Distribution of node characteristics in complex networks

Juyong Park[*†‡] and Albert-László Barabási[*†‡§]

[*]Department of Physics and Center for Complex Network Research, University of Notre Dame, South Bend, IN 46556; [†]Department of Physics and Center for Complex Network Research, Northeastern University, Boston, MA 02115; and [§]Collegium Budapest, Szentháromság v.2, H-1014 Budapest, Hungary

Edited by H. Eugene Stanley, Boston University, Boston, MA, and approved September 12, 2007 (received for review May 30, 2007)

Our enhanced ability to map the structure of various complex networks is increasingly accompanied by the possibility of independently identifying the functional characteristics of each node. Although this led to the observation that nodes with similar characteristics have a tendency to link to each other, in general we lack the tools to quantify the interplay between node properties and the structure of the underlying network. Here we show that when nodes in a network belong to two distinct classes, two independent parameters are needed to capture the detailed interplay between the network structure and node properties. We find that the network structure significantly limits the values of these parameters, requiring a phase diagram to uniquely characterize the configurations available to the system. The phase diagram shows a remarkable independence from the network size, a finding that, together with a proposed heuristic algorithm, allows us to determine its shape even for large networks. To test the usefulness of the developed methods, we apply them to biological and socioeconomic systems, finding that protein functions and mobile phone usage occupy distinct regions of the phase diagram, indicating that the proposed parameters have a strong discriminating power.

assortativity | complexity | dyadic effect | graph bipartition

The pervasiveness of networked systems in biology, technology, and society (1–7) has led to a recent surge of interest in uncovering the organizing principles that govern the topology and the dynamics of various complex networks. Advances in this direction have typically focused on characterizing the topological maps that depict how the system's components connect to each other. In many real systems of scientific interest, however, the nodes themselves possess characteristics that carry important information about their role in the system. In a social network, for example, each individual can be assigned gender, race, and parameters that represent his or her preference for products or services; in a protein–protein interaction network, each protein is characterized by its biological functions (8, 9); web pages in the World Wide Web can be categorized based on their content (10–12). Often, the various node properties are not distributed at random in the network, but are correlated with the underlying network structure. At least two mechanisms may be responsible for such correlations: the placement of new links could be driven by node characteristics (people with similar interests becoming friends), or the node characteristics could be influenced by the links the node has (purchasing services used by our friends). Whatever the mechanism, there is empirical evidence that in many networks adjacent nodes show significant correlations in their properties, a phenomenon often called "assortative mixing" (13). For example, children of the same race are more likely to become friends in school (14, 15); weblogs (or "blogs") on political issues contain more hyperlinks to blogs of similar political leanings (16); proteins with similar functions have a higher chance to connect to each other (17). Consequently, in many systems, the number of links between nodes sharing a common property is larger than expected if the characteristics were distributed randomly on the network (18–22), a phenomenon called the dyadic effect.

The evidence for dyadic effect raises several fundamental questions: How many parameters are necessary to mathematically describe the statistical distribution of node characteristics in a network? Are there effects beyond dyadic; i.e., are two different configurations of node characteristics equivalent if they show the same dyadic effect? With the increasing recognition of the interplay between network structure and node characteristics (11–17, 23) helped by the ever-developing data collection abilities, these questions are becoming of great practical significance, answers to which could assist in a better understanding of complex systems.

Theory

Dyads and Dyad Counts. Assume that we are provided with a network with known node characteristics, and we wish to determine to what degree they correlate with the network structure. Consider the case where each node is characterized by a property that can take only two values, 1 or 0, for simplicity. For example, the property could capture if a molecule contributes to a specific function in the cell (1) or does not (0), or if a person belongs to a certain social group (1) or does not (0). Let us call n_1 (n_0) the number of nodes with property 1 (0) so that the total number of nodes N satisfies $N = n_1 + n_0$. This allows for three kinds of *dyads* (defined as a link and its two end nodes) in the network: $(1 - 1)$, $(1 - 0)$, and $(0 - 0)$ (Fig. 1*a*). We label the number of each dyad type m_{11}, m_{10}, m_{00}, respectively, satisfying $M = m_{11} + m_{10} + m_{00}$, where M is the total number of links in the network. Without loss of generality we choose m_{11} and m_{10} as independent parameters, representing the dyads containing nodes with property 1.

