Data storage by individuals: The structure of directory trees

Konstantin Klemm

Interdisciplinary Centre for Bioinformatics
University of Leipzig, Germany

in collaboration with
Víctor M. Eguíluz and Maxi San Miguel

University of the Balearic Islands, Spain

Outline

1. Introduction / Motivation

2. Growth model for directory trees

3. Comparing model and data: degree distribution, distances, communities

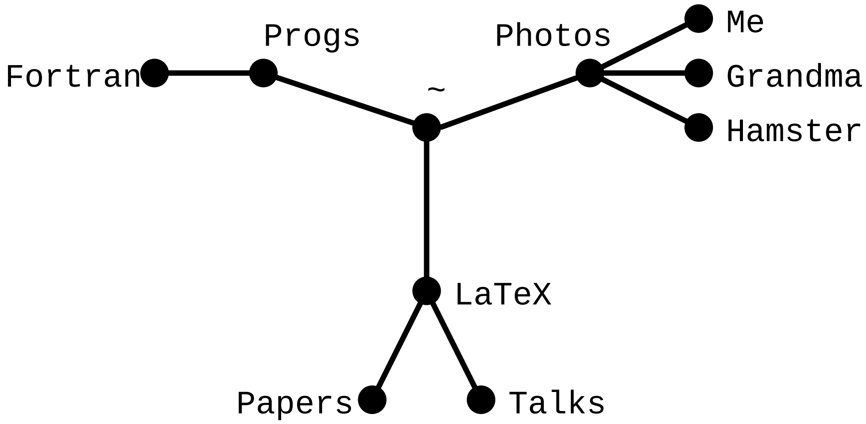
4. Conclusion

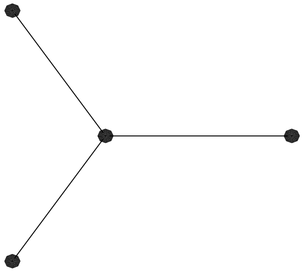
5. Appendix: Relevance for RNA secondary structure

Directory trees: What?

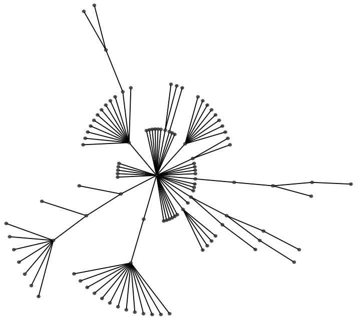
Construction:

```
> mkdir Progs
> cd Progs
> mkdir Fortran
>
```

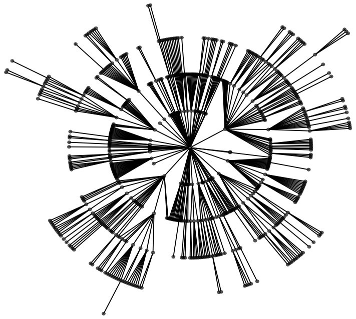




$N = 4$



$N = 107$



$N = 645$

Directory trees: Why care?

- hierarchical structures “self-organized by individuals”

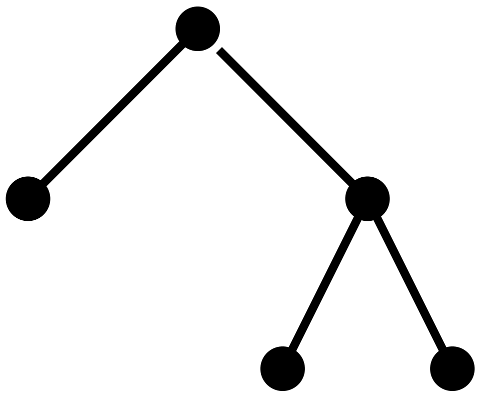
- may reflect hierarchy of concepts in human minds

- possible application in optimization information storage / retrieval

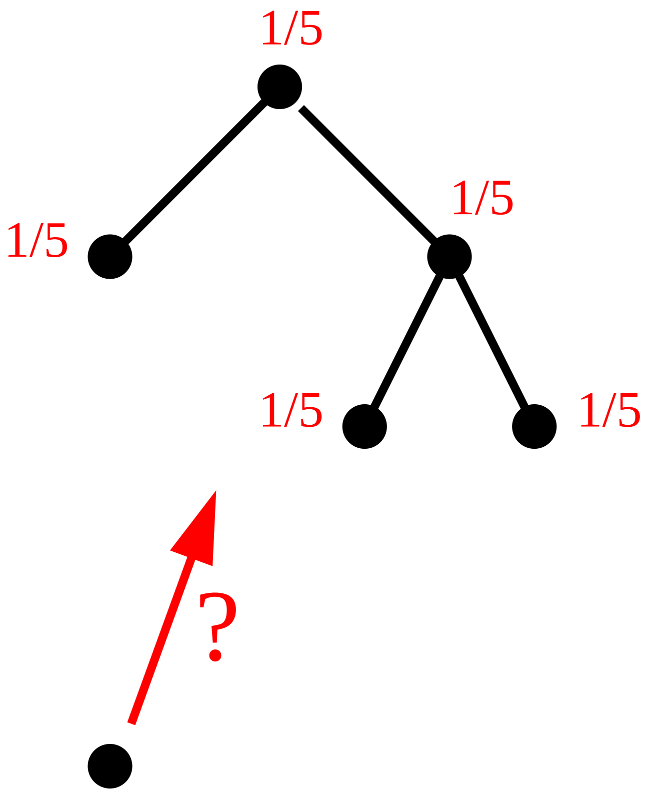
- many realizations available
  $\Rightarrow$ statistics

