Empirical Analysis of an Evolving Social Network

Gueorgi Kossinets[1]* and Duncan J. Watts[1,2]*

Social networks evolve over time, driven by the shared activities and affiliations of their members, by similarity of individuals' attributes, and by the closure of short network cycles. We analyzed a dynamic social network comprising 43,553 students, faculty, and staff at a large university, in which interactions between individuals are inferred from time-stamped e-mail headers recorded over one academic year and are matched with affiliations and attributes. We found that network evolution is dominated by a combination of effects arising from network topology itself and the organizational structure in which the network is embedded. In the absence of global perturbations, average network properties appear to approach an equilibrium state, whereas individual properties are unstable.

Social networks have attracted great interest in recent years, largely because of their likely relevance to various social processes, such as information processing (*1*), distributed search (*2*), and diffusion of social influence (*3*). For many years, however, social scientists have also been interested in social networks as dynamic processes in themselves (*4*): Over time, individuals create and deactivate social ties, thereby altering the structure of the networks in which they participate. Social network formation is a complex process in which many individuals simultaneously attempt to satisfy their goals under multiple, possibly conflicting, constraints. For example, individuals often interact with others similar to themselves—a tendency known as homophily (*5*, *6*)—and attempt to avoid conflicting relationships (*7*, *8*) while exploiting cross-cutting circles of acquaintances (*9*). However, the realization of these intentions is subject to spatial and social proximity of available others (*9*, *10*). In circumstances where individuals may benefit from cooperative relationships, they may emphasize embedded ties—those belonging to locally dense clusters (*11*). For example, they may choose new acquaintances who are friends of friends—a process known as triadic closure (*12*). They may, however, also seek access to novel information and resources and hence benefit from access to bridges (*13*)—connections outside their circle of acquaintances—or by spanning structural holes (*14*) precisely between others who do not know one another. Finally, social ties may dissolve for various reasons, such as when they are not supported by other relations (*15*), or else conflict with them (*16*).

To what extent each of these individually plausible mechanisms manifests itself in various social and organizational contexts is largely an empirical matter, requiring longitudinal (i.e., collected over time) network data (*4*) combined with information about individuals' attributes and group affiliations (*6*, *10*, *17*). Yet longitudinal network data are rare, and the best known examples are for small groups (*4*, *18*). Recent studies of much larger networks, by contrast, have tended to focus on cross-sectional (i.e., static) analysis (*19*, *20*), or they have emphasized either the interactions between individuals (*21*, *22*) or their group affiliations (*17*), but not both.

We analyzed a longitudinal network data set created by merging three distinct but related data structures. First, we compiled a registry of e-mail interactions in a population of 43,553 undergraduate and graduate students, faculty, and staff of a large university over the course of one academic year. For each e-mail message, the timestamp, sender, and list of recipients (but not the content) were recorded. Second, for the same population, we gathered information specifying a range of personal attributes (status, gender, age, departmental affiliation, and number of years in the community). Third, we obtained complete lists of the classes attended and taught, respectively, by students and instructors in each semester. For privacy protection, all individual and group identifiers were encrypted; we can determine, for example, whether two individuals were in the same class together but not which class that was. Because in a university setting class attendance provides essential opportunities for face-to-face interaction (at least for students), we used classes to represent the changing affiliation structure.

Our use of e-mail communication to infer the underlying network of social ties is supported by recent studies reporting that use of e-mail in local social circles is strongly correlated with face-to-face and telephone interactions (*23*, *24*). Individuals and groups of individuals may differ in their e-mail usage; thus, inferences drawn on a small sample of communicating pairs may be confounded by the idiosyncrasies of particular personalities and relationships. However, by averaging over thousands of such relationships, we expect that our results will represent only the most general regularities (at least within the environment of a university community) governing the initiation and progression of interpersonal communication. To ensure that our data do indeed reflect interpersonal communication as opposed to ad hoc mailing lists and other mass mailings, we filtered out messages with more than four recipients (95% of all messages had four or fewer addressees). After filtering, there were 14,584,423 messages exchanged by the users during 355 days of observation.

