Origins of Homophily in an Evolving Social Network[1]

Gueorgi Kossinets
Google Inc.

Duncan J. Watts
Yahoo! Research

The authors investigate the origins of homophily in a large university community, using network data in which interactions, attributes, and affiliations are all recorded over time. The analysis indicates that highly similar pairs do show greater than average propensity to form new ties; however, it also finds that tie formation is heavily biased by triadic closure and focal closure, which effectively constrain the opportunities among which individuals may select. In the case of triadic closure, moreover, selection to "friend of a friend" status is determined by an analogous combination of individual preference and structural proximity. The authors conclude that the dynamic interplay of choice homophily and induced homophily, compounded over many "generations" of biased selection of similar individuals to structurally proximate positions, can amplify even a modest preference for similar others, via a cumulative advantage–like process, to produce striking patterns of observed homophily.

INTRODUCTION

The "homophily principle"—the observed tendency of "like to associate with like"—is one of the most striking and robust empirical regularities of social life (Lazarsfeld and Merton 1954; Laumann 1966; Verbrugge 1977; McPherson and Smith-Lovin 1987; Marsden 1988; Burt 1991; Mc-

[1] We thank the *AJS* reviewers for helpful comments. This research was supported in part by the Institute for Social and Economic Research and Policy at Columbia University, the James S. McDonnell Foundation, and the National Science Foundation (grant no. SES 0339023). Portions of this research were completed while the authors were in the Department of Sociology at Columbia University. Direct correspondence to Gueorgi Kossinets, Google Inc., 1600 Amphitheatre Parkway, Mountain View, California 94043. E-mail: gk297@columbia.edu

0002-9602/2009/11502-0003$10.00

Pherson, Smith-Lovin, and Cook 2001). Friends, spouses, romantic partners, co-workers, colleagues, and other professional and recreational associates all tend to be more similar to each other than randomly chosen members of the same population with respect to a variety of dimensions, including race, age, gender, socioeconomic status, and education. The ubiquitous presence of homophily, moreover, presents an important set of questions to sociologists, on account of its relation to issues such as segregation, inequality, and social mobility (Blau 1977; Blau and Schwartz 1984; Moody 2001). In this article, however, it is not the presence or absence of homophily that we investigate, but rather its origins. Over time, that is, individuals selectively form new ties, while allowing other, existing ties to lapse, and through these microlevel processes, macro patterns of association emerge. To the extent that these emergent patterns are relevant to theorists and policy makers alike, therefore, the mechanisms by which they are generated ought to be of interest as well. On what grounds, then, do individuals selectively make or break some ties over others, and how do these choices shed light on the observation that similar people are more likely to become acquainted than dissimilar people?

Intuitively, the answer may seem obvious: people form ties with similar others because, rightly or wrongly, they prefer to. There are many reasons why this might be so. Similarity of attributes and experience arguably simplifies the process of evaluating, communicating with, and even predicting the behavior of others (Festinger 1957; Werner and Parmelee 1979; Hamm 2000). One should therefore expect that trust and solidarity would be easier to establish with similar than with dissimilar counterparts (Portes and Sensenbrenner 1993; Banks and Carley 1996; Mollica, Gary, and Trevino 2003), thereby significantly reducing the risks associated with forming new ties—a phenomenon that has been invoked to explain, for example, the role of cultural similarity in fostering trade and labor networks (Lincoln, Gerlach, and Takahashi 1992; Mouw 2003). Correspondingly, one would also expect that the ongoing cost of maintaining ties would be lower between similar than between dissimilar alters, and the benefits possibly greater as well, implying that homophilous ties should be more stable and should last longer, as has in fact been claimed previously (Felmlee, Sprecher, and Bassin 1990; Leenders 1996). In other words, the observation that individuals interact preferentially with similar others is easily explained in terms of their individual, psychological preference for doing so.

Against this intuitively plausible explanation, however, stands an equally striking fact of social life—that an individual's choice of relations is heavily constrained by other aspects of his or her life, such as geographical location, choice of occupation, place of work, and so on, that

expose him or her to some potential acquaintances, while effectively excluding many others (Feld 1981, 1982; Ibarra 1993). Individuals, moreover, are not uniformly distributed in terms of race, ethnicity, wealth, gender, or age either in space (Liben-Nowell et al. 2005) or across organizations (McPherson and Smith-Lovin 1982), but rather are sorted into shared environments, such as schools, workplaces, or neighborhoods, that are frequently more homogeneous than the population at large. Thus, even if individuals select into these environments for reasons that are unrelated to whom they would like to meet, the combination of structural constraint on the available choices that are plausibly available to them and the concomitant homogeneity of those choices will nevertheless generate strong patterns of homophily. If, for example, high school teachers are disproportionately female, and investment bankers are disproportionately male, then the resulting pattern of interactions in their workplaces will generally exhibit gender homophily, even if individuals in both professions select among their available work colleagues without regard to gender.