- system sizes vary over several orders of magnitude
  $\Rightarrow$ study system size scaling

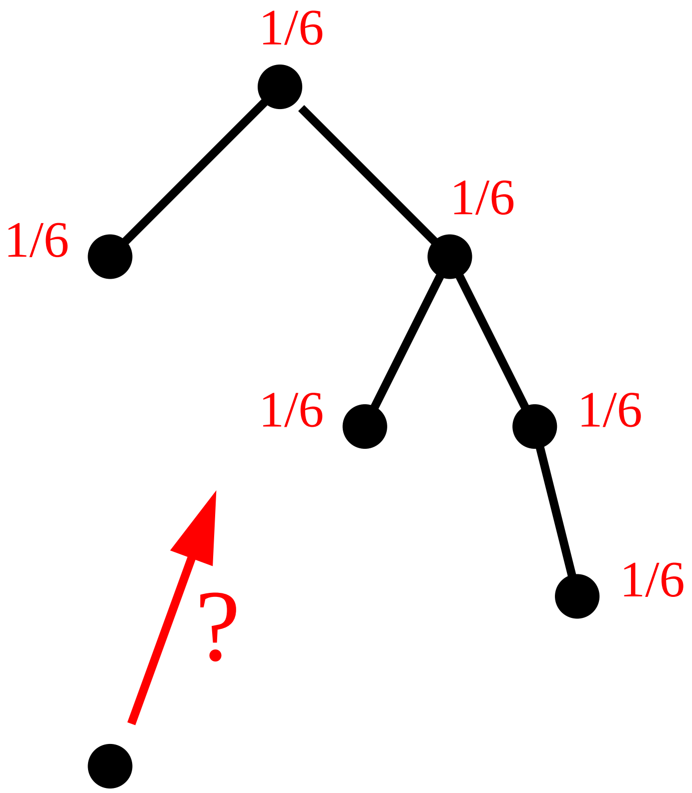
Model: homogeneous attachment



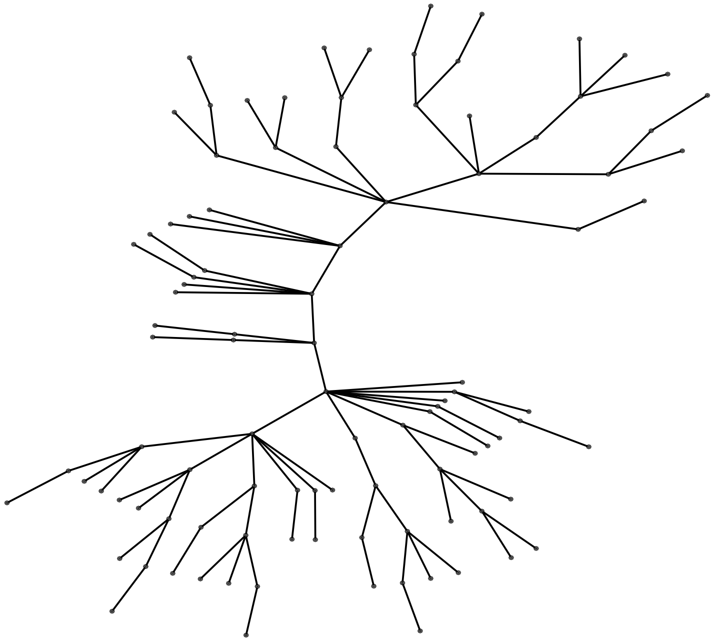
Model: homogeneous attachment



