UNIVERSITY OF OXFORD

Systems Biology Doctoral Training Centre

Oxford Physics

Comparative Network Analysis

Sumeet Agarwal, Gabriel Villar, Nick Jones

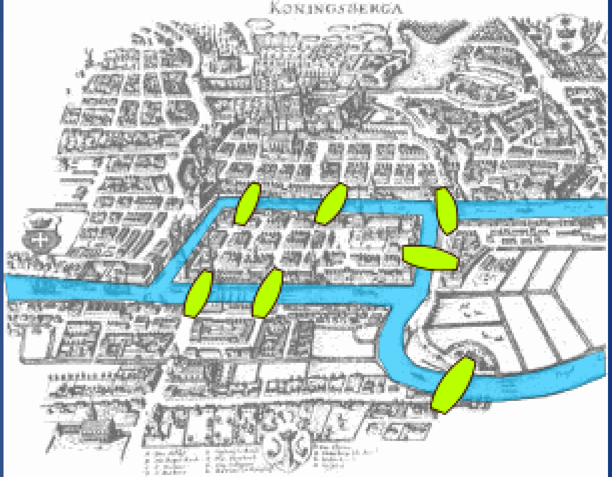
Motivation: Consolidation of network science

- 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...

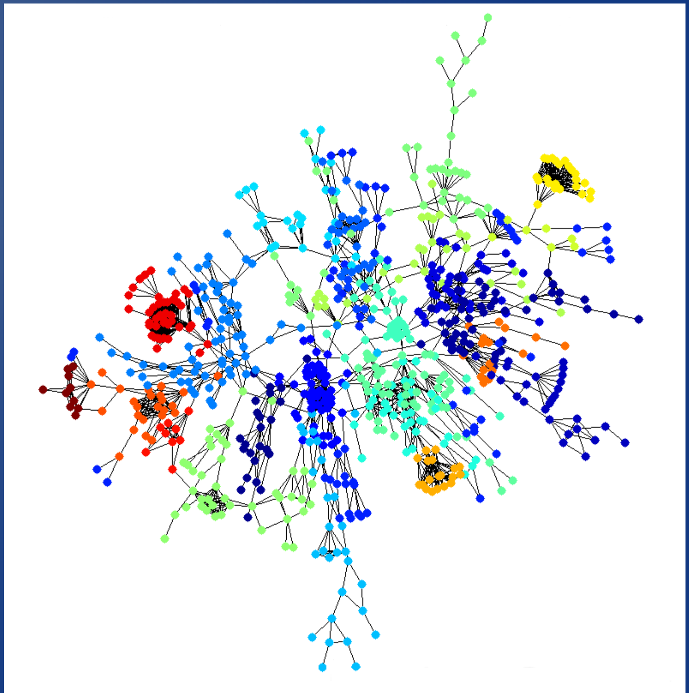
- This has led to a fragmented literature, with inconsistent terminology and frequent reinvention of concepts and methodologies

- 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?

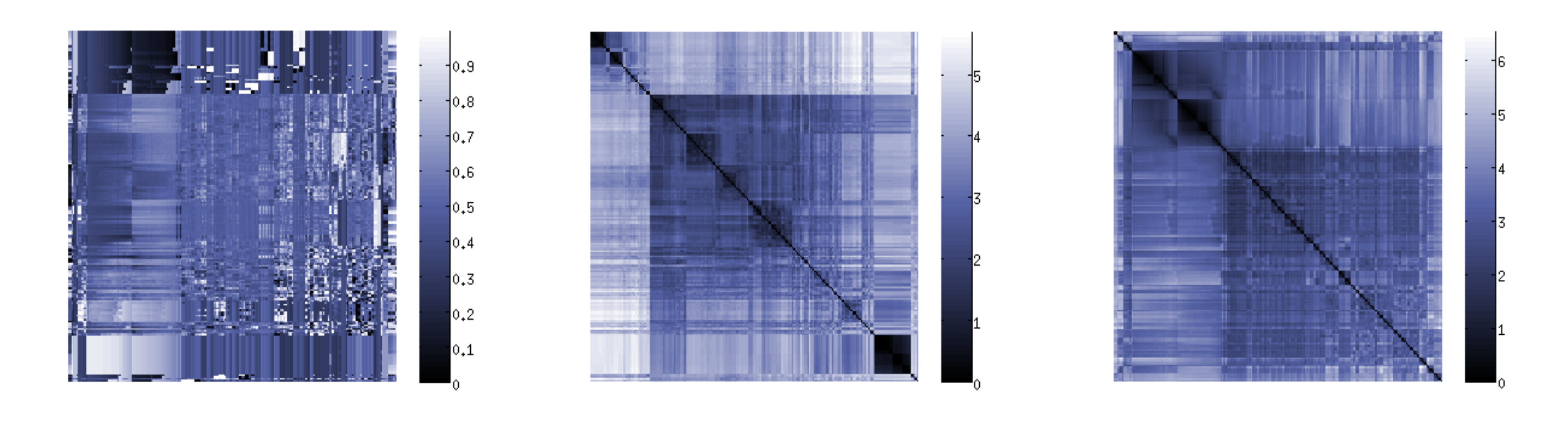


Courtesy: Wikipedia



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

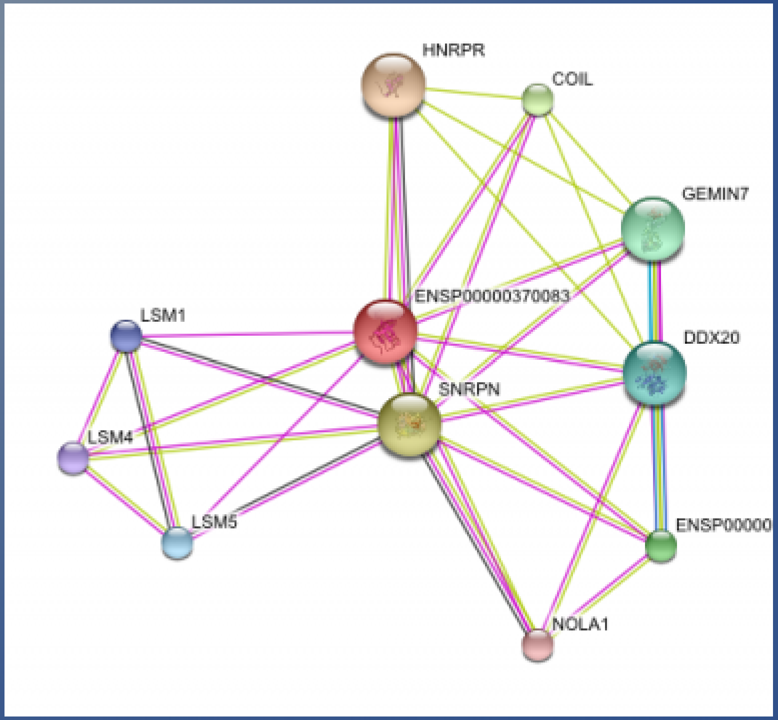
What kinds of networks do we study?

