EXTENDED PDF FORMAT SPONSORED BY
Tools for Your
Biological Target
sigma.com/biotarget2

Revisiting the Foundations of Network Analysis

Carter T. Butts

Science **325**, 414 (2009);
DOI: 10.1126/science.1171022

This copy is for your personal, non-commercial use only.

If you wish to distribute this article to others, you can order high-quality copies for your colleagues, clients, or customers by clicking here.

Permission to republish or repurpose articles or portions of articles can be obtained by following the guidelines here.

The following resources related to this article are available online at www.sciencemag.org (this information is current as of May 24, 2012):

Updated information and services, including high-resolution figures, can be found in the online version of this article at:
http://www.sciencemag.org/content/325/5939/414.full.html

A list of selected additional articles on the Science Web sites **related to this article** can be found at:
http://www.sciencemag.org/content/325/5939/414.full.html#related

This article **cites 28 articles**, 8 of which can be accessed free:
http://www.sciencemag.org/content/325/5939/414.full.html#ref-list-1

This article has been **cited by** 5 article(s) on the ISI Web of Science

This article has been **cited by** 7 articles hosted by HighWire Press; see:
http://www.sciencemag.org/content/325/5939/414.full.html#related-urls

This article appears in the following **subject collections:**
Sociology
http://www.sciencemag.org/cgi/collection/sociology

Downloaded from www.sciencemag.org on May 24, 2012

Science (print ISSN 0036-8075; online ISSN 1095-9203) is published weekly, except the last week in December, by the American Association for the Advancement of Science, 1200 New York Avenue NW, Washington, DC 20005. Copyright 2009 by the American Association for the Advancement of Science; all rights reserved. The title *Science* is a registered trademark of AAAS.

PERSPECTIVE

Revisiting the Foundations of Network Analysis

Carter T. Butts

Network analysis has emerged as a powerful way of studying phenomena as diverse as interpersonal interaction, connections among neurons, and the structure of the Internet. Appropriate use of network analysis depends, however, on choosing the right network representation for the problem at hand.

The past decade has seen a dramatic surge of interest in the study of networks, with much of it in fields outside the "traditional" areas of mathematics, computer science, and the social sciences (*1*, *2*). By providing a formal mechanism for representation, measurement, and modeling of relational structure, the use of network analytic methods in these new domains (including physics, biology, and medicine) has arguably paved the way for a range of advances. On the other hand, this rapid expansion creates the risk that existing methods may be misapplied or misinterpreted, leading to inappropriate conclusions and generally poor results.

Standard Framework and Core Assumptions

Most network research is based on a representational formalism borrowed from graph theory. Researchers begin with a finite set of identifiable entities, which are represented via a vertex set. Each element of this set, commonly called a node, represents a single entity that potentially may take part in the relation under study. Relationships themselves are represented via edges, which conventionally are either unordered pairs of nodes (in which case the relation is said to be undirected) or ordered pairs of nodes (in which case the relation is said to be directed). The network is represented by a graph, which is defined as the set of nodes together with the set of pairwise relationships among them (Fig. 1A).

This representational framework is quite restrictive. To represent a system in this way, we must be able to reduce it to a well-defined set of discrete components whose interactions are strictly dyadic in nature. For any given (possibly ordered) pair of such components, the relationship is dichotomous, either present or absent; although such a framework may seem so restrictive as to be useless, its typical purpose is to serve as an approximation to the structure of a more complex system, for purposes of studying a particular property (such as the diffusion of a disease in a community over a specific time scale). Moreover, it is precisely the reductive nature of graphical structure that has facilitated its rich mathematical development (*3*) and associated scientific applications (*4*, *5*).

Extensions and relaxations of this basic framework designed to accommodate more complex situations are many and varied. We may avoid the assumption of dichotomous relationships by allowing edges to carry different weights [such as the differing connection strengths among neurons in *Caenorhabditis elegans* (Fig. 1B) (*6*)]. Multilateral relationships (such as group memberships) may be represented by means of "hyperedges," which can involve arbitrarily many nodes (*7*). Temporal aspects of relationships may be handled by treating them as time series (*8*), such as with repeated cross-sectional sampling of group structure or Internet topology; as time intervals (*9*), such as with life history data on marital and employment relationships; or as effectively instantaneous events (*10*), such as with e-mail exchange or radio communications (Fig. 1C).

