Satellite Workshop on
"Complex Networks: Dynamics and Topology Interplay"
(The European Conference on Complex Systems, Dresden (Germany)
(4-5 October 2007)

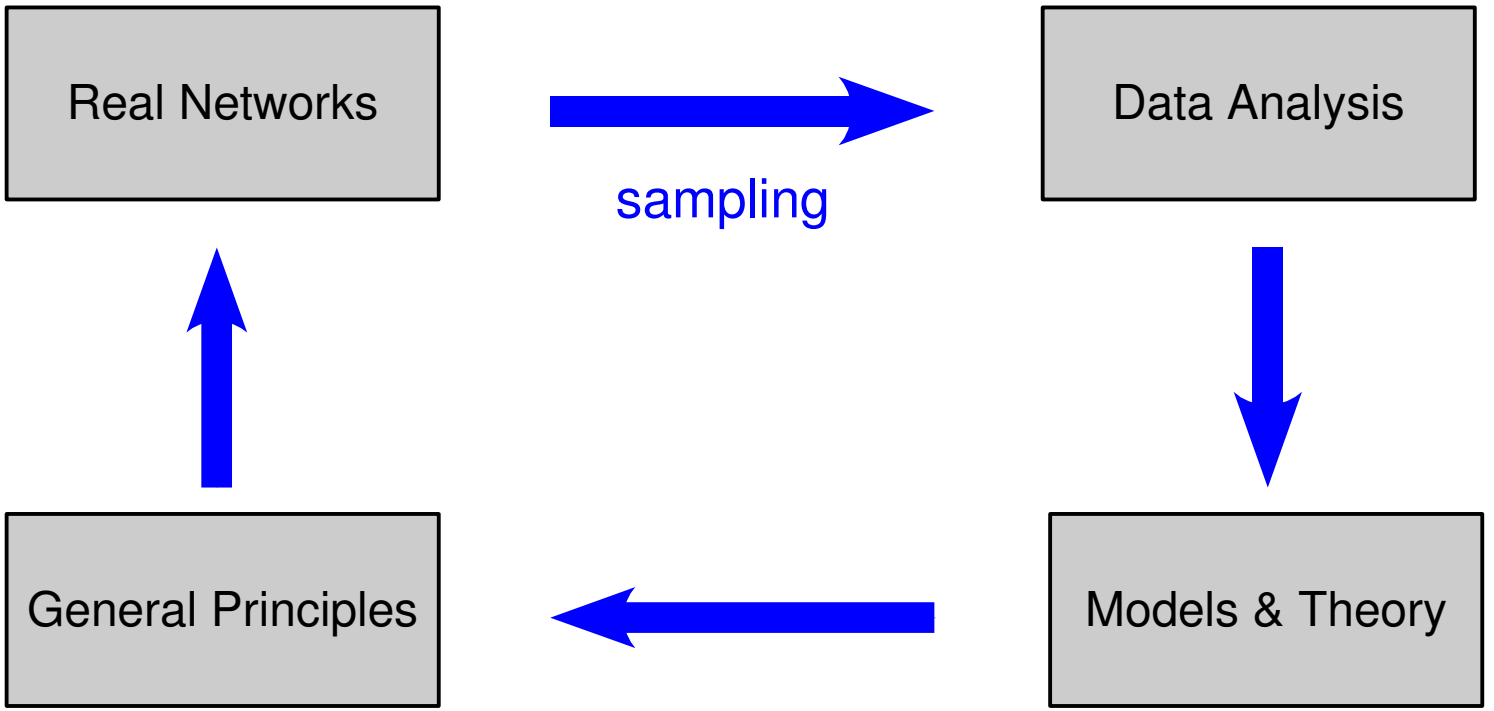
New insights on the traceroute process of networks explorations

Luca Dall'Asta, ICTP-Trieste (Italy)

preprint:
L. Dall'Asta, http://arxiv.org/abs/0706.3768 (2007)

Motivation

The science of Complex Networks is phenomenologically based :



Observed (statistical) features common to many real networks:

- degree heterogeneity
- small-world
- high transitivity
- non-trivial correlations
- modularity, hierarchical structure, ...

Influence on structural and dynamical properties of networks:
(phase transitions, percolation, robustness, spreading, synchronization, etc.)

Are network sampling techniques reliable???

1) Biological Networks:

experiments → (edge) random sampling (?)

Petermann & De Los Rios, *EPJB* **38** (2004); Stumpf, Wiuf & May, *PNAS* **102** (2005);
Stumpf & Wiuf, *PRE* **72** (2005); Lee, Kim, & Jeong, *PRE* **73** (2006);
Han, Dupuy, Bertin, Cusick, & Vidal, *Nature Biotech.* **23** (2005).

2) Social Networks:

surveys, interviews → snowball sampling

Frank, *Math. Sci. Hum.,* **104** (1988); Goodman, Ann. Math. Stat. **32** (1961);
Granovetter, *Am. J. Soc.*, **81**, (1976); Heckathorn, *Soc. Prob.*, **44** (1997);
Newman, *Soc. Net.* **25** (2003); Kossinets, *Soc. Net.* **28** (2006).