Broadly speaking, therefore, we can identify two theoretically distinct mechanisms by which homophily arises—namely, *choice homophily* and *induced homophily* (McPherson and Smith-Lovin 1987)—corresponding to what Mayhew (1980) called "individualistic" and "structuralist" views of the world, respectively.[2] That is, to the extent that some observed prevalence of homophilous ties can be attributed to individual, psychological preferences, it should be called *choice homophily*, and to the extent that it can be shown to arise as a consequence of the homogeneity of structural opportunities for interaction, as in neighborhoods, schools, workplaces, voluntary organizations, and even friendship circles (Feld 1981), it should be labeled induced homophily.[3] Although clear in principle, however, differentiating between these two mechanisms is complicated by a third fact of social life: the relevant social environments are rarely, if ever, determined exogenously, but rather arise (at least in part) out of choices made by the very individuals whose subsequent friendship choices the environments then constrain (McPherson and Ranger-Moore 1991; Emirbayer and Goodwin 1994). On what basis, then—individual-

[2] A third possibility is that individuals who are acquainted will become more similar over time via a process of social influence. However, the attributes we consider here are either not malleable (e.g., gender and age) or else do not change appreciably on the time scale of interest (e.g., academic major); thus, although social influence is no doubt important in many contexts, we do not consider it here.

[3] Formally, induced homophily is the level of homophily expected from random mixing within groups given group assignments, and choice homophily is the level of homophily in excess of induced homophily (McPherson and Smith-Lovin 1987).

istic or structural—are these prior choices made? As with the question of tie formation itself, either possibility can be argued on theoretical grounds.

On the one hand, it is plausible to assert that people select into the environments they choose precisely *in order* to meet the kinds of people that also participate in those environments. Students choose Ivy League universities in order to benefit from high-status mentors and peers alike; celebrities and socialites express strong preferences to attend only the "right" parties and functions, based on who is hosting or attending; business school students flock to mixers and "networking events," where they hope to gain access to future employers; and academics invited to conferences have been known to ask who else has agreed to attend before deciding themselves. Thus, even if, hypothetically speaking, all observed homophily could be accounted for in terms of the homogeneity of opportunities presented to individuals on account of their group memberships, it is possible that the effects of structural proximity are really just a proxy for unobserved individual preferences.[4] Indeed, precisely this sort of explanation underlies the "latent variables" approach to network formation (Handcock, Hoff, and Raftery 2002).

On the other hand, it is also plausible to assert that the biased sorting of individuals to structurally proximate positions is itself a consequence of structural constraints—just constraints on some prior round of decision making. That is, choices about which activities to undertake, organizations to join, and social events to attend are themselves constrained by still further elements of the social and organizational environment. Not everyone, for example, can make the choice to attend an Ivy League school, and the only individuals who are in a position to pick and choose between high-profile gatherings are already members of an elite minority. For some people, these choices are next to impossible, whereas for others, they require almost no thought or effort. Once again, in other words, what seems like a choice—albeit this time a choice regarding selection into some prior "risk set," rather than the choice of some particular alter—may be, in effect, just another manifestation of structural constraint at work.

Clearly this interplay between structural constraint, on the one hand,

[4] In order to account for the effects of biased sorting of individuals among groups, McPherson et al. (2001) have introduced a distinction between *inbreeding* and *baseline* homophily, which is subtly but importantly different from the choice-vs.-induced dichotomy: they define baseline homophily as the level of homophily expected from random mixing in the population and inbreeding homophily as the level of homophily in excess of that baseline. Inbreeding homophily, therefore, includes instances that would be classified as choice homophily (which clearly generates homophily in excess of the baseline set by demographic opportunity), but also includes some amount of induced homophily for precisely the reason that group homogeneity may be an outcome of some "inbreeding" process over and above what is determined by the overall demographic distribution.

and individuality intentionality, on the other, can propagate backward over generations of friends and social contexts, potentially without limit. Whom one meets, in other words, depends in part on what position one occupies in social, physical, and organizational "space"; but one's position in that space depends in turn on choices one has made in the past—including choices of previous relationships (McPherson 2004). Understanding the origins of homophily therefore requires nothing less than unwinding multiple generations of choices: choices of friends, which are biased by prior choices of environments, which are in turn determined by prior choices of both friends and other environments, and so on. An essential requirement for addressing the origins of homophily through empirical analysis, therefore, is longitudinal network data (McPherson et al. 2001, p. 437), which historically have been difficult to obtain. Fortunately, however, interest in collecting and analyzing longitudinal network data has been growing recently (Doreian and Stokman 1997; Suitor, Wellman, and Morgan 1997; Snijders 2001; Moody, McFarland, and Bender-deMoll 2005; Goodreau 2007), spurred in part by advances in computing and communications technology that increasingly permit real-time observation of dyadic interactions, even for very large populations (Cortes, Pregibon, and Volinsky 2003; Kossinets and Watts 2006; Onnela et al. 2007). In addition, electronic databases offer the potential to track a range of affiliations and activities that serve as "social foci" (Feld 1981) for the population in question, thus permitting, in principle at least, detailed examination of coevolving interactions and social-organization structure.

