UNIVERSITY OF OXFORD

Systems Biology Doctoral Training Centre

High Throughput Analysis of Networks

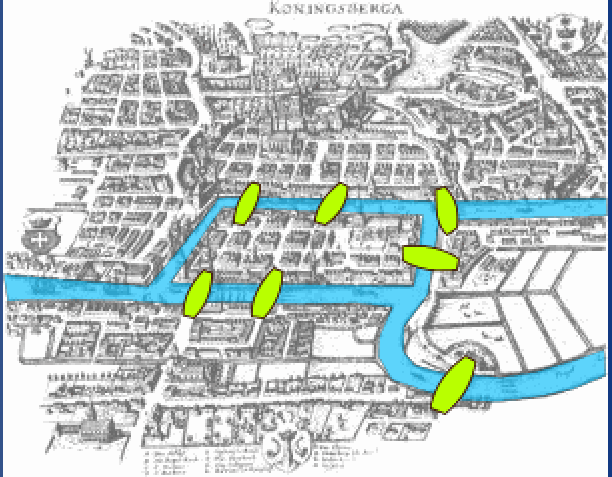
Sumeet Agarwal, Gabriel Villar, Nick Jones

20 September 2010

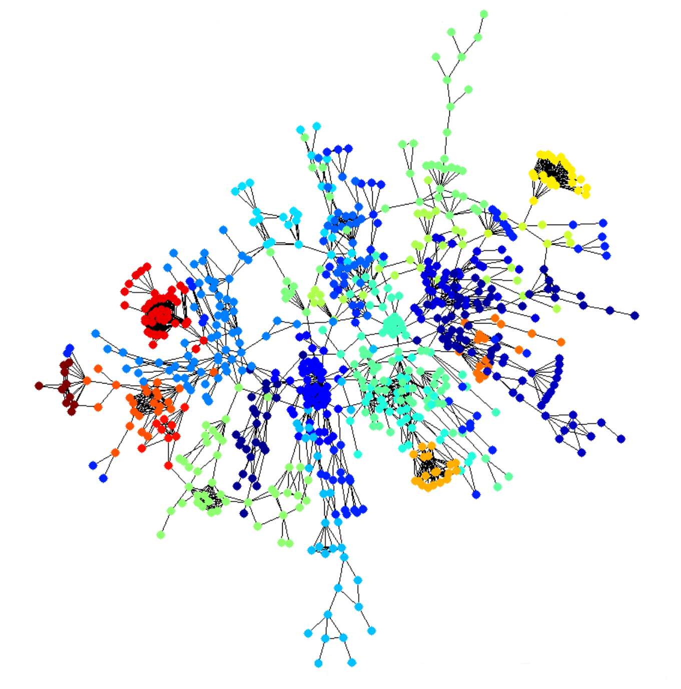
Workshop on Analysis of Complex NEtworks (ACNE)
ECML/PKDD 2010, Barcelona

Motivation: Consolidation of network science

- The study of graphs and networks goes back at least to Euler. People from a wide range of disciplines have contributed: Mathematicians, Computer Scientists, Electrical Engineers, Sociologists, Physicists, Statisticians...
- This has led to a fragmented literature, with inconsistent terminology and frequent reinvention of concepts and methodologies
- Our aim is to utilise the power of computing and data mining techniques to construct a comprehensive database of networks and network algorithms, and use this to systematically investigate patterns of relationships between different kinds of networks and metrics/features
- This kind of data-driven approach may allow us to choose the most relevant features for a given task, motivate appropriate network models, and in general answer the question: What are the best ways of thinking about networks?

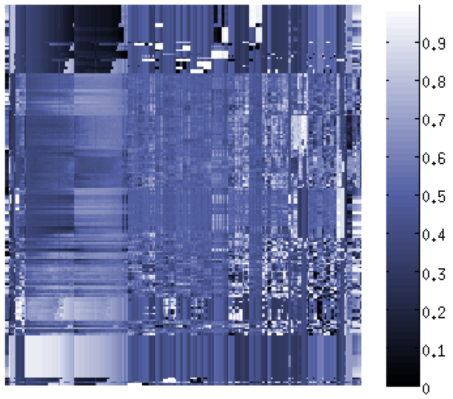


Courtesy: Wikipedia

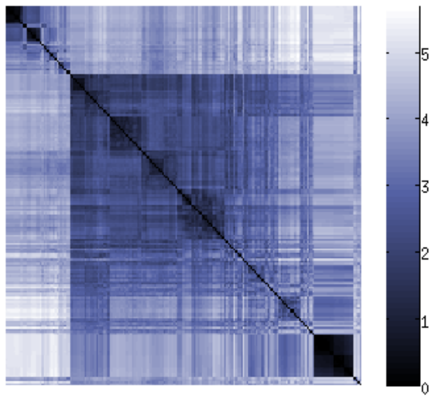


What is “high throughput network analysis”?