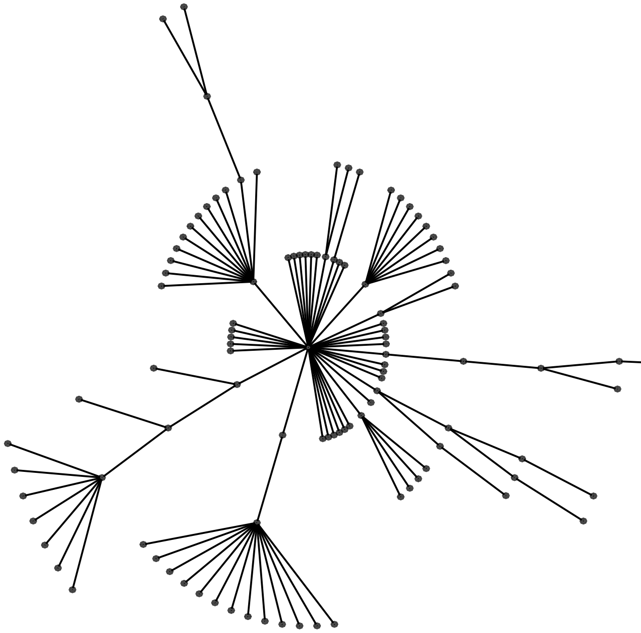
Model: homogeneous attachment



Comparing the model with the data

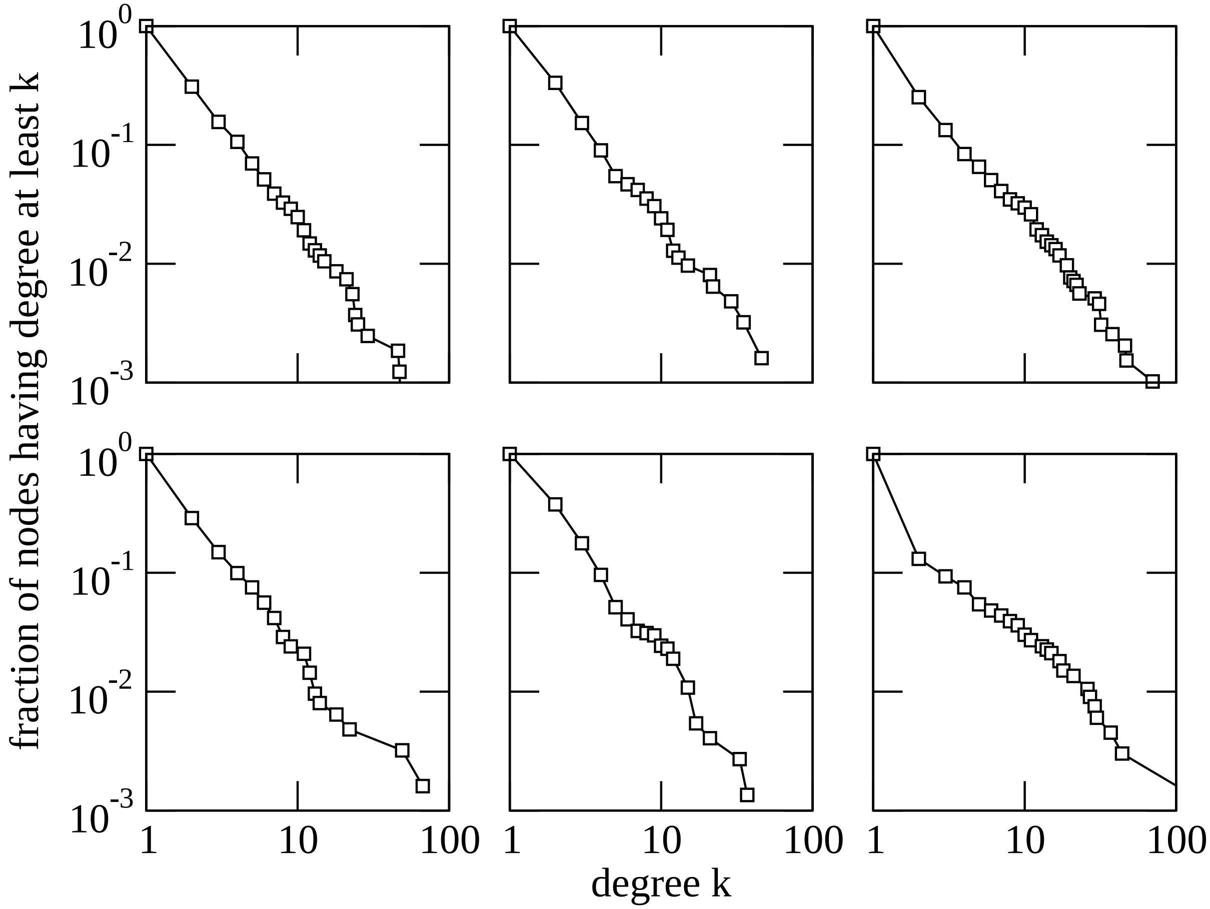


homogeneous attachment (model)

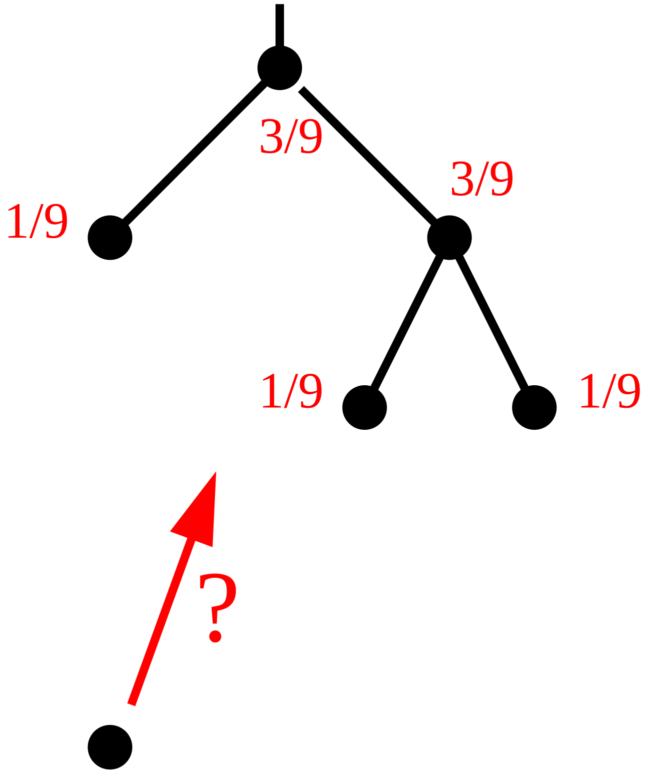


directory tree (data)