If property 1 is distributed randomly among the N nodes, i.e., if any node has an equal chance of possessing it, the expected values of m_{11} and m_{10} are (18, 24)

$$\bar{m}_{11} = \binom{n_1}{2} \times p = \frac{n_1(n_1 - 1)}{2} p, \tag{1a}$$

$$\bar{m}_{10} = \binom{n_1}{1} \binom{n_0}{1} \times p = n_1 (N - n_1)p, \tag{1b}$$

where $p \equiv 2M/N(N - 1)$ is the *connectance*, representing the average probability that two nodes are connected. Statistically significant deviations of m_{11} and m_{10} from their expected values $\bar{m}_{11}$ and $\bar{m}_{10}$ imply that property 1 is not distributed randomly.

Author contributions: J.P. and A.-L.B. designed research; J.P. performed research; J.P. analyzed data; and J.P. and A.-L.B. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Abbreviation: PPI, protein–protein interaction.

[‡]To whom correspondence may be addressed. E-mail: fyzixx@gmail.com or alb@nd.edu.

This article contains supporting information online at www.pnas.org/cgi/content/full/0705081104/DC1.

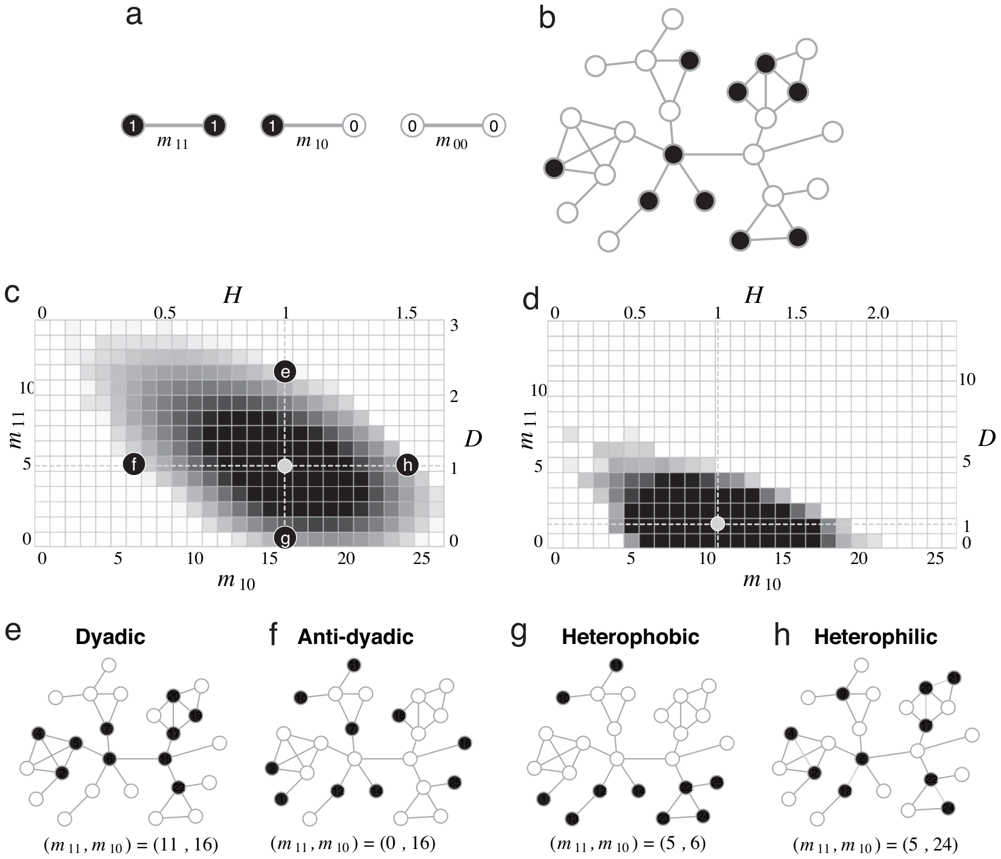

Fig. 1. Definitions and examples. (*a*) When nodes of a network can be classified into two classes, with labels 1 and 0 (shown as black and white, respectively), we have three types of dyads, as shown in the figure. The number of each dyad $\{m_{11}, m_{10}, m_{00}\}$ satisfies $M = m_{11} + m_{10} + m_{00}$, where M is the total number of links in the network. (*b*) A network of $N = 25$ nodes and $M = 32$ links, on which $n_1 = 10$ black nodes are distributed randomly. The complete phase diagram of possible values of (m_{11}, m_{10}) for $n_1 = 10$ (*c*) and for $n_1 = 5$ (*d*). (*e–h*) Configurations of four extreme points indicated on the phase diagram (*c*) for $n_1 = 10$.

We quantify the magnitude of such effects via *dyadicity* D and *heterophilicity* H defined as

$$D \equiv \frac{m_{11}}{\bar{m}_{11}} \quad \text{and} \quad H \equiv \frac{m_{10}}{\bar{m}_{10}}. \qquad [2]$$

We call property 1 to be *dyadic* if $D > 1$ (antidyadic if $D < 1$), indicating that the nodes with property 1 tend to connect more (less) densely among themselves than expected for a random configuration. Similarly we define property 1 to be heterophilic if $H > 1$ (heterophobic if $H < 1$), meaning that the nodes with property 1 have more (fewer) connections to nodes with property 0 than expected randomly. Note that Dp is the connection probability between a (1–1) node pair, whereas Hp is that between a (1–0) node pair.