3) WWW & P2P Networks:

crawling → random walks

Stutzbach, Rejaie, Duffield, Sen, Willinger, *IMC '06* (2006);
Lee, Kim, & Jeong, *PRE* **73** (2006); Ahn, Han, Kwak, Moon, & Jeong, *WWW '07* (2007).

Sampling the Internet: known results

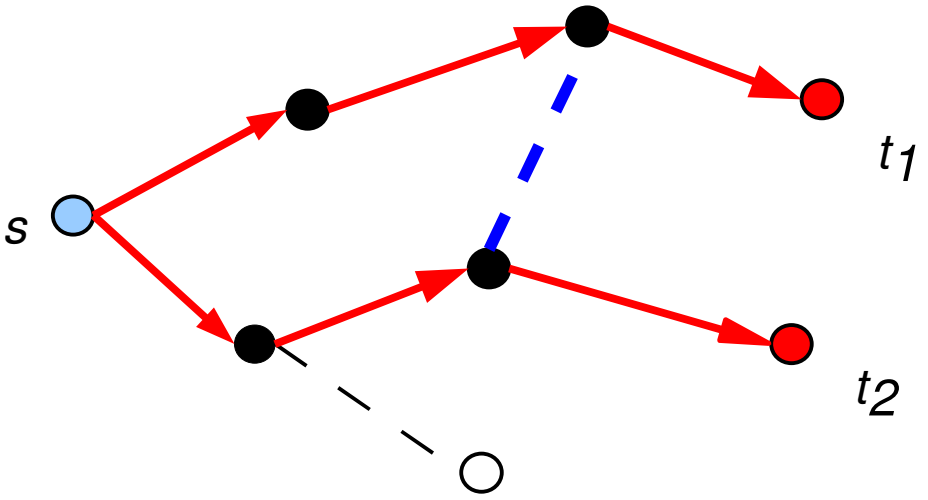
Internet maps (both *routers* and *Autonomous Systems* levels) present:

- power-law degree distributions $P(k) \sim k^{-\gamma}$ with $2.1 < \gamma < 2.5$
  (M. Faloutsos & al. 1999);

- other properties: small world, high clustering, hierarchical structure, degree correlations, ...

(see e.g. R. Pastor-Satorras & al. *PRL* **87,** 2001 and *PRE* **65** 2002;
S. Yook & al. *PNAS* **99**, 2002; S. Carmi & al. *PNAS* **104**, 2007;
J.I. Alvarez-Hamelin & al., *NIPS* **18**, 2006)

Sampling the Internet: traceroute-like methods



Traceroute probes follow the routing path from the source to the destination.

First approximation: a routing path is the shortest path between nodes.

Tree from a *source s* to a set of *targets* $\{t_i\}$ following given routing paths.

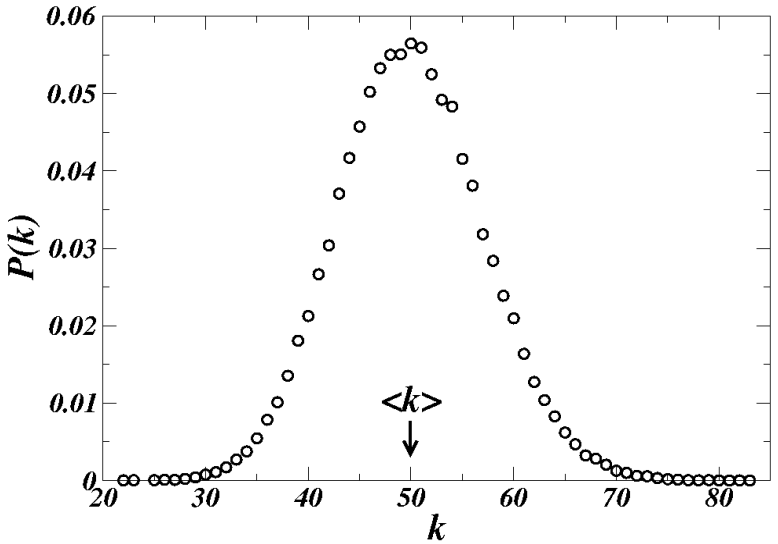
BIASES!!

- *Node sampling is incomplete*
- *Lateral connectivity is missed*
  (edges are systematically underestimated)