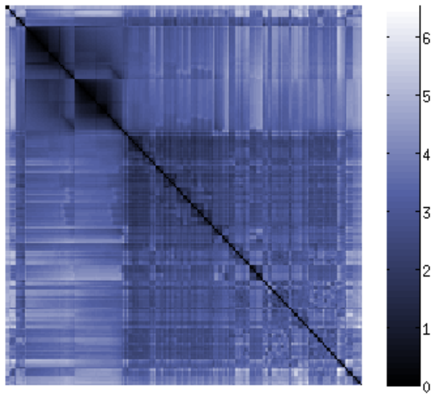
- An attempt to study network properties at a rather abstract level, using computing power to automate many different analytic procedures across many different networks
- This gives us a matrix of networks versus metrics/features, which can be mined to identify features and networks of interest, cluster them into ‘families’, learn predictive models for system phenotype etc.
- It is a way of organising and systematising the diverse range of network analysis techniques to give us a better sense of the current state of the field



Data matrix:
networks vs. metrics



Correlation matrix:
networks vs. networks



Correlation matrix:
metrics vs. metrics

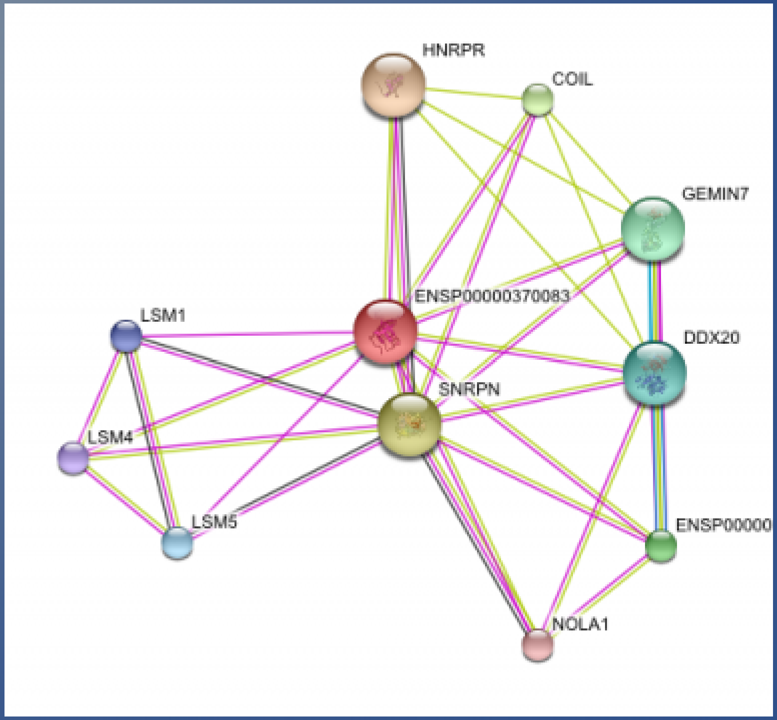
What kinds of networks do we study?

- Network representations have been used to study a wide variety of data:
  - Technological networks (railways, telephone lines, internet)
  - Information networks (WWW, cell phones, e-mail)
  - Social networks (friendship/kinship, Facebook, Twitter)
  - Biological networks:
    - Ecological
    - Neural
    - Subcellular (metabolic, protein-protein, gene regulation)
- We attempt to gather as many data sets as we can from different sources, and also construct synthetic data sets for comparative purposes

What kinds of metrics do we study?

Simple numeric features: size, assortativity (degree correlations), mean path length

Summaries of feature distributions over nodes/links: degree, centrality measures, clustering coefficient