To understand the significance of D and H, we must first realize that m_{11} and m_{10} cannot assume arbitrary values. For example, m_{11} can never be larger than $\min(M, \binom{n_1}{2})$, and m_{10} cannot exceed $\min(M, n_1 n_0)$. Yet, there are subtler constraints determined by the interplay between network structure and n_1. We illustrate this by using a network of $n = 25$ nodes and $m = 32$ links in Fig. 1*b*. In Fig. 1 *c* and *d*, we present the phase diagrams characterizing the distribution of an arbitrary property on the network when $n_1 = 10$ (Fig. 1*c*) and $n_1 = 5$ (Fig. 1*d*). The left and bottom axes show m_{11} and m_{10}, whereas the right and the upper axes show the corresponding D and H calculated from Eq. **1**. The darkness of each square represents the "degeneracy," i.e., the number of ways of placing n_1 nodes on the network while maintaining (m_{11}, m_{10}). An open square means that there is no configuration consistent with (m_{11}, m_{10}), indicating that not all (D, H) configurations are available for the network. As shown by the difference in the shapes of the phase diagrams in Fig. 1 *c* and *d*, the available configurations are highly dependent on n_1. The dotted lines $m_{11} = \bar{m}_{11}$ $(D = 1)$ and the $m_{10} = \bar{m}_{10}$ $(H = 1)$ intersect at the random expectation $(\bar{m}_{11}, \bar{m}_{10})$.

The phase diagrams (Fig. 1 *c* and *d*) are helpful in illustrating the properties of node groups showing various values of D and H. We first observe that a random distribution of a property ($D = H = 1$, Fig. 1*b*) represents a region of high degeneracy point in the phase diagram, i.e., the most "typical" of all configurations. Atypical configurations, such as those shown in Fig. 1 *e–h*, are visibly different from the random configuration: in a dyadic configuration ($D \gg 1$, Fig. 1*e*), black nodes are pushed into the highly interlinked central clusters of the network to maximize m_{11}, whereas in an antidyadic configuration ($D \ll 1$, Fig. 1*f*), the black nodes tend to avoid linking to one another; in a heterophobic configuration ($H \ll 1$, Fig. 1*g*), the black nodes are pushed into the periphery to avoid contact with white nodes and minimize m_{10}, whereas in a heterophilic configuration ($H \gg 1$, Fig. 1*h*), the black nodes occupy the hubs so that contact with white nodes is maximized.