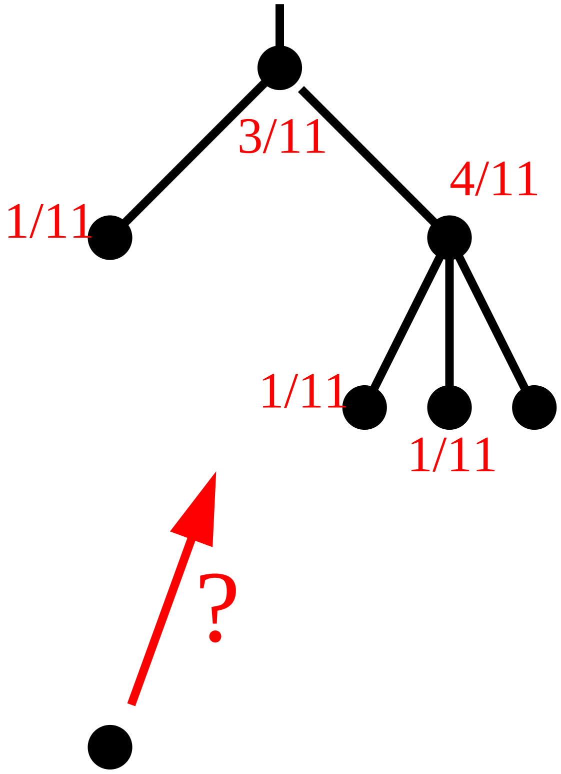
Degree distributions: Power laws



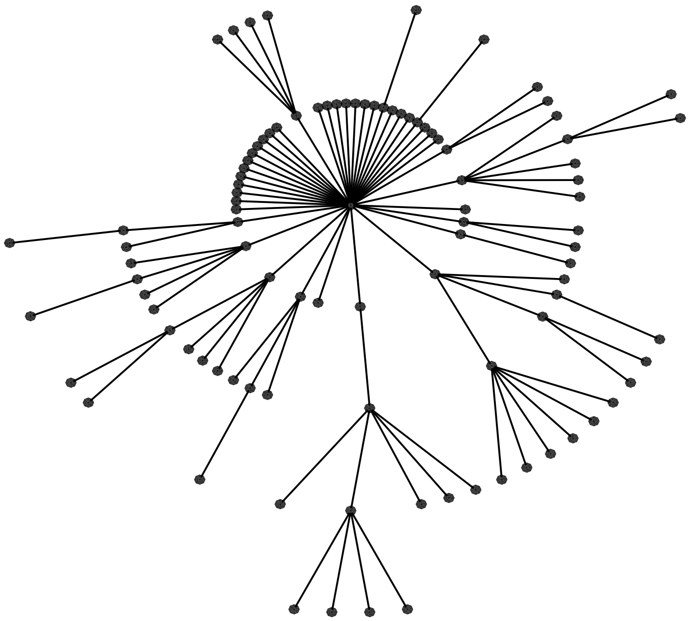
Model: preferential attachment



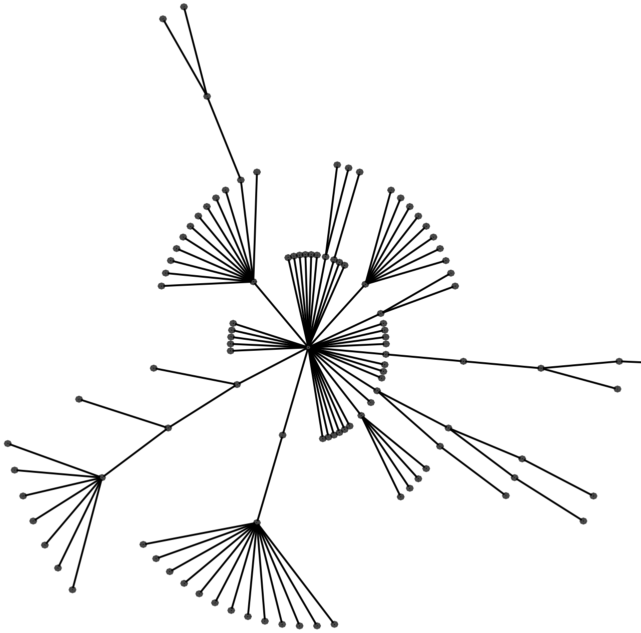
Model: preferential attachment



Comparing the model with the data



preferential attachment (model)



directory tree (data)

