# Theory and simulations of measurement biases

A.Vespignani

**Lecture Notes in Computer Science, 3405, 140-153 (2005).**

**Theoretical Computer Science, 355, 6-24 (2006).**

Indiana University School of informatics

the biocomplexity institute
Indiana University

# Collaborators

- Luca Dall’Asta
- Ignacio Alvarez-Hamelin
- Alain Barrat
- Alexei Vazquez

# Internet mapping

## Deployement of measurement tools

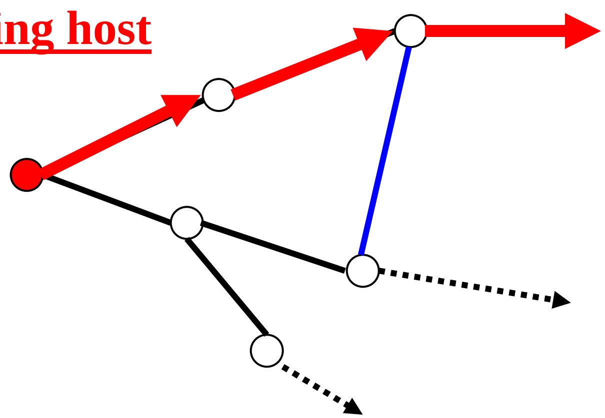
**Passive**

**Inspection of routing tables and paths stored in routers**
**Packet sampling**

## Active

**Traceroute-like based tools map the paths to selected IP address from a testing host.**

Testing host



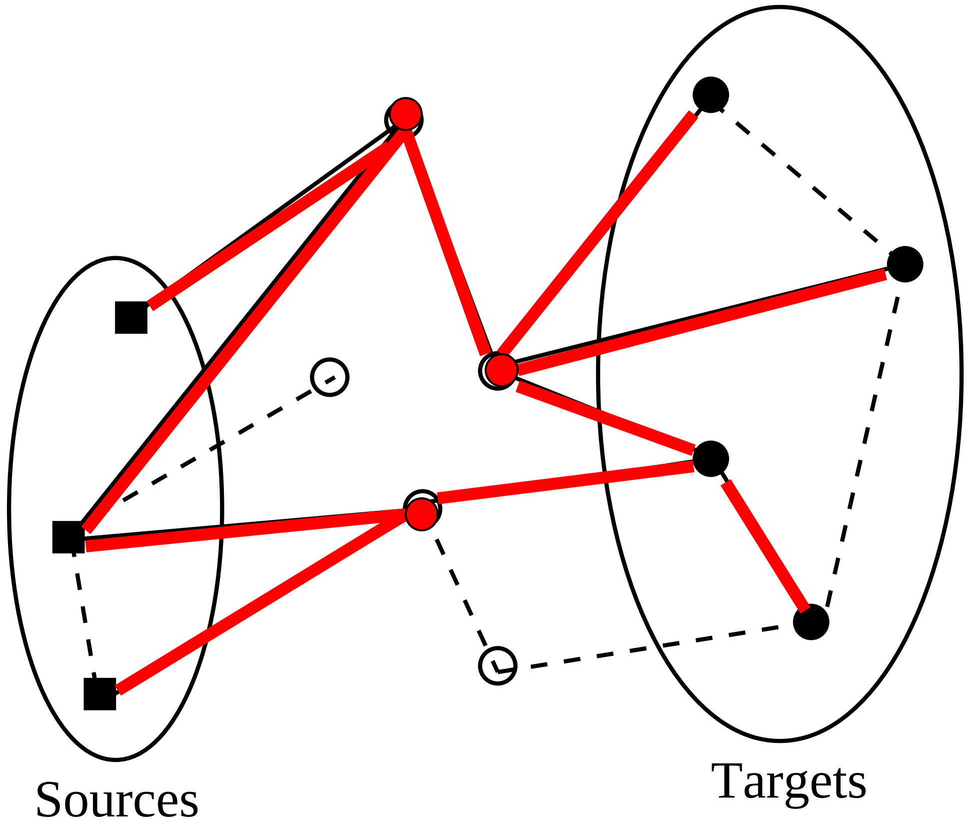
**= spanning tree**

- **One path to each node**
- **NO cross-paths**

**Burch & Cheswick (1999)**

# Measurements infrastructures

**Merging partial spanning tress from multiple sources**




**Internet tomography**

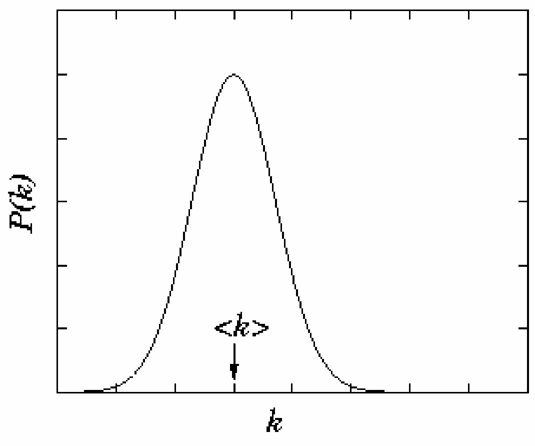