Phase Diagrams for Large Networks. For sufficiently small networks (Fig. 1*b*), the phase diagram can be obtained via an exhaustive enumeration of each way in which the n_1 black nodes can be placed on the network. This task becomes infeasible for large networks, because the number of possible configurations $\binom{N}{n_1}$ increases expo-

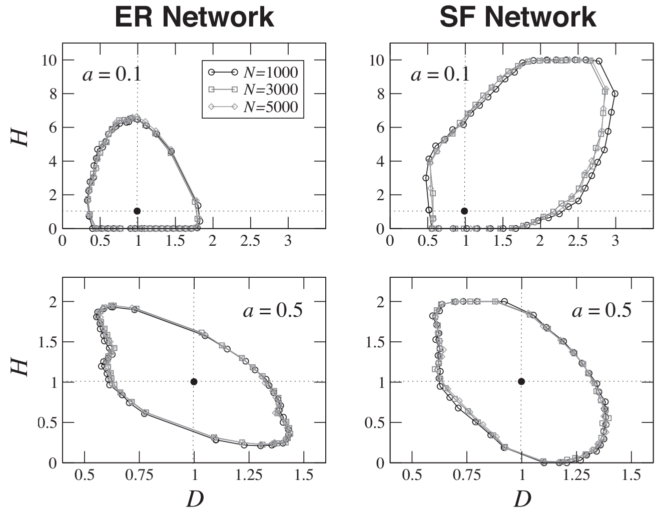

Fig. 2. Phase boundaries calculated by the Fiduccia-Mattheyses algorithm for Erdös–Rényi (ER) and scale-free (SF) networks of varying sizes N = 1,000 (open circles), 3,000 (open squares), and 5,000 (open diamonds), for $a \equiv n_1/N$ = 0.1 and a = 0.5. Average degree was set to $\langle k \rangle = 6$. The points corresponding to the random configuration ($D = H = 1$) are indicated by filled circles.

nentially with N. However, we can try to find only the boundary of the phase diagram through optimizing (i.e., maximizing and minimizing) m_{10} given m_{11} (or vice versa). This task maps into graph bipartition, which is an NP-complete problem (25, 26). We can speed up the process by using the Mattheyses and Fiduccia heuristic with $O(M)$ running time for a single optimization run (27) [see supporting information (SI) *Text*, SI Fig. 4].

This procedure allows us to ask an important question: does the boundary of the phase diagram depend on the network size N, assuming that the coverage ratio $a \equiv n_1/N$ is constant (i.e., the fraction of nodes with property 1 is independent of N)? The limiting case of $N \to \infty$ is called the *thermodynamic limit* in statistical physics, and it is of interest for most analytic approaches (1, 2, 28). We study networks constructed by using two canonical models: (*i*) the Erdös–Rényi network, in which two nodes are connected randomly with equal probability (29, 30), and (*ii*) the scale-free network with a power-law degree distribution $p_k \propto k^{-3}$ (31). Their $D-H$ phase diagrams (Fig. 2) indicate that the phase boundary is *stationary*, or system-size independent (see also *SI Text*). Therefore, if we need to construct the phase diagram for a prohibitively large network, we may be able to extract a considerably smaller sample of the network (with statistical properties sufficiently similar to those of the full network) whose phase diagram in the D–H space should be the same as that of the original network.

Applications to Real Networks

To demonstrate the relevance of the proposed methodology to real systems, we study the distribution of node characteristics in selected biological and socioeconomic systems.

