LETTERS

Hierarchical structure and the prediction of missing links in networks

Aaron Clauset[1,3], Cristopher Moore[1,2,3] & M. E. J. Newman[3,4]

Networks have in recent years emerged as an invaluable tool for describing and quantifying complex systems in many branches of science[1–3]. Recent studies suggest that networks often exhibit hierarchical organization, in which vertices divide into groups that further subdivide into groups of groups, and so forth over multiple scales. In many cases the groups are found to correspond to known functional units, such as ecological niches in food webs, modules in biochemical networks (protein interaction networks, metabolic networks or genetic regulatory networks) or communities in social networks[4–7]. Here we present a general technique for inferring hierarchical structure from network data and show that the existence of hierarchy can simultaneously explain and quantitatively reproduce many commonly observed topological properties of networks, such as right-skewed degree distributions, high clustering coefficients and short path lengths. We further show that knowledge of hierarchical structure can be used to predict missing connections in partly known networks with high accuracy, and for more general network structures than competing techniques[8]. Taken together, our results suggest that hierarchy is a central organizing principle of complex networks, capable of offering insight into many network phenomena.

Much recent work has been devoted to the study of clustering and community structure in networks[5,6,9–11]. Hierarchical structure goes beyond simple clustering, however, by explicitly including organization at all scales in a network simultaneously. Conventionally, hierarchical structure is represented by a tree, or dendrogram, in which closely related pairs of vertices have lowest common ancestors that are lower in the tree than those of more distantly related pairs (see Fig. 1). We expect the probability of a connection between two vertices to depend on their degree of relatedness. Structure of this type can be modelled mathematically by using a probabilistic approach in which we endow each internal node r of the dendrogram with a probability p_r and then connect each pair of vertices for which r is the lowest common ancestor independently with probability p_r (Fig. 1).

This model, which we call a hierarchical random graph, is similar in spirit to (although different in realization from) the tree-based models used in some studies of network search and navigation[12,13]. Like most work on community structure, it assumes that communities at each level of organization are disjoint. Overlapping communities have occasionally been studied (see, for example, ref. 14) and could be represented with a more elaborate probabilistic model; however, as we discuss below, the present model already captures many of the structural features of interest.

Given a dendrogram and a set of probabilities p_r, the hierarchical random graph model allows us to generate artificial networks with a specified hierarchical structure, a procedure that might be useful in certain situations. Our goal here, however, is a different one. We wish to detect and analyse the hierarchical structure, if any, of networks in the real world. We accomplish this by fitting the hierarchical model to observed network data by using the tools of statistical inference, combining a maximum-likelihood approach[15] with a Monte Carlo sampling algorithm[16] on the space of all possible dendrograms. This technique allows us to sample hierarchical random graphs with probability proportional to the likelihood that they generate the observed network. To obtain the results described below we combine information from a large number of such samples, each of which is a reasonably likely model of the data.

The success of this approach relies on the flexible nature of our hierarchical model, which allows us to fit a wide range of network structures. The traditional picture of communities or modules in a network, for example, corresponds to connections that are dense within groups of vertices and sparse between them—a behaviour called 'assortativity' in the literature[17]. The hierarchical random

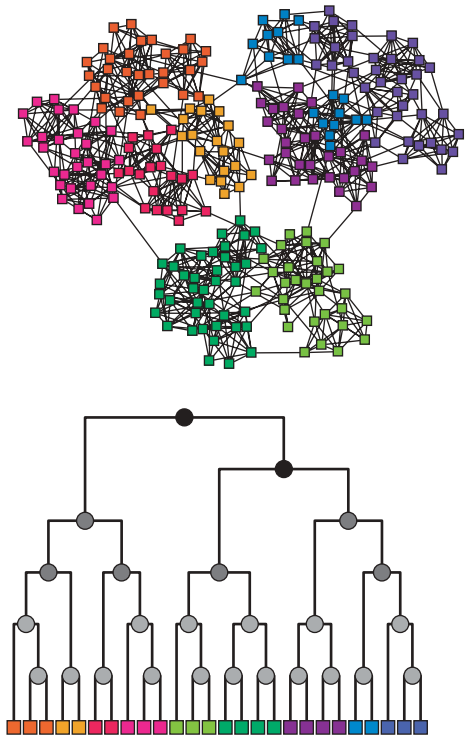

Figure 1 | A hierarchical network with structure on many scales, and the corresponding hierarchical random graph. Each internal node r of the dendrogram is associated with a probability p_r that a pair of vertices in the left and right subtrees of that node are connected. (The shades of the internal nodes in the figure represent the probabilities.)