- 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

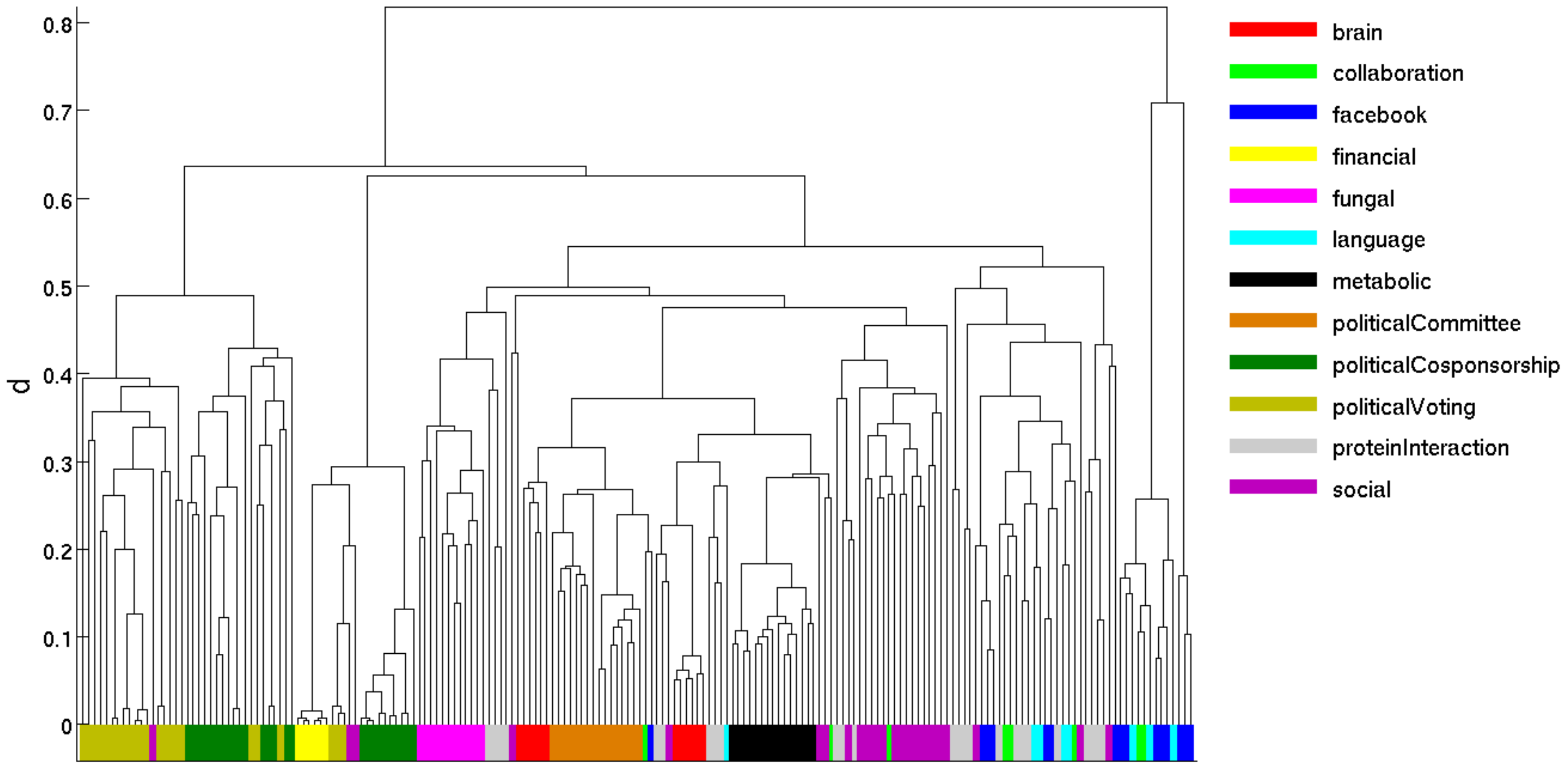


Community structure: partition entropy, modularity, coarse-grained networks

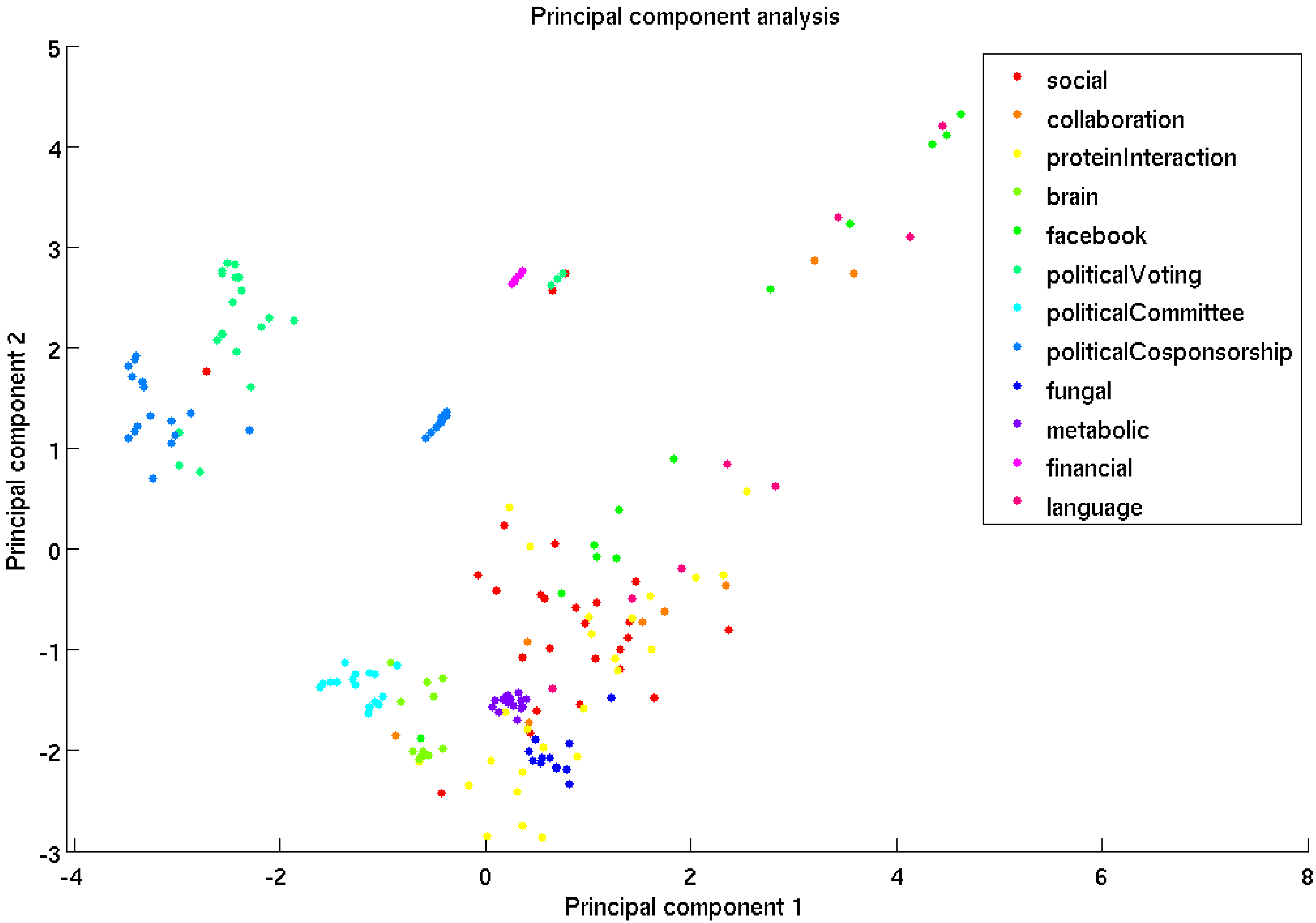
