

Data storage by individuals: The structure of directory trees

Konstantin Klemm

Bioinformatics Group, Leipzig University, Germany

Víctor M. Eguíluz and Maxi San Miguel

University of the Balearic Islands, Spain

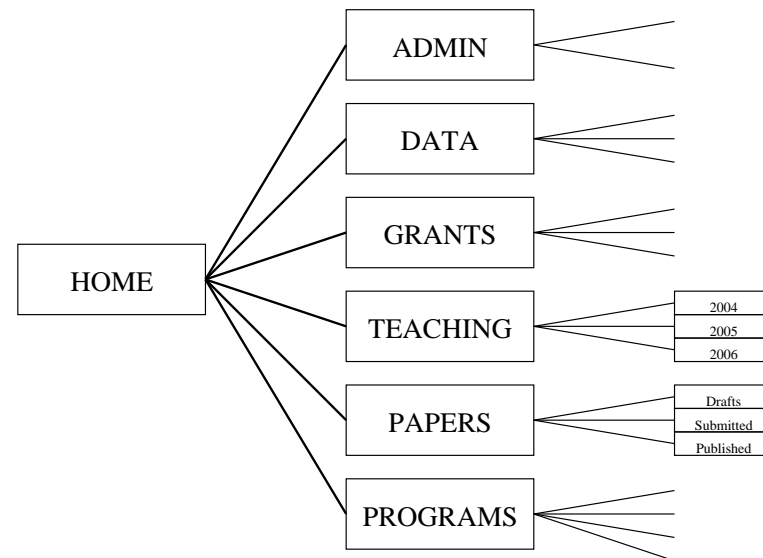
How do humans structure information?

... and, first of all, where and how can we observe this?