Protein–Protein Interaction (PPI) Network of *S. cerevisiae*. The filtered yeast interactome PPI network of *S. cerevisiae* (32) consists of N = 1,379 proteins and M = 2,493 links, each representing an experimentally documented interaction between two proteins (33). The MIPS (http://mips.gsf.de) classification places each protein into one or several of 16 functional classes, depending on whether it does (1) or does not (0) participate in some well characterized cellular function. The D and H parameters for each class are shown in Fig. 3*a* (see also SI Table 1). We find that all functional classes are dyadic and heterophobic, showing a highly

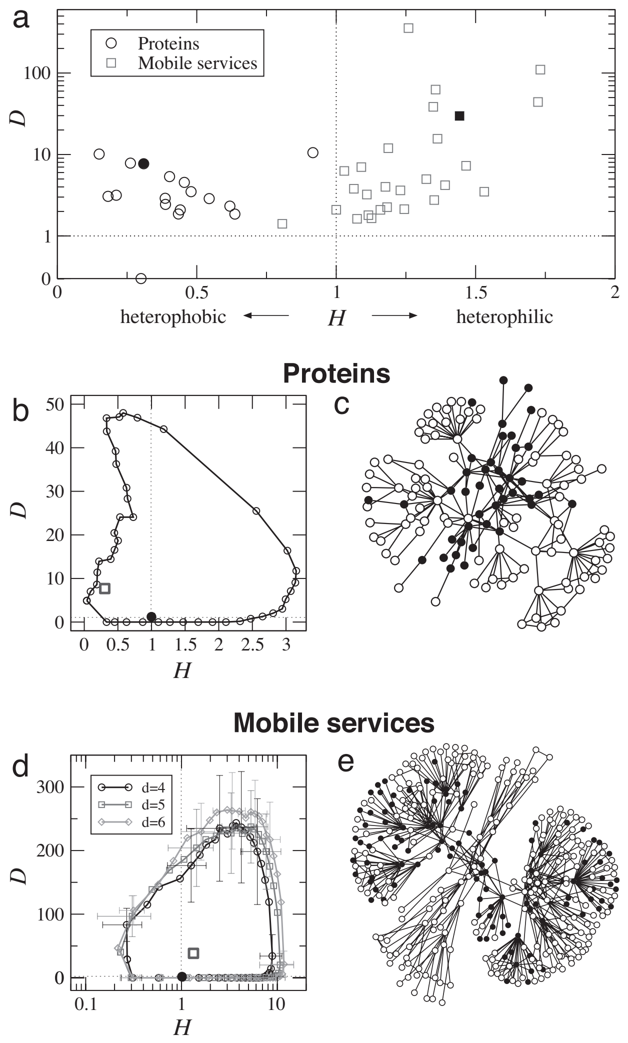

Fig. 3. A study of node characteristics distributions on two networks. (*a*) Values of the D and H parameters for proteins belonging to the main MIPS functional classes in the protein interaction network of *S. cerevisiae* (circles) and service usage in the mobile communication network (squares). Functions and names of each class and service are given in *SI Text*. In *b*–*e*, we study two node properties, one from each network (indicated by two filled shapes in *a*). (*b*) The phase diagram for a node group of size n_1 = 95 in the *S. cerevisiae* PPI network, corresponding to the MIPS functional class 30, i.e., proteins involved in cellular communications and signal transduction. The (D, H) values of this class of proteins indicate that they are more dyadic and less heterophilic than expected for a random configuration (filled circle), highlighting the modular structure of the functional class. Furthermore, the proximity of the real (D, H) to the phase boundary shows that its heterophilicity is almost as small as it can be given its dyadicity and the network structure. (*c*) A portion of the *S. cerevisiae* PPI network showing the MIPS class-30 proteins (filled nodes) and their neighbor proteins. (*d*) The phase diagram for the mobile Chat service. Due to the prohibitively large size of the mobile phone network, the boundary-finding algorithm was applied on network samples (i.e., egocentric networks of radii d = 4, d = 5, and d = 6 centered on random vertices) with constant coverage ratio. The solid lines connect the average boundary positions. (*e*) A portion of the mobile network showing users of Chat (filled nodes) and their neighbors. Unlike MIPS class-30 proteins of *c*, Chat users are heterophilic as well as dyadic, being more spread out over the network.