Ongoing social relationships produce spikes of e-mail exchange that can be observed and counted (*20*, *21*). The stronger the relationship between two individuals, the more spikes will be observed for this particular pair, on average, within a given time interval. We approximate instantaneous strength w_{ij} of a relationship between two individuals i and j by the geometric rate of bilateral e-mail exchange within a window of $\tau = 60$ days (*25*). The instantaneous network at any point in time includes all pairs of individuals that sent one or more messages in each direction during the past 60 days. Using daily network approximations, we calculated (i) shortest path length d_{ij} and (ii) the number of shared affiliations s_{ij} for all pairs of individuals in the network on 210 consecutive days spanning most of the fall and spring semesters (*25*). By identifying new ties that appear in the network over time, we can compute two sets of measures: (i) cyclic closure and (ii) focal closure biases. For some specified value of d_{ij}, cyclic closure bias is defined as the empirical probability that two previously unconnected individuals who are distance d_{ij} apart in the network will initiate a new tie. Thus cyclic closure naturally generalizes the notion of triadic closure (*12*), i.e., formation of cycles of length three. By analogy, we define focal closure bias as the empirical probability that two strangers who share an interaction focus (in the present case, a class) will form a new tie. Because class attendance is relevant mostly for students, the results on focal and cyclic closure are presented here for a subset of 22,611 graduate and undergraduate students (*25*).

Figure 1A (triangles) shows that in the absence of a shared focus (i.e., class), cyclic closure diminishes rapidly in strength with d_{ij}, implying that individuals who are far apart in the network have no opportunity to interact and hence are very unlikely to form ties. For example, individuals who are separated by two intermediaries ($d_{ij} = 3$) are about 30 times less likely to initiate a new tie

[1]Department of Sociology and Institute for Social and Economic Research and Policy, Columbia University, 420 West 118th Street, MC 3355, New York, NY 10027, USA. [2]Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501, USA.

*To whom correspondence should be addressed. E-mail: gk297@columbia.edu (G.K.); djw24@columbia.edu (D.J.W.)

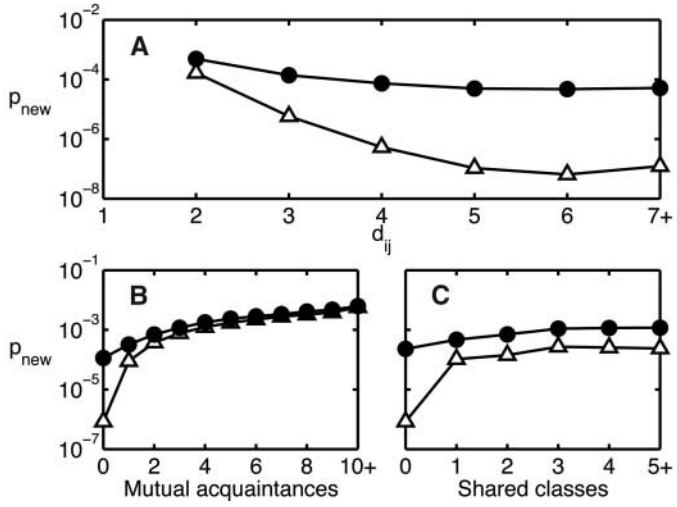

Fig. 1. Cyclic and focal closure. (**A**) Average daily empirical probability p_{new} of a new tie between two individuals as a function of their network distance d_{ij}. Circles, pairs that share one or more interaction foci (attend one or more classes together); triangles, pairs that do not share classes. (**B**) p_{new} as a function of the number of mutual acquaintances. Circles, pairs with one or more shared foci; triangles, pairs without shared foci. (**C**) p_{new} as a function of the number of shared interaction foci. Circles, pairs with one or more mutual acquaintances; triangles, pairs without mutual acquaintances. Lines are shown as a guide for the eye; standard errors are smaller than symbol size.

