High Throughput Network Analysis

Sumeet Agarwal[1,2], Gabriel Villar[1,2,3], and Nick S Jones[2,4,5]

[1] Systems Biology Doctoral Training Centre, University of Oxford, Oxford OX1 3QD, United Kingdom
[2] Department of Physics, University of Oxford, Oxford OX1 3PU, United Kingdom
[3] Department of Chemistry, University of Oxford, Oxford OX1 3TA, United Kingdom
[4] Oxford Centre for Integrative Systems Biology, University of Oxford, OX1 3QU, United Kingdom
[5] CABDyN Complexity Centre, University of Oxford, Oxford OX1 1HP, United Kingdom

Introduction

Gene regulatory systems, metabolic pathways, neuronal connections, food webs, social structures and the Internet are all naturally represented as networks; indeed, this may be said of any collection of distinct, interacting entities. Sometimes the value of this mathematical abstraction is clear; for instance, to minimise the spread of an epidemic it may be important to prioritise the immunisation of individuals with high centrality. In many cases, however, one may not know beforehand how a network representation could increase ones understanding of its real-world counterpart.

It may be that abstracting a real-world system as a network discards all of the relevant information, but this seems unlikely for such a high-dimensional representation. Here, we presume that there is some valuable information encoded in the network; the problem is simply to find it. One approach for doing so is to draw a full diagram of the network, since this can, if clearly drawn, contain all of the recorded information. However, an unambiguous diagram is only feasible for very small networks, in which case it is unlikely that the mathematical abstraction will return any surprising results. To learn about a network of any significant size it is therefore necessary to characterise it by summary descriptions, which we will refer to as *metrics*.

A great variety of metrics exist in the literature, but studies that aim to characterise a particular network typically employ a small subset of these, and the choice of metrics is not systematic. Similarly, when a new model for generating synthetic networks is presented, the synthetic networks are compared to real networks in only a few characteristics. This may be justified if one is interested only in the behaviour of a particular metric; but if the goal is to develop synthetic networks that are statistically indistinguishable from real networks, it is important to look at these networks in as many ways as possible. The same is true of exploratory network analysis. Finally, it is typical for a new metric to be introduced with a comparison to only a few existing metrics. The lack of a systematic comparison makes it difficult to tell which metrics give genuinely novel

information about a network, and which pairs of metrics might be redundant or complementary.

Efforts to address this have recently been made [2], but it remains true that there is as yet no systematic program for characterising network structure [7] that can be used to compare the diverse ways in which networks are analysed. We introduce a more systematic framework, in the form of a matrix whose rows correspond to networks, and columns to metrics; we term this the *data matrix*. Each element of the data matrix contains the value of one metric as applied to one network. In this paper we show that this framework enables the systematic comparison of networks and metrics, and demonstrate its utility in the objective selection of metrics for a given purpose; in model fitting; in the analysis of evolving networks; and to determine the robustness of metrics to variations in network size, network damage and sampling effects.

Networks

We collected approximately 1,200 real network data sets. These included several types of biological networks (such as trophic, brain connectivity, protein interaction and metabolic networks), social networks, computer networks and miscellaneous others (including word adjacency and transportation networks). In addition to these real networks, we generated synthetic networks from the Erdős-Rényi, Watts-Strogatz, Barabási-Albert, fitness and graphlet arrival models.

Metrics

This study included approximately 60 base metrics taken from the literature. In order to obtain single numbers from metrics that return distributions (over nodes or links), we generated a number of summary statistics of these distributions, including measures of central tendency and skewness and also likelihoods of certain model fits. Additionally, we include graph clustering or community detection [3, 8] metrics, which return a partition of the network into subnetworks. We then summarise this in a number of ways, such as computing partition entropy and coarse-grained measures on the network of subnetworks.

Selected Results

Given that a large number of metrics exist for describing a network, selecting appropriate subsets for particular tasks is important. Here we demonstrate two applications of feature selection in a supervised learning setting.