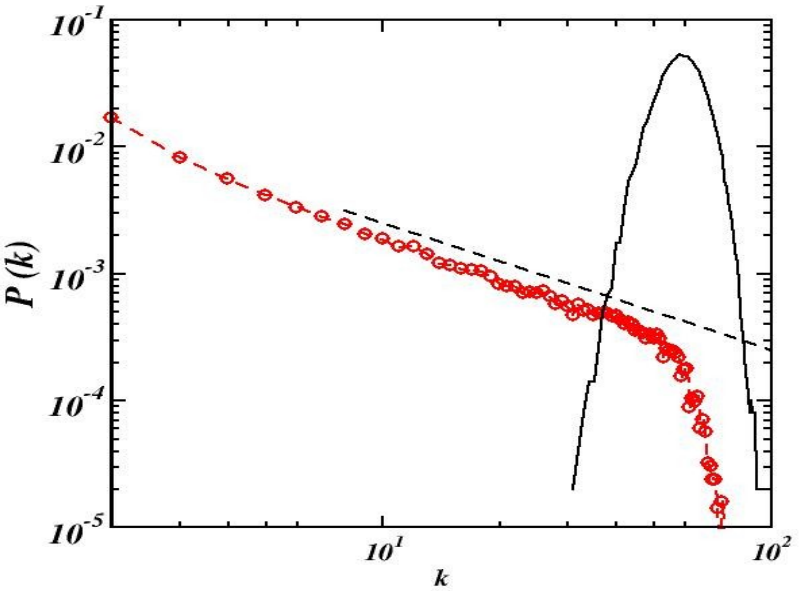
Statistical properties of sampled graphs may sharply differ from the original ones (in particular the degree distribution)

1) single-source problem (theory&experiments):

power-law degree distributions can be observed even if the underlying networks is homogeneous!!!



BAD SAMPLING



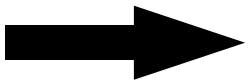
(A. Lakhina et al. INFOCOM (2003); T. Petermann & P. De Los Rios, *EPJB* **38** (2004); A. Clauset & C. Moore *PRL* **94**, (2005); D. Achlioptas & al., *STOC'05* (2005); Z. Toroczkai & K. Bassler, *Nature* **428** (2004))

2) multiple-source problem (theory&experiments):

- MF approximate expressions for node and edge discovery probabilities;

- relation with betweenness centrality (central nodes are better sampled);

- high-degree tails are sampled with higher accuracy;



qualitative behavior observed in multi-source traceroute-based sampling methods are reliable !!

(LD, J.I. Alvarez-Hamelin, A. Barrat, A. Vazquez, & A. Vespignani, *PRE* **71** (2005), LD, J.I. Alvarez-Hamelin, A. Barrat, A. Vazquez, & A. Vespignani, *LNCS* **3405** (2005), LD & al., *Theor. Com. Sci.* **355** (2006); J.L. Guillaume & al., *Com. Net.* **50** (2006);)

Partial conclusions:

1) Single-source experiments are not totally reliable;

2) Power-laws observed in mapping projects using a large number of sources are reliable (e.g. DIMES project) ;

Open Issues:

1) understanding single-source vs. multi-source results,

2) how to correct biases:
network inference to get unbiased estimators

(F. Viger & al. PRE **75** (2007), A. Flaxman & J. Vera, cond-mat/0705.3253 (2007))

Unifying single-source and multi-source views

Method single-source: **MF-like differential equations formalism**

(inspired by A.Clauset & C.Moore *PRL* **94**, (2005))

Every dynamic tree-like process can be written as a set of coupled differential equations for mean-field densities:

- of unreached nodes *u(t)*;

- of interfacial nodes *i(t)*;

- of bulk nodes *b(t)*; with $u(t)+i(t)+b(t)=1$.

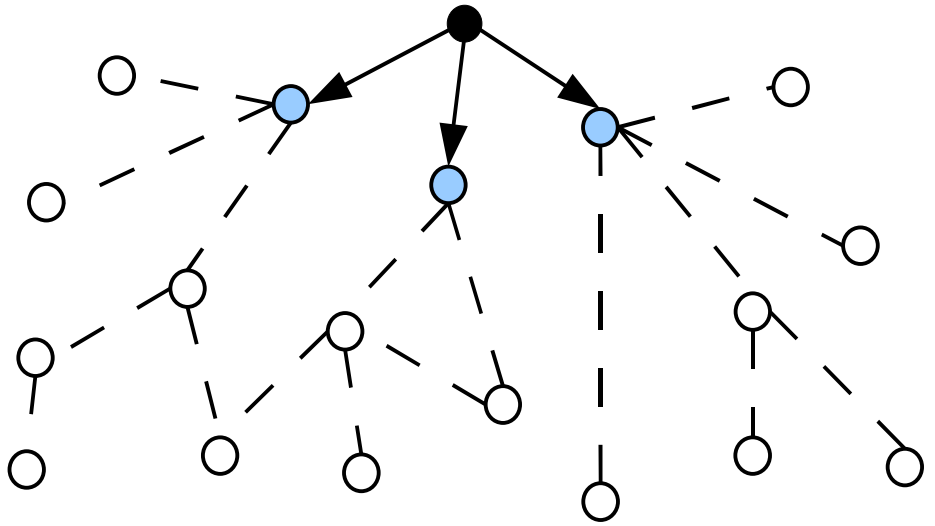
NOTE: the densities can depend on the degree as well as on hidden variables

e.g. $\{ u_k(t), i_k(t), b_k(t) \}$ with $\sum_k u_k(t) = u(t)$

Single-source traceroute sampling generates a growing (spanning) tree :

