High Throughput Network Analysis

Gabriel Villar[1,2,3], Sumeet Agarwal[1,2], 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

1 Introduction

Gene regulatory systems, metabolic pathways, neuronal connections, food webs, social structures and the Internet are all naturally represented as networks. However, it is not always clear how such a representation aids the understanding of its real-world counterpart. It may be that abstracting a complex 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. To learn about networks of any significant size it is generally necessary to characterise them 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. It is typical for a new metric to be introduced with comparison to only a few existing ones. The lack of a systematic comparison makes it difficult to know which metrics capture some genuinely novel aspect of network structure, and which might be redundant. Efforts to address this have recently been made [3], but it remains true that there is as yet no systematic program for characterising network structure [8] that can be used to compare the diverse ways in which networks are analysed. We attempt to introduce such a framework, in the form of a matrix whose rows correspond to networks, and columns to metrics; we term this the *data matrix*. Here we seek to demonstrate some applications of this approach to network classification, finding redundant metrics, model-fitting, using synthetic networks to contextualise and motivate generative models for real-world networks, studying evolving networks, relating network features to evolutionary phylogenies, and determining the robustness of metrics to network damage and sampling effects.

2 Materials and Methods

2.1 Networks

We collected about 1,200 real networks. These included several types of biological networks, social networks, computer networks and miscellaneous others. The networks ranged in size from a few tens of nodes upto tens of thousands of nodes.

2.2 Metrics

We included about 60 base metrics taken from the literature. Whilst some metrics, such as the diameter, are scalars, others return a vector: for example, the degree distribution returns a number for each degree in the network. In order add such vector metrics to our data matrix, we generated a number of summary statistics of these distributions. Additionally, we include graph clustering or community detection [4, 9] metrics: these return a partition of the network into subnetworks, which can also be summarised in various ways. A fourth kind of metric is goodness-of-fit of some simple generative models: e.g., how well the network is explained by a preferential attachment process [1,10]. In sum, the combination of scalar metrics and summary statistics gives us about 400 features in our data matrix.

3 Selected Results

To demonstrate the utility of this approach for feature selection, we are looking at the problem of detecting phylogenetic signal in biological networks. Rather than just grouping data points into independent classes, we would like to take into account the entire structure of evolutionary relationships between species, which may be represented by means of a phylogenetic tree. Given such a tree, the objective is to find features of network structure that co-vary with the phylogeny. We are currently working on this using ideas from the area of phylogenetic comparative methods [2, 6, 7]. As a rough preliminary step, we obtained a set of 43 metabolic networks [5] and grouped them at the leaves of a highly simplified phylogeny. Each network was represented by its feature vector, and features were ranked based on information gain at each of the three branching points in our phylogeny. We found that features based on *closeness*, a measure of node centrality, were amongst the most informative ones at all of the branching points. This suggests that closeness may be a biologically relevant network property, and it should be of interest to study this in greater detail.

As an example to show how unsupervised learning can be used for network characterisation, 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 1. We see that even in this 2-dimensional mapping, certain kinds of networks seem to fall into very cohesive groupings, suggesting that our feature vectors are capturing functionally important properties of those systems.

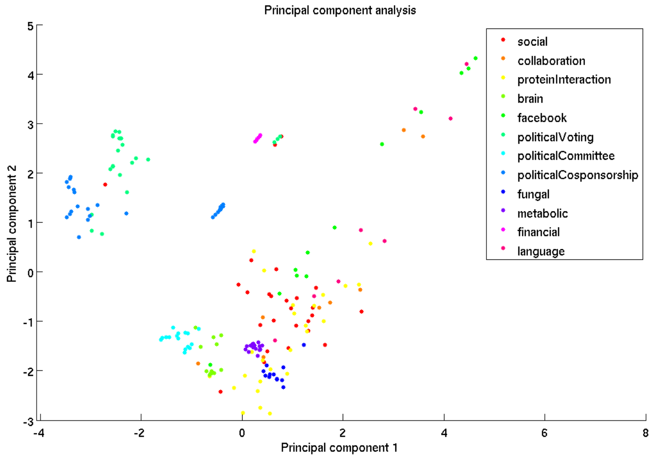

Fig. 1. Results of PCA on a set of 192 networks, using 433 features. The two largest principal components are shown.

4 Discussion

In some ways, the approach taken here is complementary to standard perspectives in network science. When a new metric is introduced in the literature, it may be motivated by what aspects of a network it is expected to capture, or by some distinguishing feature of its calculation. Similarly, new network models are assessed by how closely they match particular metrics. Here, we simply apply all of the available metrics to a set of networks, and use the results to explore the networks or metrics in an unprejudiced manner. This framework as a way of systematically comparing metrics should be valuable for both exploratory network analysis, and for finding the best way to answer a particular question in a data-driven manner. Our early results, based on relatively simple models and metrics, suggest that our framework may serve as a powerful tool for a variety of feature selection and network characterisation tasks. It continues to be work in progress, but we hope that in due course, public distribution of the software and database built for this project will benefit the community and see new applications of the framework.

References

1. Barabási, A.L., Albert, R.: Emergence of scaling in random networks. Science 286(5439), 509–512 (October 1999), http://dx.doi.org/10.1126/science.286.5439.509
2. Felsenstein, J.: Phylogenies and the comparative method. The American Naturalist 125(1), 1–15 (January 1985), http://dx.doi.org/10.1086/284325
3. 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
4. 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
5. 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
6. 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
7. 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
8. 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
9. 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
10. Yule, G.U.: A mathematical theory of evolution, based on the conclusions of Dr. J. C. Willis, F.R.S. Philosophical Transactions of the Royal Society of London, Ser. B 213, 21–87 (1925)