***Claffy et al (1999).***

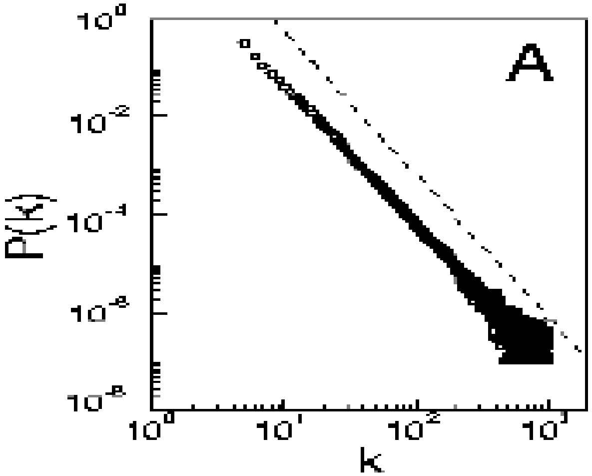
# Introduction of Biases

- Missing lateral connectivity
- Vertices and edges best sampled in the proximity of sources
- Number of sources and target is important (total traceroute probes)
- Location of sources and target in the graph
- Technical issues…..(interface resolution, security, etc.etc.)

# In case of strong biases….

…..the statistical properties of the sampled graph could be sharply different from the original one




**Crovella et al. 2002**

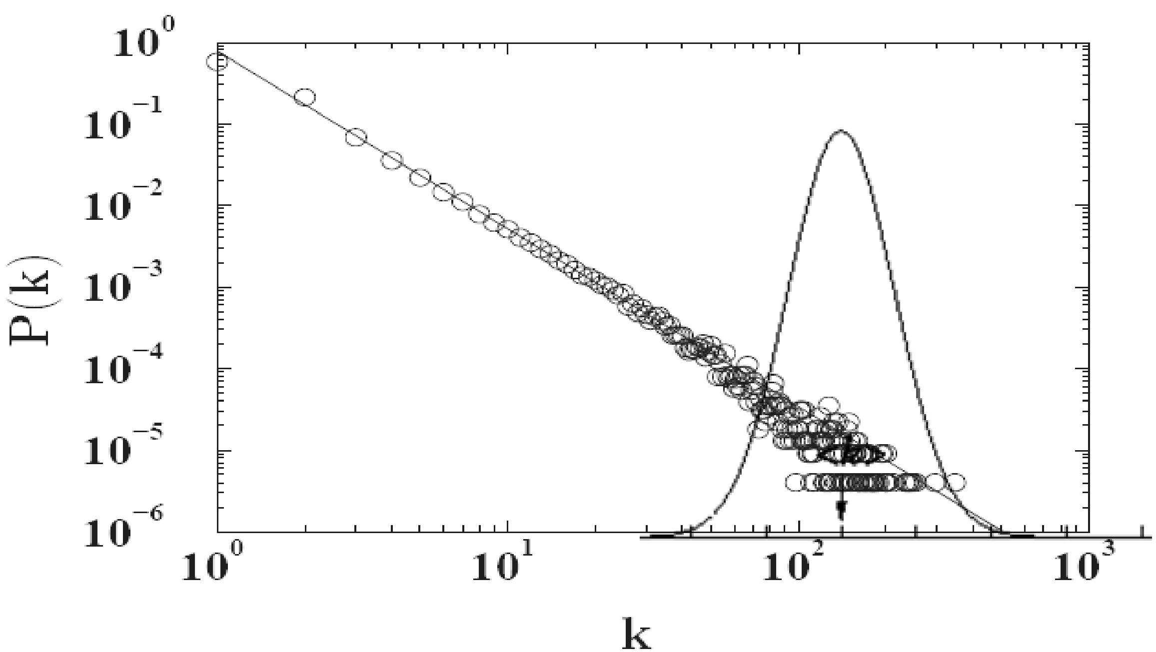
**Clauset & Moore 2004**

**De Los Rios & Petermann 2004**

# Be cautious….

- Theory for a single or very few sources.

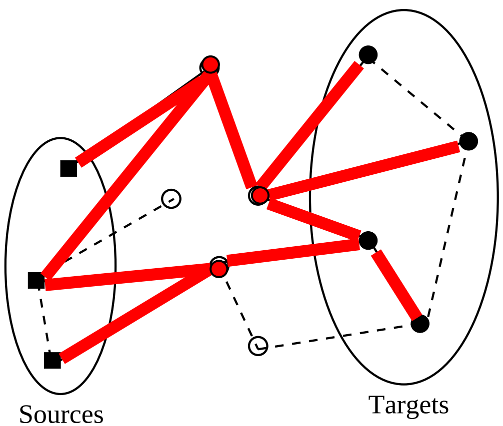
- Theoretical results are at odd with reality:




A poissonian distribution compatible with the data should have <k>= 102-103
Not realistic!