In this article, we study the origins of homophily in a particular university community, using a network data set comprising over 30,000 students, faculty, and staff, in which interactions are recorded in real time along with individual attributes and features of the relevant organizational structure. We exploit the dynamic nature of the data to consider (*a*) the interplay between structural proximity and individual preference for similarity in accounting for observed choices of interaction partners and (*b*) the interplay between the same two forces also in accounting for the observed homogeneity of structurally proximate positions themselves (which we label the "risk set"). In brief, we find that neither of the stylized theoretical views that we have outlined above—individualistic versus structural—can adequately account for the striking levels of homophily observed in our population; rather, both play an important but partial role, where each reinforces the other. Moreover, this picture remains largely constant whether we are considering the homogeneity of ties themselves or the homogeneity of social contexts—whether groups or friendship circles—that act as the risk sets from which ties are overwhelmingly selected.

DATA AND METHODS

Our analysis is based on the population of 30,396 undergraduate and graduate students, faculty, and staff in a large U.S. university, who used their university e-mail accounts to both send and receive messages during one academic year.[5] The data set, which comprises interaction, affiliation, and attribute-type longitudinal data, was constructed by merging three different databases: (1) the logs of e-mail interactions within the university over one academic year, (2) a database of individual attributes (status, gender, age, department, number of years in the community, etc.), and (3) records of course registration, in which courses were recorded separately for each semester. For privacy protection purposes, all individual and group identifiers were encrypted (i.e., each person's e-mail address, each department name, and each course number were replaced with a random string of characters). Critically, however, common identifiers were used for the same individuals across databases; thus it is possible, for example, to tell if two persons with certain individual characteristics were in the same class together without knowing either the real names of the individuals or the class title.

The available variables could be categorized into four groups: personal characteristics (age, gender, home state, formal status, years in school); organizational affiliations (primary department, school, campus, dormitory, academic field); course-related variables (courses taken, courses taught); and e-mail-related variables (days active, messages sent, messages received, in-degree, out-degree, reciprocated degree).[6] As indicated in table 1, the population of 30,396 selected individuals is a mix of undergraduate students (21%), graduate and professional students (27%), faculty members (13%), administrators and staff (13.4%), and finally "affiliates" (25%)—a category that includes postdoctoral researchers, visiting scholars, exchange students, and recent alumni.[7] For each e-mail message sent within the university community we obtained the time stamp (in minutes

[5] There were 43,553 individuals who used university e-mail to both send and receive messages during the academic year. To make sure that our analysis was unaffected by population turnover, we identified 34,574 users who were active throughout both semesters (i.e., they sent and received e-mail in both the first and the last months of the academic year). Of those, 30,396 individuals exchanged messages with others in the subset throughout the year. We therefore selected these 30,396 active e-mail users as our population of interest.

[6] The precise definitions of all variables are provided in app. A, and a note about missing data appears in app. B.

[7] Although the university provides an option for alumni to forward e-mail to a different address indefinitely, the ability to send messages from the university account is typically terminated six months after graduation. Because we limited the population to individuals who both sent and received e-mail using their university address, only a small fraction of recent graduates are present in our data set.

TABLE 1
DISTRIBUTION OF INDIVIDUALS BY STATUS

	FALL		SPRING	
STATUS	*N*	% of Total	*N*	% of Total
Administrator	2,981	9.8	2,945	9.7
Affiliate	7,464	24.6	7,461	24.5
Faculty	3,956	13.0	3,970	13.1
Graduate	4,532	14.9	4,547	15.0
Instructor	378	1.2	386	1.3
Nondegree	109	.4	109	.4
Professional	3,519	11.6	3,528	11.6
Staff	1,135	3.7	1,126	3.7
Undergraduate ...	6,322	20.8	6,324	20.8
Total	30,396	100.0	30,396	100.0

since the start of data collection), the sender ID, and the IDs of all recipients of the message, extracted from the mail server logs and appropriately anonymized (the contents of messages were not recorded).[8] To ensure that the data represent interpersonal communication, we included only messages that were sent to a single recipient (other than the sender—i.e., excluding self-addressed e-mails)—a category that accounted for 82% of all e-mail.[9]

After we cleaned the data in this fashion, the resulting data set comprised 7,156,162 messages exchanged by 30,396 stable e-mail users during 270 days of observation.[10] Table 2 shows average values of attribute variables as well as e-mail volume, broken down by status—for example, an

[8] Some e-mail clients and servers split messages with long recipient lists and "blind carbon copy" e-mails into several messages with identical contents but different recipients. Such messages have the same sender and time-stamp and only differ in size, inasmuch as the respective recipient lists differ (Malmgren et al. 2008). To deal with such artifacts, we grouped simultaneous messages from the same sender that differed in size by less than 100 bytes and considered them to be instances of the same multirecipient e-mail.