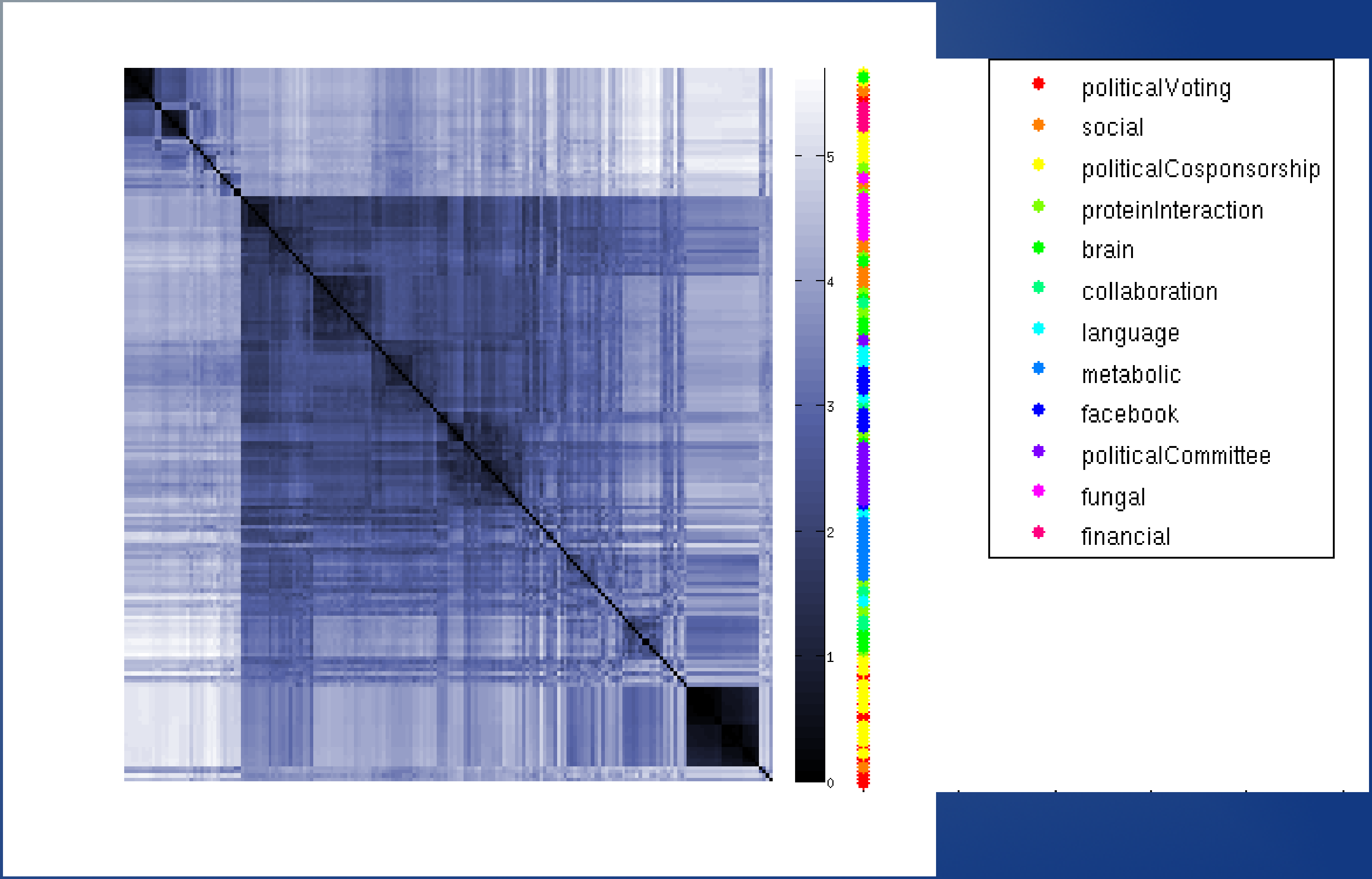


Community structure: partition entropy, modularity, coarse-grained networks

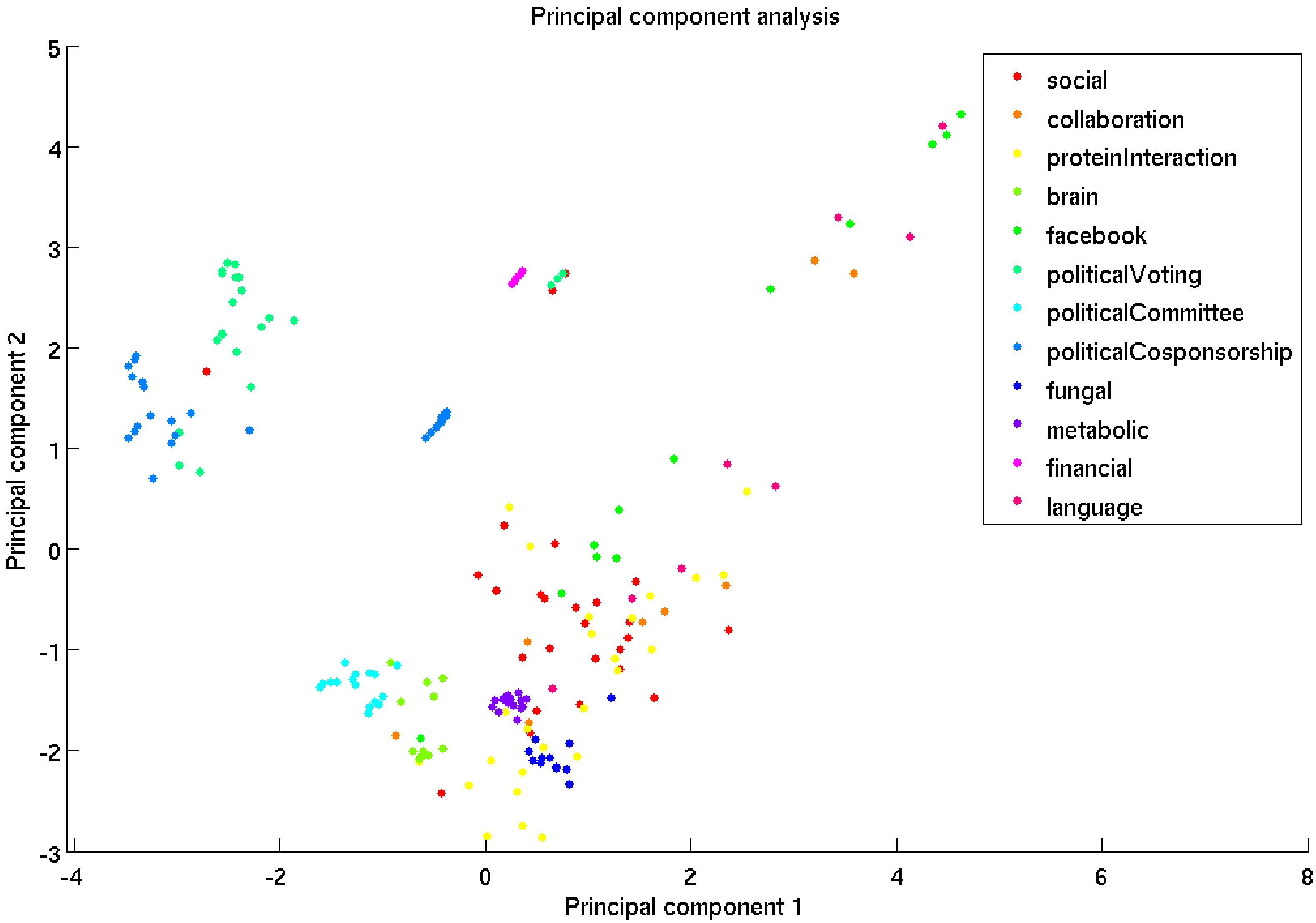
Model fits: how well the network is explained by a certain generative model (preferential attachment, duplication and divergence)