# Homogeneous approximation (mean-field) theory of Internet exploration

Let us define the quantity $\sigma_{i,j}^{(l,m)}$ that takes the value 1 if the edge $(i, j)$ belongs to the selected $\mathcal{M}$-path between vertices $l$ and $m$, and 0 otherwise. For a given set of sources and targets $\Omega = \{\mathcal{S}, \mathcal{T}\}$, the indicator function that a given edge $(i, j)$ will be discovered and belongs to the sampled graph is simply $\pi_{i,j} = 1$ if the edge $(i, j)$ belongs to at least one of the $\mathcal{M}$-paths connecting the source–target pairs, and 0 otherwise. We can obtain an exact expression for $\pi_{i,j}$ by noting that $1 - \pi_{i,j}$ is 1 if and only if $(i, j)$ does not belong to any of the paths between sources and targets, i.e. if and only if $\sigma_{i,j}^{(l,m)} = 0$ for all $(l, m) \in \Omega$. This leads to




**One configuration**

$$\pi_{i,j} = 1 - \prod_{l \neq m} \left(1 - \sum_{s=1}^{N_S} \delta_{l,i_s} \sum_{t=1}^{N_T} \delta_{m,j_t} \sigma_{i,j}^{(l,m)}\right)$$

**Let's consider an average discovery probability…………**

# Homogeneous theory of traceroute-like exploration

$$\varepsilon = \frac{N_s N_T}{N} = \rho_T N_s$$

$N_s$ = # sources

$N_T$ = # targets ($\rho_T$ -> density of targets)

$$\langle \pi_{i,j} \rangle \simeq 1 - \exp\left(-\varepsilon \widetilde{b_{ij}}\right)$$

**Edge detection probability**

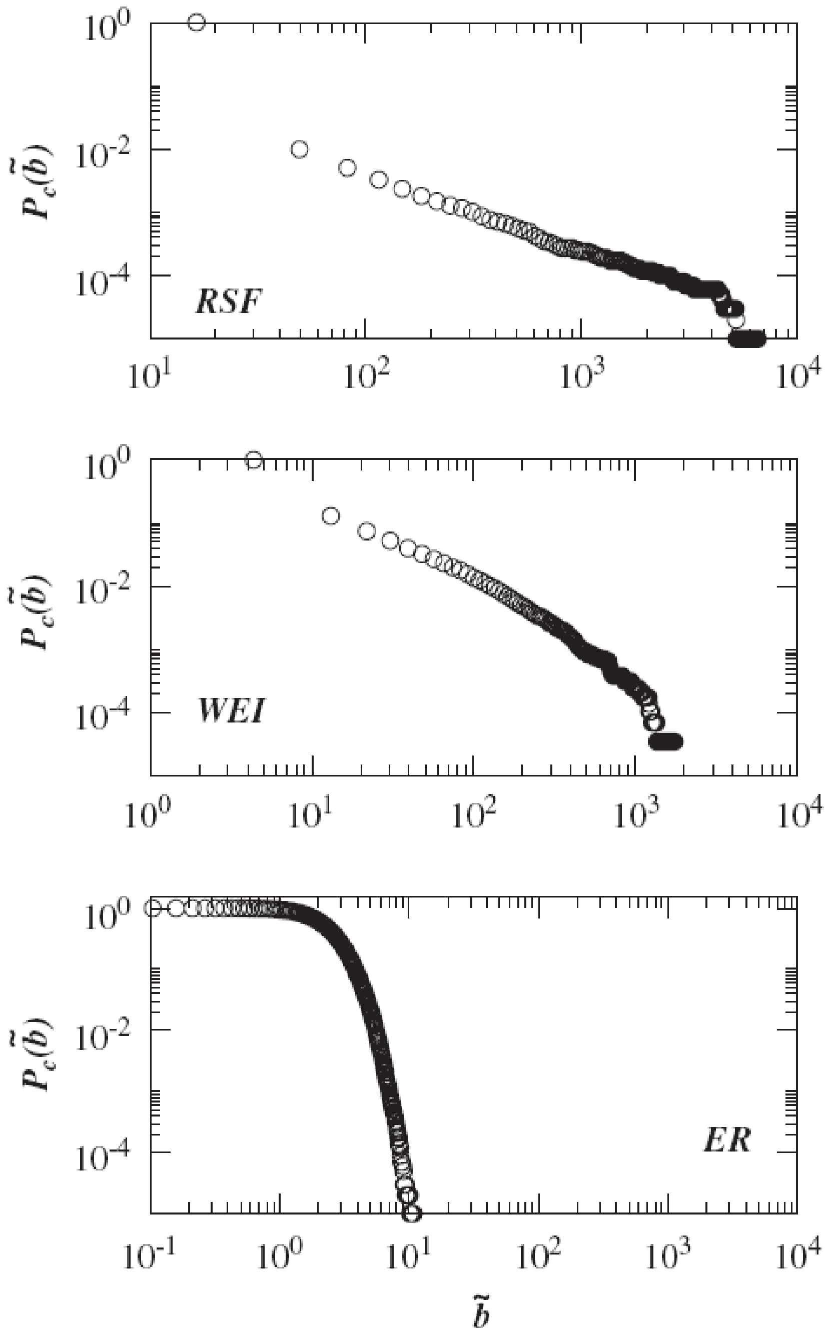
$$\langle \pi_i \rangle \simeq 1 - (1 - \rho_{\mathrm{T}}) \exp\left(-\varepsilon \widetilde{b_i}\right)$$