modular structure. In Fig. 3*b*, we show the phase diagram for proteins involved in cellular communication and signal transduction (functional class 30 in SI Table 1). Note that the real configuration sits near the phase boundary, indicating that the proteins in this class display the most heterophobic configuration given their dyadicity D. In Fig. 3*d*, we show a subset of the *S. cerevisiae* network that contains some proteins of the MIPS functional class 30 and its neighbors. Their dyadic and heterophobic nature are visually apparent: the proteins cluster to maximize contact between themselves (1–1 links) while minimizing contact with proteins that do not belong to the group (1–0 links).

Mobile Phone Network. The mobile communication network consists of $N = 5$ million mobile phone users and $M = 10.7$ million links that represent calls (voice or text messages) between the users over a period of 15 days (34–36). Each user can use their handsets to access additional phone-based services such as e-mail or a real-time instant messaging service called Chat. Whether a person did (1) or did not (0) use a particular service during our observation period defines his or her classification (34) (SI Table 2).

Fig. 3*a* shows the D and H parameters for 27 mobile phone-based services, indicating that they occupy a strikingly different region of the $D-H$ space than the functional groups in the PPI network: phone services exhibit a strong tendency to be heterophilic, implying that the users of a service tend to possess a high number of contacts with nonusers as well (see Fig. 1*h*).

Given the system's extraordinary size, we could not determine the phase diagram exactly. Thus, we rely on the system size independence of the phase boundary, obtained earlier (Fig. 2). We selected several nodes at random from the network, and took their egocentric networks containing all nodes and links within geodesic distances $d = 4, 5,$ and 6. We then determined the phase boundaries of the samples (Fig. 3*d*). Although the phase boundary expands slightly as D increases, we observe a stability similar to the one observed for the Erdös–Rényi and scale-free network models.

Finally, placing the values of D and H for Chat in the phase diagram (Fig. 3*d*), we find that the usage of Chat is more dyadic and heterophilic than randomly expected, although not extreme in either aspect, staying far from the boundary. This is illustrated in Fig. 3*f*, representing a small portion of the network, which indicates that Chat users are often connected to each other but not clustered to the degree that the proteins were in Fig. 3*e*, given the relatively weak heterophilic effect compared with what could have been obtained if the real configuration were closer to the phase boundary.

Discussion and Conclusion

Here, we explored a question of increasing importance in network characterization: how the node properties correlate with the underlying network topology. Four findings stand out: (*i*) Dyadicity alone is not sufficient to characterize the statistical features of a node property: two parameters, D and H, are necessary. (*ii*) To understand the degree of departure of a node characteristic distribution from random, we need the full phase diagram. For large systems, this can be obtained by using a heuristic algorithm for graph bipartition. (*iii*) We found that in the $D-H$ space the phase diagram is independent of the system size N, which was put to use when we studied the service usage patterns in a prohibitively large mobile communication network. However, note that, for some systems, we do observe a slower convergence of the phase diagram with increasing system size, which we attribute to the role of the hubs (see *SI Text*). Thus, the convergence of the phase diagram to a limiting shape requires further study. (*iv*) The D and H parameters have strong discriminating power: although mobile phone service usage and functional groups of proteins in the PPI network of *S. cerevisiae* are both dyadic ($D > 1$), they show distinctly different H, the former being heterophilic ($H > 1$) and the latter heterophobic ($H < 1$). This distinction between the two network types, along with the sampling method introduced above, may also assist in the practical applications of network theory to new classes of problems. For example, the tools developed here may help us gain a more detailed understanding of social segregation problems, such as the social effects present in the obesity epidemic (23, 37).