[1]Department of Computer Science, and [2]Department of Physics and Astronomy, University of New Mexico, Albuquerque, New Mexico 87131, USA. [3]Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, New Mexico 87501, USA. [4]Department of Physics and Center for the Study of Complex Systems, University of Michigan, Ann Arbor, Michigan 48109, USA.

graph can capture behaviour of this kind using probabilities p_r that decrease as we move higher up the tree. Conversely, probabilities that increase as we move up the tree correspond to 'disassortative' structures in which vertices are less likely to be connected on small scales than on large ones. By letting the p_r values vary arbitrarily throughout the dendrogram, the hierarchical random graph can capture both assortative and disassortative structure, as well as arbitrary mixtures of the two, at all scales and in all parts of the network.

To demonstrate our method we have used it to construct hierarchical decompositions of three example networks drawn from disparate fields: the metabolic network of the spirochaete *Treponema pallidum*[18], a network of associations between terrorists[19], and a food web of grassland species[20]. To test whether these decompositions accurately capture the important structural features of the networks, we use the sampled dendrograms to generate new networks, different in detail from the originals but, by definition, having similar hierarchical structure (see Supplementary Information for more details). We find that these 'resampled' networks match the statistical properties of the originals closely, including their degree distributions, clustering coefficients, and distributions of shortest path lengths between pairs of vertices, despite the fact that none of these properties is explicitly represented in the hierarchical random graph (Table 1, and Supplementary Fig. 3). It therefore seems that a network's hierarchical structure is capable of explaining a wide variety of other network features as well.

The dendrograms produced by our method are also of interest in themselves, as a graphical representation and summary of the hierarchical structure of the observed network. As discussed above, our method can generate not just a single dendrogram but a set of dendrograms, each of which is a good fit to the data. From this set we can, by using techniques from phylogeny reconstruction[21], create a single consensus dendrogram, which captures the topological features that appear consistently across all or a large fraction of the dendrograms and typically is a better summary of the network's structure than any individual dendrogram. Figure 2a shows such a consensus dendrogram for the grassland species network, which clearly reveals communities and subcommunities of plants, herbivores, parasitoids and hyperparasitoids.

Another application of the hierarchical decomposition is the prediction of missing interactions in networks. In many settings, the discovery of interactions in a network requires significant experimental effort in the laboratory or the field. As a result, our current pictures of many networks are substantially incomplete[22–28]. An alternative to checking exhaustively for a connection between every pair of vertices in a network is to try to predict, in advance and on the basis of the connections already observed, which vertices are most likely to be connected, so that scarce experimental resources can be focused on testing for those interactions. If our predictions are good, we can in this way substantially reduce the effort required to establish the network's topology.

The hierarchical decomposition can be used as the basis for an effective method of predicting missing interactions as follows. Given an observed but incomplete network, we generate, as described above, a set of hierarchical random graphs—dendrograms and the associated probabilities p_r—that fit that network. Then we look for pairs of vertices that have a high average probability of connection within these hierarchical random graphs but are unconnected in the observed network. These pairs we consider the most likely candidates for missing connections. (Technical details of the procedure are given in Supplementary Information.)

We demonstrate the method by using our three example networks again. For each network we remove a subset of connections chosen uniformly at random and then attempt to predict, on the basis of the remaining connections, which have been removed. A standard metric for quantifying the accuracy of prediction algorithms, commonly used in the medical and machine learning communities, is the AUC statistic, which is equivalent to the area under the receiver operating characteristic (ROC) curve[29]. In the present context, the AUC statistic can be interpreted as the probability that a randomly chosen missing connection (a true positive) is given a higher score by our method than a randomly chosen pair of unconnected vertices (a true negative). Thus, the degree to which the AUC exceeds 0.5 indicates how much better our predictions are than chance. Figure 2 shows the AUC statistic for the three networks as a function of the fraction of the connections known to the algorithm. For all three networks our algorithm does far better than chance, indicating that hierarchy is a strong general predictor of missing structure. It is also instructive to compare the performance of our method with that of other methods for link prediction[8]. Previously proposed methods include assuming that vertices are likely to be connected if they have many common neighbours, if there are short paths between them, or if the product of their degrees is large. These approaches work well for strongly assortative networks such as collaboration and citation