More unconventional quantities such as motif counts, linear algebra operations (eigenvectors, Laplacian) on adjacency matrix

Network Families: Single linkage clustering

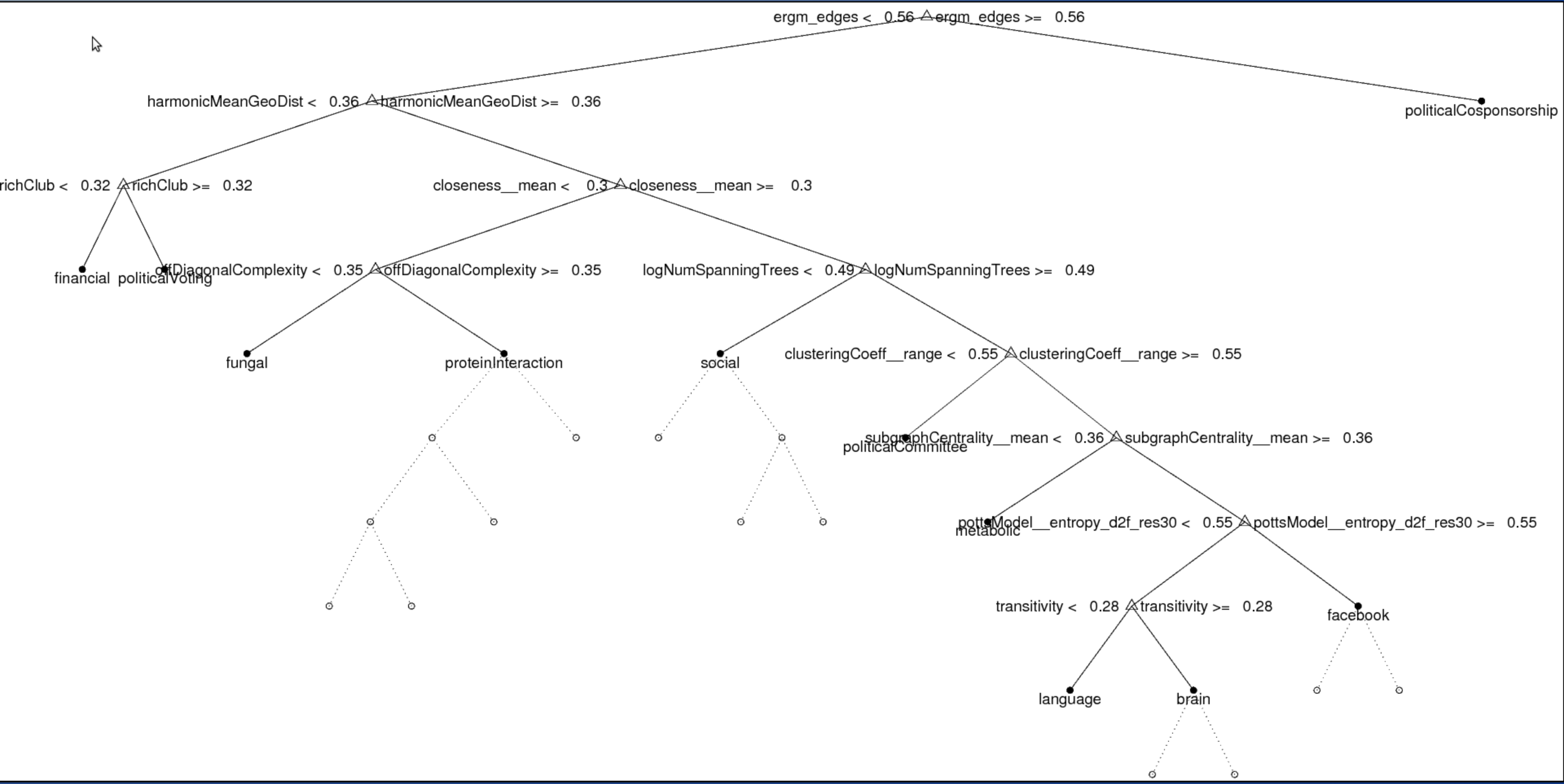


Network Families: Principal Component Analysis

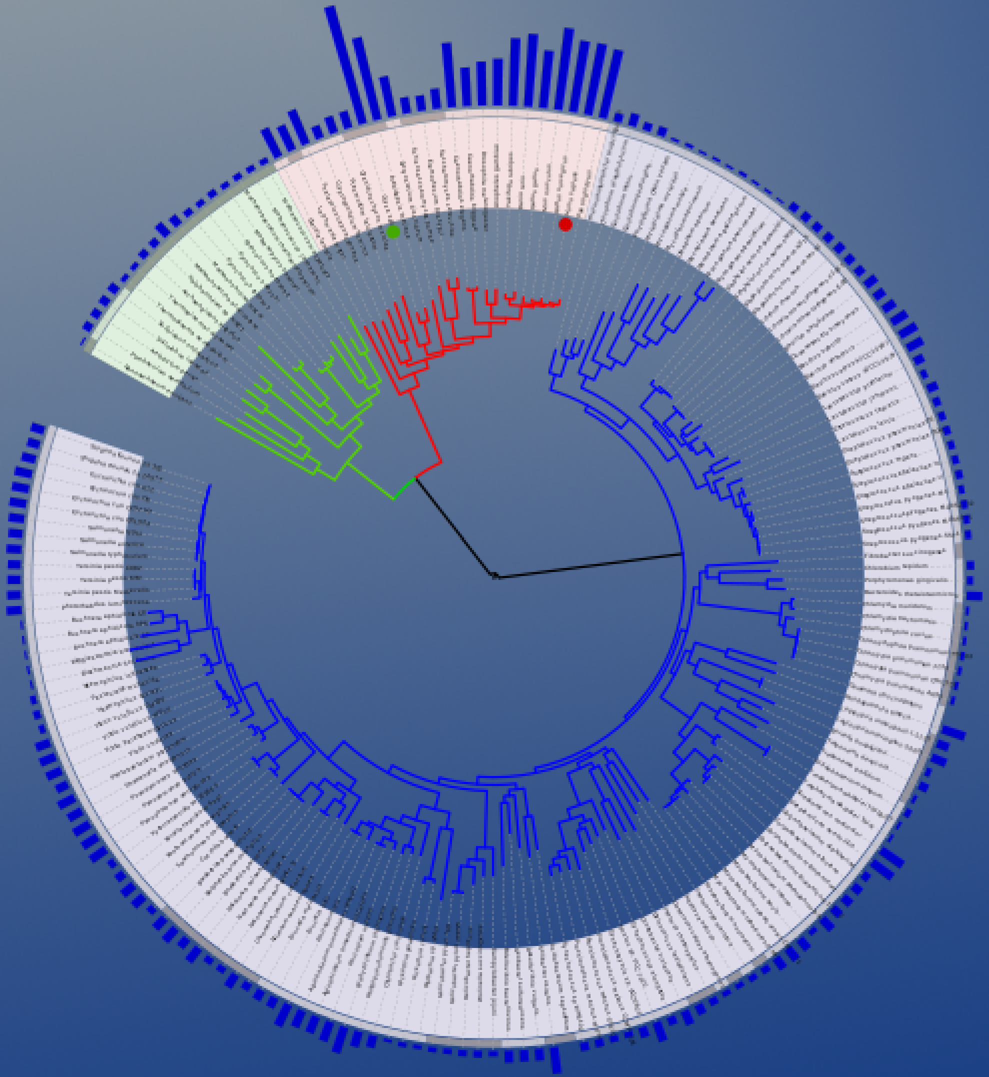


Network Classification

- Decision tree gives ~80% accuracy on a 12-class task


ergm_edges < 0.56
ergm_edges >= 0.56
harmonicMeanGeoDist < 0.36
harmonicMeanGeoDist >= 0.36
politicalCosponsorship
richClub < 0.32
richClub >= 0.32
closeness__mean < 0.3
closeness__mean >= 0.3
financial
politicalVoting
offDiagonalComplexity < 0.35
offDiagonalComplexity >= 0.35
logNumSpanningTrees < 0.49
logNumSpanningTrees >= 0.49
fungal
proteinInteraction
social
clusteringCoeff__range < 0.55
clusteringCoeff__range >= 0.55
politicalCommittee
subgraphCentrality__mean < 0.36
subgraphCentrality__mean >= 0.36
metabolic
pottsModel__entropy_d2f_res30 < 0.55
pottsModel__entropy_d2f_res30 >= 0.55
transitivity < 0.28
transitivity >= 0.28
facebook
language
brain