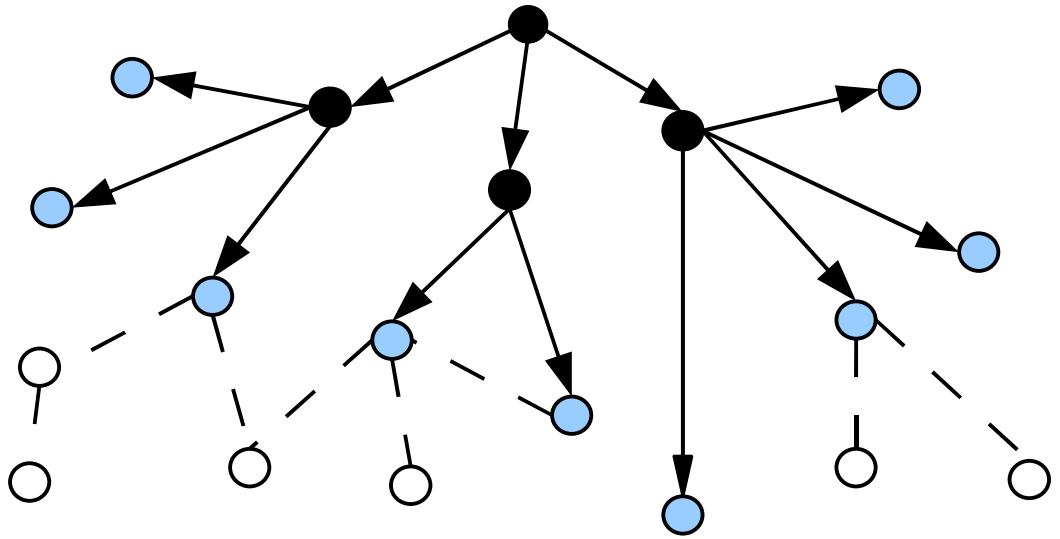
time t

1 bulk node
3 interfacial nodes
15 still unreached nodes



time t+1

4 bulk node
9 interfacial nodes
6 still unreached nodes



Dynamic Bernoulli sampling at the leaves of the growing tree:

$$\tilde{P}_1(k;t)=\sum_{l\geq k} P_t(l)Q_t(k|l)$$

with

$$P_t(l)=P(l)\frac{i_l(t)}{i(t)}$$

prob. of picking up a node of degree l at the interface of the growing cluster (main novelty!!)

$$Q_t(k|l)=\binom{l-1}{k-1}\underbrace{\left[\sum_h P(h|l)u_h(t)\right]^{k-1}}_{\approx \bar{u}(t)}\underbrace{\left[1-\sum_h P(h|l)u_h(t)\right]^{l-k}}_{\approx 1-\bar{u}(t)}$$

binomial sampling representing the conditional prob. of observing at time t a node of degree k in the subnetwork if its real degree is l.

Time-dependent single-source sampling ... sum over time steps:

$$\tilde{P}_1(k) = \sum_{l \geq k} \tilde{P}_1(k,l) \simeq \sum_{l \geq k} \frac{1}{T} \int_0^T P_t(l) Q_t(k|l)\, dt$$

$$\simeq \sum_{l \geq k} \int_0^T \frac{dt}{T} \frac{i_l(t) P(l)}{i(t)} \binom{l-1}{k-1} [\bar{u}(t)]^{k-1} [1 - \bar{u}(t)]^{l-k}$$

main difference from previous attempts

(see e.g. A.Clauset & C.Moore *PRL* **94**, (2005); D.Achlioptas & al., *STOC'05* (2005))

Degree-dependent approach:

$$\frac{d}{dt}u_k(t) = -\sum_h \frac{(h-1)}{z} P(h) \frac{i_h(t)}{i(t)} k\, u_k(t),$$

$$\frac{d}{dt}i_k(t) = +\sum_h \frac{(h-1)}{z} P(h) \frac{i_h(t)}{i(t)} k\, u_k(t) - \frac{i_k(t)}{i(t)},$$

$$\frac{d}{dt}b_k(t) = +\frac{i_k(t)}{i(t)},$$

REMARKS:

- similarity with other spreading processes (SI, SIR, rumors, etc.)

- other processes can be modeled in this way (e.g. snowball sampling)

1) Homogeneous Network ($P(k) = \delta(k-z)$ or Poisson)

Sampling relation can be solved:

$$\tilde{P}_1(k) \propto \frac{1}{k} \qquad \text{up to a cut-off at} \quad k \approx z$$

in agreement with A. Clauset & C. Moore *PRL* **94**, (2005)
D. Achlioptas, A. Clauset, D. Kempe, & C. Moore in *STOC'05*, (2005)