Table 1 | Comparison of original and resampled networks

Network	$\langle k\rangle_{real}$	$\langle k\rangle_{samp}$	C_{real}	C_{samp}	d_{real}	d_{samp}
T. pallidum	4.8	3.7(1)	0.0625	0.0444(2)	3.690	3.940(6)
Terrorists	4.9	5.1(2)	0.361	0.352(1)	2.575	2.794(7)
Grassland	3.0	2.9(1)	0.174	0.168(1)	3.29	3.69(2)

Statistics are shown for the three example networks studied and for new networks generated by resampling from our hierarchical model. The generated networks closely match the average degree $\langle k\rangle$, clustering coefficient C and average vertex–vertex distance d in each case, suggesting that they capture much of the structure of the real networks. Parenthetical values indicate standard errors on the final digits.

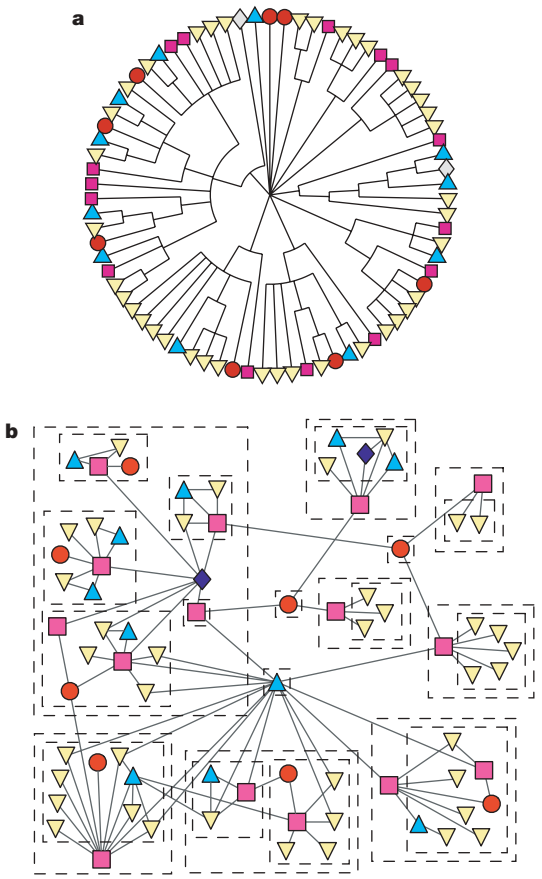

Figure 2 | Application of the hierarchical decomposition to the network of grassland species interactions. **a**, Consensus dendrogram reconstructed from the sampled hierarchical models. **b**, A visualization of the network in which the upper few levels of the consensus dendrogram are shown as boxes around species (plants, herbivores, parasitoids, hyperparasitoids and hyper-hyperparasitoids are shown as circles, boxes, down triangles, up triangles and diamonds, respectively). Note that in several cases a set of parasitoids is grouped into a disassortative community by the algorithm, not because they prey on each other but because they prey on the same herbivore.