Model fits: how well the network is explained by a certain generative model (preferential attachment, duplication and divergence)

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

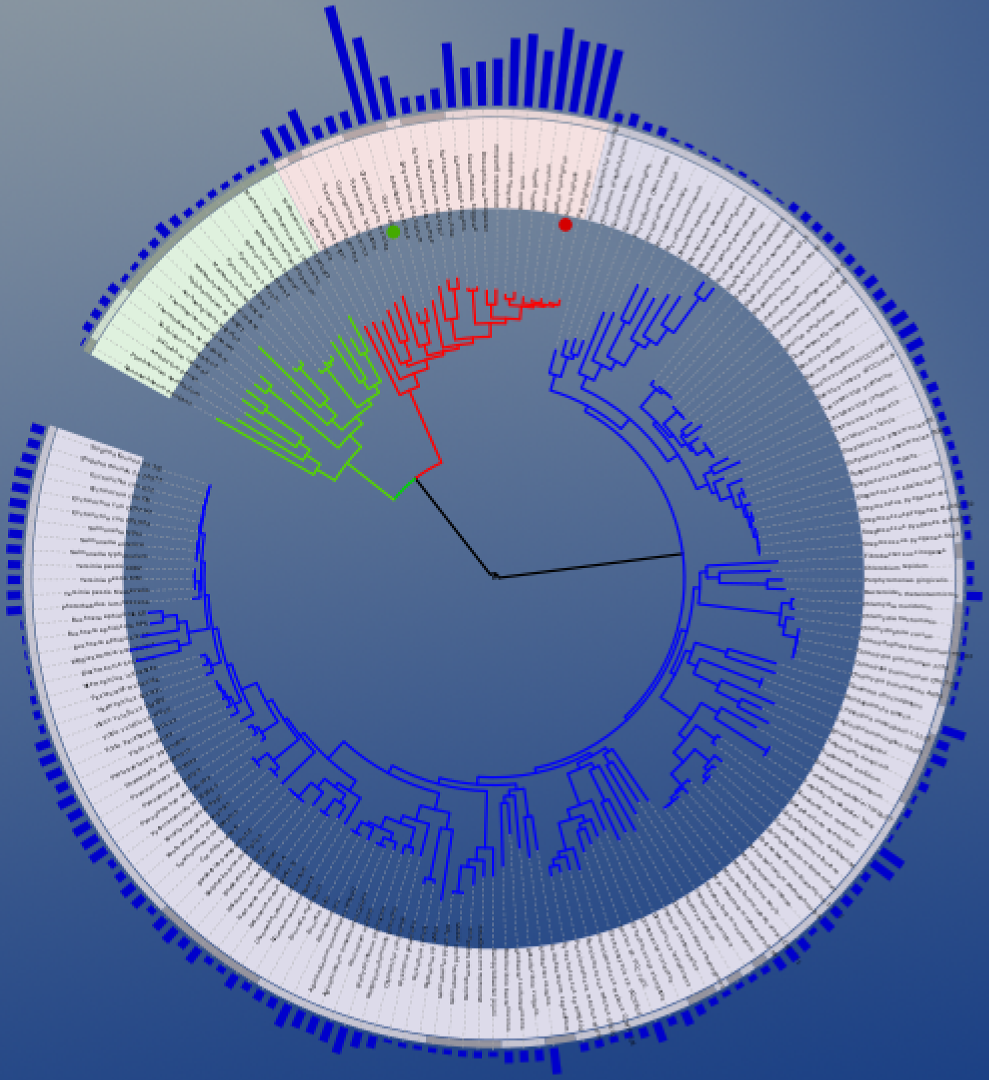
Network Families: Single linkage clustering