First, we consider two sets of networks from a study on metabolic networks [4]. The first set consists of 43 networks that each represent the full cellular network of an organism. The networks in the second set are subsets of the first, including only the metabolic part of each of the 43 networks. We used this classified network data to investigate how metabolic networks differ

from whole-cellular networks. We performed sequential feature selection to optimise the linear discriminability between the metabolic networks from eukaryotes, archaea and bacteria. A 95% classification success rate was obtained by using just three metrics (Figure 1).

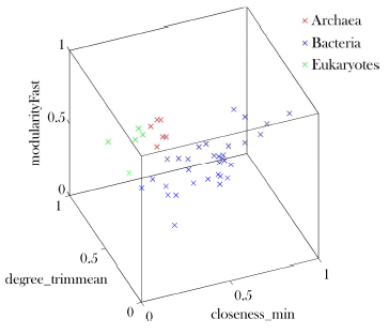
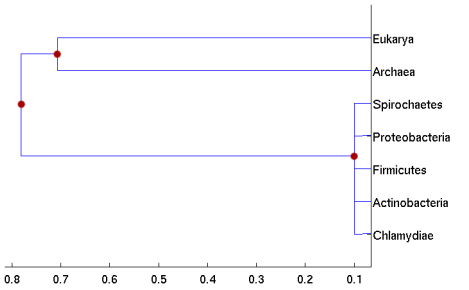


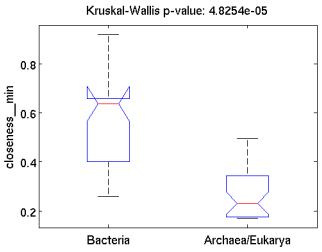
Fig. 1. Classification of metabolic networks of organisms from different kingdoms.

A natural extension of this approach is to look not only at a particular level of species classification, but instead to attempt to take into account the entire structure of evolutionary relationships between species, as represented by a phylogenetic tree. We are currently working on this using ideas from the area of phylogenetic comparative methods [1, 5, 6]: one can assume a certain statistical process (e.g., Brownian motion) underlying the variation in network characteristics along the branches of a phylogeny, and then estimate the extent to which different characteristics are constrained by the phylogenetic structure. As a rough preliminary step towards this, we have taken the 43 metabolic networks referred to above and grouped them at the leaves of a highly simplified phylogeny (Figure 2(a)). We represent each network by its feature vector of metrics, and then carry out feature selection based on information gain at each of the branching points in the phylogeny. Figure 2 shows that features based on *closeness*, a measure of node centrality, are found to be amongst the most informative ones at each of the 3 branching points. This suggests that this metric is capturing some biologically relevant network property, and it should be of interest to study this in greater detail using the approach described above.

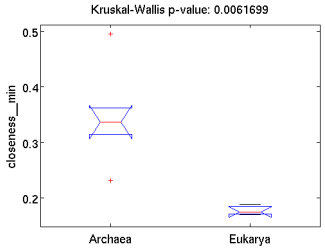
As an example to demonstrate unsupervised learning on more varied data, we took a set of 192 networks from a wide range of disciplines and carried out principal component analysis (PCA), utilising a set of 433 metrics. The results are shown in Figure 3, with each data point representing a network's position along the two largest principal components and different colours depicting the different domains from which the networks are drawn. We see that certain kinds of networks fall into very cohesive groupings, such as financial, fungal and metabolic networks. On the other hand, some types of networks such as protein inter-



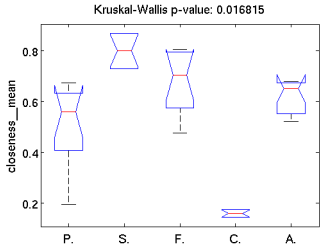
(a) Phylogenetic tree; branching points in red



(b) Boxplots for *closeness_minimum*; Bacteria vs. Archaea/Eukarya



(c) Boxplots for *closeness_minimum*; Archaea vs. Eukarya



(d) Boxplots for *closeness_mean*; for the 5 Bacterial phyla

Fig. 2. 43 metabolic networks [4] are grouped according to a simplified phylogeny **(a)**. Network features representing the closeness distribution of nodes are found to be significantly different in their distributions on either side of the 3 branching points **(b,c,d)**.

action, collaboration and social networks are much less well separated. We also attempted building a supervised classification tree for this set of networks, which resulted in a 10-fold cross-validation accuracy of nearly 80% and made use of only about 15 of the 433 features.