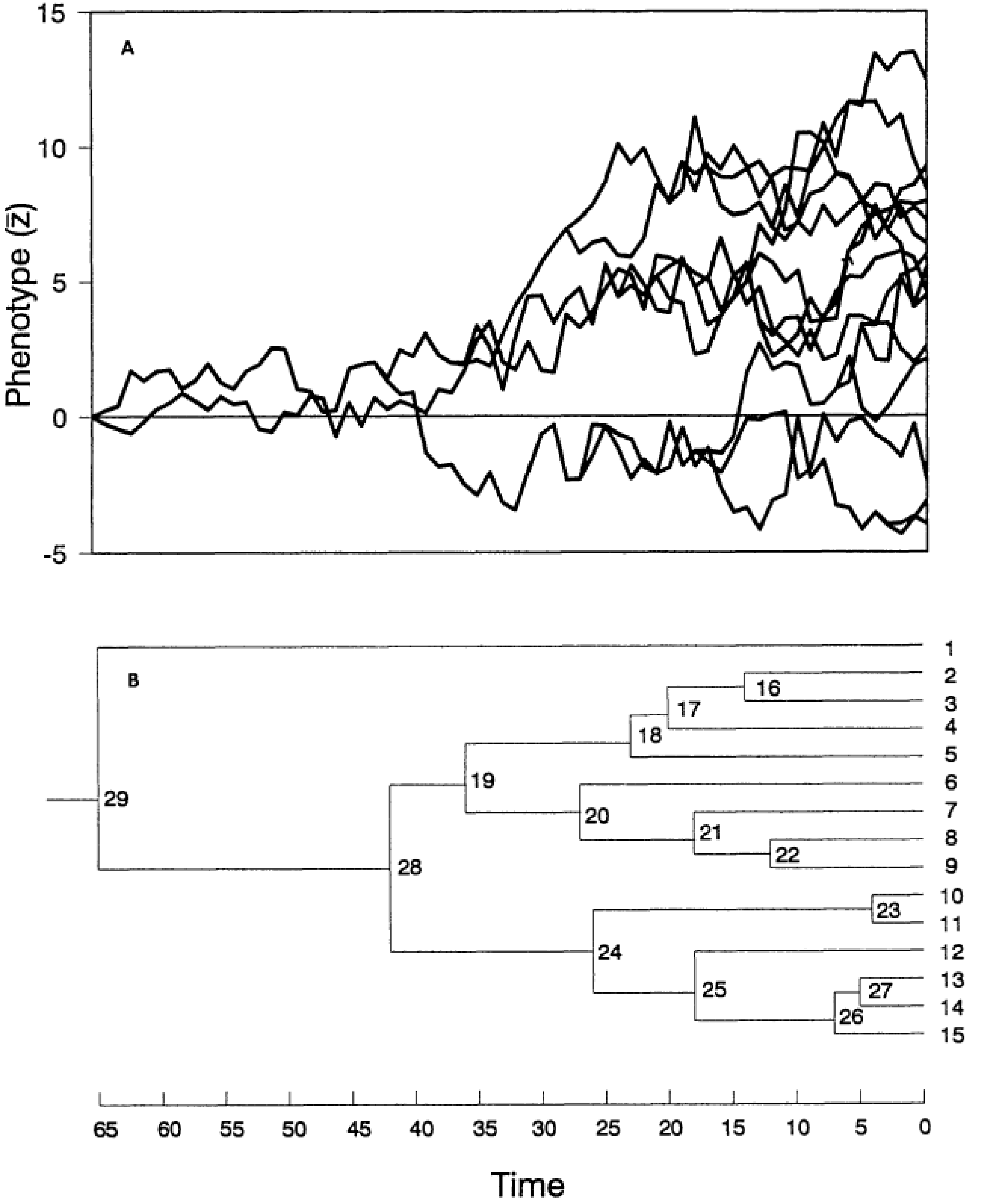
Example: Phylogenetic Comparative Methods



Courtesy: Wikipedia

- We can use features of biological networks in conjunction with independent evolutionary phylogenies to search for 'phylogenetic signals', i.e., properties that are most conserved in closely related species
- The idea is to assume a statistical process governing the evolution of any given trait (e.g., Brownian motion), and compute the likelihood of seeing the observed distribution of trait values at the leaves of the tree

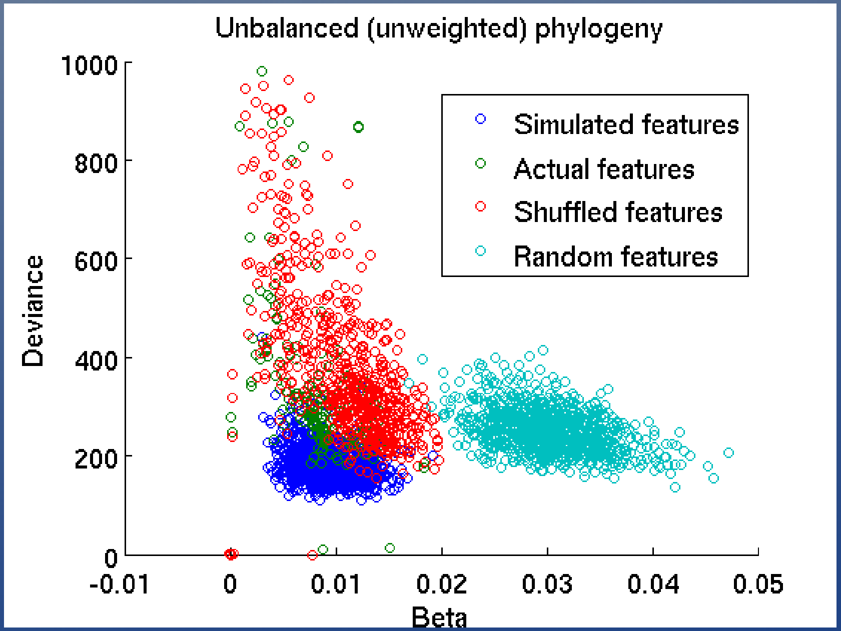
- We attempted to fit a Brownian motion model of evolution ($V = \beta t + \varepsilon$) to 272 real-valued network metrics computed on 450 metabolic networks from 158 different genuses, using a phylogeny taken from the Tree of Life