Preferential attachment model: Analytical results

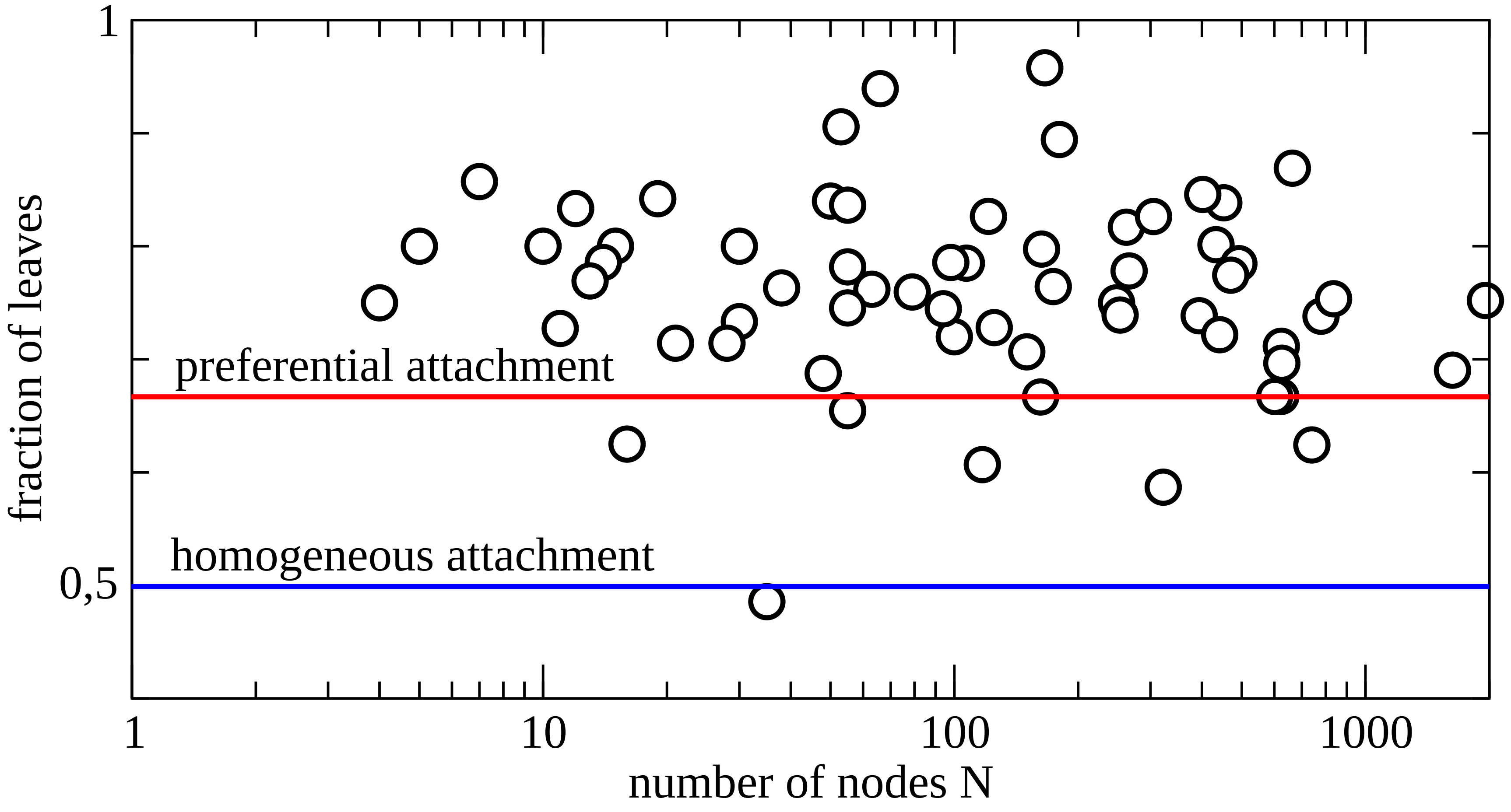
- degree distribution $P(k) \propto k^{-3}$ (power law → ok)