[9] As described in the next section, in addition to e-mails sent to individual recipients, we also retained "bulk" e-mails, defined as having more than one recipient, which we used to infer the presence of shared groups and activities that were not otherwise recorded in our data. We have also repeated our analysis including e-mails with up to five recipients as interpersonal communication, with very similar results.

[10] Only e-mail accounts on the central university server were included in the data set. However, a number of individuals also used accounts provided by their departments, such as xyz@department.university.edu (mostly the faculty and graduate students in departments such as computer science, mathematics, and physics). Unfortunately, although we can tell that such addresses are part of the university community, they cannot be matched with employee records and therefore have been excluded from this analysis.

TABLE 2

MEANS AND SDS OF NUMERIC VARIABLES BY STATUS GROUP (Spring Semester, 133 Days)

	Age	Year	Gender*	From U.S.	Courses Taught	Courses Taken	Days Active	Messages Sent	Messages Received	In-degree	Out-degree
Administrator	42.79	.76	.64	.86	0	.12	60.56	284.17	297.04	51.64	54.55
	(11.70)	(1.33)	(.48)	(.35)	(0)	(.54)	(30.03)	(326.43)	(316.24)	(52.14)	(57.14)
Affiliate	33.59	1.01	.62	.53	0	.01	36.14	89.14	87.52	20.34	20.21
	(10.65)	(1.95)	(.49)	(.50)	(0)	(.20)	(29.70)	(126.49)	(124.56)	(22.24)	(23.50)
Faculty	45.97	.25	.40	.72	.06	0	48.67	170.12	172.63	32.89	33.55
	(12.47)	(.94)	(.49)	(.45)	(.26)	(0)	(34.85)	(256.27)	(240.68)	(35.57)	(38.20)
Graduate	29.59	3.41	.55	.43	0	2.00	42.40	105.06	103.87	21.54	21.29
	(6.17)	(2.10)	(.50)	(.50)	(0)	(2.12)	(30.45)	(136.83)	(138.31)	(20.00)	(21.51)
Instructor	39.23	1.68	.59	.57	.14	.77	45.02	147.07	150.35	29.58	29.9
	(10.05)	(2.17)	(.49)	(.50)	(.35)	(1.31)	(32.07)	(240.05)	(244.82)	(38.49)	(40.80)
Nondegree	36.59	.02	.53	.74	0	1.27	14.78	34.37	27.57	11.03	8.37
	(9.93)	(.19)	(.50)	(.44)	(0)	(1.56)	(18.87)	(78.14)	(74.31)	(28.36)	(18.68)
Professional	28.06	2.59	.52	.34	0	2.45	46.05	121.60	117.82	27.59	26.61
	(4.92)	(1.05)	(.50)	(.47)	(0)	(3.02)	(30.50)	(160.47)	(154.86)	(21.99)	(22.42)
Staff	40.88	.32	.69	.70	0	.04	33.51	100.06	107.34	23.72	24.04
	(12.56)	(.79)	(.46)	(.46)	(0)	(.33)	(26.53)	(168.62)	(177.54)	(32.44)	(33.24)
Undergraduate ...	22.41	2.85	.53	.75	0	4.37	39.00	85.32	81.26	24.89	23.97
	(5.04)	(1.26)	(.50)	(.43)	(0)	(2.21)	(26.46)	(120.00)	(139.85)	(23.88)	(23.98)
Min	16	0	0	0	0	0	1	1	0	1	0
Mean	32.60	2.80	.53	.57	0	1.50	42.80	124.90	124.90	27.20	27.20
Median	29	3	1	1	0	0	38	60	58	18	18
Max	87	7	1	1	2	15	133	6,449	6,424	723	501
Valid *N*	24,523	15,870	23,064	15,763	30,396	30,396	30,396	30,396	29,019	30,396	29,019

* Female = 1; male = 0.

average individual exchanged 250 messages with other people in the selected subset during the 133 days of spring semester, where faculty and administrators had the highest average number of contacts (out-degree) inside the community (34 and 55 contacts, respectively).[11] The average out-degree for undergraduate students (24 contacts) was somewhat lower than for faculty—a pattern that might be explained by the popularity of instant messaging among undergraduates—and for nondegree students, many of whom probably do not use the university e-mail as their primary address, it was lower still (eight contacts). Although all these numbers may seem unrealistically low, we note that we have included only interactions between members of the university population itself (i.e., excluding messages sent to, or received from, outsiders); thus the numbers represent only a fraction of total e-mail volume. Moreover, large standard deviations within all categories indicate wide variation in e-mail usage between individuals of the same status, in addition to the apparent differences between status categories.

