Hierarchies created by individuals:
The structure of directory trees

Konstantin Klemm

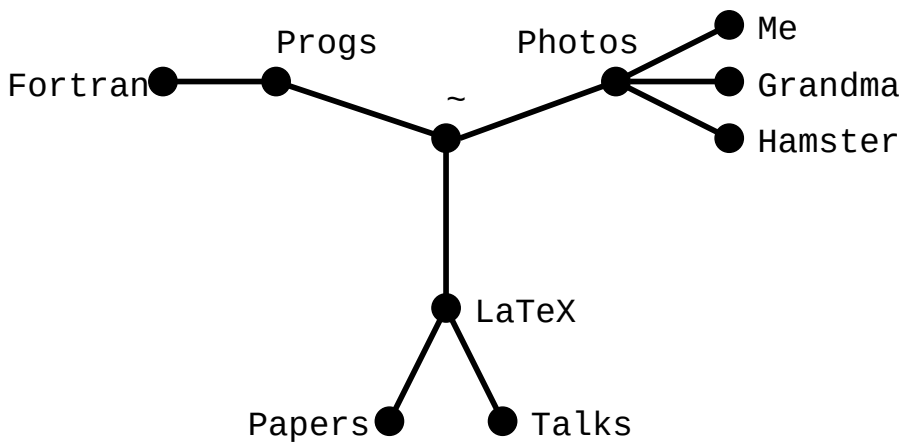
Interdisciplinary Centre for Bioinformatics
University of Leipzig, Germany

Víctor M. Eguíluz, Maxi San Miguel

Mediterranean Institute for Advanced Studies,
Palma de Mallorca, Spain

Directory trees ... what?

- tree of file folders (= directories) created by a computer user
- nodes of the tree are directories.
- a link connects a directory with its parent
- navigation by `cd` command, e.g. `cd ..` or `cd ~/Progs/`
- addition of nodes, e.g. `mkdir Fortran`



Directory trees ... why care?

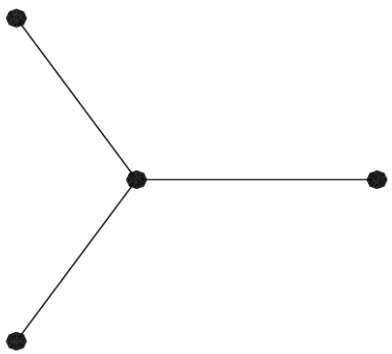
- indicate the natural (unbiased) way of organising data
- may reflect hierarchy of concepts in human minds
- possible application: improved methods for data storage / retrieval

... and especially for the statistical physicist ...

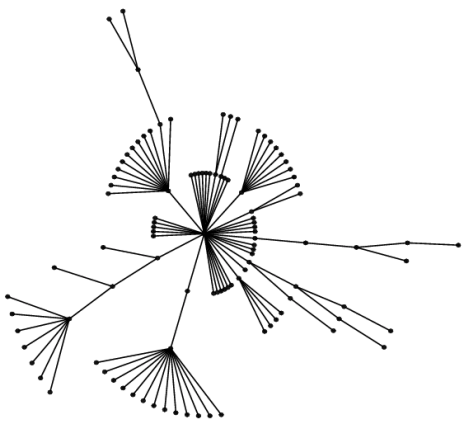
- many realizations available $\Rightarrow$ statistics
- sizes vary over orders of magnitude $\Rightarrow$ system size scaling

Data material

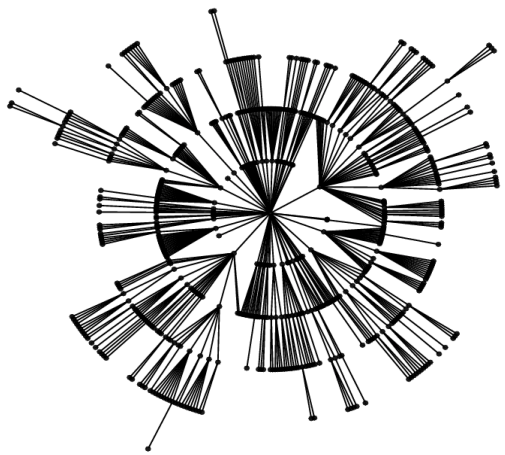
- 63 trees of sizes in the range $N = 4 \ldots 2000$
- created by faculty, postdocs, and PhD students using the UNIX / LINUX computer system at the Department of Interdisciplinary Physics of the University of the Balearic Islands.