Many measurement, analysis, and modeling techniques are rooted within the standard framework. However, when assumptions of this framework do not serve as reasonable approximations of the system of interest, alternative representations and techniques may be necessary. What factors should be considered when choosing a network representation, and what are the consequences when this choice is poorly made?

When Is a Node a Node?

Consider a biologist who wishes to examine the structure of animal parasite–plant interactions and so undertakes a network study. Given sufficient time, technology, and resources, he or she might sample some designated area and document all such interactions, perhaps constructing a network of ties between each animal within the area and the sites on which it feeds. But what should count as a potential feeding site? Treating each plant as a single site may seem reasonable for relatively small plants but would obviously obscure the potentially complex interactions associated with even a single tree. Additional detail could be accommodated by distinguishing between classes of anatomical units (such as bark versus trunk versus leaves) or within classes, but here too judgment must be applied in determining which distinctions to make. The biologist's method of defining potential feeding sites will greatly influence the structure of the interaction network.

The basic problem is the definition of the class of distinct entities on which one's relation of interest will be defined. The mere act of positing such a class, of course, smuggles in the tacit assumption that such a class can be defined (and moreover, that it is scientifically useful to do so). The choice of individual humans as nodes in studies of friendship (*11*) or kinship (*12*) networks and the use of individual publications in citation studies (*13*) are examples in which this assumption is well-justified. On the other hand, studies of interactions between aggregates such as groups (*14*), households (*15*), or organizations may encounter problems due to the fluidity of the interacting units and the fact that subunits of a larger unit may themselves interact with others both within and without the "parent."

As in the biological example, collapsing all potentially interacting elements into a single unit may be a very poor approximation of reality. For example, my research group has studied networks formed during organizational responses to disasters. If we pooled all the groups operating under the aegis of one national government, then we would obscure the difference between small units such as urban search-and-rescue teams and large government ministries or departments, and also would incorrectly suggest that the resources or collaborators of one are necessarily available to the other. Other systems that also would be obscured by pooling include the molecular architecture of protein-protein binding sites (*16*) and hierarchical structures in the topology of the Internet (*17*).

Changing the node set can substantially influence the size and density of the resulting network, with considerable implications for subsequent analysis. For instance, the behavior of basic network properties such as degree centralization (a measure of the extent to which ties are concentrated on a small number of nodes) are known to change both qualitatively and quantitatively with size (*18*), as do the properties of even fairly simple models of network formation (*19*, *20*). In hierarchical contexts, different aggregation decisions can produce networks with very different structural features (Fig. 1A). To avoid misleading conclusions, the set of nodes should be defined so as to include all distinct entities that are capable of participating in the relationship under study; this definition should be used consistently across networks. Where no such set of entities can be uniquely identified [as is sometimes true in geographic analysis in which a continuous space is modeled as a partition (*21*)], it is possible that a finite network representation will be inappropriate. An alternative framework (such as a continuous spatial representation) may prove more fruitful. In other cases [such as multilevel processes (*22*)], simultaneous analysis of the same

Department of Sociology and Institute for Mathematical Behavioral Sciences, University of California at Irvine, 3151 Social Science Plaza, Irvine, CA 92697–5100, USA. E-mail: buttsc@uci.edu

Downloaded from www.sciencemag.org on May 24, 2012

system at multiple levels of aggregation may be appropriate.

When Is an Edge an Edge?

Many legal institutions (such as marriage) are dichotomous, and few if any societies allow one to be one-third someone's mother. Even for relations with quantitative aspects, one can often usefully identify relationships as present or absent. We do or do not regard another as a friend; a given neuron does or does not connect to another. When a relationship reflects either a general tendency toward or potential for interaction, the use of a binary representation can greatly simplify both theory and measurement. It is much simpler, for instance, to study sexual relationships than to enumerate sexual encounters, and indeed the mere potential for interaction can have behavioral consequences, even when specific relationships go unused [as in the case of potential trading partners in exchange networks (*23*) or third-party observers in dominance relations (*24*)].