Network construction.—E-mail exchanges comprise discrete and intermittent "spike trains" that are often "bursty" in nature (Cortes et al. 2003; Eckmann, Moses, and Sergi 2004). Although one could study the dynamics of these spike trains as a phenomenon in itself (e.g., Barabási 2005; Malmgren et al. 2008), here we treat them simply as the observable signature of an unobservable social network that is persistent and continuous and is also evolving in time. A number of techniques have been proposed for inferring social networks from e-mail data (see, e.g., Cortes et al. 2003; Eckmann et al. 2004; Moody et al. 2005); here we employ a simple but effective method known as a *sliding window filter*—a technique often employed to analyze and visualize networks over time (Cortes et al. 2003; Moody et al. 2005; Kossinets and Watts 2006). To see how the method works, consider an interacting dyad (i, j) and let $M_{ij}^{\tau}(t)$ be the number of messages sent from actor i to actor j during the time period $(t - \tau, t)$. We define the instantaneous strength, $w_{ij}(t, \tau)$, of tie (i, j) as

$$w_{ij}(t, \tau) = \frac{1}{\tau} \sqrt{M_{ij}^{\tau}(t) M_{ji}^{\tau}(t)},$$

which is simply the geometric average of the number of messages exchanged by users i and j per unit of time, summed over the past τ time units. The geometric average serves as a conservative measure of intensity:

[11] Some of the averages reported in table 2—such as Years under faculty and staff and Courses Taught for faculty–may appear implausibly low because of missing data; the numbers are most reliable for students. Tables C1–C3 in app. C provide a more detailed breakdown of e-mails sent and received between status groups.

dyadic strength is high if and only if both message counts $M^{\tau}_{ij}(t)$ and $M^{\tau}_{ji}(t)$ are high, and it is low if either message count is low.[12]

Figure 1 illustrates the sliding window filter, showing the spike-train representations of e-mail exchange for two hypothetical dyads. Consider dyad (i, j) in the upper spike train. The spikes above and below the horizontal axis, respectively, represent messages from one individual to another and in the opposite direction. A window of length τ "slides" along the time axis in discrete steps of length δ, meaning that the edge (i, j) is active at time t if and only if at least one message has been sent in both directions within the past τ time units. By extension, the instantaneous network at time t includes all dyads with nonzero strength or, equivalently, all dyads that have exchanged messages within the interval $(t - \tau, t]$. Network approximations for times t_1 and t_2 are shown under the spike trains in figure 1: at time t_1, only dyad (i, j) has exchanged messages within the past τ time units, whereas at time t_2, both dyads (i, j) and (j, k) are active. Any given "reconstruction" of a network from a sliding window filter therefore depends on two critical parameters, τ and δ, which we estimate as follows.

In estimating a suitable value for the parameter τ, we first note that it determines, in effect, the "relevancy horizon" of past interactions—that is, the maximum time at which a past interaction is assumed to contribute to the current strength of relationship.[13] Which particular value of τ is chosen will in general depend on the substantive question of interest—for example, if we were interested in modeling the spread of information over some time scale T (e.g., a few days), we would want to set $\tau < T$ so as not to treat the network as static when it is, in fact, changing on the timescale over which the information is spreading. Because the question of interest here concerns the tie-formation process itself, we must choose τ so as to distinguish between ties that are forming as we observe the network and ties that have existed before and simply resumed communicating. That is, if two individuals have not communicated for some period of time, they may still maintain an ongoing relationship; but if two individuals do in fact cease communicating permanently, then it is arguably no longer meaningful to treat them as linked. Therefore, the value of τ that is chosen should not be too short, or some ongoing rela-

[12] Clearly, other definitions of tie strength are possible (e.g., the algebraic mean, which would assign high strength whenever one individual frequently e-mailed the other), where the appropriate choice would ultimately depend on the research question of interest.

[13] One can also use the so-called exponentially weighted moving average (EWMA) technique to progressively discount past interactions (Cortes et al. 2003); however, for the purposes of identifying the events of tie formation and dissolution EWMA is equivalent to the simple moving window method that we use.

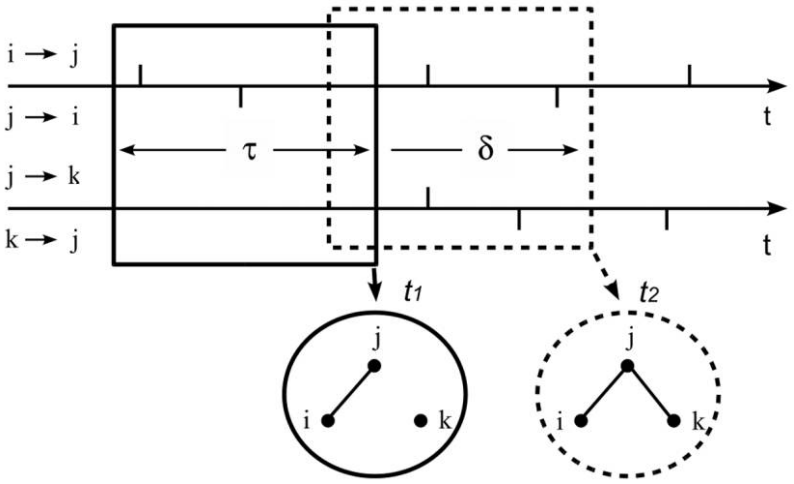

FIG. 1.—A sliding window filter to construct instantaneous network approximations from discrete dyadic interactions.