**Vertex detection probability**

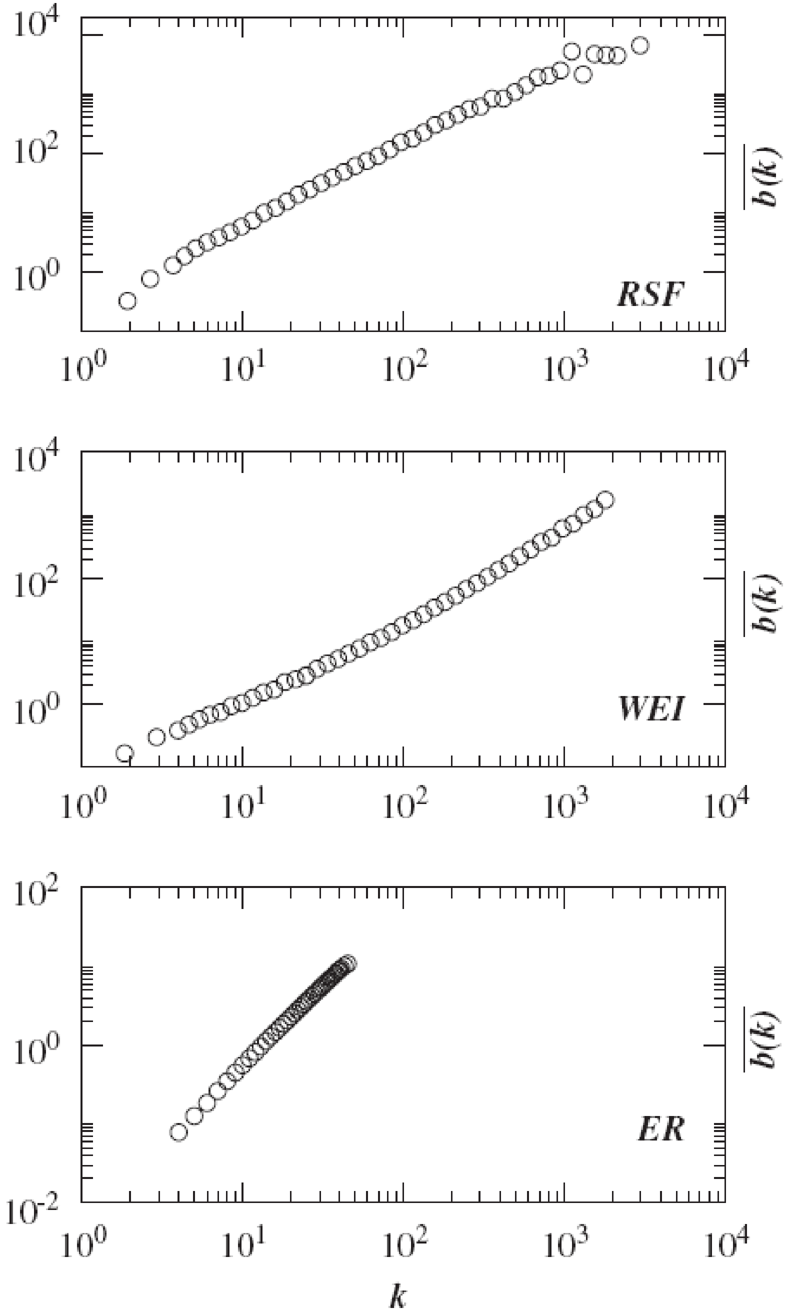
$$\langle k_i^* \rangle \simeq 2\varepsilon + 2\varepsilon \widetilde{b_i}$$