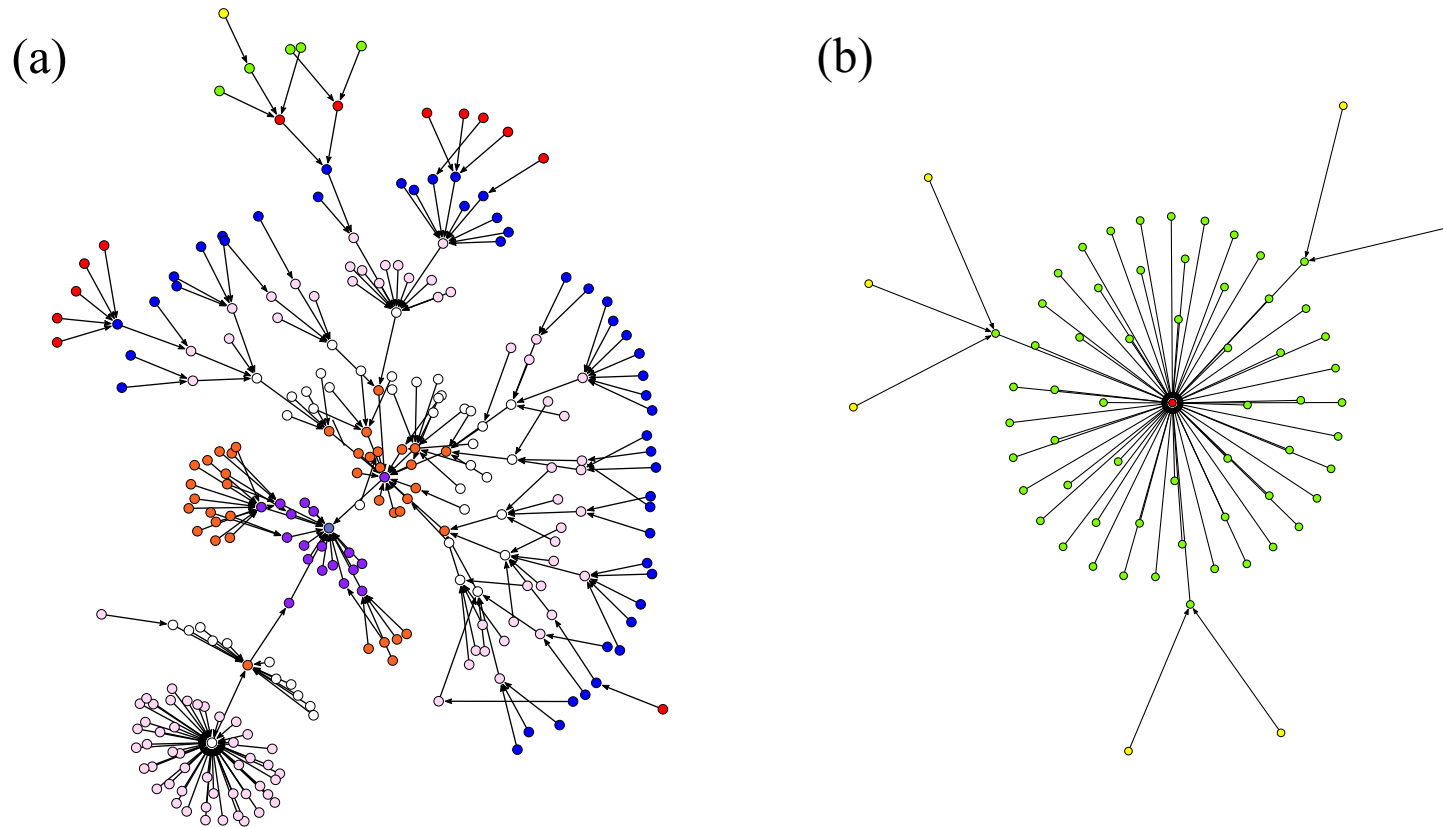
- psychological experiments (\approx interviews)
costly, potentially biased by experimental setup
- tables of contents: sectioning of books, theses etc. —
reveal how authors structure information *for others*
- structure of files and folders in user's account —
reveal how users structure information *for themselves*