tionships will be misclassified as ties that have been terminated and then reenacted. Yet τ should not be set too long either, or the calculation of relationship strength will be dominated by the past interactions (including one-off interactions) that are no longer relevant to the present state of the relationship. To account for left-censoring of the data, moreover, ties that are first observed within τ days of the onset of our data collection cannot be classified as "new" (because they may have been present and just not active); thus, longer values of τ also have the effect of discarding more data. Balancing these conflicting priorities, therefore, we have chosen $\tau = 60$ as a reasonable compromise value that correctly classifies 90% of terminating ties, while retaining as much data as possible.[14]

In addition to τ, we also need to determine the *sampling period*, δ, which determines whether events separated in time will be treated as sequential or as simultaneous with one another. As an illustration of this point, consider figure 2. Suppose that person A has two friends, B and C, and that C has two other friends, D and E. Now imagine a chain of events: (1) A and B meet C in a cafeteria one morning and A introduces B to C; (2) B and C meet for dinner the next day and C introduces B to D and E. There are three new ties created as a result (BC, BD, and BE). If we measure the network each day, we would capture the fact that event 1 preceded event 2 and would correctly determine that all three new ties

[14] As a robustness check, we have also performed our analysis for $\tau = 30$ and $\tau = 90$ days and found similar results.

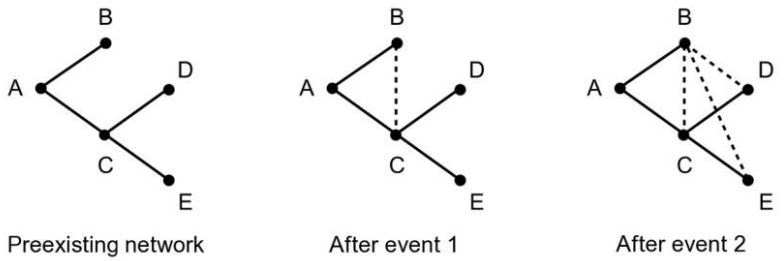

FIG. 2.—Cycle formation and the choice of sampling period

were results of triadic closure (i.e., meeting a friend of a friend). However, if we sample the network for changes less frequently—say, only once a week—all events would appear to be simultaneous and the changes would be classified as the formation of one cycle of length 3 (ABC) and two cycles of length 4 (ABDC and ABEC). Choosing too long a δ may therefore result in incorrect inferences about network processes (e.g., concluding that 4-cycles are forming when really only 3-cycles are forming). Sampling the network too frequently, however, becomes computationally unfeasible. Some measures of tie formation, for example, such as cycle closure, require computing all shortest paths in the network at every time point; thus, although in principle all calculations could be performed at the highest time resolution of our data ($\delta \approx 1$ minute), it would be wildly inefficient to do so, especially given long periods of low activity, such as nighttime. To relax the requirement of a perfect representation, we therefore calculate the median tie-formation rate, which yields $\delta_{1/2} \approx 27$ hours. We conclude that although there are periods of high rates of tie formation in the data, sampling the network for structural changes once every day produces a reasonable approximation, taking into account the natural periodicity of human activities.[15]

Applying the moving window procedure for $\tau = 60$ days and measuring the network with resolution $\delta = 1$ day we obtained 210 sequential network snapshots, which span the second half of the fall semester and the entire spring semester (by definition, the first τ days are used to approximate the network as it existed before our data collection began; thus, our dynamics "starts" at day $\tau + 1$). Descriptive statistics of the resulting 60-day average network are very similar to those reported in our previous

[15] We have checked this conclusion by comparing time series of network averages (mean and median degree, median path length, clustering coefficient) obtained with $\delta = 1$ hour and $\delta = 1$ day, for $\tau = 30$, 60, and 90 days, and all produced qualitatively similar results with respect to the properties of the approximated network.

study (Kossinets and Watts 2006) and are characteristic of those for other large social networks (see, e.g., Leskovec and Horvitz 2008). For example, the largest connected component typically occupies most of the population, varying between 93.7% and 99.0% of the 30,396 nodes (average = 95.6%), and within the giant component, all nodes can be reached from all other nodes in an average of $4 \leq d_{ij} \leq 5$ steps. The network, moreover, is very sparse, displaying an average degree varying between 13.0 and 17.5 (average = 15.9) with a skewed degree distribution.[16] Finally, the clustering coefficient (Newman, Watts, and Strogatz 2002) varies between 0.09 and 0.10 (mean = 0.096), which, in other words, is more than 2,000 times the expected value for a random graph of equivalent density (i.e., $C_{rand} \sim k/N < 15/30{,}000 = 0.0005$). As we argue in the next section, however, our main concern in this article is not structural measures of the network per se, but rather the dynamics of tie formation that drive its evolution.