2) Heterogeneous Network ($P(k) \approx k^{-\gamma}$)

neglecting logarithmic corrections for large degrees

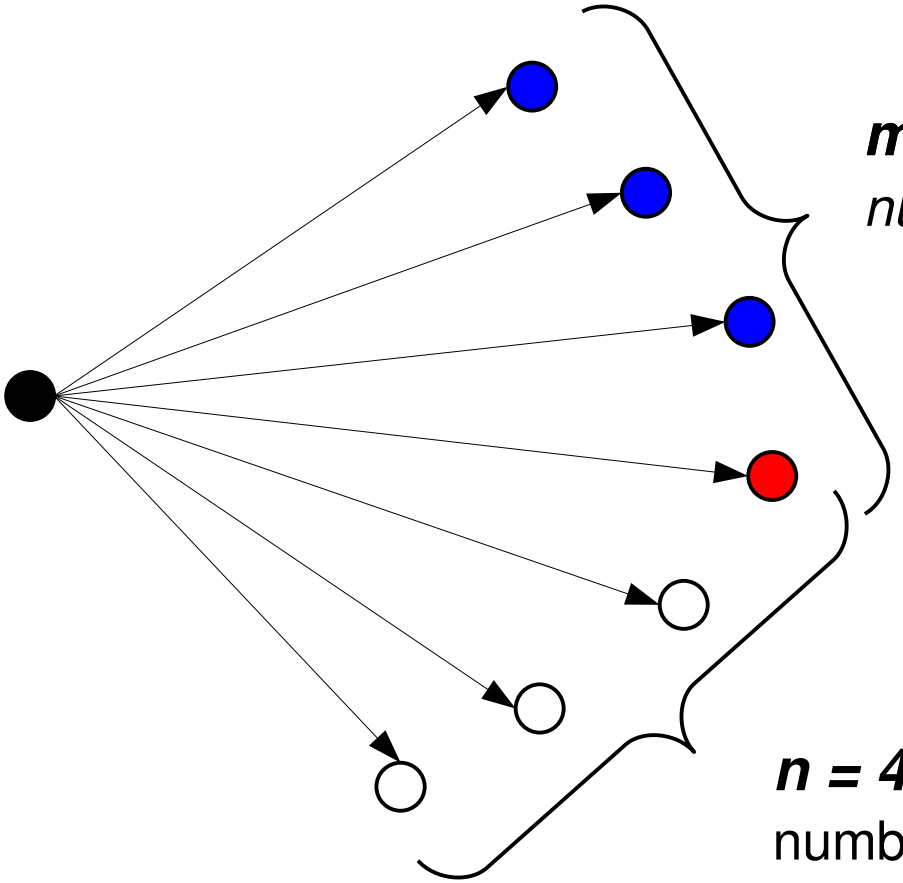
$$\tilde{P}_1(k) \;\propto\; \frac{1}{k} \sum_{l \geq k} P(l) \left[\log\left(\frac{l}{k} \right) \right]^{\xi} \;\propto\; k^{-\gamma}$$

(see also R. Cohen, M. Gonen, & A. Wool, *arXiv: cs.NI/0611157* (2006))

Merging single-source trees: **MF-like uncorrelated approach**

Recurrent relation for ν overlapping trees:

add a new tree to already existing ν-1, the local view is



m = 4,
number of neigh. in the previous ν-1 trees.

r = 1,
number of nodes in the overlapping set

n = 4,
number of nodes discovered in the last tree

Given $\tilde{R}_\nu(k,l) = \tilde{P}_\nu(k,l)/P(l)$ for a sampling from ν sources.

The recurrent relation for (uncorrelated) overlapping trees reads

$$\tilde{R}_\nu(k,l) = \sum_{m,n=1}^{l} \sum_{r=0}^{l} \delta(k-n-m+r) \times B(l,m,n,r)\tilde{R}_1(n,l)\tilde{R}_{\nu-1}(m,l)$$