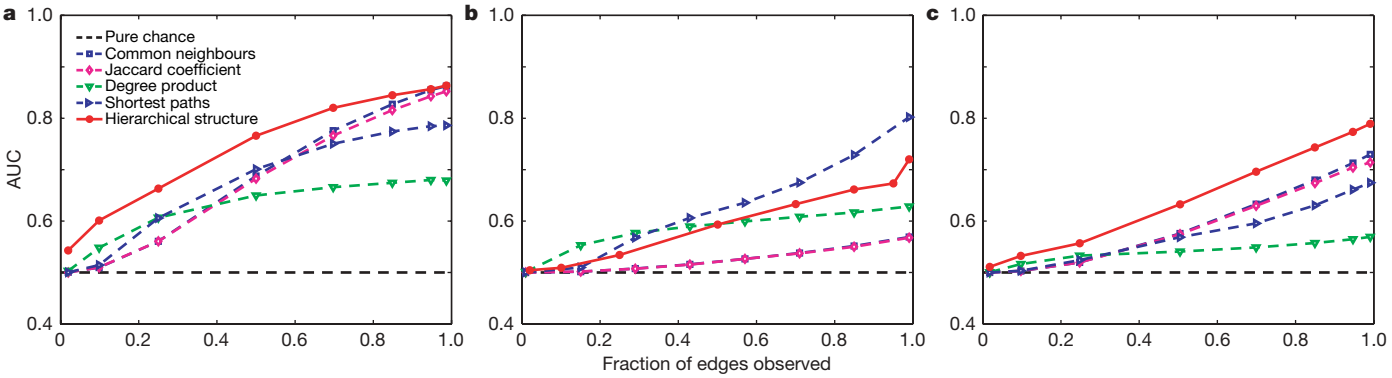

Figure 3 | Comparison of link prediction methods. Average AUC statistic—that is, the probability of ranking a true positive over a true negative—as a function of the fraction of connections known to the algorithm, for the link prediction method presented here and a variety of previously published methods. **a**, Terrorist association network; **b**, *T. pallidum* metabolic network; **c**, grassland species network.

networks[8] and for the metabolic and terrorist networks studied here (Fig. 3a, b). Indeed, for the metabolic network the shortest-path heuristic performs better than our algorithm.

However, these simple methods can be misleading for networks that exhibit more general types of structure. In food webs, for instance, pairs of predators often share prey species but rarely prey on each other. In such situations a common-neighbour or shortest-path-based method would predict connections between predators where none exists. The hierarchical model, by contrast, is capable of expressing both assortative and disassortative structure and, as Fig. 3c shows, gives substantially better predictions for the grassland network. (Indeed, in Fig. 2b there are several groups of parasitoids that our algorithm has grouped together in a disassortative community, in which they prey on the same herbivore but not on each other.) The hierarchical method thus makes accurate predictions for a wider range of network structures than the previous methods.

In the applications above, we have assumed for simplicity that there are no false positives in our network data; that is, that every observed edge corresponds to a real interaction. In networks in which false positives may be present, however, they too could be predicted by using the same approach: we would simply look for pairs of vertices that have a low average probability of connection within the hierarchical random graph but are connected in the observed network.

The method described here could also be extended to incorporate domain-specific information, such as species' morphological or behavioural traits for food webs[28] or phylogenetic or binding-domain data for biochemical networks[23], by adjusting the probabilities of edges accordingly. As the results above show, however, we can obtain good predictions even in the absence of such information, indicating that topology alone can provide rich insights.

In closing, we note that our approach differs crucially from previous work on hierarchical structure in networks[1,4–7,9,11,30] in that it acknowledges explicitly that most real-world networks have many plausible hierarchical representations of roughly equal likelihood. Previous work, by contrast, has typically sought a single hierarchical representation for a given network. By sampling an ensemble of dendrograms, our approach avoids over-fitting the data and allows us to explain many common topological features, to generate resampled networks with similar structure to the original, to derive a clear and concise summary of a network's structure by means of its consensus dendrogram, and to accurately predict missing connections in a wide variety of situations.

METHODS SUMMARY

Computer code implementing many of the analysis methods described in this paper can be found online at http://www.santafe.edu/~aaronc/randomgraphs/.

Received 13 August 2007; accepted 7 February 2008.