Dichotomous distinctions can sometimes be misleading. Many forms of interaction are inherently episodic and occur at variable rates (*25*). Dichotomization of such data not only obscures such variation but also requires selecting a threshold level, the choice of which can substantially alter the properties of the resulting network, both directly through selective tie removal (*26*) and indirectly through changes in network density (*27*). The range of structures present at different connection strengths can vary greatly (Fig. 1B). This cannot be resolved solely with better data collection or more elaborate statistical techniques. Rather, one must determine whether the relationship under study is sufficiently stable to be well-approximated by a constant function over the period of interest and whether the values taken by this function across pairs are sufficiently constrained to be approximately dichotomous. For relationships known to be highly heterogeneous (such as trade or migration rates), no single threshold may suffice; a weighted graph representation will frequently be more appropriate. More studies that assess the effectiveness of such approximations—and provide concrete, empirically validated guidelines for practice within particular problem domains—would be a welcome addition to the literature.

Time Scales and Network Processes

In determining appropriate node and edge representations, it is vital to consider the time scales on

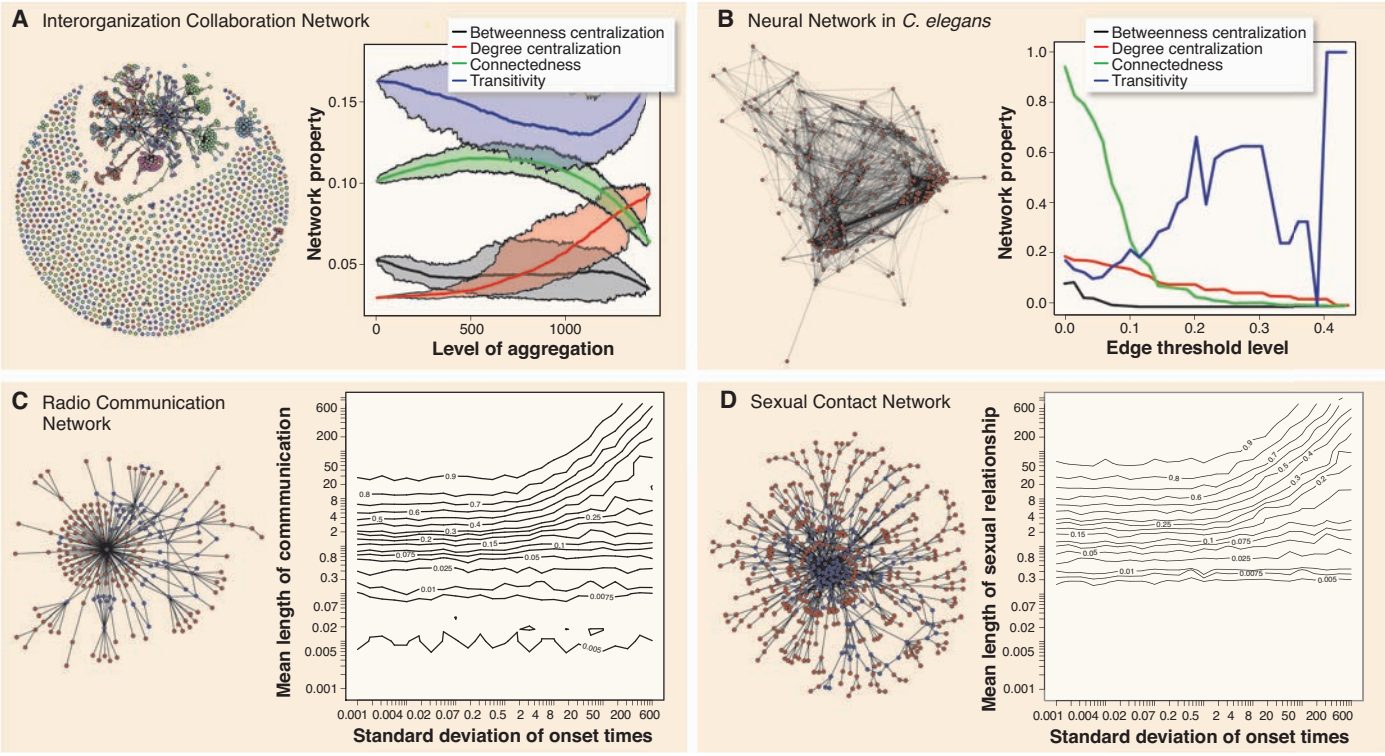

Fig. 1. Effects of changing definitions of "node" and "edge." (**A**) Network of interorganizational collaboration in the first 13 days of the Hurricane Katrina response (*39*) illustrates potential consequences of node aggregation; edges represent collaborations. Depending on the level of aggregation, each subordinate organization could be considered a node or a parent organization (containing subordinate nodes). (Left) The finest level of disaggregation, with all subordinates of a given organization having the same color. (Right) Iteratively merging randomly chosen organizations in the original network with their "parents" produces a series of increasingly aggregated structures. Shaded regions show the central 90% range of aggregated values on the basis of 100 series, and lines represent their means. The extent of aggregation affects fundamental network properties, as does the sequence of aggregation steps (indicated by width of the simulation intervals). (**B**) Neural network of *C. elegans* [based on (*6*)] shows differences in connection strength