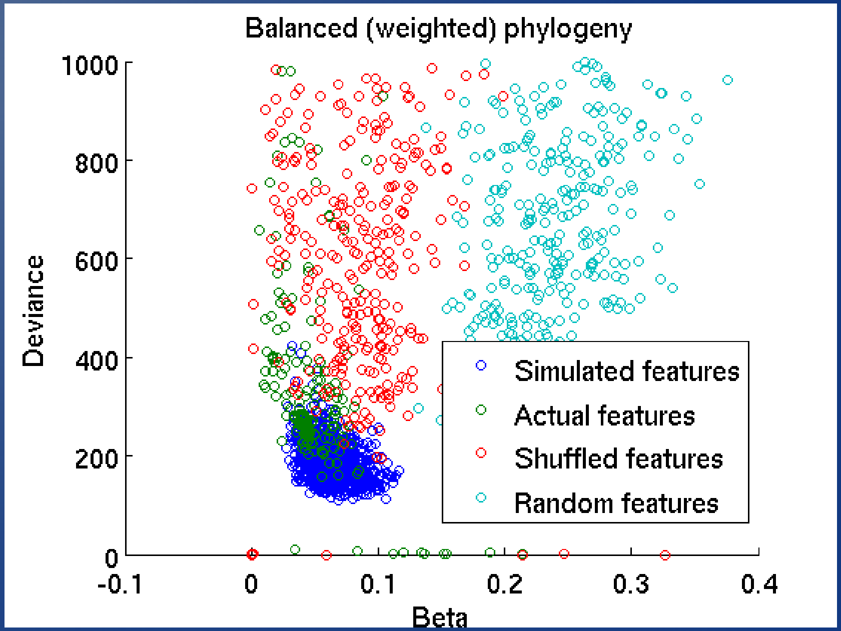


(Emilia P. Martins, *Am. Nat.* 1994)

A realistic phylogeny gives significant feature correlations

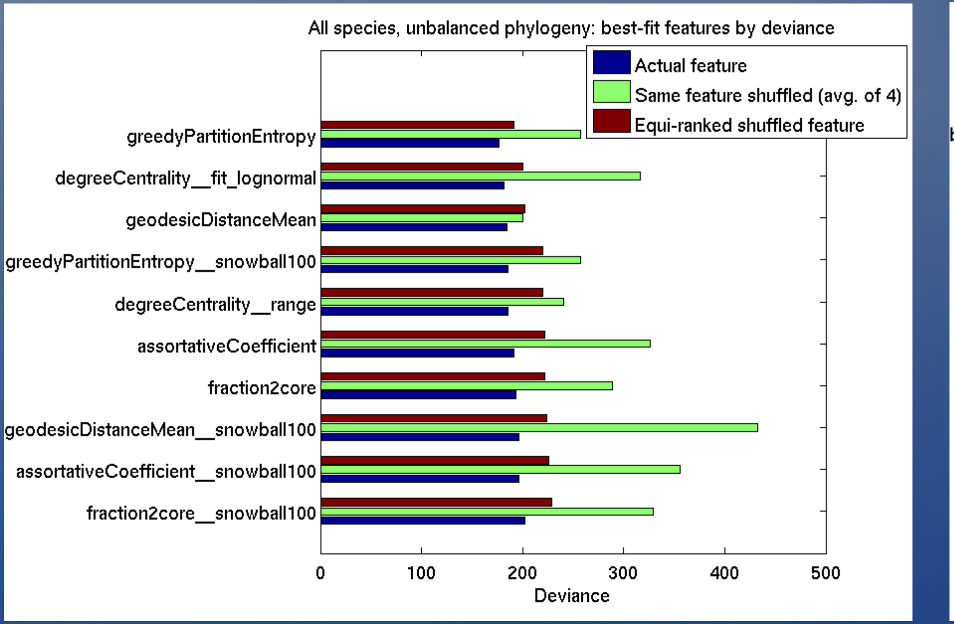
- An unbalanced version of the tree (with no branch weights) was compared with a balanced version (all leaves at the same depth)
- We used deviance (sum of sqaures of the residuals, ε) as a measure of the goodness-of-fit of the model for each metric/feature