**Effective degree observed**

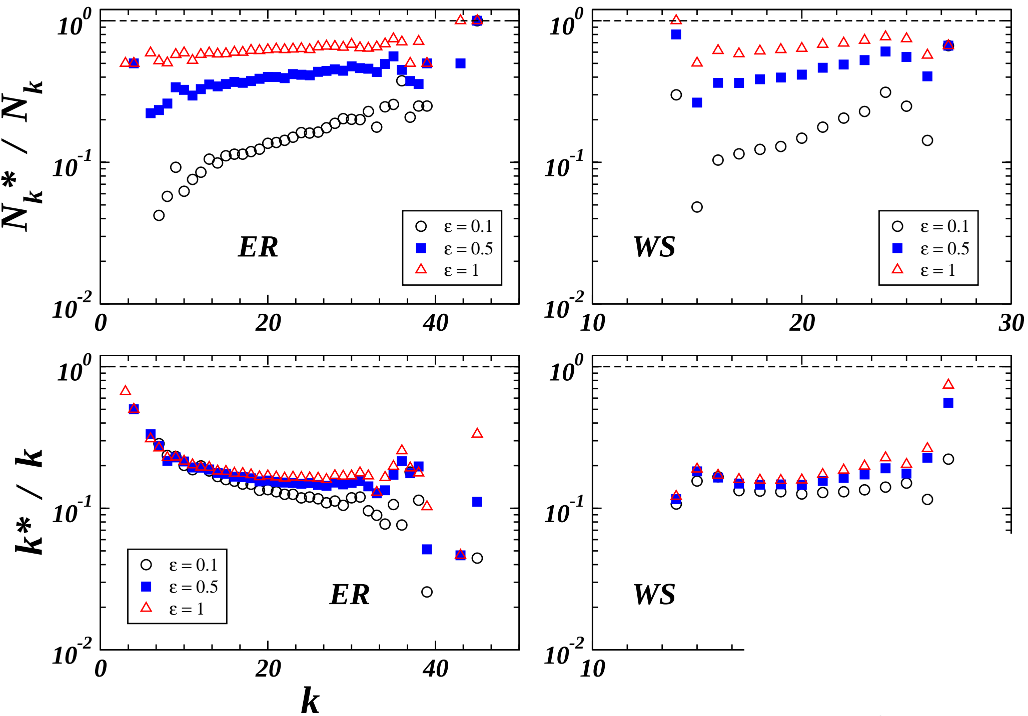
$\tilde{b}_i, \tilde{b}_{ij}$ → **Betweenness**