Directory trees: What?

- rooted trees
- each instance generated by one user alone
- node = directory, connected to its parent and its subdirectories.
- “symbolic links” ignored



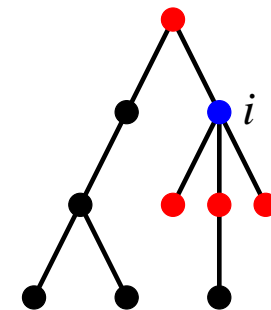
Two trees from the data set



Data set: Trees of 63 users at Physics Dept. in Palma. Sizes N in the range $4 \dots 2000$.

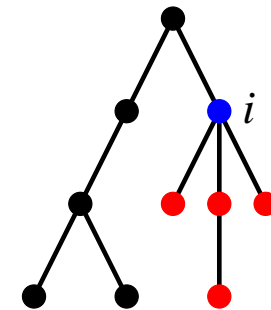
Quantities of interest

- neighborhood Γ_i : set of nodes directly linked with node i .
- degree $k_i = |\Gamma_i|$

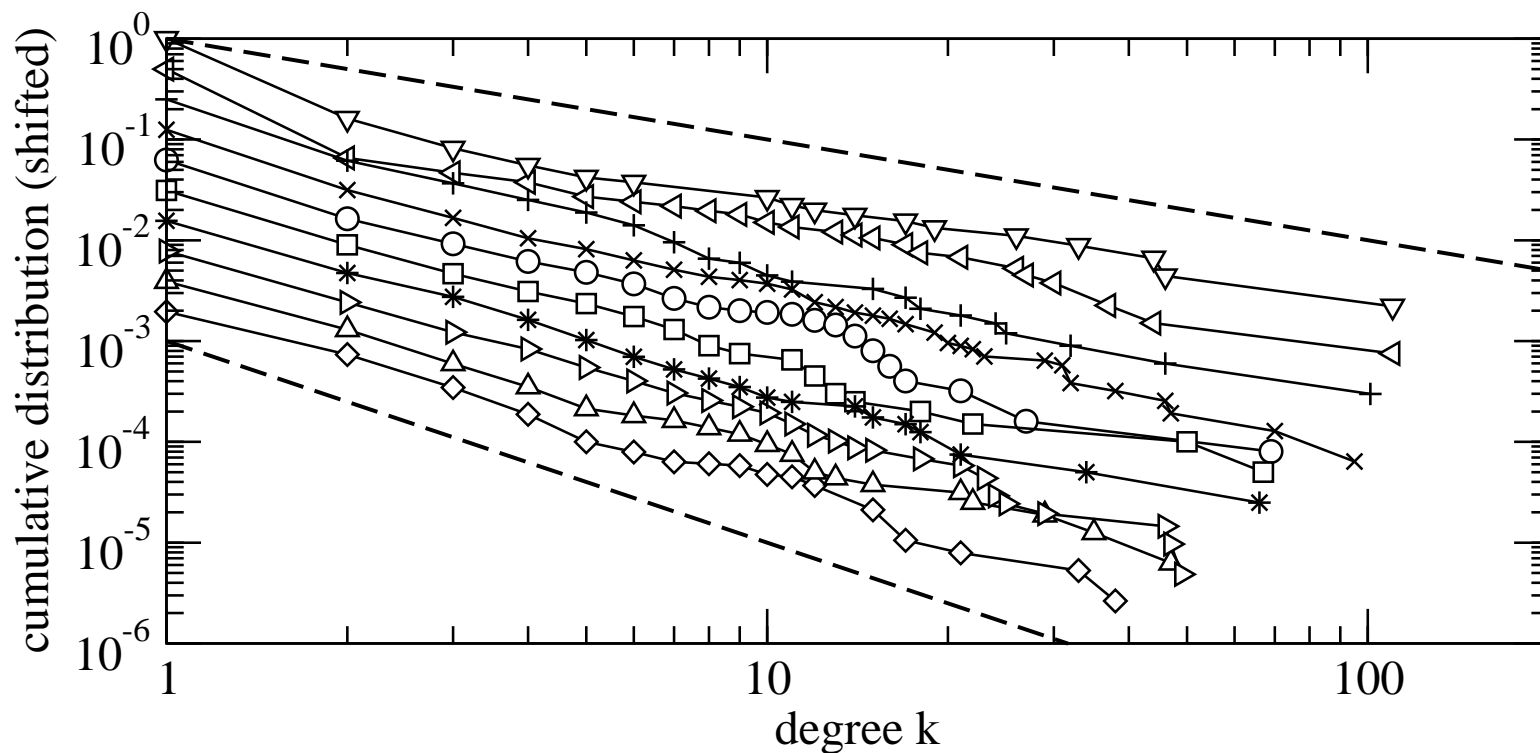


- community S_i : set of nodes “below” node i
- community size $A_i = |S_i|$

- cumulated community size $C_i = \sum_{j \in S_i} A_j$
(allometric scaling)