Network Families: Principal Component Analysis



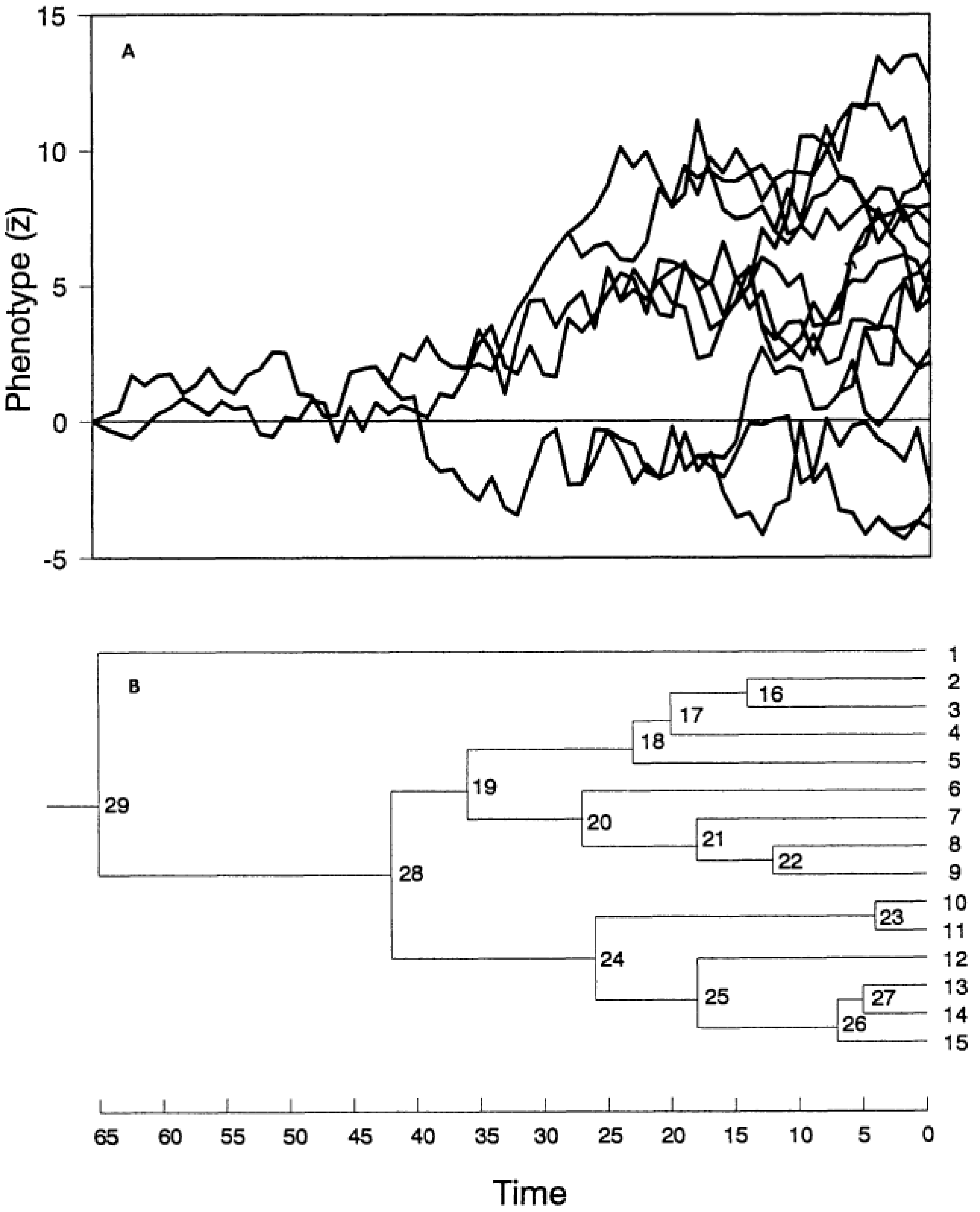
Example: Phylogenetic Comparative Methods



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

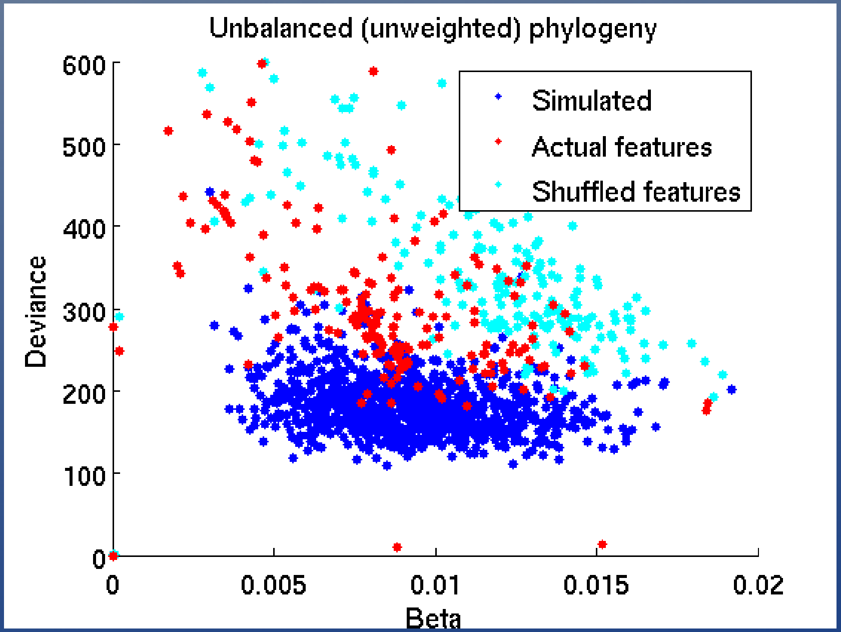
- We attempted to fit a Brownian motion model of evolution (V = βt + ε) to 272 real-valued network metrics computed on 450 metabolic networks from 158 different genuses, using a phylogeny taken from the Tree of Life