The results presented here lay the foundations for future theoretical studies on these problems. For example, one could be faced with systems where a node belongs to several classes at once (36, 38), or nodes can be divided into several classes, not just two, as in the case explored in this paper. For the latter case, an extension of our method requires introducing more dyad types depending on the number of possible pairing of node types. In general, with x distinct node types we obtain $x(x + 1)/2$ dyads. As a consequence, the phase diagram will have to be drawn in the $(x + 2)(x - 1)/2$-dimensional space. Note that the methodology developed in this study generalizes into this higher dimension as well. One may also investigate changes in the phase diagram induced by specific network properties such as degree assortativity or module size distribution. In a degree-assortative network (39, 40), where nodes with similar degrees are connected more densely than in a network with no such assortativity, a group of n_1 nodes can achieve a higher D by occupying positions of similar degrees, or a lower D by occupying positions of dissimilar degrees. Therefore, assortativity deforms the phase boundary so that points with larger D and smaller H values are included. In a similar fashion, the existence of topologically distinct groups or communities can effectively raise D and lower H. Indeed, if a module exists whose size s is equal to n_1, D can be maximal and H minimal when all n_1 nodes occupy the module. However, if $s < n_1$, then $(n_1 - s)$ nodes from the group are forced out of the module and have to connect to dissimilar nodes, thereby lowering D and raising H. On the other hand, if $s > n_1$, the nodes are forced to share the positions in the module with $(s - n_1)$ dissimilar nodes, again lowering D and raising H. Therefore, the existence of distinct modules will transform the phase boundaries so that they include larger D and smaller H values for n_1 that coincide with the module sizes.

Lastly, we note that we have only studied unweighted networks, ignoring the fact that in reality some links are far stronger than others in various contexts. For example, in the phone communication network we can measure the strength of the interaction between two individuals either as the number of minutes spent on the phone talking to each other, or the number of calls placed between them. Although including such weights will not change the phase diagram, it would be desirable to extend the proposed methodology to account for this additional layer of information. Most important, additional work is needed to understand the origin of the observed correlations between node characteristics and network structure. What mechanisms make the protein interaction network heterophobic, the service usage in mobile communication network heterophilic, while both are dyadic? Can we predict from microscopic mechanisms the nature of the observed node-network correlations? Answering these questions could lead to a deeper understanding of the interplay between structure and functions in complex systems.

We thank Deok-Sun Lee and Zoltán Toroczkai for useful discussions, and Gábor Szabó and Kwang-Il Goh for providing the network data. This work was supported by the James S. McDonnell Foundation 21st Century Science Initiative in Studying Complex Systems Research Award, National Science Foundation (NSF) Grants ITR DMR-0426737, CNS-0540348, and IIS-0513650, and National Institutes of Health Grants 1P50HG4233-01, U56 CA 113004, and U01 A1070499-01 (subaward no. 111620-2), and the NAP Project sponsored by the National Office for Research and Technology (KCKHA005). Calculations were performed on the Notre Dame Biocomplexity Cluster supported in part by NSF Major Research Instrumentation Program Grant DBI-0420980.