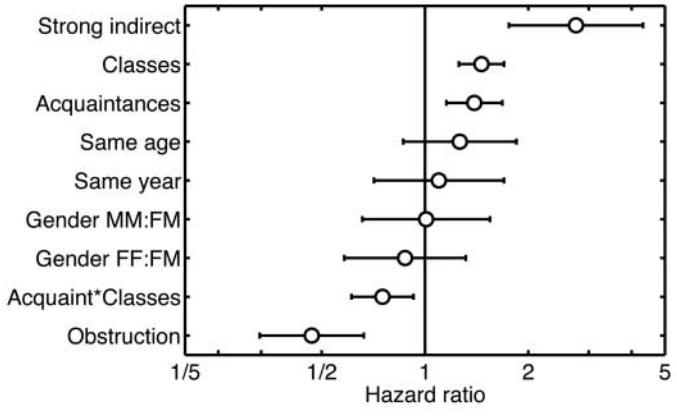

Fig. 2. Results of multivariate survival analysis of triadic closure for a sample of 1190 pairs of graduate and undergraduate students. Shown are the hazard ratios and 95% confidence intervals from Cox regression of time to tie formation between two individuals since their transition to distance $d_{ij} = 2$. Hazard ratio g means that the probability of closure changes by a factor of g with a unit change in the covariate or relative to the reference category. We treat a covariance as significant if the corresponding 95% confidence interval does not contain $g = 1$ (no effect). Predictors, sorted by effect magnitude: strong indirect (1 if indirect connection strength is above sample median, 0 otherwise), classes (number of shared classes), acquaintances (number of mutual network neighbors less 1), same age (1 if absolute difference in age is less than 1 year, 0 otherwise), same year (1 if absolute difference in number of years at the university is less than 1, 0 otherwise), gender [effects of male-male (MM) and female-female (FF) pair, respectively, relative to a female-male (FM) pair], acquaint*classes (interaction effect between acquaintances and classes), and obstruction (1 if no mutual acquaintance has the same status as either member of the pair, 0 otherwise) (*25*).

than individuals who are separated by only one intermediary ($d_{ij} = 2$). Figure 1A (circles), however, demonstrates that when two individuals share at least one class, they are on average 3 times more likely to interact if they also share an acquaintance ($d_{ij} = 2$), and about 140 times more likely if they do not ($d_{ij} > 2$). In addition, Fig. 1B shows that the empirical probability of tie formation increases with the number of mutual acquaintances both for pairs with (circles) and without (triangles) shared classes, becoming independent of shared affiliations for large numbers of mutual acquaintances (six and more). Figure 1C displays equivalent information for shared classes, indicating that while the effect of a single shared class is roughly interchangeable with a single mutual acquaintance, the presence of additional acquaintances has a greater effect than additional foci in our data set. These findings imply that even a minimally accurate, generative network model would need to account separately for (i) triadic closure, (ii) focal closure, and (iii) the compounding effect of both biases together.

Our data can also shed light on theoretical notions of tie strength (*13*) and attribute-based homophily (*6*, *26*). We found (Fig. 2) that the likelihood of triadic closure increases if the average tie strength between two strangers and their mutual acquaintances is high, which supports commonly accepted theory (*6*, *13*). By contrast, homophily with respect to individual attributes appears to play a weaker role than might be expected. Of the attributes we considered in this and other models (*27*)—status (undergraduate, graduate student, faculty, or staff), gender, age, and time in the community—none has a significant effect on triadic closure. The significant predictors are tie strength, number of mutual acquaintances, shared classes, the interaction of shared classes and acquaintances, and status obstruction, which we define as the effect on triadic closure of a mediating individual who has a different status than either of the potential acquaintances. For example, two students connected through a professor are less likely to form a direct tie than two students connected through another student, ceteris paribus. We suspect, however, that status obstruction may be an indicator of unobserved focal closure beyond class attendance. Thus, although homophily has often been observed with respect to individual attributes in cross-sectional data (*6*, *26*), these effects may be mostly indirect, operating through the structural constraint of shared foci (*10*), such as selection of courses or extracurricular activities.

Our results also have implications for the utility of cross-sectional network analysis, which relies on the assumption that the network properties of interest are in equilibrium (*4*). Figure 3 shows that different network measures exhibit varying levels of stability over time and with respect to the smoothing window τ. Average vertex degree $\langle k \rangle$, fractional size of the largest component S, and mean shortest path length L all exhibit seasonal changes and produce different measurements for different choices of τ, where $\langle k \rangle$ is especially sensitive to τ. The clustering coefficient C (*28*), however, stays virtually constant as $\langle k \rangle$ changes, suggesting, perhaps surprisingly, that averages of local network properties are more stable than global properties such as L or S. Nevertheless, these results suggest that as long as the smoothing window τ is chosen appropriately and care is taken to avoid collecting data in the vicinity of exogenous changes (e.g., end of semester), average network measures remain stable over time and thus can be recovered with reasonable fidelity from network snapshots.