**Betweenness distribution of some models**

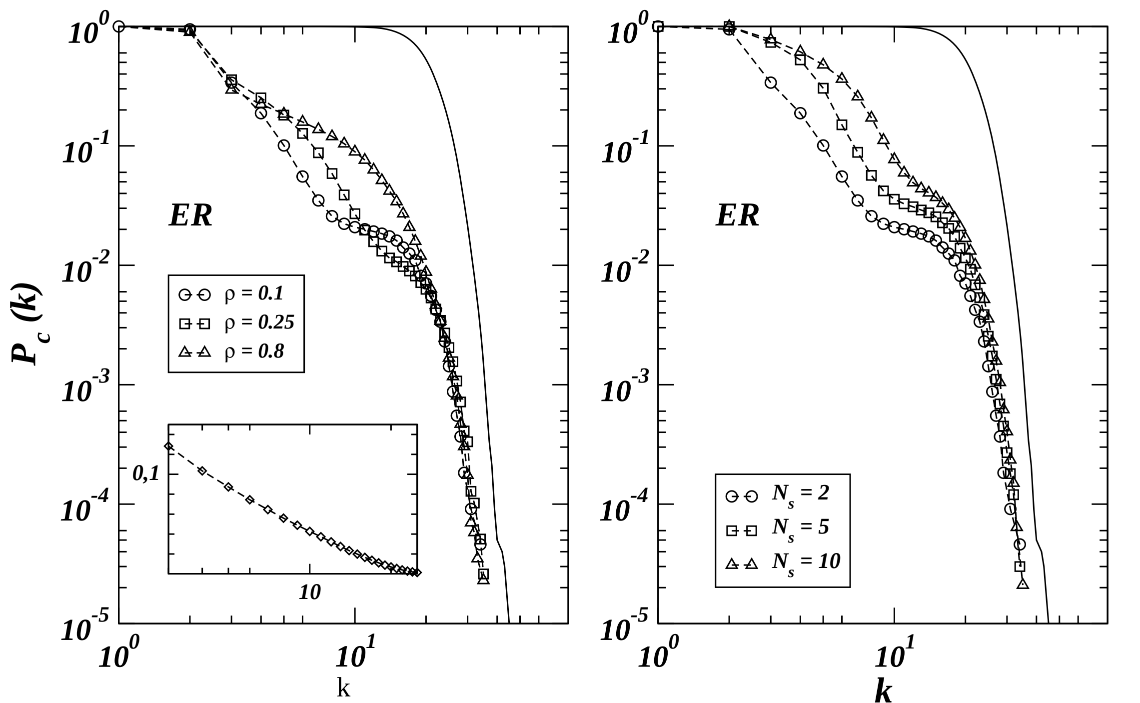


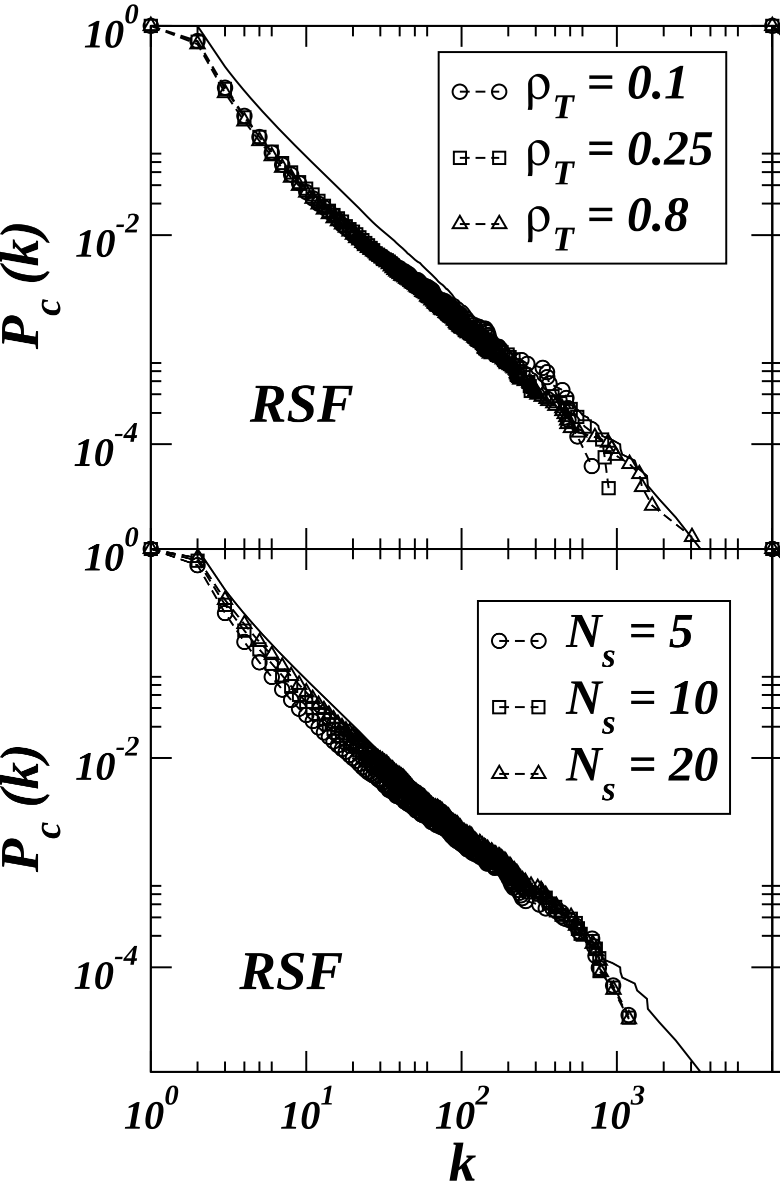

**Betweenness and degree are statistically related**

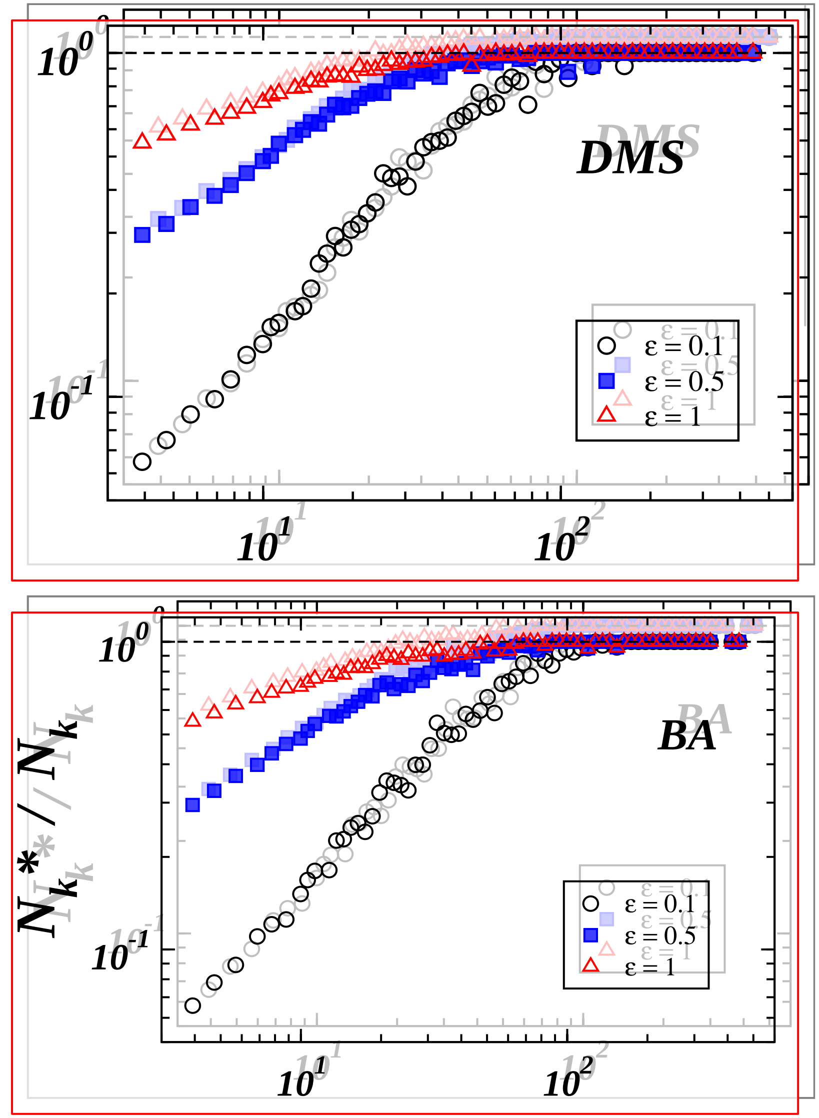



**Homogeneous graphs give rise to spurious effects**

**Average