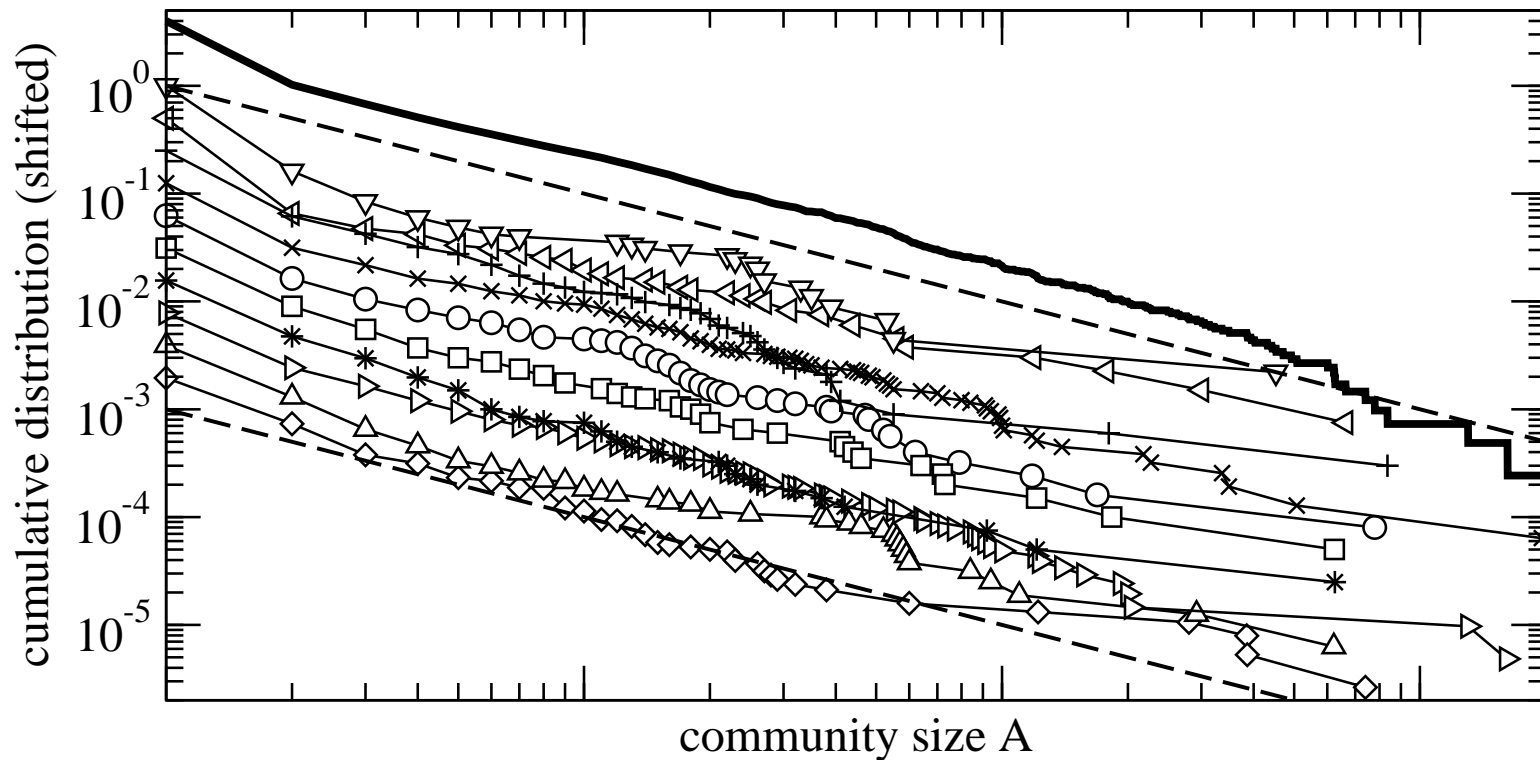


Degree distributions



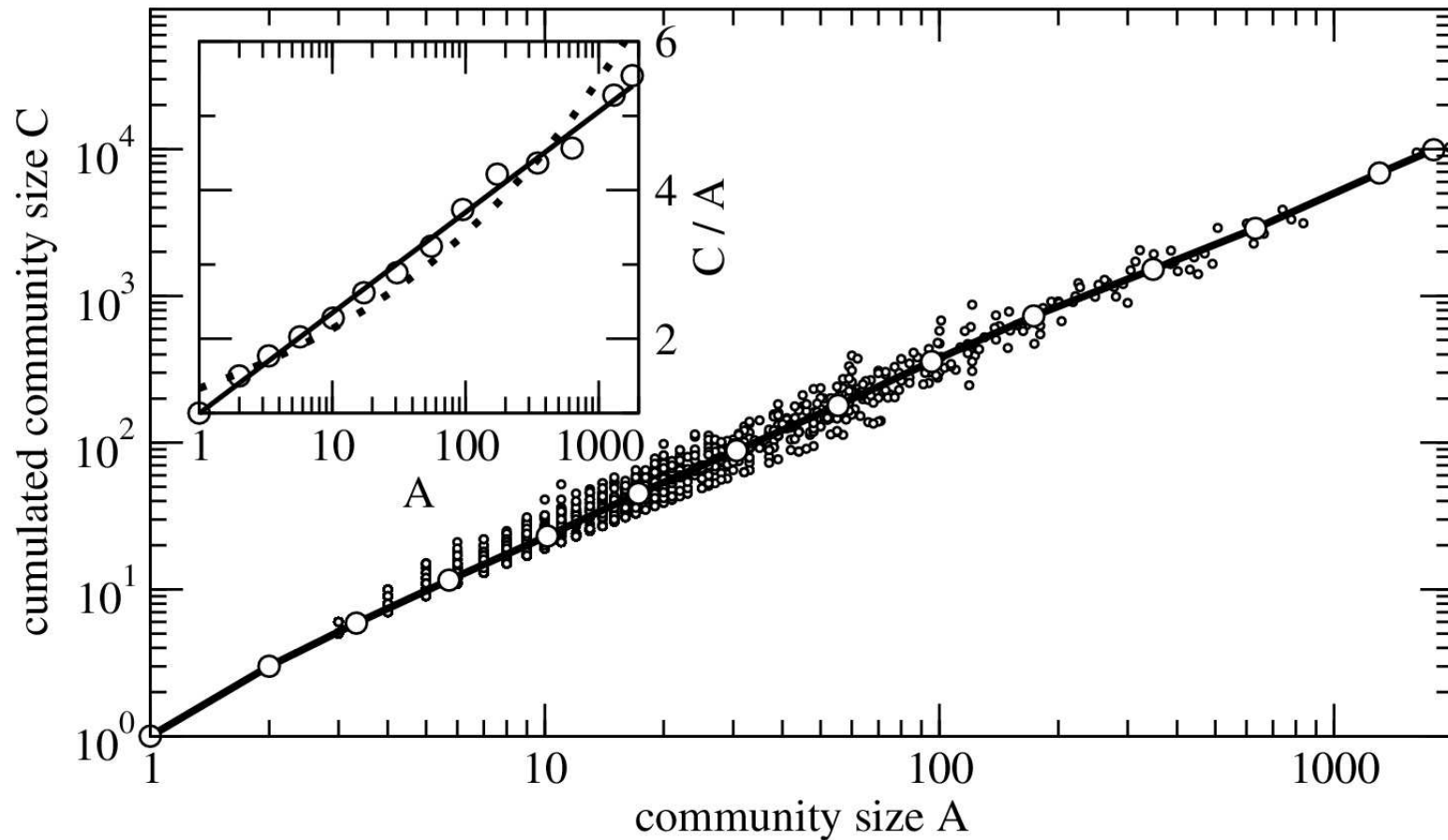
Cumulative degree distributions of each of the 10 largest trees (thin lines with symbols). Dashed lines indicate degree exponents $\gamma = 2$ and $\gamma = 3$.

Community size distributions



Cumulative community size distributions of each of the 10 largest trees (thin lines with symbols) and all 63 trees together (thick curve). Dashed lines indicate community size exponent $\tau = 2$.

Allometric scaling, “radius vs. volume”



Cumulated community size $C_i = \sum_{j \in S_i} A_j$ against community size A_i for all 16452 communities in the data set.