The relative stability of average network properties, however, does not imply equivalent stability of individual network properties, for which the empirical picture is more complicated. On the one hand, we find that distributions of individual-level properties are stable, with the same caveats that apply to averages. For example (Fig. 4, A to C), the shape of the degree distribution $p(k)$ is relatively constant across the duration of our

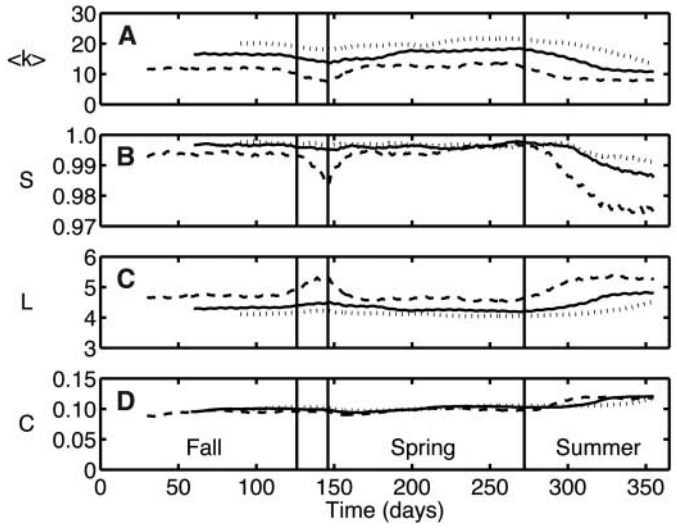

Fig. 3. Network-level properties over time, for three choices of smoothing window τ = 30 days (dashes), 60 days (solid lines), and 90 days (dots). (**A**) Mean vertex degree $\langle k \rangle$. (**B**) Fractional size of the largest component S. (**C**) Mean shortest path length in the largest component L. (**D**) Clustering coefficient C.

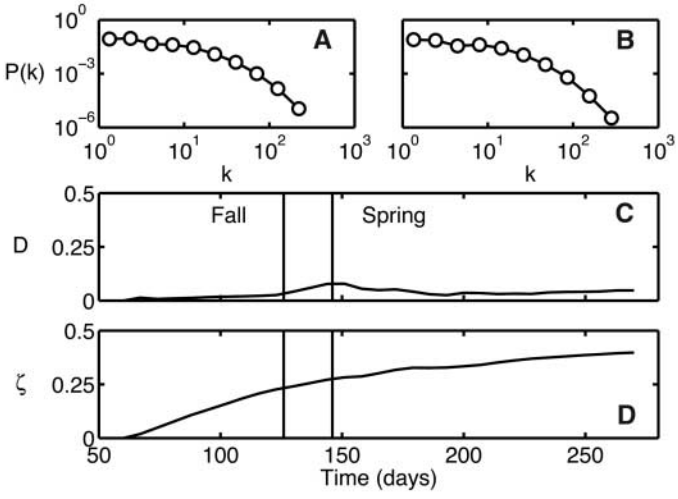

Fig. 4. Stability of degree distribution and individual degree ranks. (**A**) Degree distribution in the instantaneous network at day 61, logarithmically binned. (**B**) Same at day 270. (**C**) The Kolmogorov-Smirnov statistic D comparing degree distribution in the instantaneous network at day 61 and in subsequent daily approximations. (**D**) Dissimilarity coefficient for degree ranks $\zeta = 1 - r_S^2$, where r_S is the Spearman rank correlation between individual degrees at day 61 and in subsequent approximations. ζ varies between 0 and 1 and measures the proportion of variance in degree ranks that cannot be predicted from the ranks in the initial network.

data set except during natural spells of reduced activity, such as winter break (Fig. 4C). On the other hand, as Fig. 4D illustrates, individual ranks change substantially over the duration of the data set. Analogous results (*27*) apply to the concept of "weak ties" (*13*): The distribution of tie strength in the network is stable over time, and bridges are, on average, weaker than embedded ties [consistent with (*13*)]. However, they do not retain their bridging function, or even remain weak, indefinitely.

Our results suggest that conclusions relating differences in outcome measures such as status or performance to differences in individual network position (*14*) should be treated with caution. Bridges, for example, may indeed facilitate diffusion of information across entire communities (*13*). However, their unstable nature suggests that they are not "owned" by particular individuals indefinitely; thus, whatever advantages they confer are also temporary. Furthermore, it is unclear to what extent individuals are capable of strategically manipulating their positions in a large network, even if that is their intention (*14*). Rather, it appears that individual-level decisions tend to "average out," yielding regularities that are simple functions of physical and social proximity. Sharing focal activities (*10*) and peers (*26*), for example, greatly increases the likelihood of individuals becoming connected, especially when these conditions apply simultaneously.