connectivity always dominate**






**Heavy-tailed graph are better discriminated**

**Tail is sampled very effectively**

# Redundancy

- # of discoveries of the same edge or vertex

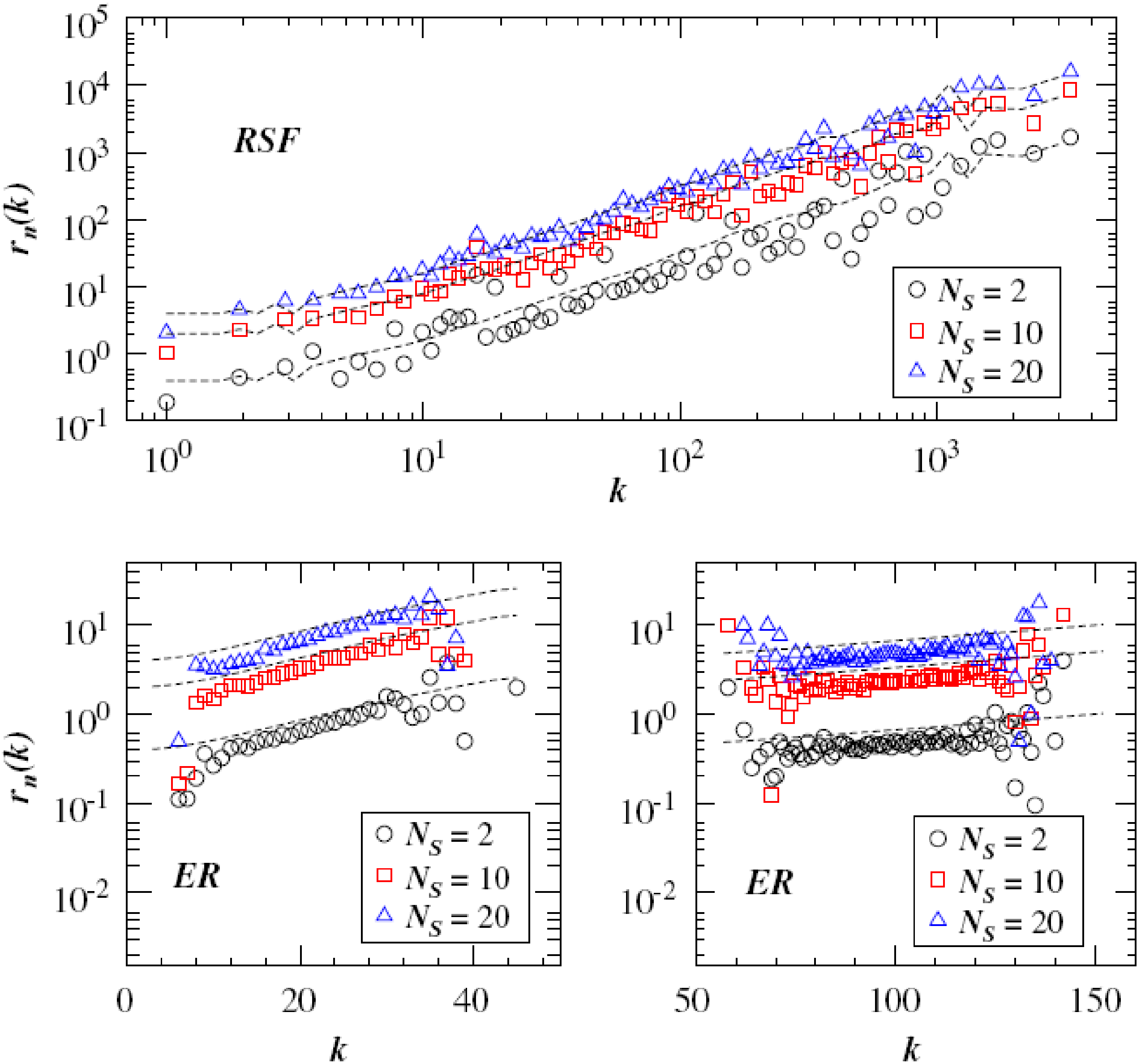
$$\langle r_e(i,j)\rangle \simeq \rho_{\mathrm{T}}\rho_{\mathrm{S}} b_{ij}$$