$$B(l,m,n,r) = \frac{\binom{l-m}{n-r}\binom{m}{r}}{\binom{l}{n}}$$

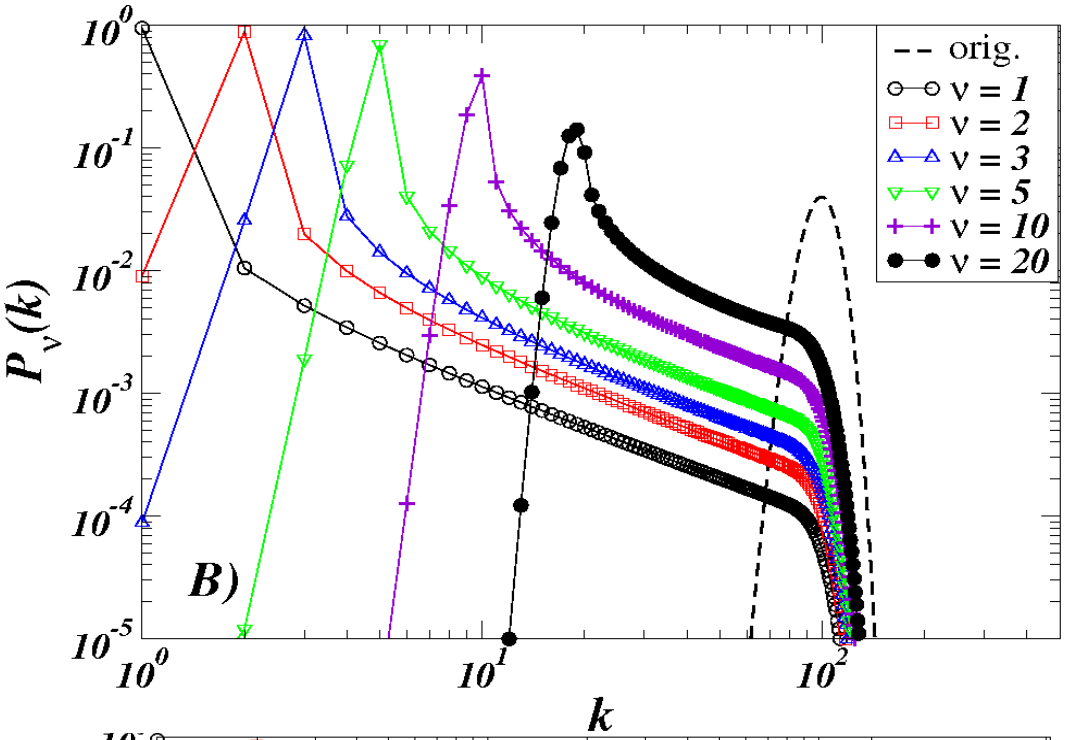
hypergeometric distribution

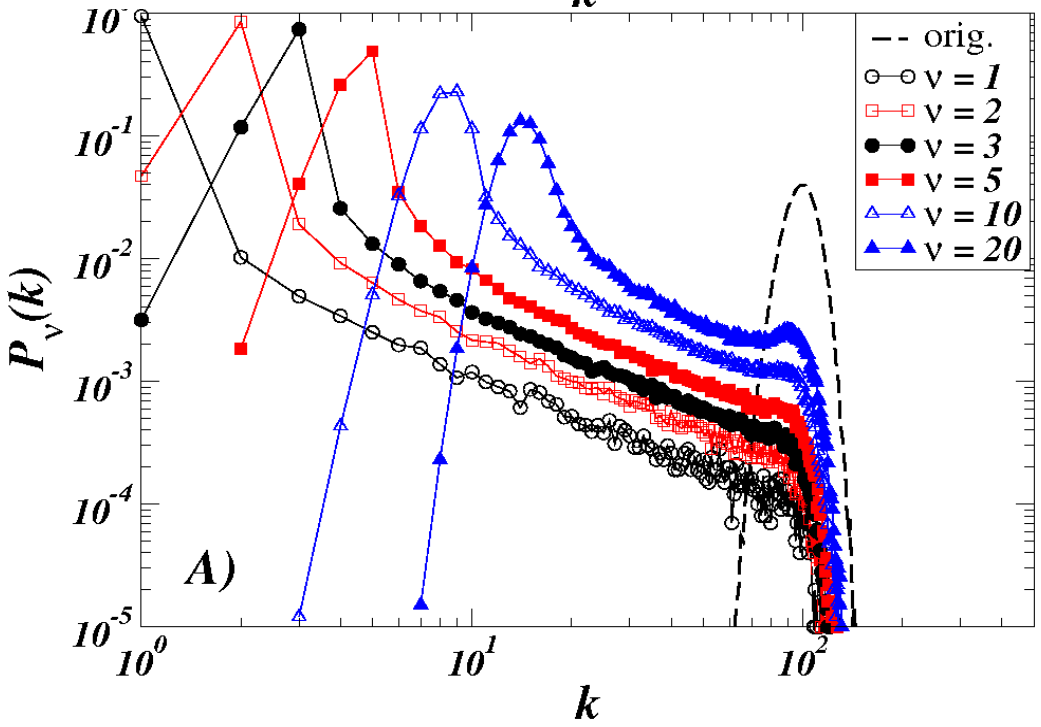
Poisson random graph

- numerical solution of recurrent relation with ν sources

- simulation of tree-like sampling from ν sources.

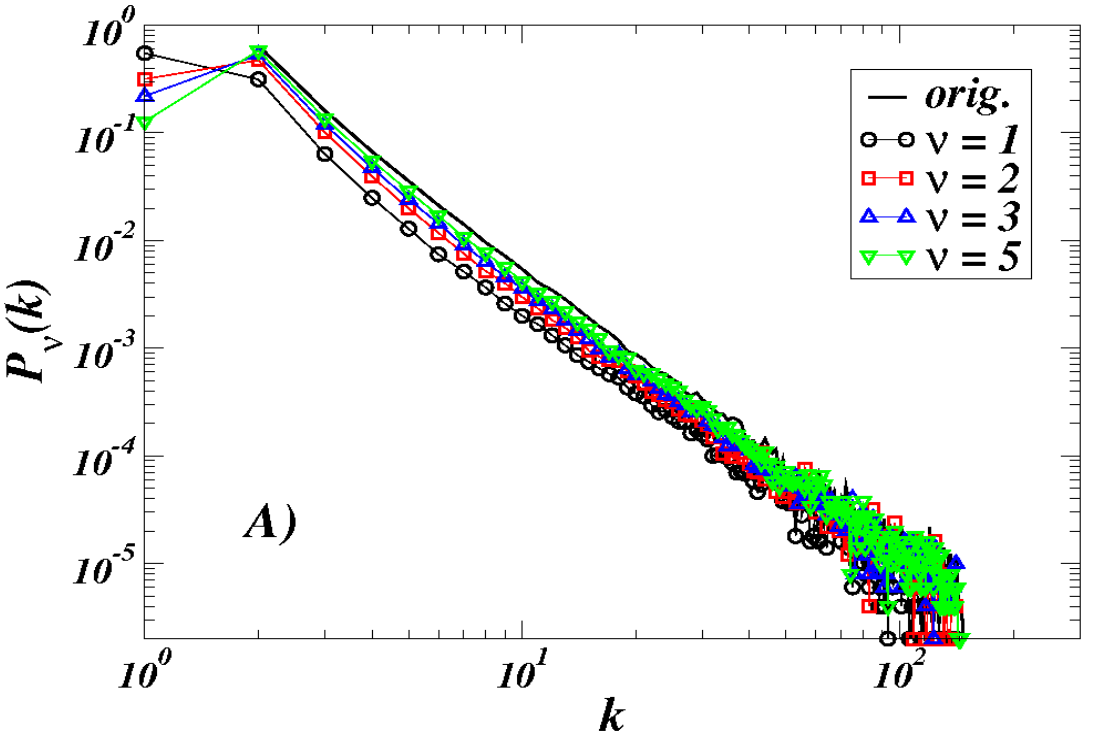
Good qualitative agreement!!