N=4

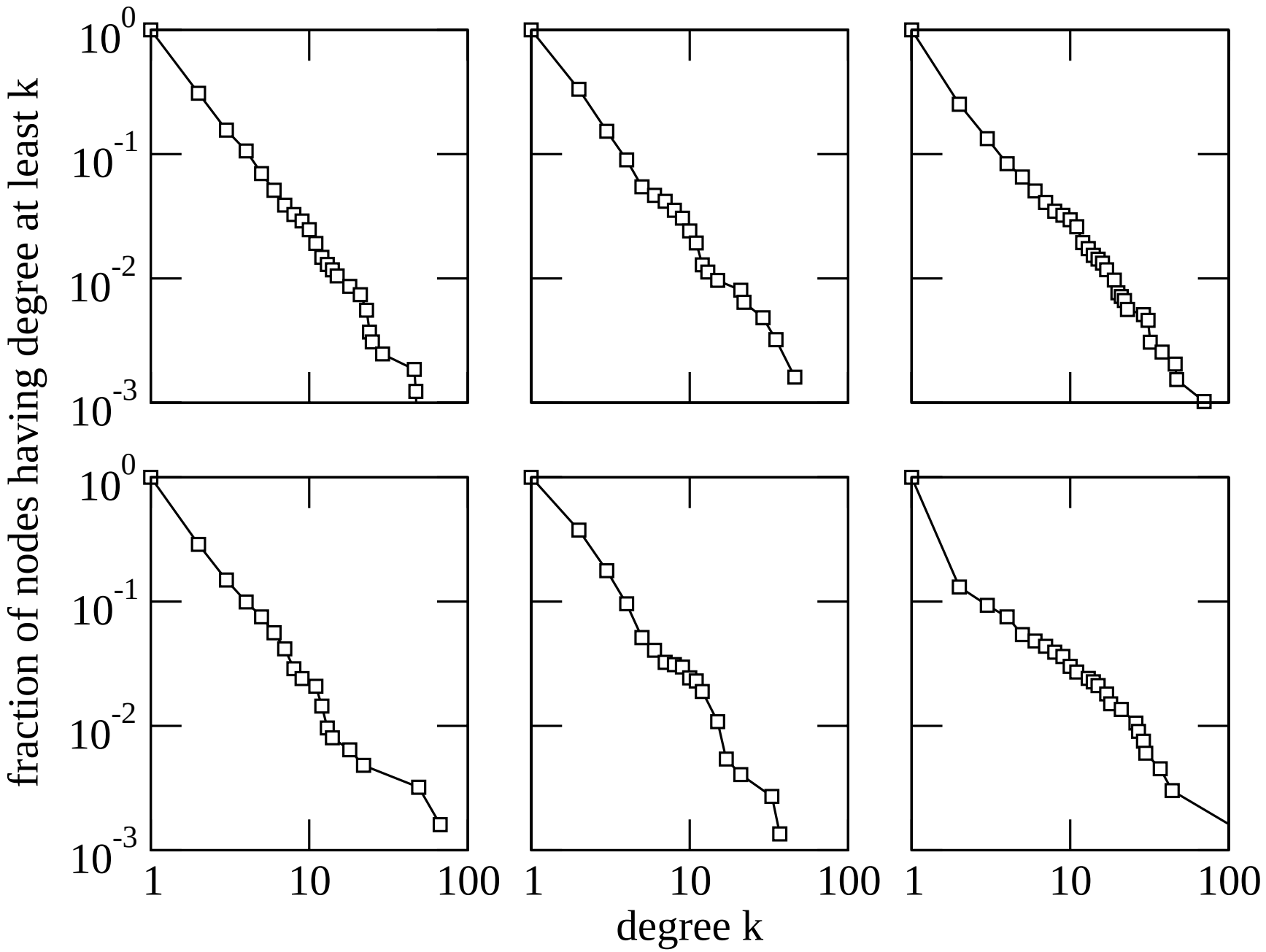


N=107

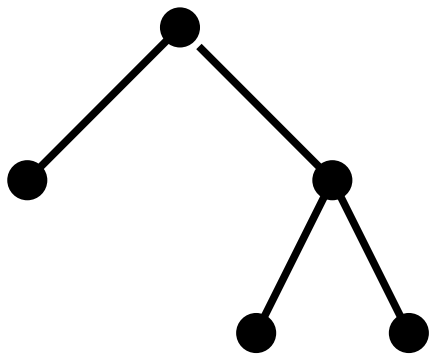


N=645

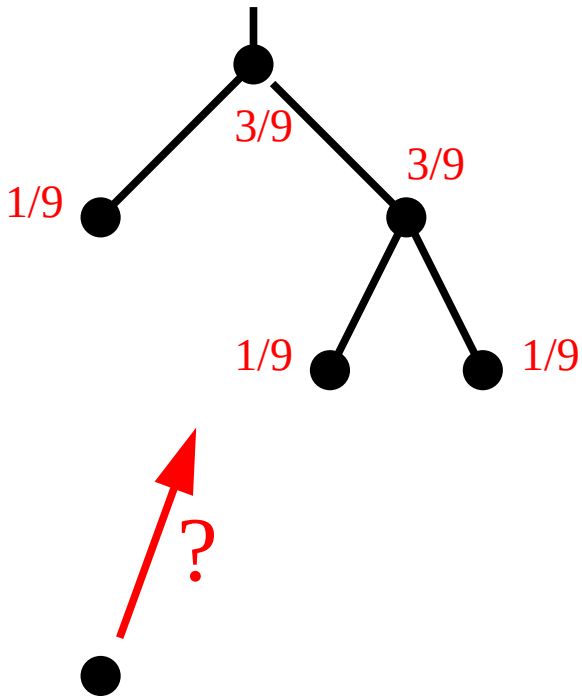
Degree distributions



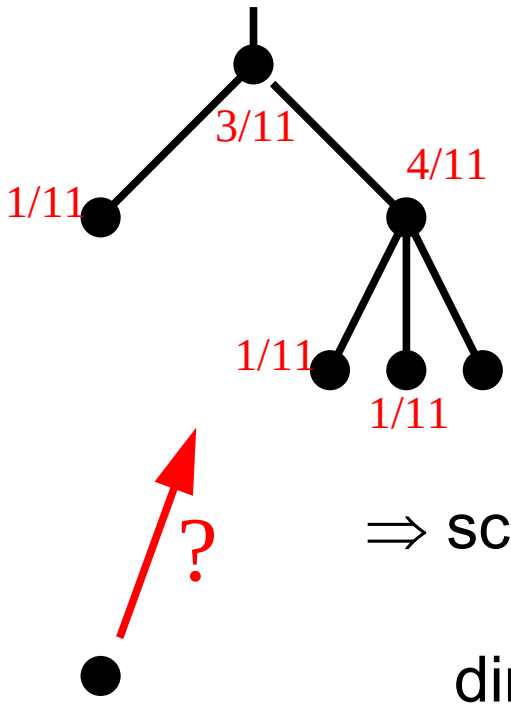
Model: preferential attachment