Overview: observed scaling

degree distribution	$P(k) \sim k^{-\gamma}, \quad \gamma > 2$ non-universal
community size distr.	$Q(A) \sim A^{-\tau}, \quad \tau = 2$ universal(?)
allometric scaling	$C \sim A \ln A$

Growth model — definition

- Trees generated by iterative attachment of nodes.
- Probability to attach new node to node of degree k in a tree of size N

$$\Pi(k) = q \frac{k-1}{N} + (1-q) \frac{1}{N}$$

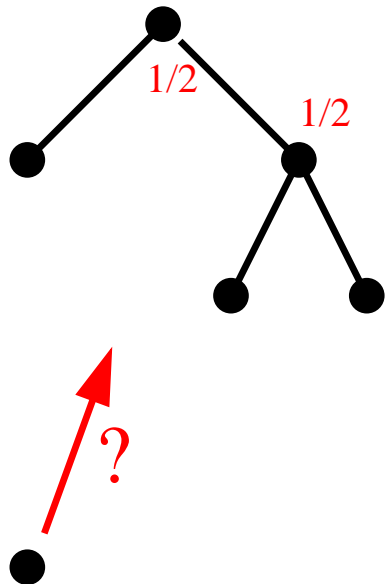
- Combines preferential and homogeneous attachment at tunable ratio q

Growth model — illustration

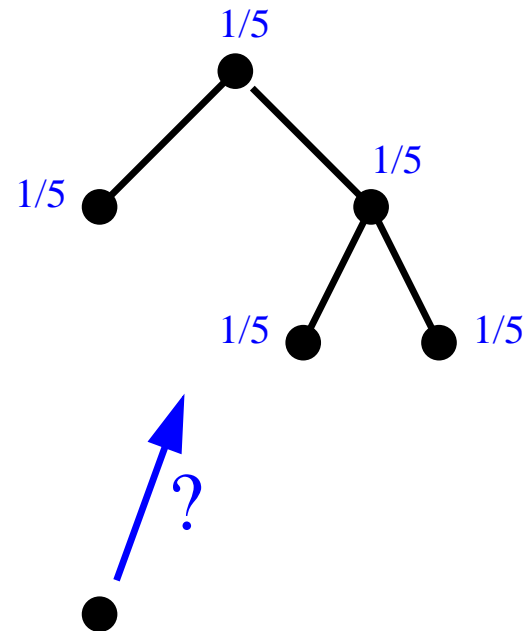
preferential attachment (copying)

homogeneous attachment

$$\Pi(k) \propto k - 1$$



$$\Pi(k) = 1/N$$



Growth model — analytical results

degree distribution	$P(k) \sim k^{-\gamma}, \gamma = 1 + q^{-1}$ (non-universal)
community size distr.	$Q(A) \sim A^{-\tau}, \tau = 2$ (universal)
allometric scaling	$C(A) = A[(1 - q) \ln A + 1]$