- fraction of leaves $P(1) \approx 2/3$

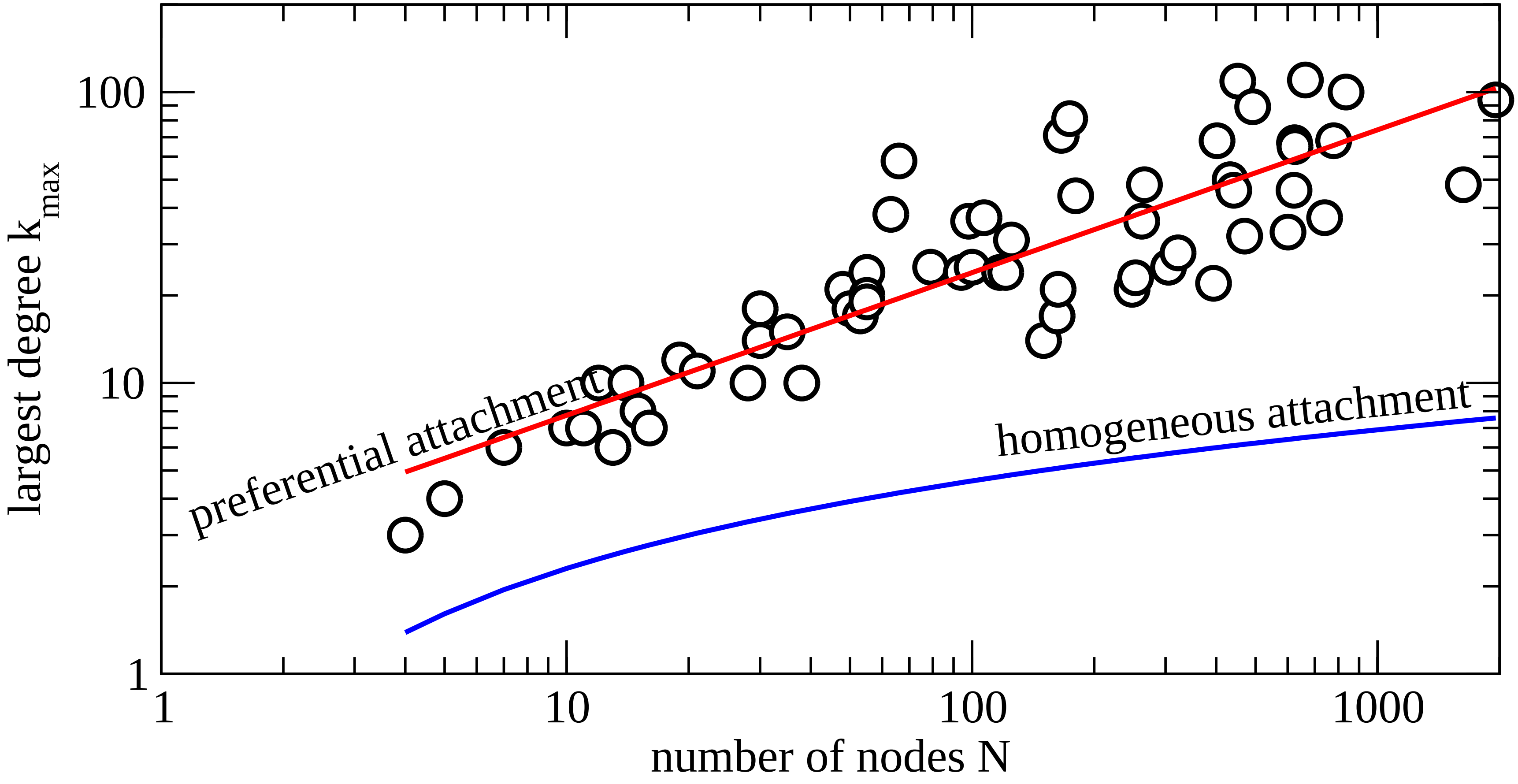
- maximum degree $k_{\text{max}} \propto N^{1/2}$ for system size N