**Edge redundancy**

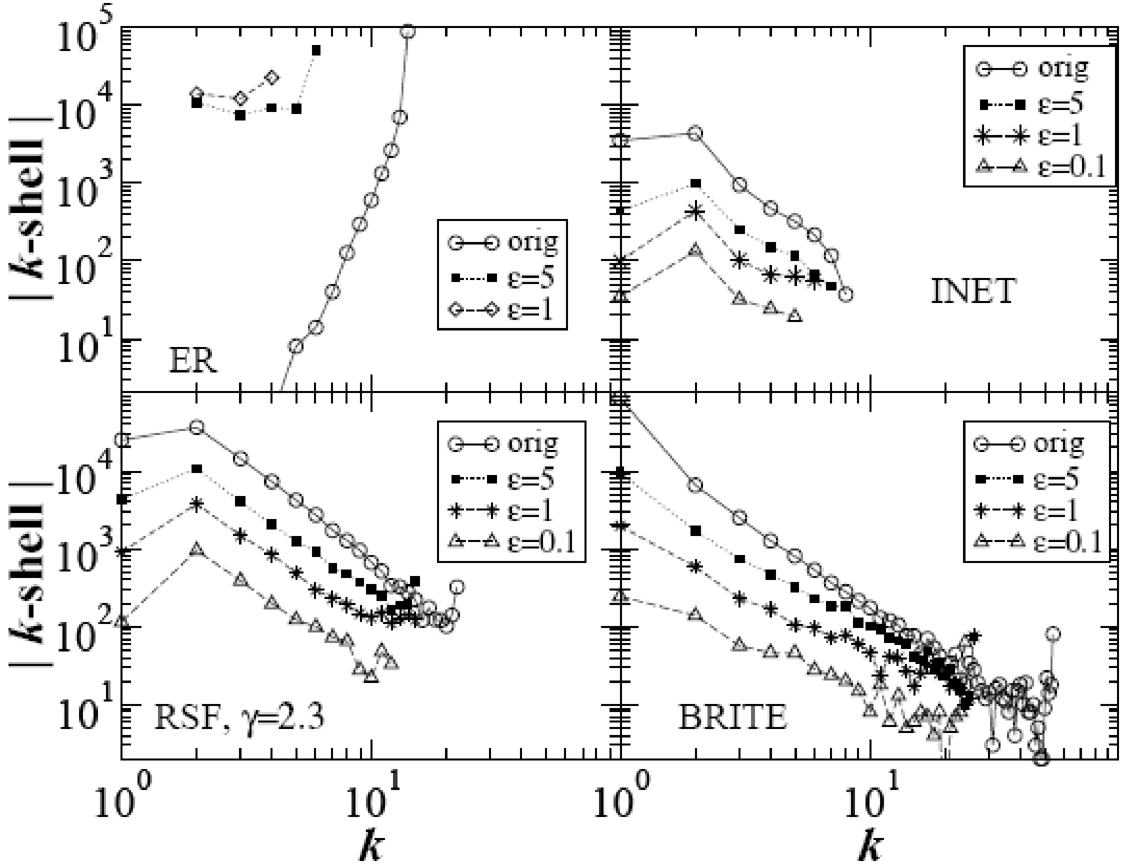
$$\langle r_n(i)\rangle \simeq 2\varepsilon + \rho_{\mathrm{S}}\rho_{\mathrm{T}} b_i$$

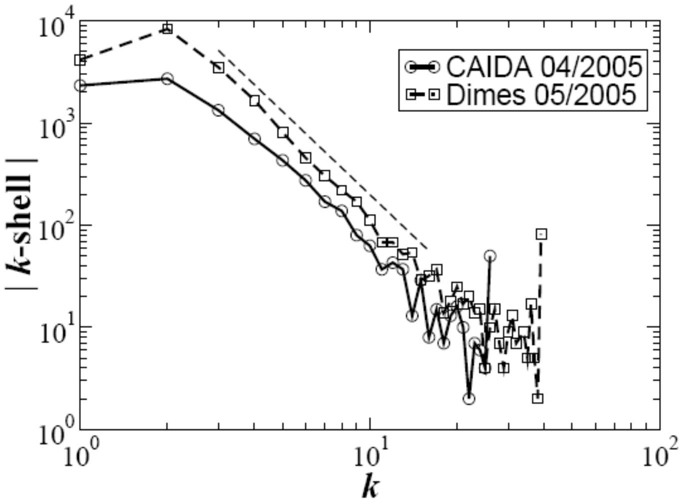
**Vertex redundancy**

# Discovery redundancy



# K-core structure….



# What do we learn….

- The more the better………
- The more the graph is heavy-tailed and the more it is clearly discriminated…
- The heavy tail is what is measured the first and the better…..

- The results concern qualitative features:
    - Heavy-tails
    - Structure of the k-cores
    - Assortative/disassortative behavior

- Quantitative features are however affected
    - Exact functional forms
    - Exponents
    - Outliers