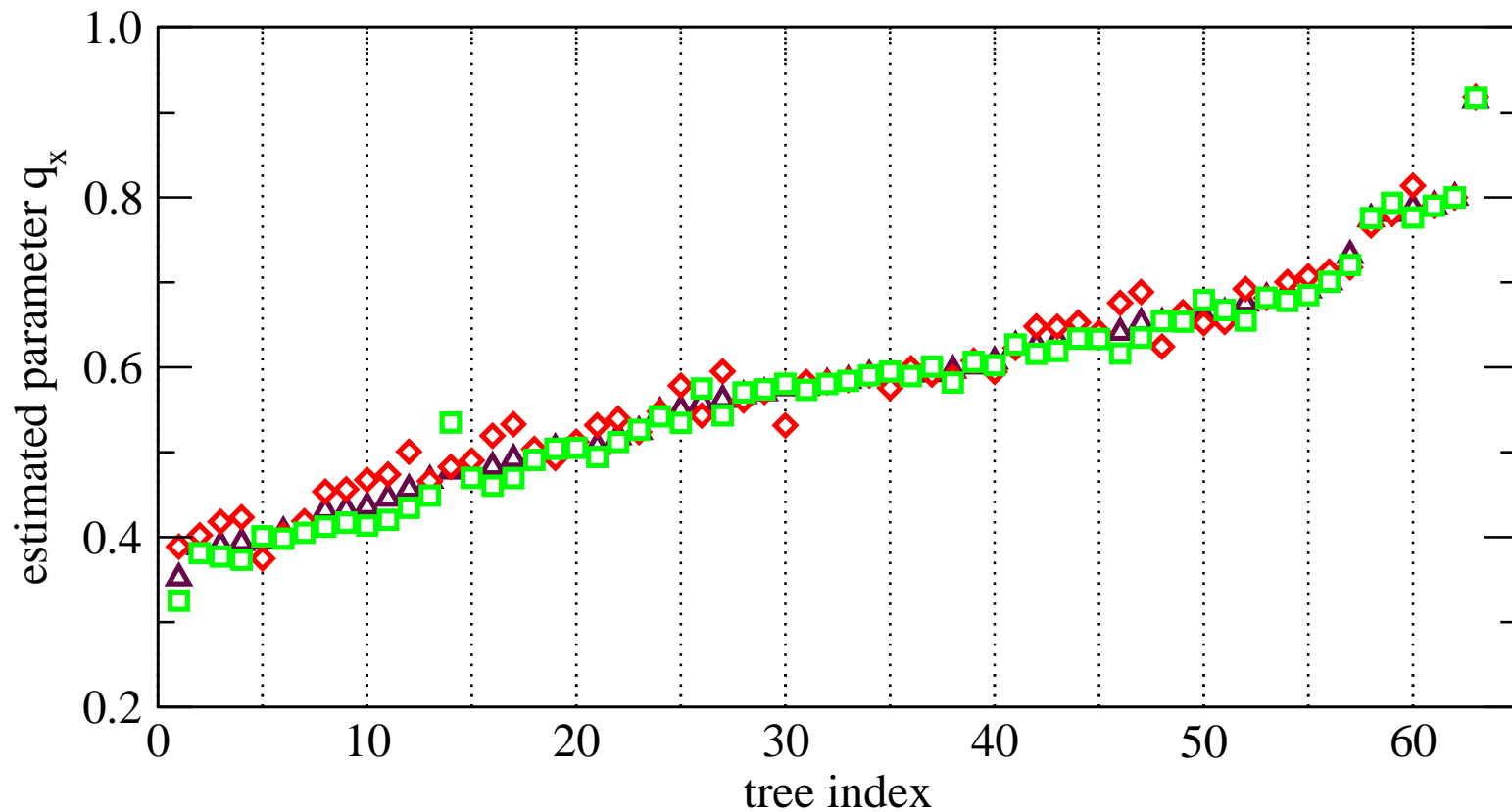
Growth model — analytical results

degree distribution	$P(k) \sim k^{-\gamma}, \gamma = 1 + q^{-1}$ (non-universal)	✓
community size distr.	$Q(A) \sim A^{-\tau}, \tau = 2$ (universal)	✓
allometric scaling	$C(A) = A[(1 - q) \ln A + 1]$	✓