- average distance of nodes from root $\lambda = \frac{1}{2} \ln N$

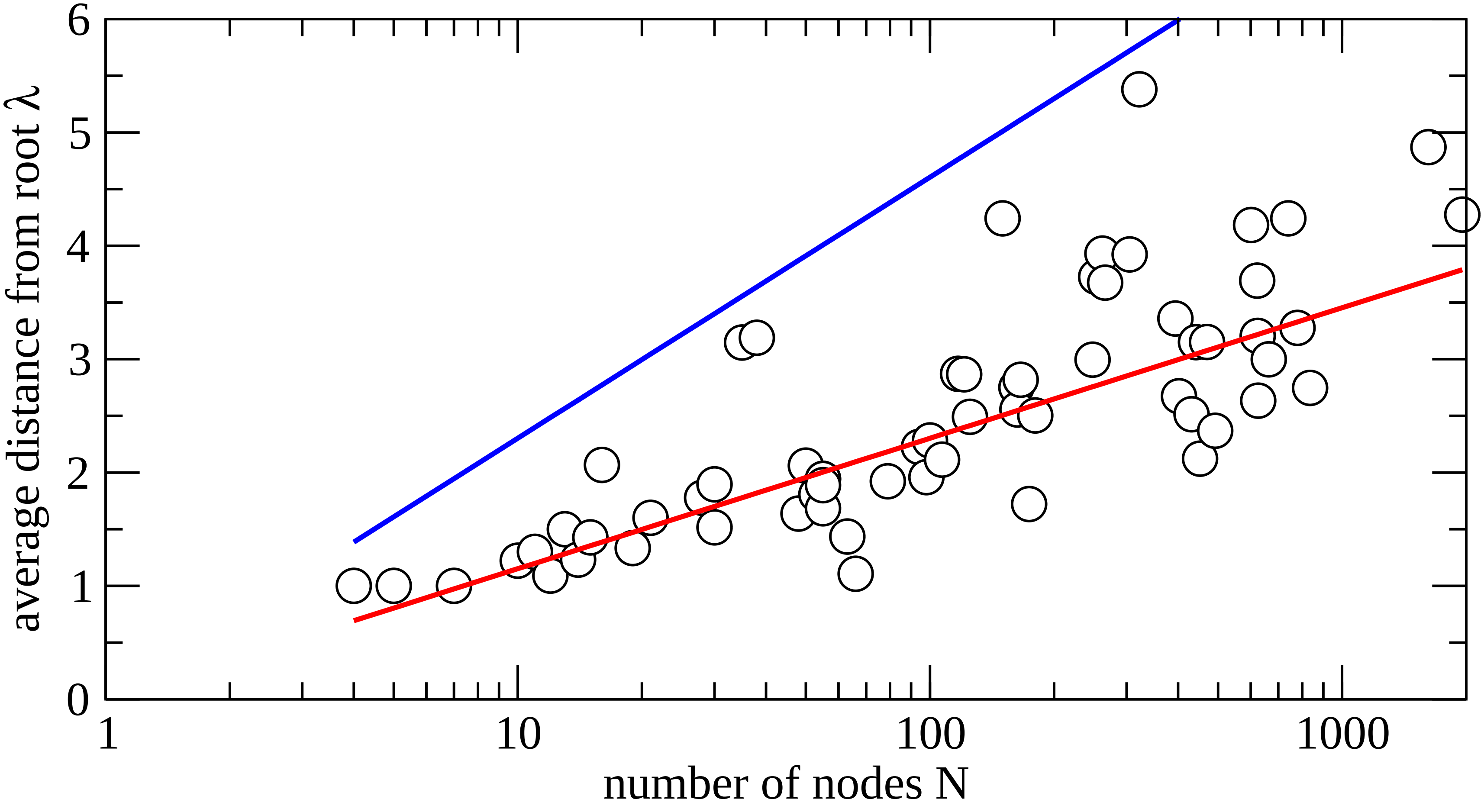
Fraction of leaves



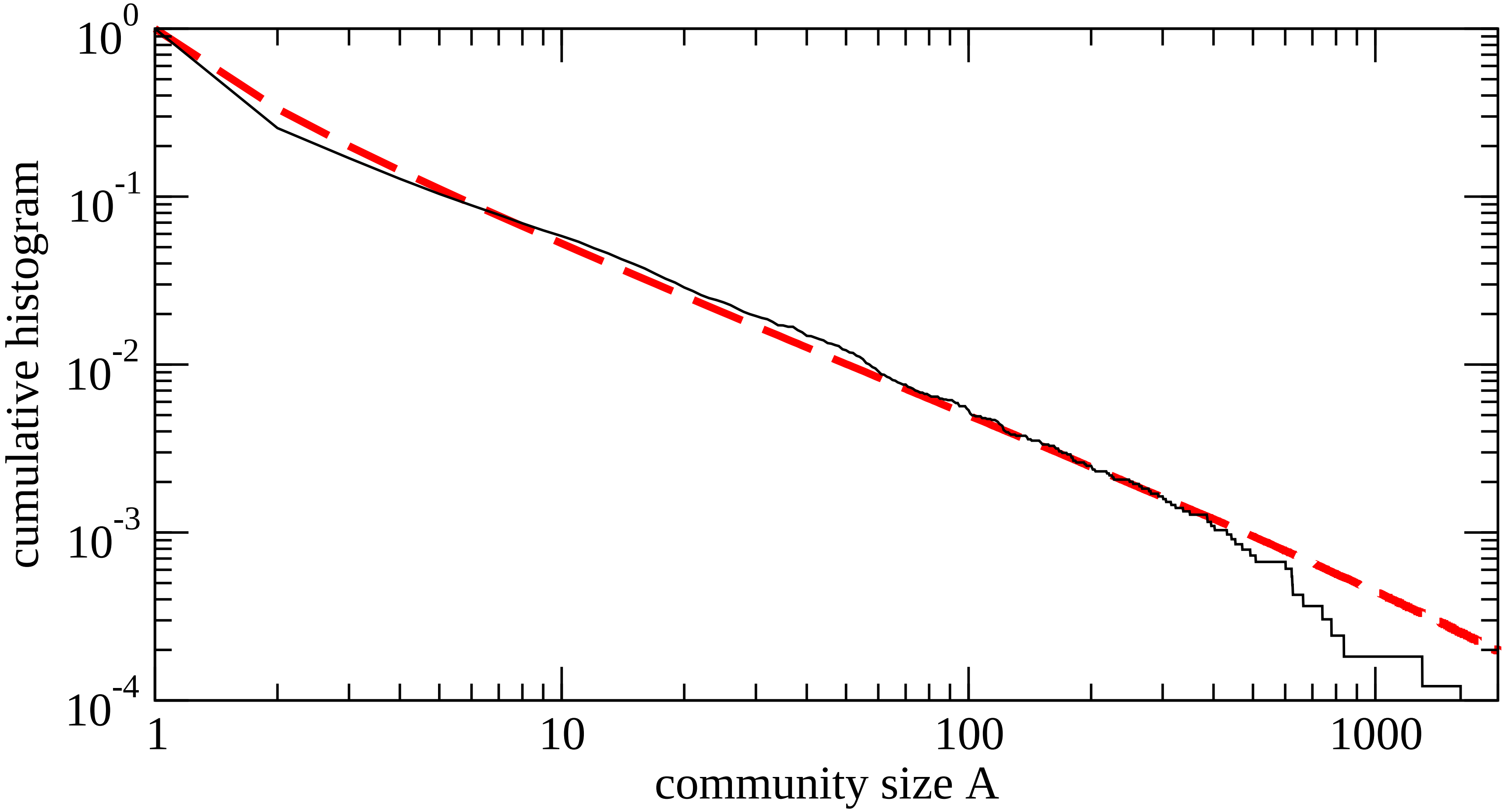
Largest degree



Nodes' distance from the root



Community structure



Conclusion

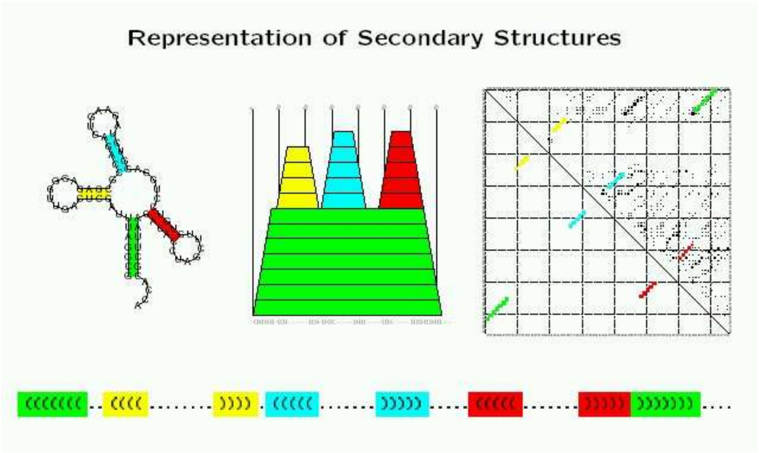
- Directory trees have interesting non-trivial structure