among neurons indicated as line (edge) shading in the figure at left. Nodes represent neurons, and edges reflect direct connections (arrows indicate directionality). All possible connections are shown regardless of their strength (threshold level = 0). Taking different thresholds for the edges (from 0 to 50% of maximum observed edge strength) leads to networks with different structural properties (right). (**C**) and (**D**) show the effects of edge timing, depicted as contour plots (right), in systems in which the edges are not static; each line represents the fraction of the population reached by the diffusion process. (Left) Time aggregates for the network being studied, including all relationships occurring during the observation period. (C) A radio communication network from the World Trade Center disaster, containing the largest set of people described in (*29*) from Port Authority Trans-Hudson channel 26 that were connected to each other by any chain of calls (left). Numbers within each contour line (right) indicate the mean fraction of the network that could receive information from a randomly chosen individual through an exponential diffusion process with the indicated edge parameters over 250 simulation runs. Static properties have not changed, but edge-timing variation (how long communication lasts and/or when it starts relative to the observation period) leads to variation in diffusion potential. (D) Diffusion simulation on the largest component of a sexual contact network described in (*40*) produces similar results (right) as in (C), although the degree distribution and cohesion properties (proportion of people connected by multiple common paths) differ. Each line indicates the proportion ultimately infected by a random individual (averaged over 250 trials) given the parameters of the diffusion process.

Downloaded from www.sciencemag.org on May 24, 2012

which the processes of interest unfold. For processes such as information diffusion, which unfold over hours or days, stable relationships such as kinship or friendship ties [with turnover times on the order of years (*28*)] may be approximated as essentially static. Such networks cannot be fixed in a life-cycle context, however, in which one's time scale of interest may span several decades. Likewise, the dynamics of rapidly evolving networks [such as radio communications during emergencies (Fig. 1C) (*29*)] are of potential importance even for fast-moving processes, such as information exchange. Failure to consider dynamics can lead to extremely misleading results.