Model: preferential attachment



Model: preferential attachment

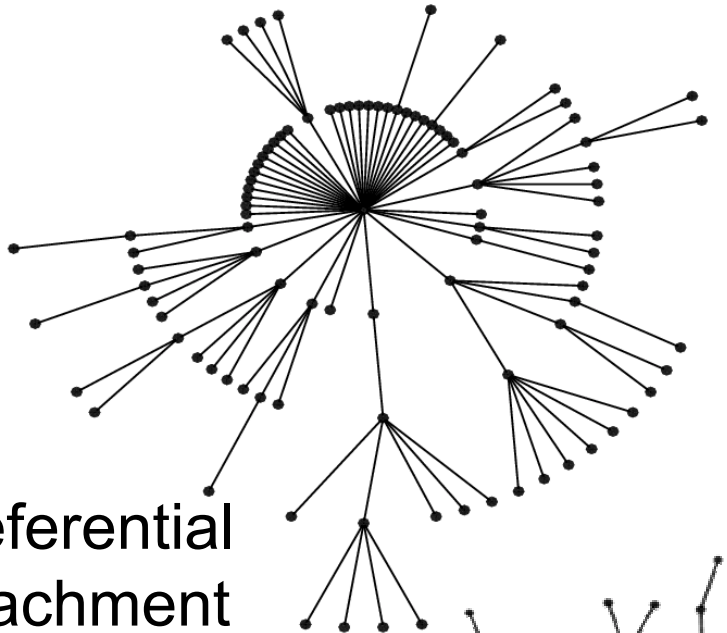


$\Rightarrow$ scale-free trees: $P(k) \sim k^{-\gamma}$

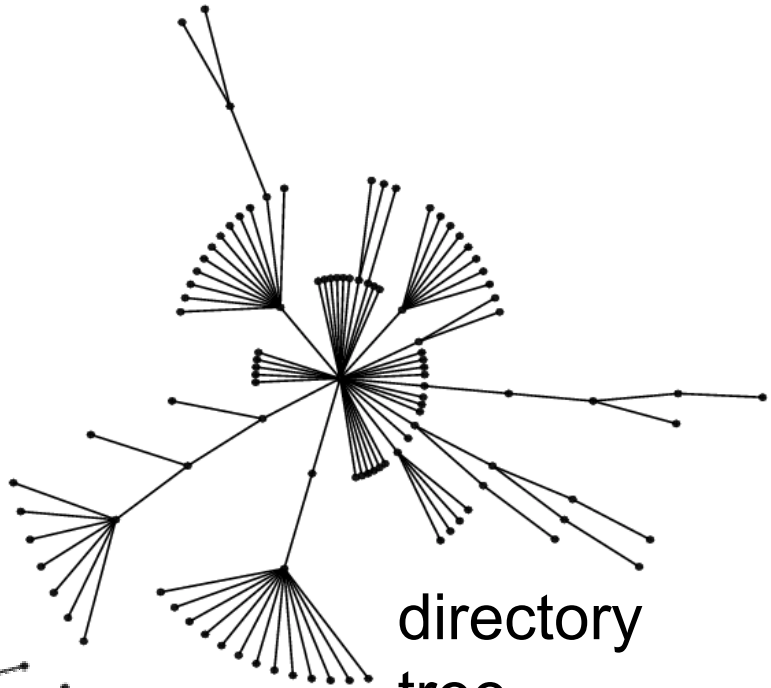
with $\gamma = 3$