NETWORK EVOLUTION

Because our primary interest in this article is to understand how homophily emerges over time as a function of the decisions of individuals to make and break ties, our focus is largely on the formation of new ties, as well as to a lesser extent on the dissolution of existing ties—that is, on processes of network evolution rather than network structure itself. To model the evolution of our network, we study in detail two kinds of tie formation mechanisms: *cyclic closure* and *focal closure* (Kossinets and Watts 2006). Cyclic closure is premised on the theoretical notion of transitivity (Rapoport 1953; Holland and Leinhard 1971), which suggests that if two individuals are connected to a mutual third party, they will tend to become connected themselves, as illustrated in part A of figure 3. Because such a process results in the formation of cycles of length 3, or "triads," it is often called *triadic closure* or *triad completion* (Rapoport 1953; Banks and Carley 1996).[17] Cyclic closure, therefore, is a generalization of triadic closure, in which cycles of length greater than 3 can also

[16] As with the total number of e-mails sent per person, this range of average degree may seem low; i.e., one might expect that a typical student or university employee would communicate with a larger number of people on a regular basis. These values, however, are in fact reasonable because they only include e-mail interactions within the selected subset of consistent e-mail users that were reciprocated within 60 days (some e-mails may have been reciprocated by other types of communication, including face-to-face).

[17] We note that what we have defined as triadic closure is often called *transitive closure* in the social networks literature (Wasserman and Faust 1994). Because we wish to distinguish between the transitive closure of cycles of different lengths, we refer to the process of closing an incomplete triad as triadic closure.

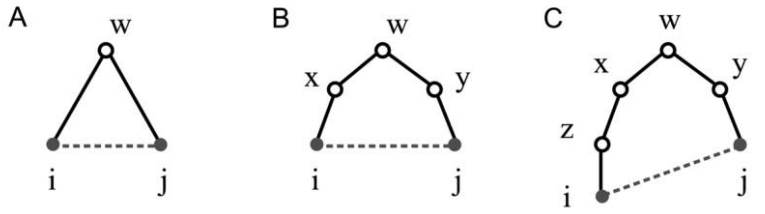

FIG. 3.—Tie formation as closure of network cycles. Part A depicts triadic closure; B illustrates focal closure of a long cycle; C shows closure of a long cycle as a chain of referrals (strategic search).

be formed in evolving networks as a consequence of social processes that operate over longer ranges. As shown in part B of figure 3, for example, i and j may also form a tie even when neither is acquainted with w, as might occur at some group activity organized by w to which the invitees (here, x and y) are asked to bring their own friends. The acquaintanceship (i, j) can then be made, with or without the assistance of any intermediaries, thereby closing a cycle of length 4. Longer cycles still (fig. 3, pt. C) may be formed as the result of a chain of referrals—representing what has recently been dubbed "strategic search" (Kleinberg and Raghavan 2005).

Focal closure, by contrast, follows from an alternative theory of tie formation—that of "social interaction foci" (Feld 1981), which are defined as the various groups, contexts, and activities around which social life is organized and which in turn facilitate interpersonal interactions. In a university setting, class attendance provides essential opportunities for face-to-face interaction between students, and therefore we treat all officially recorded classes as social foci. Because all courses (3,537 in the fall semester and 3,141 in the spring) are recorded explicitly in our data and we know when any pair of individuals shared a class affiliation, we refer to these opportunities as *explicit foci*. In most cases classes are shared between students of the same status (undergraduate or graduate), but it is also possible that graduate students can take undergraduate courses, and vice versa, leading to connections across different groups within the university community. Faculty are also associated with classes, primarily as instructors, where graduate students (and occasionally even undergraduates—e.g., in physical education classes) may serve as instructors too. Finally, staff members are eligible to take classes as well, although they do not take nearly as many as full-time students. Thus classes, in principle, may serve as important forums for interaction for most subgroups.

Classes, however, are certainly not the only foci of interaction, even for students, and for many members of the university community, including most staff and even some faculty, they are probably not the most important. Ideally, therefore, we would like to have a record of all possible focal activities—not only classes, but also social groups, sporting and cultural organizations, shared housing, and so forth—so that we could study separately their effects on social interactions over time. Although one can easily imagine a database in which all student and nonstudent groups are explicitly recorded and regularly updated, possibly even in real time, our data set explicitly codes only for classes administered by the university registrar. Fortunately, it is possible to overcome this practical obstacle in part by mining the available data in more creative ways. Specifically, we make use of the "bulk" messages that we discarded earlier (when constructing the network of dyadic interactions), treating them as indicators of social foci, defined broadly as any kind of shared affiliation, group, or activity that generates a demand for group-oriented communication. Because these social foci are inferred indirectly from the e-mail communication patterns and not recorded explicitly, we call them *implicit foci*. We note that implicit foci are considerably more general in scope than classes, including students and nonstudents alike, and may in fact represent either of two distinct kinds of groups. First, they may signal the presence of organized groups—for example, sporting clubs, seminar series, or student associations—that facilitate regular opportunities for interaction in the same manner as classes and departments, but that are simply not recorded in the available data. And second, they may represent what Mayer (1966) has called "quasi groups"—collectives such as alumni clubs or electronic mailing lists—which, although not implying regular face-to-face contact in the manner of social foci, nevertheless connote a shared group identity and therefore may facilitate interaction between their members when some other opportunity or a need for direct communication arises.