A useful example of where static representations can go awry is provided by the case of HIV diffusion. Studies of sexual behavior generally find that the number of sexual partners possessed by a given individual over a fixed period of time is skewed (the mean is farther out in the long tail of the distribution than is the median) (*30*). Early studies of the behavior of simple diffusion processes on networks with extremely skewed [specifically, power-law (*31*)] degree distributions strongly suggested that epidemic potentials for HIV and similar sexually transmitted diseases were primarily governed by the behavior of a small number of individuals with large numbers of sexual contacts (*32*, *33*). This conclusion was of considerable practical import because it implied that only hub-targeted strategies were likely to prove efficacious in reducing epidemic thresholds (*31*, *32*). Although the applicability of the power-law degree model to these networks has since been questioned (*30*, *34*), equally important is the assumption that the time-aggregated network of sexual contacts was an effective model for HIV diffusion. The timing and duration of relationships are critical factors in the susceptibility of the dynamic network to disease transmission (*35*), factors that are hidden by the time-aggregated representation. This can be seen in Fig. 1D; for a given network, everyone may become infected or no one may be infected, depending on the edge duration and time of onset.

Studies of diffusion on dynamic networks suggest that partnership concurrency is also an important predictor of epidemic potential; uniformly low-degree networks potentially support epidemics when relationships are long and coterminous, and networks with high-degree nodes often fail to support epidemics when relationships are short and sequential (*35*–*37*). Interventions aimed at minimizing concurrent links are not necessarily the same as hub-targeted strategies, and thus the public health recommendations that follow from a dynamic network analysis may differ from those that would seem reasonable based on the assumption of a static, time-aggregated network.

Although HIV diffusion is a compelling example, it should be emphasized that similar issues can arise in systems as apparently different as radio communication (Fig. 1C) and peer-to-peer networks. Recent work in the latter area, for instance, has emphasized the impact of the entry and exit of network members (or "churn") on system performance (*38*); in this case, edge dynamics (potential and actual data transfers) can be understood only by taking into account the dynamic nature of the set of nodes.

Conclusion

To represent an empirical phenomenon as a network is a theoretical act. It commits one to assumptions about what is interacting, the nature of that interaction, and the time scale on which that interaction takes place. Such assumptions are not "free," and indeed they can be wrong. Whether studying protein interactions, sexual networks, or computer systems, the appropriate choice of representation is key to getting the correct result.