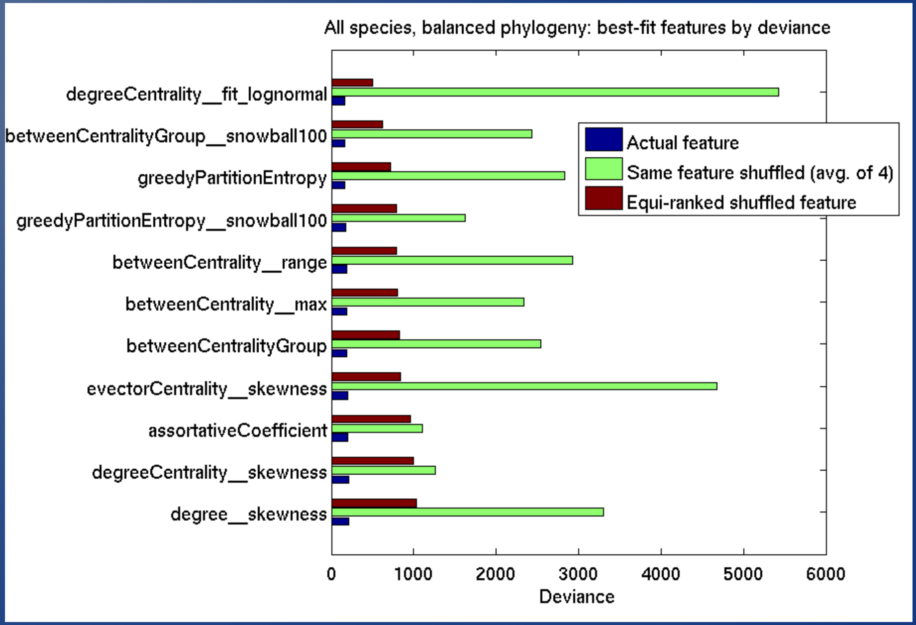




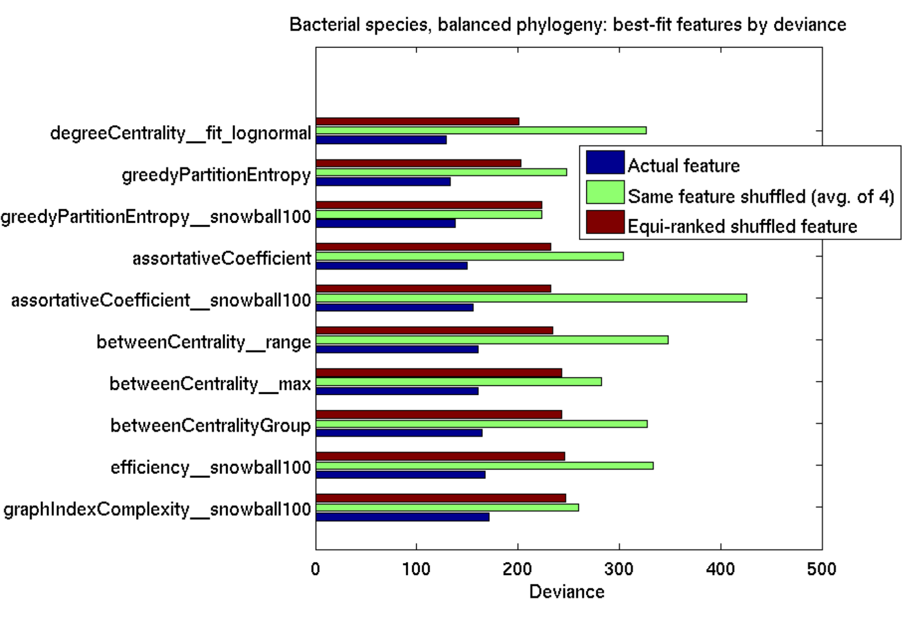
Metabolic Networks: Best-fit features on varying phylogenies

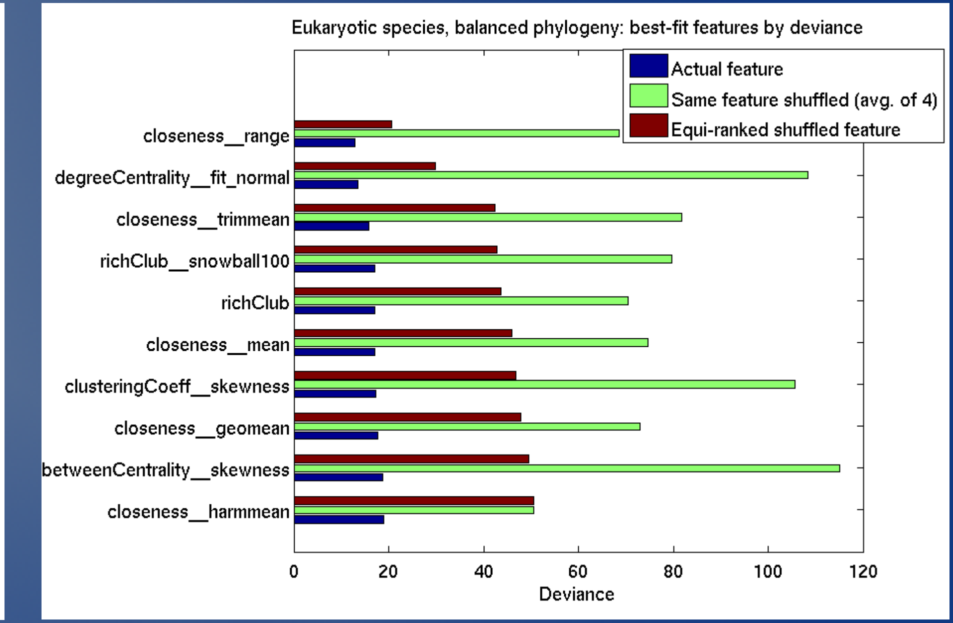
- We compare the quality of fit (deviance) for the best-fit real network features with the best-fit shuffled features
- The difference is significant only on the balanced phylogeny





Signals are weaker on just bacterial or eukaryotic trees





- The best features are still significantly better-fit than shuffled versions, but the difference appears to be most pronounced when bacteria and eukaryotes are both present in the phylogeny

Conclusions and Further Work

- Our approach is an attempt at systematically comparing and categorising a variety ways of measuring network structure and properties, and also looking at robustness and scaling properties of different metrics
- A data-driven approach to examining large numbers of networks and metrics is useful for feature selection in classification tasks, identifying redundant metrics and matching real-world networks to appropriate generative models
- Quantifying the significance of biological network features in the context of evolutionary phylogenies provides one approach towards the problem of establishing relationships between network structure and function
- Our focus in the coming few months will be to carry out specific case studies along these lines to demonstrate the value of the project; ultimately it provides a tool which can give meaningful results only in the context of an appropriately framed scientific question

Acknowledgments

- Charlotte Deane
- Mason Porter
- Dan Fenn
- Ben Fulcher
- Anna Lewis
- Max Little

Systems Biology
Doctoral Training Centre
SYS bio

UNIVERSITY OF
OXFORD

Φxford
hysics

CLARENDON FUND

OCISB

EPSRC
Pioneering research
and skills

BBSRC
bioscience for the future