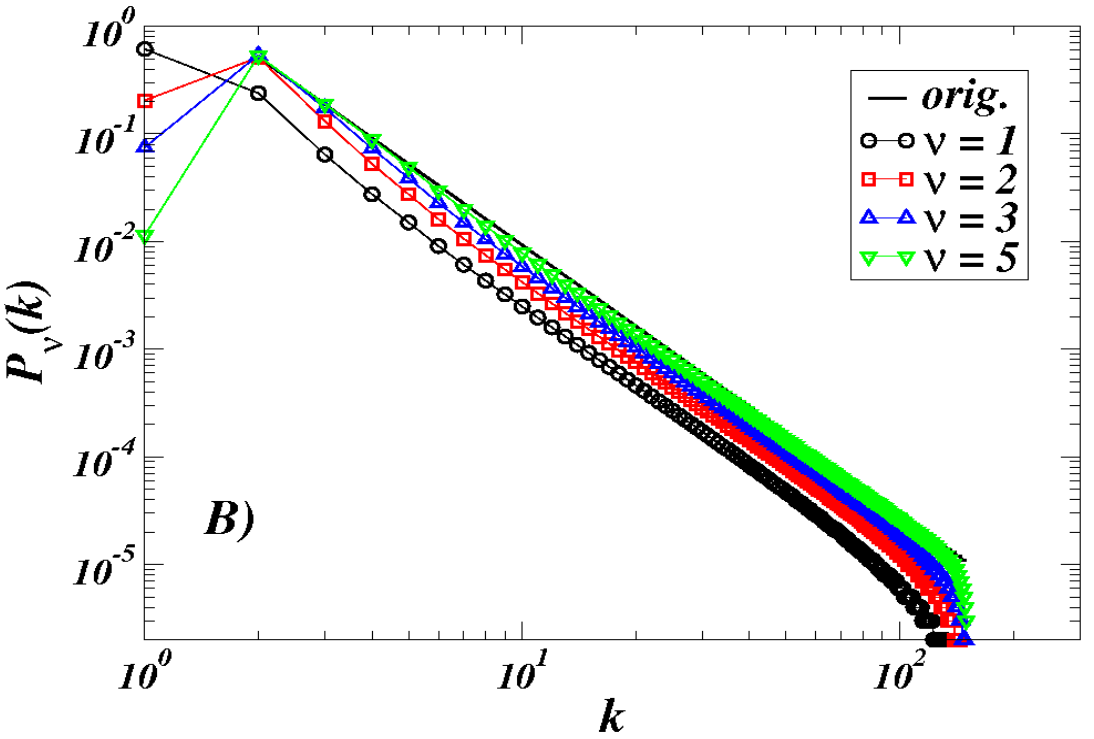


Scale-free random graph

- numerical solution of recurrent relation with ν sources



- simulation of tree-like sampling from ν sources.



Again good qualitative agreement!!

Conclusions

Poisson random graph:

- single-source gives a power-law with exponent -1 up to a cut-off *O(z)*;
- fair sampling requires *O(z)* sources;
- “metric” correlations improve the accuracy

Scale-free random graph

- high-degree nodes are preferentially sampled at the beginning of the process;
- negligible bias even for single-source process;
- very accurate sampling with just few sources;

Outlook:

- study other quantities (redundancy, triangles, effect of correlations)
- extend this formalism to other sampling processes (snowball sampling, random walks, etc.)