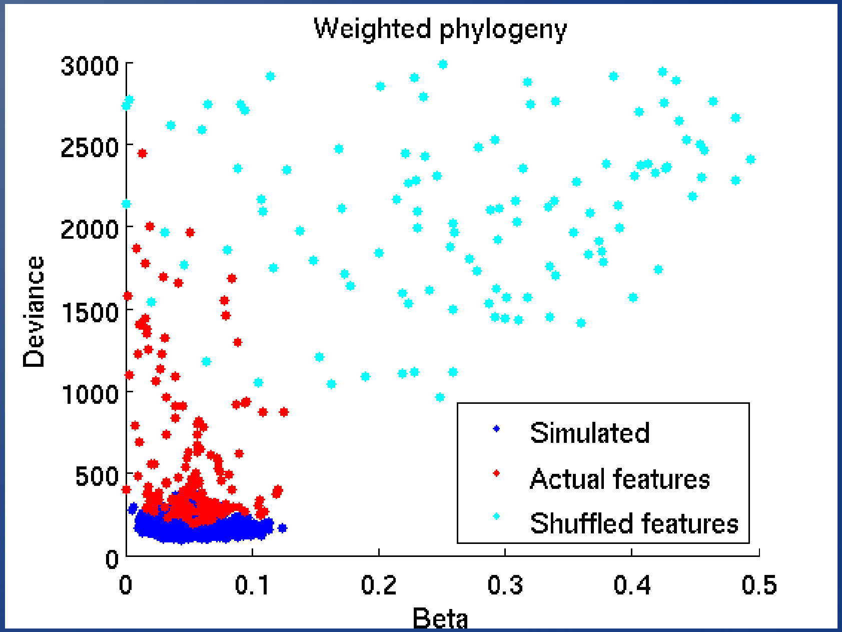


(Emilia P. Martins, *Am. Nat.* 1994)

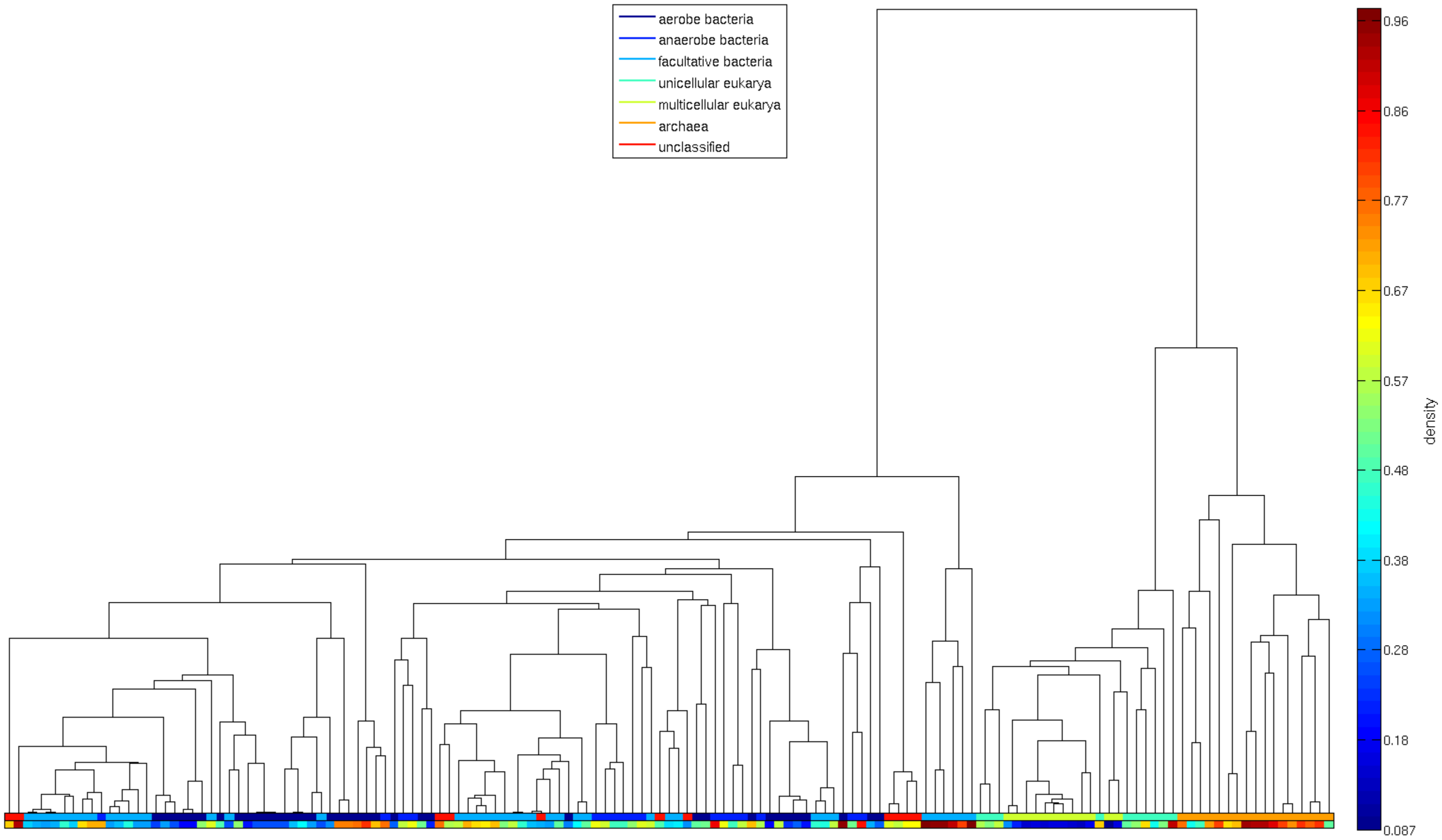
A realistic phylogeny gives significant feature correlations

- An unbalanced version of the tree (with no branch weights) was compared with a weighted version (based on actual estimates of evolution times)
- 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