References and Notes

1. S. P. Borgatti, A. Mehra, D. J. Brass, G. Labianca, *Science* **323**, 892 (2009).
2. M. Newman, A. Barabasi, D. J. Watts, Eds., *The Structure and Dynamics of Networks* (Princeton Univ. Press, Princeton, NJ, 2006).
3. B. Bollobás, *Modern Graph Theory* (Springer, New York, 1998).
4. M. E. J. Newman, *SIAM Rev.* **45**, 167 (2003).
5. U. Brandes, T. Erlebach, Eds., *Network Analysis: Methodological Foundations* (Springer-Verlag, Berlin, 2005).
6. D. J. Watts, S. H. Strogatz, *Nature* **393**, 440 (1998).
7. S. Wasserman, K. Faust, *Social Network Analysis: Methods and Applications* (Cambridge Univ. Press, Cambridge, 1994).
8. T. A. B. Snijders, *J. Math. Sociol.* **23**, 149 (1996).
9. C. T. Butts, J. E. Pixley, *J. Math. Sociol.* **28**, 81 (2004).
10. C. T. Butts, *Sociol. Methodol.* **38**, 155 (2008).
11. K. M. Carley, D. Krackhardt, *Soc. Networks* **18**, 1 (1996).
12. J. P. Boyd, *J. Math. Psychol.* **6**, 139 (1969).
13. N. P. Hummon, P. Doreian, *Soc. Networks* **11**, 39 (1989).
14. R. L. Breiger, *Soc. Forces* **53**, 181 (1974).
15. M. Murphy, *Eur. J. Popul.* **12**, 363 (1996).
16. D. Reichmann, O. Rahat, M. Cohen, H. Neuvirth, G. Schreiber, *Curr. Opin. Sys. Biol.* **17**, 67 (2007).
17. L. Subramanian, S. Agarwal, J. Rexford, R. H. Katz, *Proceedings of IEEE INFOCOM* (2002).
18. C. T. Butts, *Soc. Networks* **28**, 283 (2006).
19. D. Strauss, *SIAM Rev.* **28**, 513 (1986).
20. M. S. Handcock, *Dynamic Social Network Modeling and Analysis*, R. Breiger, K. M. Carley, P. Pattison, Eds. (National Academies, Washington, DC, 2003), pp. 229–240.
21. S. Openshaw, *The Modifiable Areal Unit Problem* (Geo Books, Norwich, 1984).
22. P. R. Monge, N. S. Contractor, *Theories of Communication Networks*. (Oxford Univ. Press, Oxford, 2003).
23. D. Willer, Ed., *Network Exchange Theory* (Praeger, Westport, CN, 1999).
24. I. D. Chase, C. Tovey, D. Spangler, M. Manfredonia, *Proc. Natl. Acad. Sci. U.S.A.* **99**, 5744 (2002).
25. H. Whitehead, S. Dufault, *Adv. Stud. Behav.* **28**, 33 (1999).
26. J. P. Onnela *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **104**, 7332 (2007).
27. K. Faust, *Sociol. Methodol.* **37**, 209 (2007).
28. D. L. Morgan, M. B. Neal, P. Carder, *Soc. Networks* **19**, 9 (1997).
29. C. T. Butts, M. Petrescu-Prahova, B. R. Cross, *J. Math. Sociol.* **31**, 121 (2007).
30. D. Hamilton, M. S. Handcock, M. Morris, *Sex. Transm. Dis.* **35**, 30 (2008).
31. F. Liljeros, C. R. Edling, L. A. N. Amaral, H. E. Stanley, Y. Aberg, *Nature* **411**, 907 (2001).
32. Z. Dezso, A. Barabási, *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* **65**, 055103 (2002).
33. R. Pastor-Satorras, A. Vespignani, *Phys. Rev. Lett.* **86**, 3200 (2001).
34. J. H. Jones, M. S. Handcock, *Proc. R. Soc. London Ser. B* **270**, 1123 (2003).
35. J. Moody, *Soc. Forces* **81**, 25 (2002).
36. M. Morris, M. Kretzschmar, *AIDS* **11**, 641 (1997).
37. M. Morris, S. Goodreau, J. Moody, *Sexually Transmitted Diseases*, K. K. Holmes, *et al.*, Eds. (McGraw-Hill, New York, ed. 4, 2007), chap. 7.
38. D. Stutzbach, R. Rejaie, *Proceedings of ACM SIGCOMM* (2006).
39. C. T. Butts *et al.*, *31st Annual Hazards Research and Applications Workshop, Boulder, CO* (2006).
40. J. J. Potterat *et al.*, *Sex. Transm. Infect.* **78** (suppl. 1), i152 (2002).

41. The author would like to thank K. Faust, M. Morris, J. Moody, C. Marcum, A. Markopoulou, and R. Martin for helpful comments, and J. Potterat and S. Muth for making their data available. Supported in part by NSF awards BCS-0827027 and CMS-0624257 and by Office of Naval Research award N00014-08-1-1015.

10.1126/science.1171022

PERSPECTIVE

Disentangling the Web of Life

Jordi Bascompte

Biodiversity research typically focuses on species richness and has often neglected interactions, either by assuming that such interactions are homogeneously distributed or by addressing only the interactions between a pair of species or a few species at a time. In contrast, a network approach provides a powerful representation of the ecological interactions among species and highlights their global interdependence. Understanding how the responses of pairwise interactions scale to entire assemblages remains one of the great challenges that must be met as society faces global ecosystem change.

Network approaches to ecological research emphasize the pattern of interactions among species (the way links are arranged within the network) rather than the identity of the species composing a community (the nodes of the network of interactions). The idea of a complex network of interactions among species is as old as Darwin's contemplation of the tangled bank, showing the importance of networks in ecology (*1*). Despite this early realization, however, networks have only recently been incorporated into mainstream ecological theories. The "web of life"

Integrative Ecology Group, Estación Biológica de Doñana, Consejo Superior de Investigaciones Científicas, Calle Américo Vespucio s/n, E-41092 Sevilla, Spain. E-mail: bascompte@ebd.csic.es

Downloaded from www.sciencemag.org on May 24, 2012