It may be the case, of course, that the individuals in our population—mostly students and faculty—do not strategically manipulate their networks because they do not need to, not because it is impossible. Thus, our conclusions regarding the relation between local and global network dynamics may be specific to the particular environment that we have studied. Comparative studies of corporate or military networks could help illuminate which features of network evolution are generic and which are specific to the cultural, organizational, and institutional context in question. We note that the methods we introduced here are generic and may be applied easily to a variety of other settings. We conclude by emphasizing that understanding tie formation and related processes in social networks requires longitudinal data on both social interactions and shared affiliations (*4*, *6*, *10*). With the appropriate data sets, theoretical conjectures can be tested directly, and conclusions previously based on cross-sectional data can be validated or qualified appropriately.

References and Notes

1. P. S. Dodds, D. J. Watts, C. F. Sabel, *Proc. Natl. Acad. Sci. U.S.A.* **100**, 12516 (2003).
2. J. M. Kleinberg, *Nature* **406**, 845 (2000).
3. T. W. Valente, *Network Models of the Diffusion of Innovations* (Hampton Press, Cresskill, NJ, 1995).
4. P. Doreian, F. N. Stokman, Eds., *Evolution of Social Networks* (Gordon and Breach, New York, 1997).
5. P. Lazarsfeld, R. Merton, in *Freedom and Control in Modern Society*, M. Berger, T. Abel, C. Page, Eds. (Van Nostrand, New York, 1954), pp. 18–66.
6. M. McPherson, L. Smith-Lovin, J. M. Cook, *Annu. Rev. Sociol.* **27**, 415 (2001).
7. J. A. Davis, *Am. J. Sociology* **68**, 444 (1963).
8. T. M. Newcomb, *The Acquaintance Process* (Holt Rinehart and Winston, New York, 1961).
9. P. M. Blau, J. E. Schwartz, *Crosscutting Social Circles* (Academic Press, Orlando, FL, 1984).
10. S. L. Feld, *Am. J. Sociology* **86**, 1015 (1981).
11. J. S. Coleman, *Sociol. Theory* **6**, 52 (1988).
12. A. Rapoport, *Bull. Math. Biophys.* **15**, 523 (1953).
13. M. S. Granovetter, *Am. J. Sociology* **78**, 1360 (1973).
14. R. S. Burt, *Am. J. Sociology* **110**, 349 (2004).
15. M. Hammer, *Soc. Networks* **2**, 165 (1980).
16. M. T. Hallinan, E. E. Hutchins, *Soc. Forces* **59**, 225 (1980).
17. M. E. J. Newman, *Proc. Natl. Acad. Sci. U.S.A.* **98**, 404 (2001).
18. S. Wasserman, K. Faust, *Social Network Analysis: Methods and Applications* (Cambridge Univ. Press, Cambridge, 1994).
19. M. E. J. Newman, *SIAM Review* **45**, 167 (2003).
20. J. P. Eckmann, E. Moses, D. Sergi, *Proc. Natl. Acad. Sci. U.S.A.* **101**, 14333 (2004).
21. C. Cortes, D. Pregibon, C. Volinsky, *J. Comp. Graph. Stat.* **12**, 950 (2003).
22. P. Holme, C. R. Edling, F. Liljeros, *Soc. Networks* **26**, 155 (2004).
23. B. Wellman, C. Haythornthwaite, Eds., *The Internet in Everyday Life* (Blackwell, Oxford, 2003).
24. N. K. Baym, Y. B. Zhang, M. Lin, *New Media Soc.* **6**, 299 (2004).
25. Materials and methods are available as supporting material on *Science* Online.
26. H. Louch, *Soc. Networks* **22**, 45 (2000).
27. G. Kossinets, D. J. Watts, data not shown.
28. M. E. J. Newman, S. H. Strogatz, D. J. Watts, *Phys. Rev. E* **6402**, 026118 (2001).

29. We thank P. Dodds and two anonymous reviewers for helpful comments and B. Beecher and W. Bourne for assistance with data collection and anonymization. This research was supported by NSF (SES 033902), the James S. McDonnell Foundation, Legg Mason Funds, and the Institute for Social and Economic Research and Policy at Columbia University.

Supporting Online Material

www.sciencemag.org/cgi/content/full/311/5757/88/DC1
Materials and Methods
References

1 July 2005; accepted 29 November 2005
10.1126/science.1116869