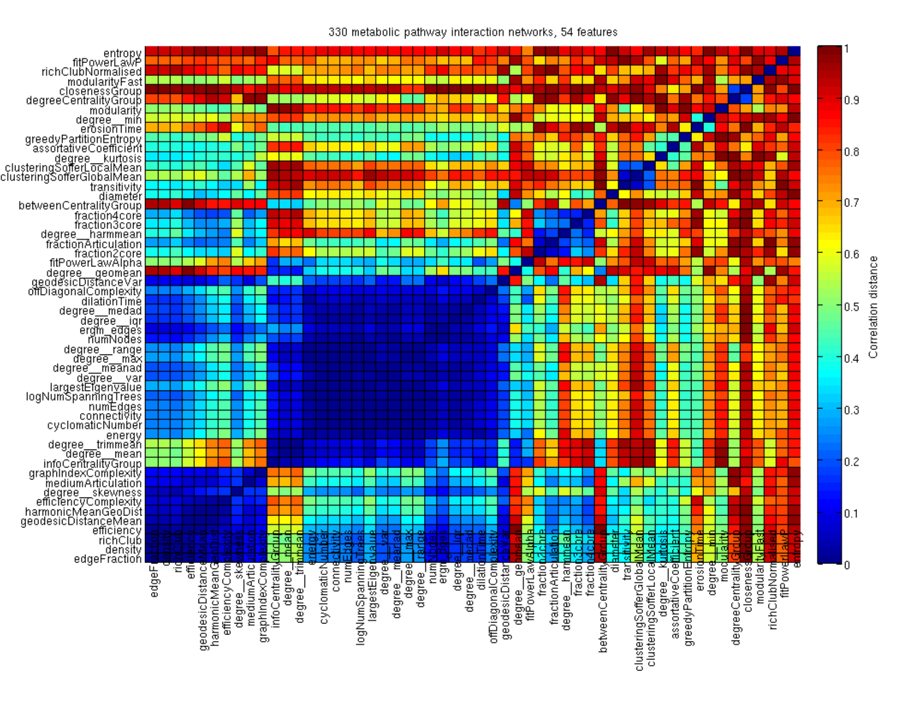


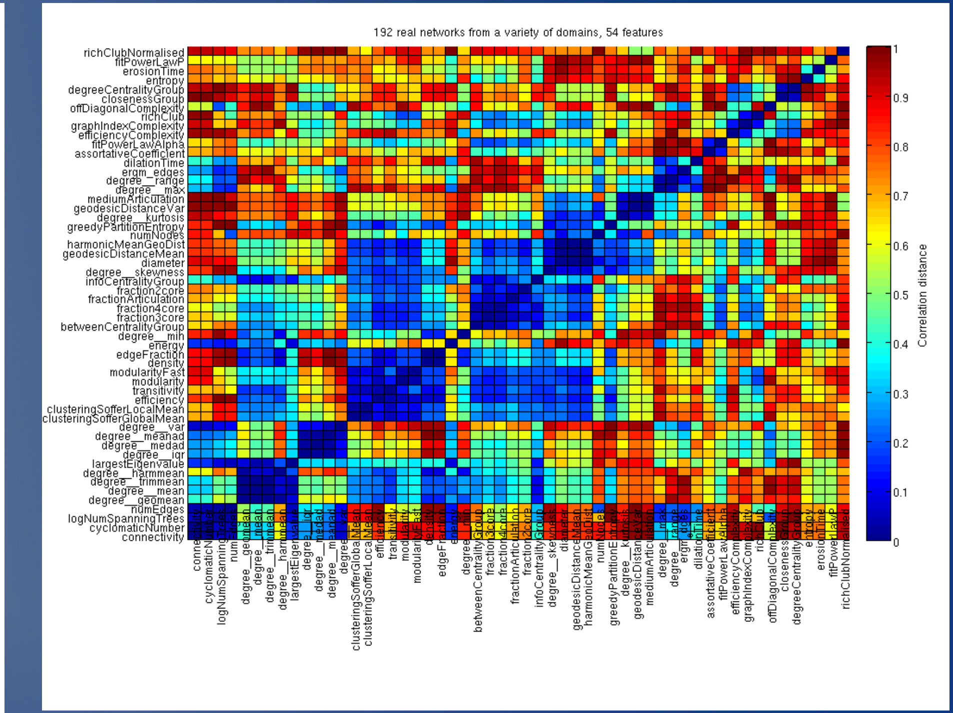
How do networks features vary across the phylogeny?



- Such approaches can be thought of as one way of resolving a debate over the nature of biological taxonomy: pheneticism (Linnæus) vs. cladism (Darwin)

Feature correlations: pointers to 'simplicity' in nature?





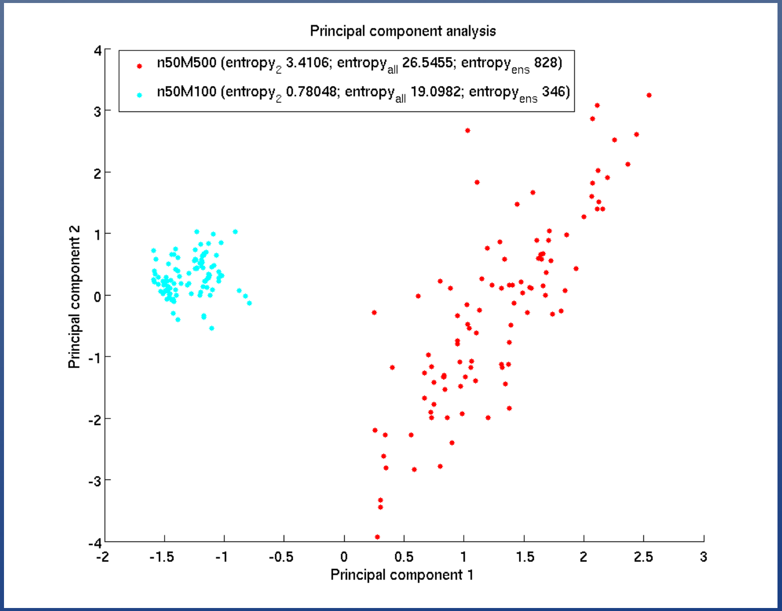
- For restricted classes of networks, many generically different ways of thinking about or characterising networks appear to become degenerate
- Perhaps functional network classes sit on low-dimensional manifolds in the high-dimensional structure space
- One way to think of this is that real-world network categories have relatively low entropy, because they have evolved under entropy-lowering constraints. Can we use such observations to actually recover the underlying generative constraints or mechanisms?