1. Wasserman, S. & Faust, K. *Social Network Analysis* (Cambridge Univ. Press, Cambridge, 1994).
2. Albert, R. & Barabási, A.-L. Statistical mechanics of complex networks. *Rev. Mod. Phys.* **74,** 47–97 (2002).
3. Newman, M. E. J. The structure and function of complex networks. *SIAM Rev.* **45,** 167–256 (2003).
4. Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai, Z. N. & Barabási, A.-L. Hierarchical organization of modularity in metabolic networks. *Science* **30,** 1551–1555 (2002).
5. Clauset, A., Newman, M. E. J. & Moore, C. Finding community structure in very large networks. *Phys. Rev. E* **70,** 066111 (2004).
6. Guimera, R. & Amaral, L. A. N. Functional cartography of complex metabolic networks. *Nature* **433,** 895–900 (2005).
7. Lagomarsino, M. C., Jona, P., Bassetti, B. & Isambert, H. Hierarchy and feedback in the evolution of the *Escherichia coli* transcription network. *Proc. Natl Acad. Sci. USA* **104,** 5516–5520 (2001).
8. Liben-Nowell, D. & Kleinberg, J. M. The link-prediction problem for social networks. *J. Am. Soc. Inform. Sci. Technol.* **58,** 1019–1031 (2007).
9. Girvan, M. & Newman, M. E. J. Community structure in social and biological networks. *Proc. Natl Acad. Sci. USA* **99,** 7821–7826 (2002).
10. Krause, A. E., Frank, K. A., Mason, D. M., Ulanowicz, R. E. & Taylor, W. W. Compartments revealed in food-web structure. *Nature* **426,** 282–285 (2003).
11. Radicchi, F., Castellano, C., Cecconi, F., Loreto, V. & Parisi, D. Defining and identifying communities in networks. *Proc. Natl Acad. Sci. USA* **101,** 2658–2663 (2004).
12. Watts, D. J., Dodds, P. S. & Newman, M. E. J. Identity and search in social networks. *Science* **296,** 1302–1305 (2002).
13. Kleinberg, J. in *Proc. 2001 Neural Inform. Processing Systems Conf.* (eds Dietterich, T. G., Becker, S. & Ghahramani, Z.) 431–438 (MIT Press, Cambridge, MA, 2002).
14. Palla, G., Derényi, I., Farkas, I. & Vicsek, T. Uncovering the overlapping community structure of complex networks in nature and society. *Nature* **435,** 814–818 (2005).
15. Casella, G. & Berger, R. L. *Statistical Inference* (Duxbury, Belmont, 2001).
16. Newman, M. E. J. & Barkema, G. T. *Monte Carlo Methods in Statistical Physics* (Clarendon, Oxford, 1999).
17. Newman, M. E. J. Assortative mixing in networks. *Phys. Rev. Lett.* **89,** 208701 (2002).
18. Huss, M. & Holme, P. Currency and commodity metabolites: Their identification and relation to the modularity of metabolic networks. *IET Syst. Biol.* **1,** 280–285 (2007).
19. Krebs, V. Mapping networks of terrorist cells. *Connections* **24,** 43–52 (2002).
20. Dawah, H. A., Hawkins, B. A. & Claridge, M. F. Structure of the parasitoid communities of grass-feeding chalcid wasps. *J. Anim. Ecol.* **64,** 708–720 (1995).
21. Bryant, D. in *BioConsensus* (eds Janowitz, M., Lapointe, F.-J., McMorris, F. R., Mirkin, B. & Roberts, F.) pp. 163–184 (Series in Discrete Mathematics and Theoretical Computer Science, Vol. 61, American Mathematical Society-DIMACS, Providence, RI, 2003).
22. Dunne, J. A., Williams, R. J. & Martinez, N. D. Food-web structure and network theroy: The role of connectance and size. *Proc. Natl Acad. Sci. USA* **99,** 12917–12922 (2002).
23. Szilágyi, A., Grimm, V., Arakaki, A. K. & Skolnick, J. Prediction of physical protein–protein interactions. *Phys. Biol.* **2,** S1–S16 (2005).

24. Sprinzak, E., Sattath, S. & Margalit, H. How reliable are experimental protein-protein interaction data? *J. Mol. Biol.* **327,** 919–923 (2003).
25. Ito, T. *et al.* A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl Acad. Sci. USA* **98,** 4569–4574 (2001).
26. Lakhina, A., Byers, J. W., Crovella, M. & Xie, P. in *INFOCOM 2003: Twenty-Second Annual Joint Conf. IEEE Computer and Communications Societies* (ed. Bauer, F.) Vol. 1 332–341 (IEEE, Piscataway, New Jersey, 2003).
27. Clauset, A. & Moore, C. Accuracy and scaling phenomena in Internet mapping. *Phys. Rev. Lett.* **94,** 018701 (2005).
28. Martinez, N. D., Hawkins, B. A., Dawah, H. A. & Feifarek, B. P. Effects of sampling effort on characterization of food-web structure. *Ecology* **80,** 1044–1055 (1999).
29. Hanely, J. A. & McNeil, B. J. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* **143,** 29–36 (1982).
30. Sales-Pardo, M., Guimerá, R., Moreira, A. A. & Amaral, L. A. N. Extracting the hierarchical organization of complex systems. *Proc. Natl Acad. Sci. USA* **104,** 15224–15229 (2007).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank J. Dunne, M. Gastner, P. Holme, M. Huss, M. Porter, C. Shalizi and C. Wiggins for their help, and the Santa Fe Institute for its support. C.M. thanks the Center for the Study of Complex Systems at the University of Michigan for hospitality while some of this work was conducted.

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to A.C. (aaronc@santafe.edu).