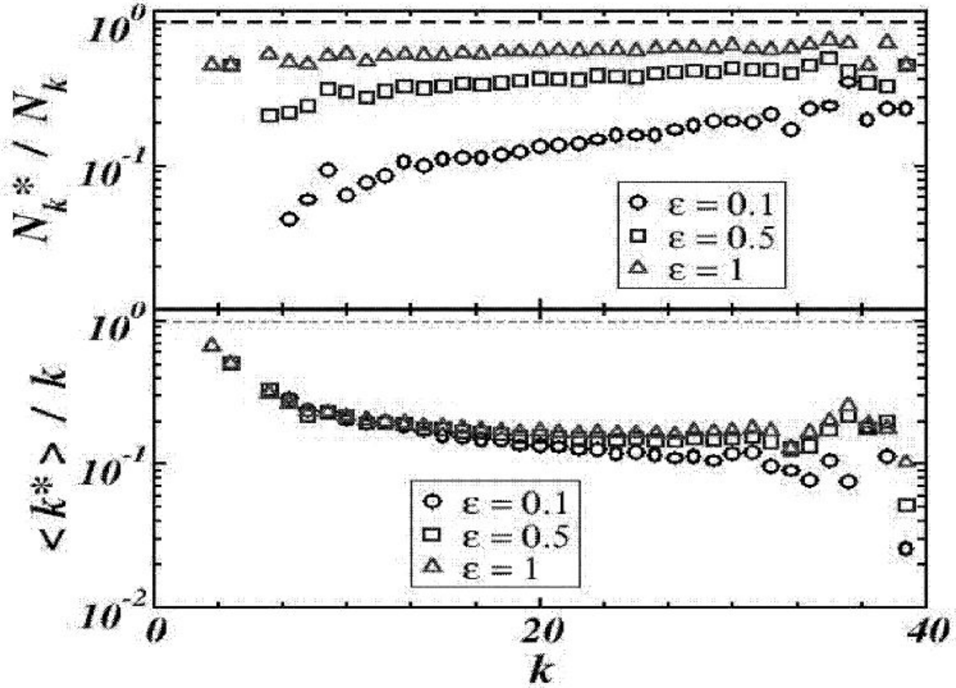


=> Expected: Hubs well-sampled independently of ε

$$\boxed{k \gg \varepsilon^{-1/\beta}} \qquad (b(k) \sim k^{\beta})$$

Numerical simulations

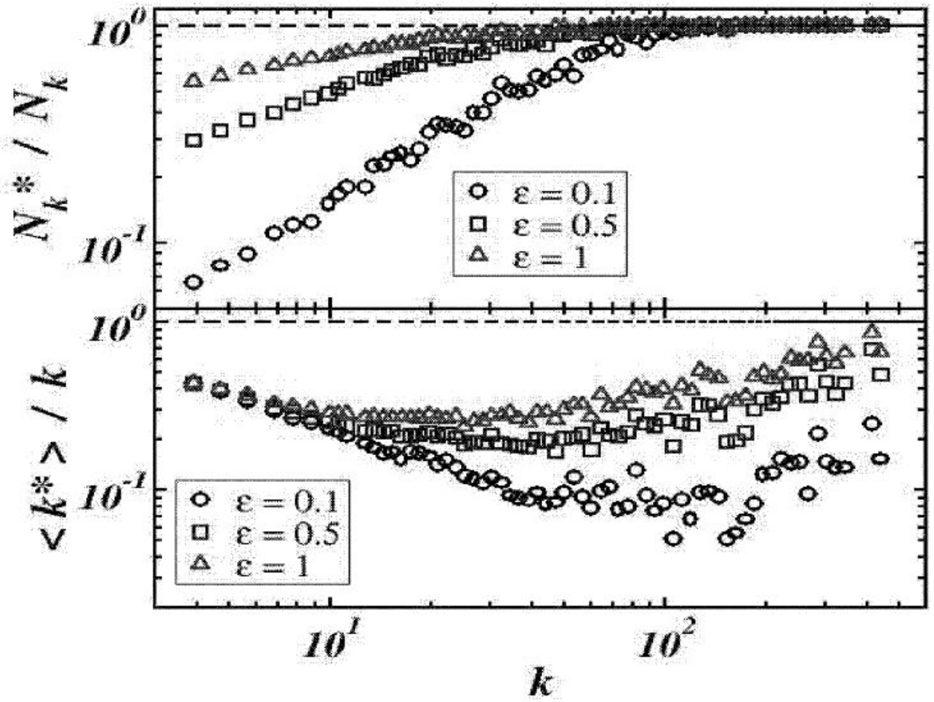
Homogeneous graphs



Homogeneously pretty badly sampled

Numerical simulations

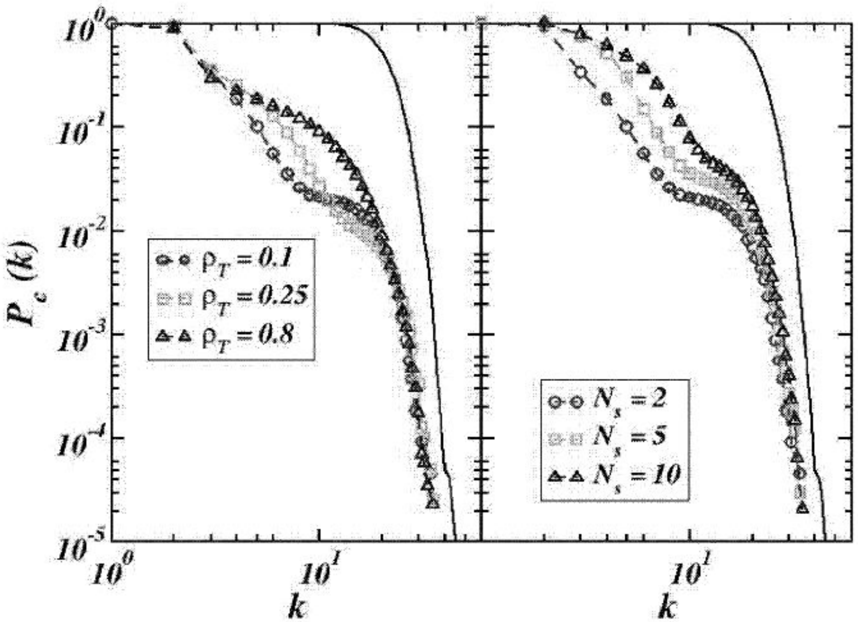
Scale-free graphs

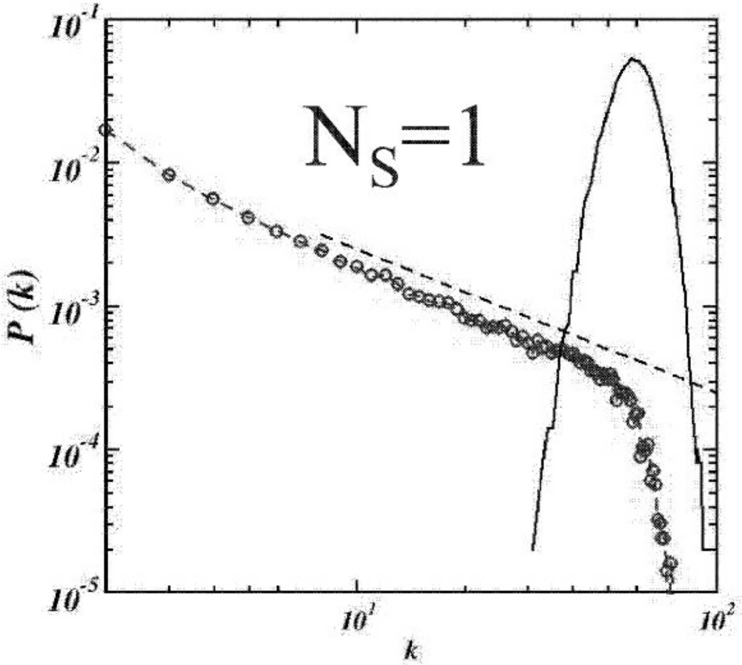


Hubs are well discovered

Numerical simulations

Homogeneous graphs

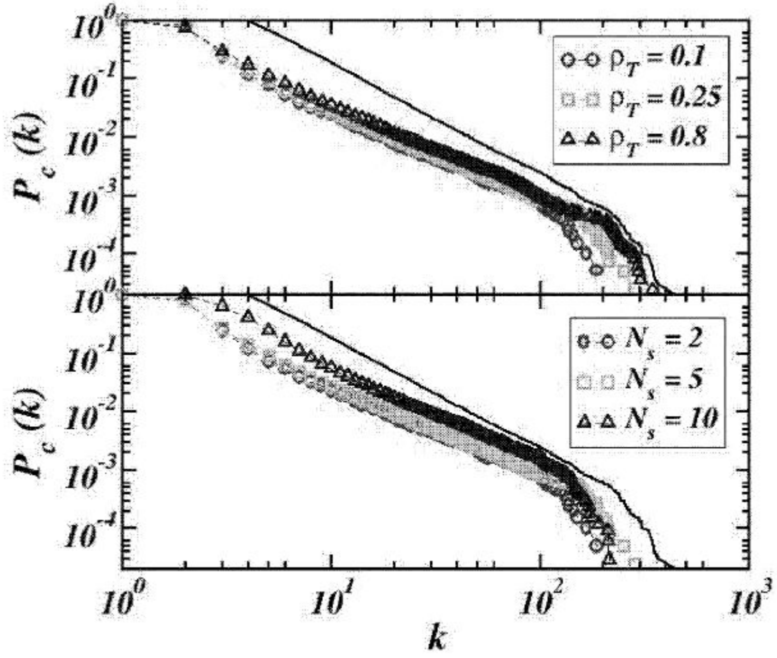




- *heavy-tailed $P^*(k)$ only for $N_S = 1$ (cf Clauset and Moore 2005)*
- ***cut-off** around <k> => large, unrealistic <k> needed*
- *bad sampling of $P(k)$*

Numerical simulations

Scale-free graphs



- *good sampling,* ***especially of the heavy-tail****;*
- *almost independent of N_S;*
- *slight bending for low degree (less central) nodes => bad evaluation of the exponents.*

Summary

- **Analytical** approach to a traceroute-like sampling process
- Link with the topological properties, in particular the betweenness
- Usual random graphs more "difficult" to sample than heavy-tails
- Heavy-tails well sampled
- Bias yielding a sampled scale-free network from a homogeneous network:
  - only in few cases
  - <k> has to be unrealistically large

Take-home message

Heavy tails properties are a genuine feature of the Internet

however

Quantitative analysis might be strongly biased (wrong exponents...)

Perspectives

- **Optimized strategies:**
  - **separate influence of ρ_T, ρ_S**
  - **location of sources, targets** cf also Guillaume and Latapy 2004
  - **investigation of other networks**
- **Results on redundancy issues**
- **Estimation of the real size of a network from a sampling ?**
- **Massive deployment**

www.tracerouteathome.net ; www.netdimes.org

and...

- **The internet is a weighted networks**

  bandwidth, traffic, efficiency, routers capacity

- **Data are scarse and on limited scale**
- **Interaction among topology and traffic**