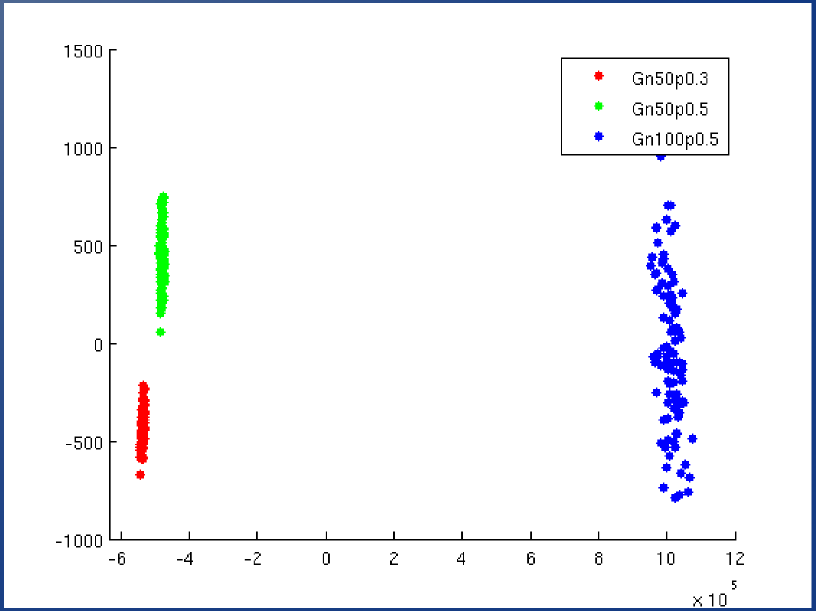
An 'empirical' measure for network entropy?

- We can think of a model or ensemble of networks as specifying a probability distribution over all possible networks; and thus we can define the entropy of this distribution in the standard way. For simple models this can be computed analytically. E.g., for the ensemble G(N,L) (networks with N nodes and L links), the entropy is given by

$$H = -\Sigma\, p_i \log p_i = \log {}^{N(N-1)/2}C_L$$

- Using our method we can also generate a sample from a given ensemble, embed it in a feature space and compute its empirical entropy that way. How do these two measures of entropy match up?



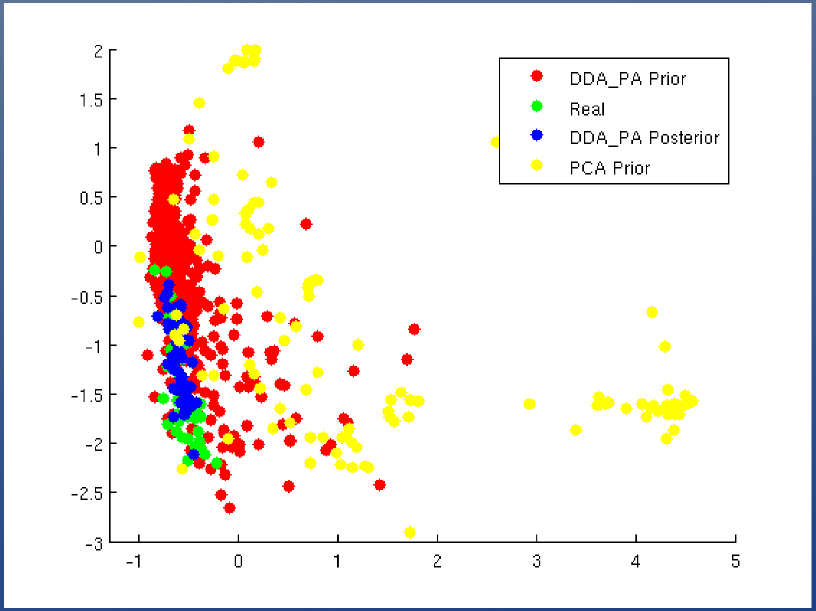


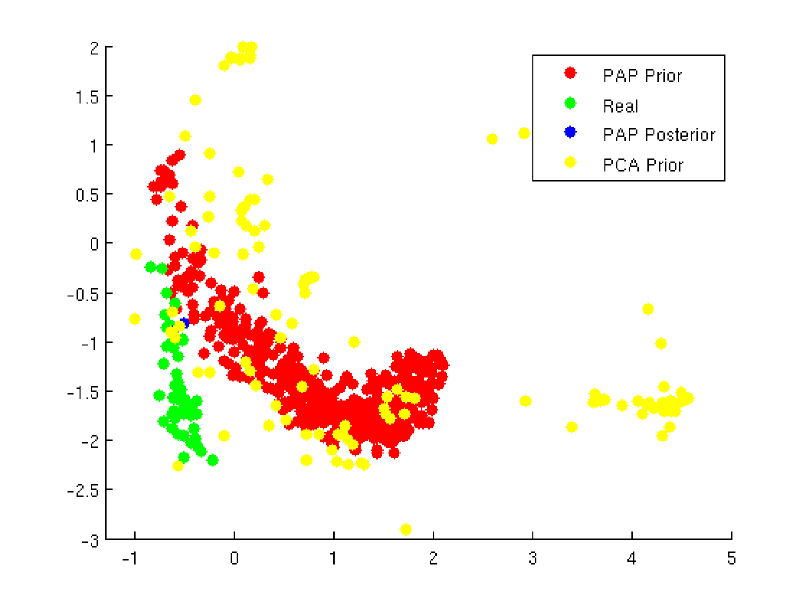
Recovering network models

- The fact that our low-dimensional network embedding allows us to estimate entropy suggests that we could use this for fitting appropriate models to real networks, using the related approach of Approximate Bayesian Computation:

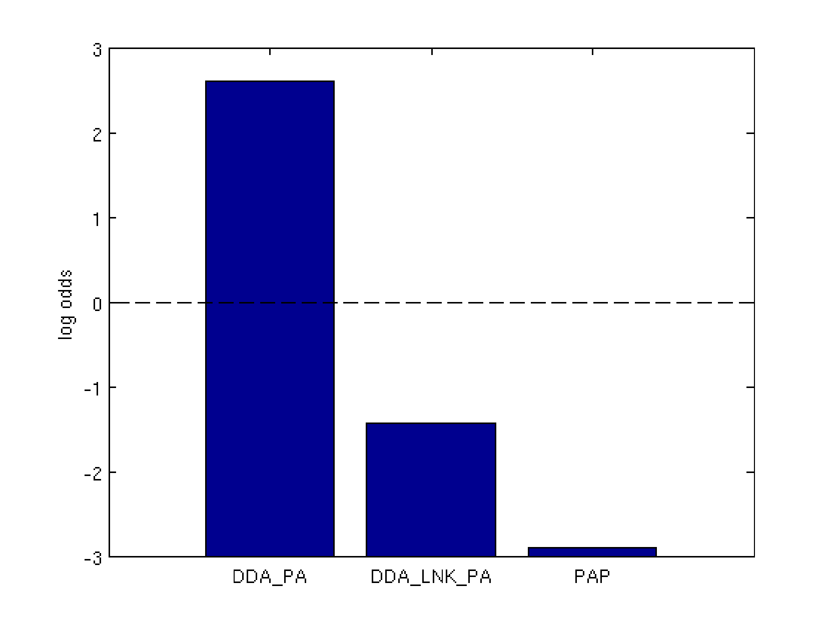
$$P(M|D) \sim P(D|M).P(M)$$

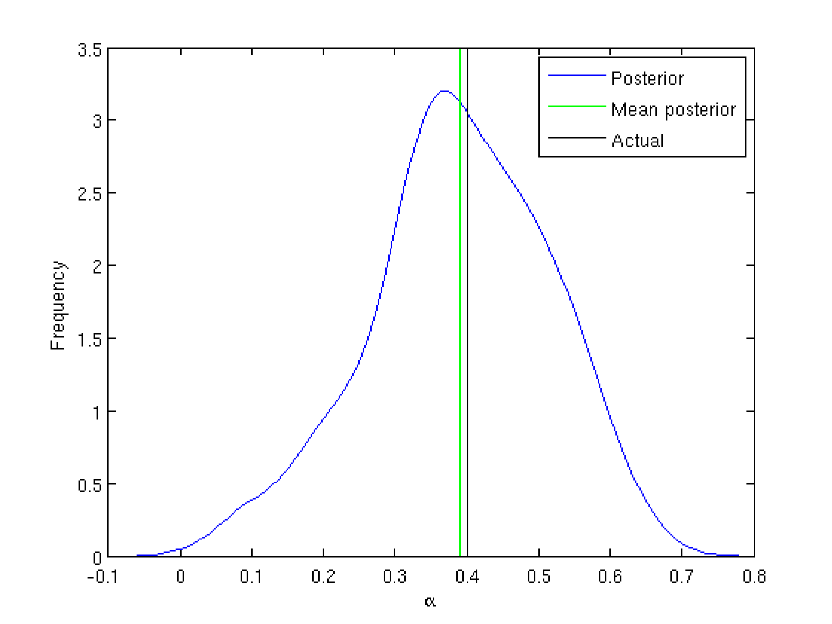
- We have tried generating synthetic networks using a model proposed for the evolution of protein-protein interaction networks, to see how well we can recover the model and its parameters

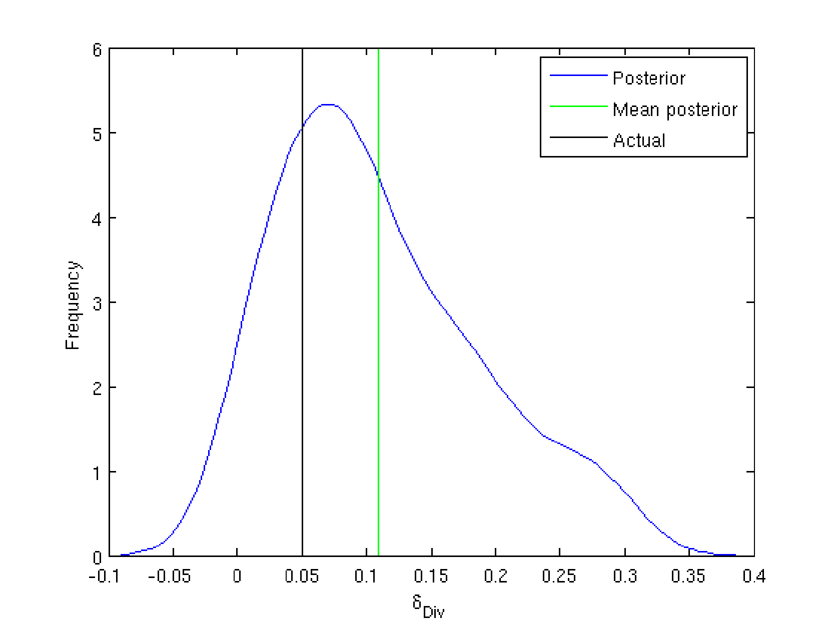


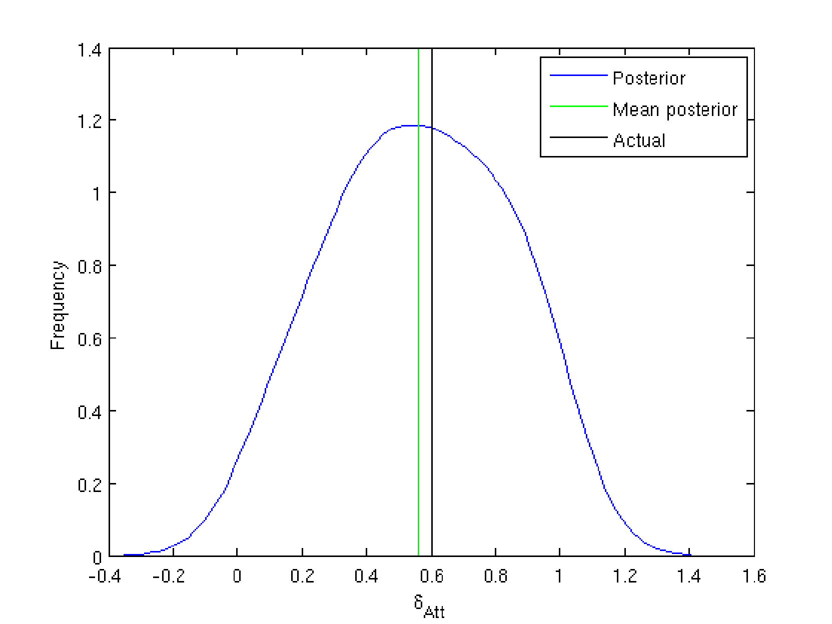


Recovering models with parameters









Conclusions

- 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
- We have demonstrated several different applications of the framework, corresponding to different ways of relating network structure to behaviour/complexity; ultimately it provides a tool which can give meaningful results only in the context of an appropriately framed scientific question

Acknowledgements

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