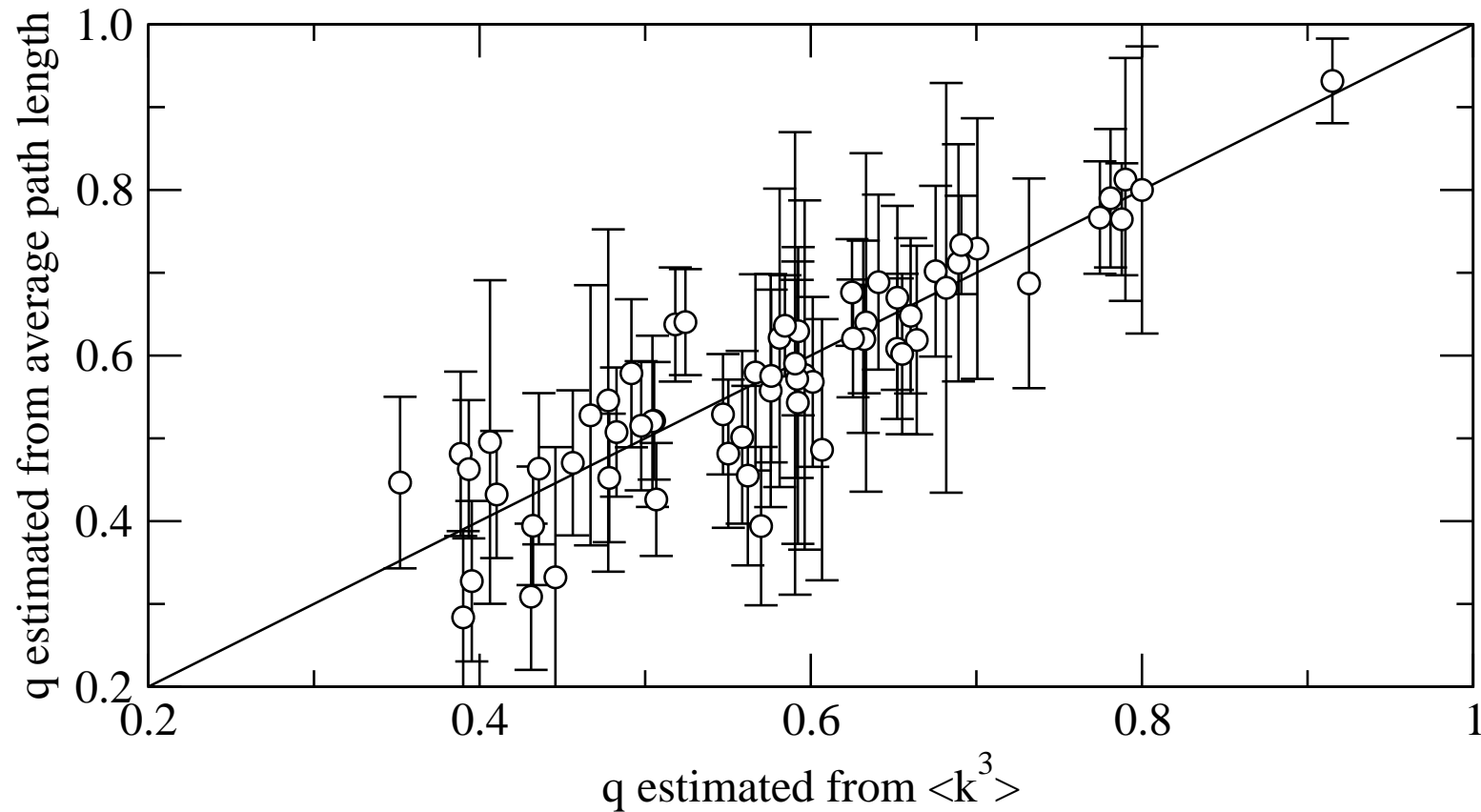
Agreement between model and empirical trees

Estimating q from the empirical trees (1)

Maximum likelihood estimates of q based on the second (\diamond), third (\triangle) and fourth (\square) moment of the degree distribution.



Estimating q from the empirical trees (2)



⇒ Values of q estimated consistently from different observables.

Summary / Outlook

- Directory trees have interesting non-trivial structure with scaling properties.
- Properties are largely reproduced by a simple growth model.
- Model assigns each user a parameter value q that can be extracted from her/his tree (max. likelihood estimates).
- Future work: long-term observation of directory trees and comparison with dynamics of the model.

Klemm, Eguíluz, and San Miguel, *Phys. Rev. Lett.* **95**, 128701 (2005).