directory trees $\gamma = 2.2 \ldots 2.8$

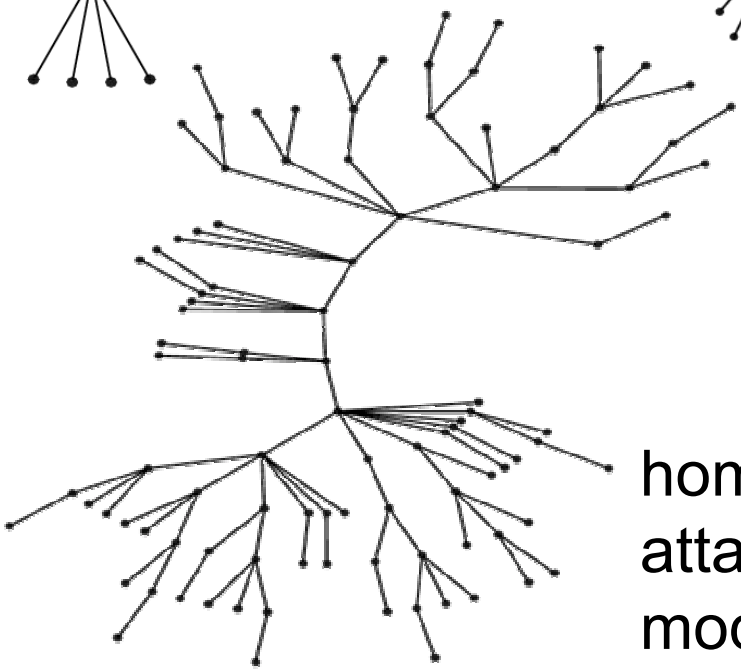
Comparing model and data



preferential attachment model

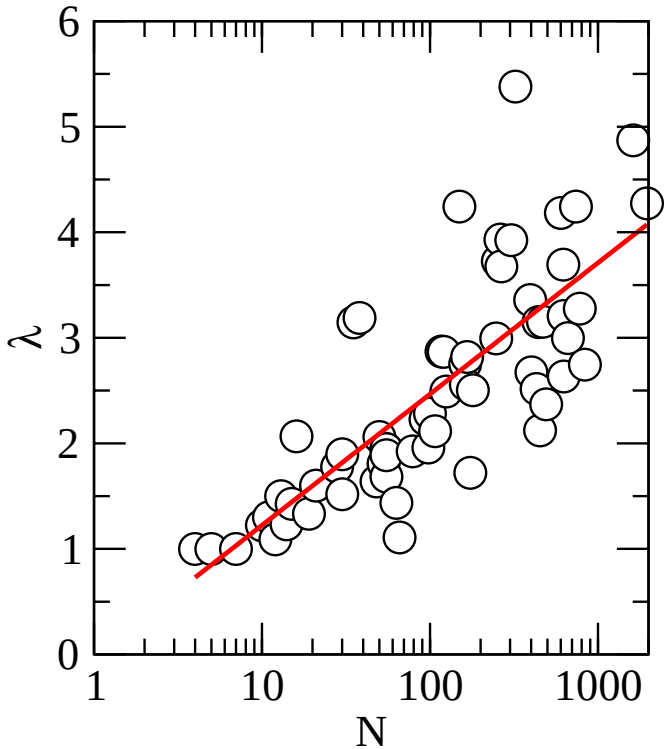


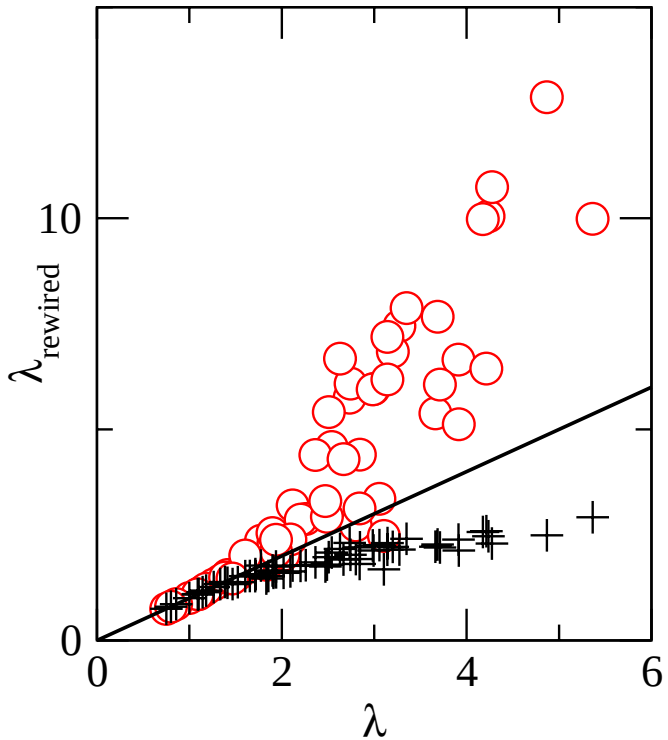