- Statistical properties of the ensemble can be explained by preferential attachment model:
  - degree distribution, fraction of leaves, maximum degree
  - distances on the tree
  - community structure

- Apparently all users follow the same rules for tree construction

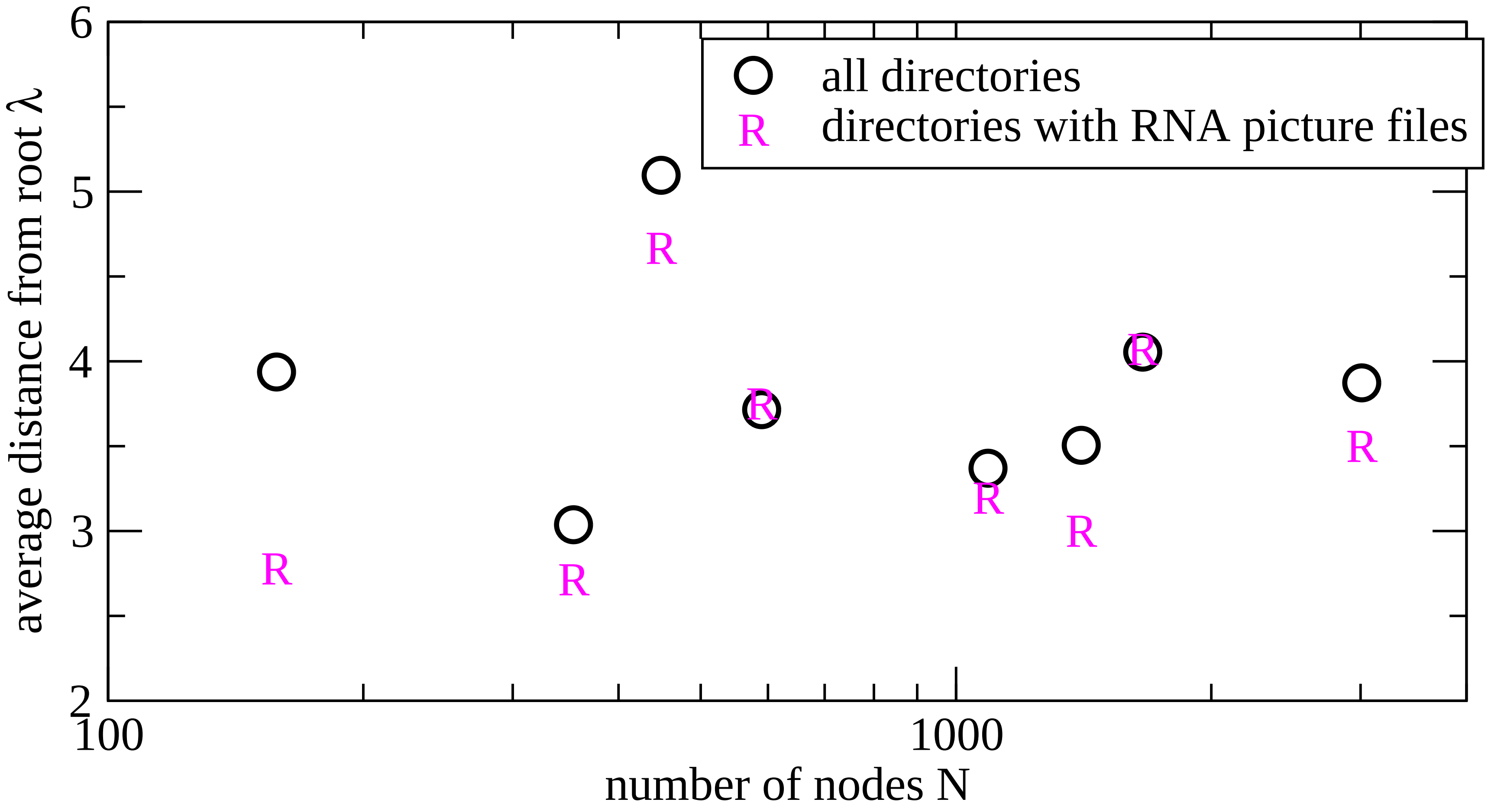
RNA picture files

- tRNA_phe_dp.*ps*
- tRNA_phe_ss.*ps*
- tRNAmnt.*ps*
- tRNA_phe_circ.*ps*
- bxf.*ps*
- bracket.*ps*



Directory trees of the users caro, ingrid, ivo, martin, roman, studla, xtina, xtof. Filesystem data provided by Sonja — Thanks!

Where are the RNA pictures?


average distance from root λ
number of nodes N
○ all directories
R directories with RNA picture files
R