Power-law uncorrelated random network:

Remark 1: while in homogeneous networks the statistics of interfacial nodes is almost time-independent, i.e.

$$\frac{i_k(t)}{i(t)} \approx \frac{k}{z}$$

in heterogeneous networks, high degree nodes are sampled first

A scaling law holds $$\frac{i_k(t)}{i(t)} \approx \frac{k}{z} F\left[t \left(\frac{k}{z} \right)^{1/\beta} \right]$$

with $$F(x) \approx x^{-\beta} \quad for\ x \gg 1, \quad F(x) \approx 1 \quad for\ x \ll 1$$

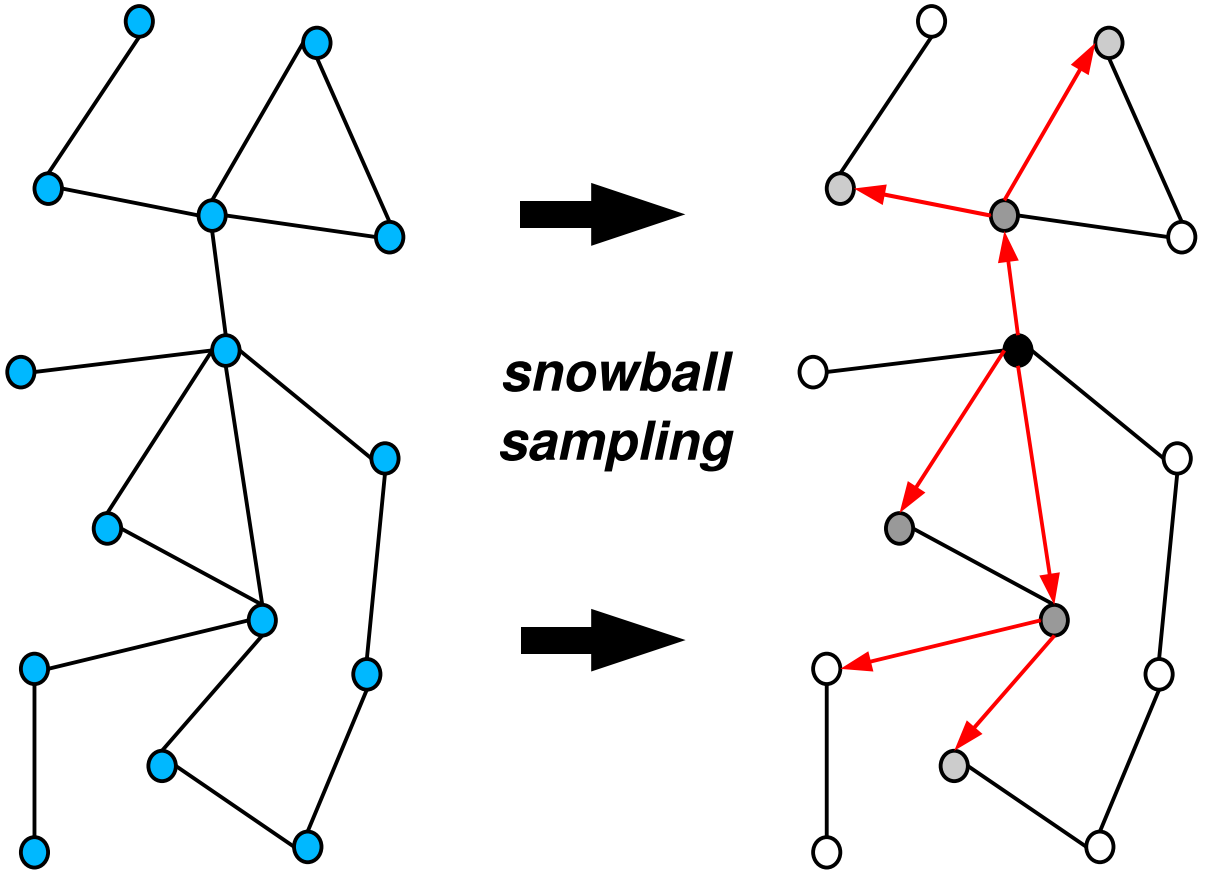
Remark 2: $$\bar{u}(t) \approx \exp\left(-a t^{\alpha}\right), \qquad \alpha < 1$$

2nd. Example – Social Networks

- sampling methods:
  surveys and interviews

- random sampling
  (correlations among
  actors? inbreeding?)

- snowball sampling
  (subnetwork generated
  from a single node)



Related works:

Frank, *Math. Sci. Hum.,* **104** (1988); Goodman, Ann. Math. Stat. **32** (1961);
Granovetter, *Am. J. Soc.*, **81**, (1976); Heckathorn, *Soc. Prob.*, **44** (1997);
Newman, *Soc. Net.* **25** (2003); Kossinets, *Soc. Net.* **28** (2006).

3rd. Example – WWW & P2P Networks

sampling methods: - breadth-first search / snowball sampling

- random walks ($r \approx log\ N$)

sample peers with prob. $p(i \to j) = \frac{1}{degree(i)}$

biased toward the hubs

- Metropolis-Hastings method

$$q(i \to j) = p(i \to j)\, min\left(\frac{\mu(j)\, p(j \to i)}{\mu(i)\, p(i \to j)}, 1 \right)$$

almost unbiased search (if $\mu(j) = \mu(i)$)

Related works:

Stutzbach, Rejaie, Duffield, Sen, Willinger, *IMC '06* (2006);
Lee, Kim, & Jeong, *PRE* **73** (2006); Ahn, Han, Kwak, Moon, & Jeong, *WWW '07* (2007).