Discussion

In some ways, the approach taken here is complementary to standard perspectives in network science. When a new metric is introduced in the networks lit-

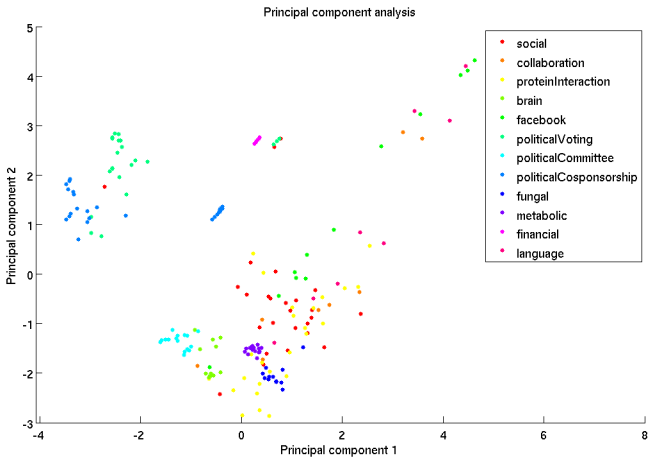


Fig. 3. Results of PCA on a set of 192 networks, using 433 features. The two largest principal components are shown.

erature, it may be motivated by an expectation of what aspects of a network it will capture, or by some distinguishing feature of its calculation. Similarly, new network models are assessed by how closely they match certain particular metrics. Here, we simply apply all of the available metrics to a set of networks, and use the resulting data structure to explore the networks or metrics in an unprejudiced manner. This framework as a way of systematically comparing metrics should be valuable for both explorative network analysis, and for finding the best way to answer a particular question in a data-driven manner. It continues to be work in progress, but we hope that once complete, public distribution of the software and database built for this project will benefit users and see new applications of the framework.

References

1. Felsenstein, J.: Phylogenies and the comparative method. The American Naturalist 125(1), 1–15 (January 1985), http://dx.doi.org/10.1086/284325
2. Filkov, V., Saul, Z.M., Roy, S., D'Souza, R.M., Devanbu, P.T.: Modeling and verifying a broad array of network properties. EPL (Europhysics Letters) 86(2), 28003 (April 2009), http://dx.doi.org/10.1209/0295-5075/86/28003
3. Fortunato, S.: Community detection in graphs. Physics Reports 486(3-5), 75–174 (2010), http://www.sciencedirect.com/science/article/B6TVP-4XPYXF1-1/2/99061fac6435db4343b2374d26e64ac1
4. Jeong, H., Tombor, B., Albert, R., Oltvai, Z.N., Barabasi, A.L.: The large-scale organization of metabolic networks. Nature 407(6804), 651–654 (October 2000), http://dx.doi.org/10.1038/35036627

5. Macholán, M.: The mouse skull as a source of morphometric data for phylogeny inference. Zoologischer Anzeiger 247(4), 315–327 (October 2008), http://dx.doi.org/10.1016/j.jcz.2008.06.001
6. Martins, E.P.: Estimating the rate of phenotypic evolution from comparative data. The American Naturalist 144(2), 193–209 (August 1994), http://dx.doi.org/10.1086/285670
7. Newman, M.E.J.: The structure and function of complex networks. SIAM Review 45(2), 167–256 (2003), http://link.aip.org/link/?SIR/45/167/1
8. Porter, M.A., Onnela J-P, Mucha, P.J.: Communities in networks. Notices of the American Mathematical Society 56(9), 1082–1097, 1164–1166 (September 2009), http://arxiv.org/abs/0902.3788