1. Newman MEJ, Barabási A-L, Watts DJ (2006) *The Structure and Dynamics of Networks* (Princeton Univ Press, Princeton).
2. Albert R, Barabási A-L (2002) *Rev Mod Phys* 74:47–97.
3. Newman MEJ (2003) *SIAM Rev* 45:167–256.
4. Dorogovtsev SN, Mendes JFF (2003) *Evolution of Networks: From Biological Nets to the Internet and WWW* (Oxford Univ Press, Oxford).
5. Pastor-Satorras R, Vespignani A (2004) *Evolution and Structure of the Internet: A Statistical Physics Approach* (Cambridge Univ Press, Cambridge, UK).
6. Caldarelli G (2007) *Scale-Free Networks: Complex Webs in Nature and Technology* (Oxford Univ Press, Oxford).
7. Boccaletti S, Latora V, Moreno Y, Chavez M, Hwang D-U (2006) *Phys Rep* 424:175–308.
8. Barabási A-L, Oltvai ZN (2004) *Nat Rev Genet* 5:101–113.
9. Albert R (2005) *J Cell Sci* 118:4947–4957.
10. Adamic L (1999) in *Proceedings of ECDL* (Springer, Heidelberg), pp 443–452.
11. Dumais S, Chen H (2000) in *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (ACM Press, New York), pp 256–263.
12. Menczer F (2002) *Proc Natl Acad Sci USA* 99:14014–14019.
13. Newman MEJ (2002) *Phys Rev Lett* 89:208701.
14. Shrum WNHC, Jr, Hunter SMD (1988) *Sociol Edu* 61:227–239.
15. Moody J (2001) *Am J Soc* 107:679–716.
16. Adamic LA, Glance N (2005) in *Workshop on Link Discovery: Issues, Approaches and Applications (LinkKDD-2005)* (ACM Press, New York), pp 36–43.
17. Yook S-H, Oltvai ZN, Barabási A-L (2004) *Proteomics* 4:928–942.
18. Alba RD (1973) *J Math Soc* 3:113–126.
19. Seidman SB (1983) *Social Networks* 5:97–107.
20. Sailer LD, Gaulin SJC (1984) *Am Anthropol* 86:91–98.
21. Borgatti SP, Everett MG, Shirey PR (1990) *Social Networks* 12:337–358.
22. White DR, Harary F (2001) *Sociol Methods* 31:305–359.
23. Christakis NA, Fowler JH (2007) *N Engl J Med* 357:370–379.
24. Fienberg SE, Meyer MM, Wasserman S (1985) *J Am Stat Assoc* 80:51–67.
25. Garey MR, Johnson DS, Stockmeyer L (1976) *Theor Comput Sci* 1:267–283.
26. Feder T, Hell P, Klein S, Motwani R (1999) in *Proceedings of the Thirty-First Annual ACM Symposium on Theory of Computing* (Assoc Comput Machinery Press, Washington, DC), pp 464–472.
27. Fiduccia CM, Mattheyses RM (1982) *Proceedings of the 19th Conference on Design Automation* (IEEE, Los Alamitos, CA), pp 171–181.
28. Park J, Newman MEJ (2004) *Phys Rev E* 70:066117.
29. Erdös P, Rényi A (1959) *Pub Math* 6:290–297.
30. Bollobás B (1998) *Modern Graph Theory*, Graduate Texts in Mathematics (Springer, New York).
31. Barabási A-L, Albert R (1999) *Science* 286:509–512.
32. Han J-DJ, Bertin N, Hao T, Goldberg DS, Berriz GF, Zhang LV, Dupuy D, Walhout AJM, Cusick ME, Roth FP, Vidal M (2004) *Nature* 430:88–83.
33. Jeong H, Mason SP, Barabási A-L, Oltvai ZN (2001) *Nature* 411:41–42.
34. Szabó G, Barabási A-L (2006) www.arxiv.org/abs/physics/0611177.
35. Onnela J-P, Saramäki J, Hyvönen J, Szabó, G., Lazer D, Kaski K, Kertész J, Barabási A-L (2007) *Proc Natl Acad Sci USA* 104:7332–7336.
36. Palla G, Barabási A-L, Vicsek T (2007) *Nature* 446:664–667.
37. Barabási A-L (2007) *N Engl J Med* 357:404–407.
38. Palla G, Derényi I, Farkas I, Vicsek T (2005) *Nature* 435:814–818.
39. Newman MEJ, Park J (2004) *Phys Rev E* 68:036122.
40. Pastor-Satorras R, Vázquez A, Vespignani A (2001) *Phys Rev Lett* 87:258701.