directory tree



homogeneous attachment model

Path length from root

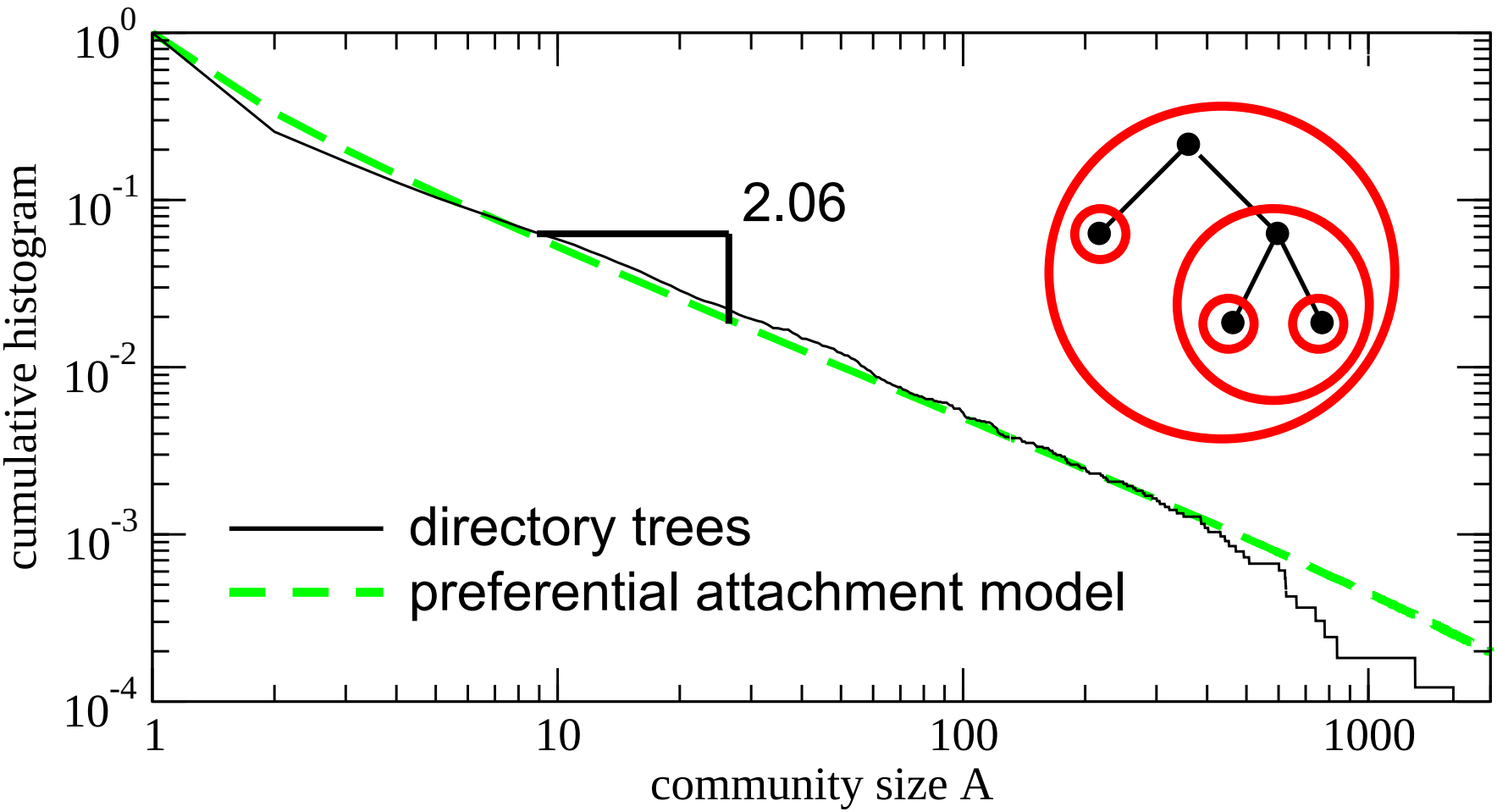




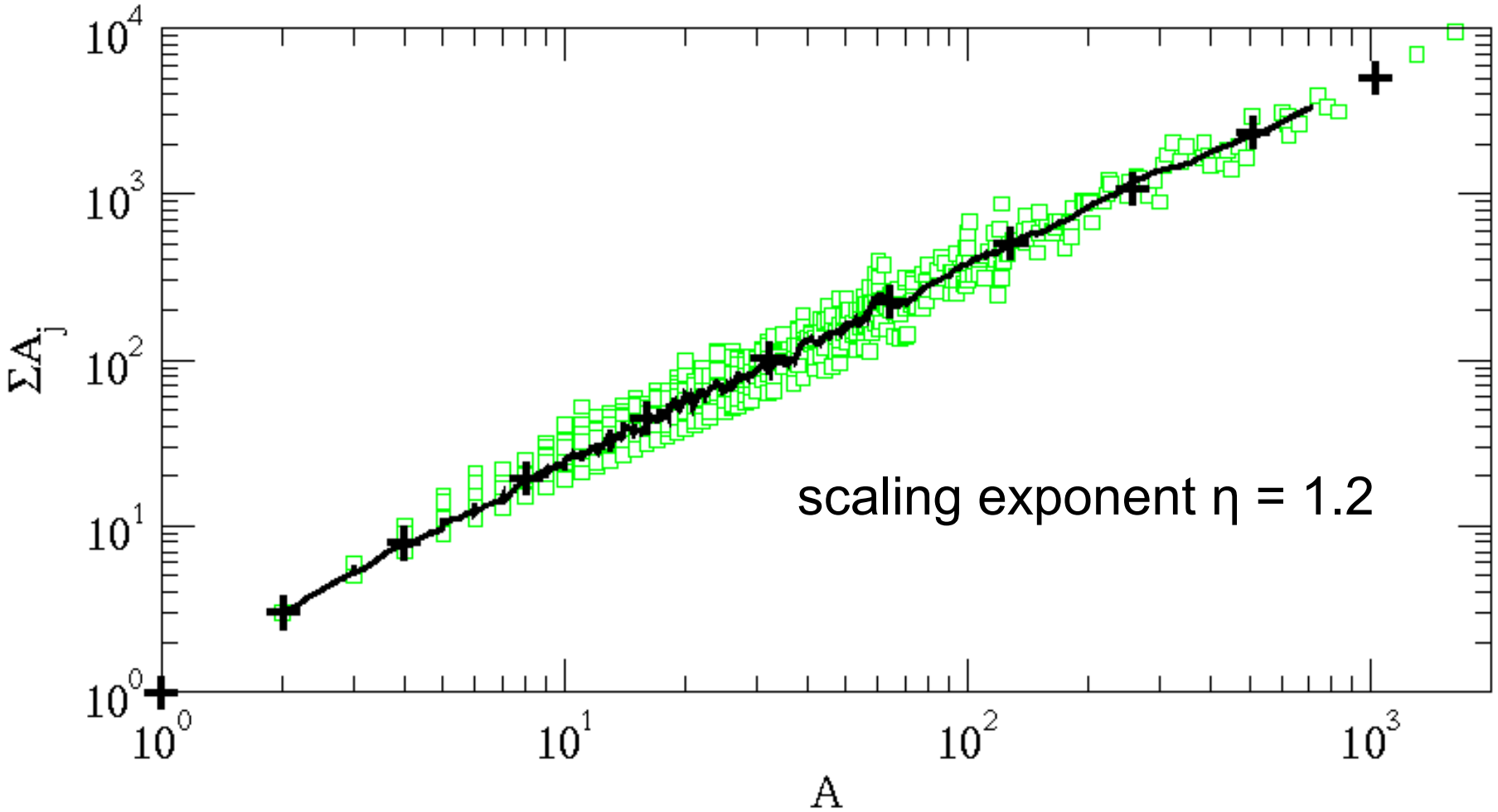
fit: $\lambda = 0.54 \ln N$

model: $\lambda^{(pref)} = 0.5 \ln N$

Community structure



Allometric scaling



Conclusions

- directory trees have interesting non-trivial structure
- scale-free degree distribution
- logarithmic increase of path length with system size
- scale-free community size distribution, exponent $\tau \approx 2$.
- allometric scaling, exponent $\eta \approx 1.2$
- all found structural properties are well explained by the preferential attachment model.

preprint http://www.arXiv.org/abs/cond-mat/0403239