One might expect these two kinds of groups to have different effects on tie formation, and for this reason it would be preferable to identify them separately. However, the nature of our data does not permit us to do so, and hence we treat all implicit foci as indistinguishable. Furthermore, in contrast with our data on explicit foci, where distinct classes have unique identifiers, there are no explicit labels associated with different implicit foci. Therefore we quantify the "strength of shared membership" for every pair of individuals i and j simply as the number of times, g_{ij}, that i and j appear together among the recipients of a "bulk" message (i.e., a message with more than one recipient) and compute this quantity separately for each semester. Because g_{ij} may vary across a range of values, there is no theoretical equivalent to "sharing a class," which is a binary distinction. As figure 4 shows, however, the empirical cumulative

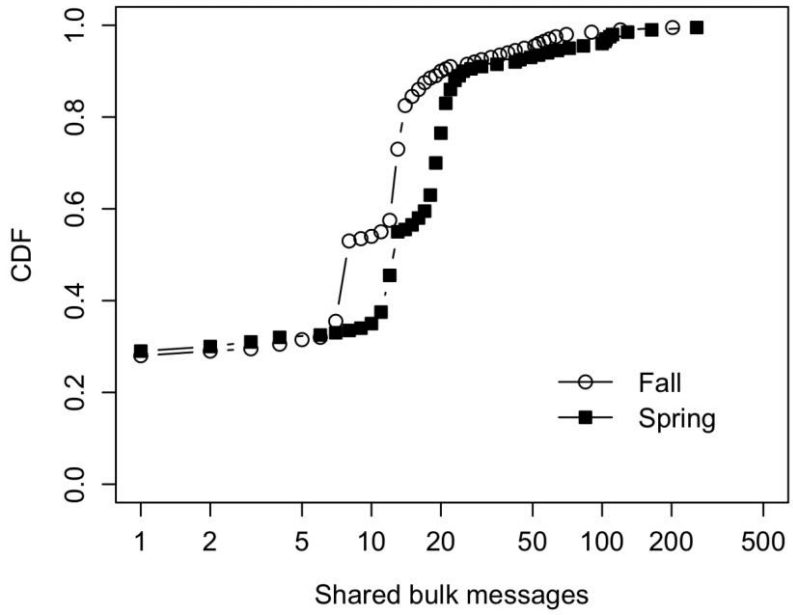

FIG. 4.—Cumulative distribution function (CDF) of shared bulk messages over all pairs of individuals, for fall and spring semesters.

distribution of g_{ij} is strongly S-shaped, showing little growth between one and five or above 20 bulk messages, but rapid change in the interval between five and 20 messages. An advantage of S-shaped curves like this one is that they define natural "threshold" conditions, which can be used to transform continuous variables into binary ones (i.e., either above or below the threshold value). We therefore define a new dummy variable, $q_{ij} \in \{0, 1\}$, such that a pair of individuals i and j shares an implicit focus ($q_{ij} = 1$) if and only if the number of bulk messages jointly addressed, g_{ij}, exceeds some critical value g_*, and not otherwise ($q_{ij} = 0$). In this manner, we effectively divide the population into "strongly" and "weakly" related pairs in a way that is analogous to the explicit condition of pairs that share "at least one class."[18]

[18] Clearly it would be desirable to be able to separately count multiple implicit foci in the same way that we distinguish between multiple classes; however, doing so raises a number of conceptual and technical difficulties associated with inferring communities from equivalence measures and matching them over time. Blockmodeling (White, Boorman, and Breiger 1976) or, alternatively, a suite of recently proposed partitioning methods (Girvan and Newman 2002; Moody and White 2003; Palla, Barabási, and Vicsek 2007) may be useful in this regard, but these methods also introduce serious computation and interpretability issues that remain to be resolved.

Even after we accept the notion of shared implicit foci defined in terms of a threshold value g_*, our one remaining problem is how to choose g_* appropriately—for example, setting $g_* \approx 20$ would count only pairs above the 90th percentile, whereas $g_* \approx 10$ would be equivalent to the 50th percentile. Rather than choosing between these possibilities arbitrarily, we consider instead the effects on tie formation of shared implicit foci for different values of the threshold g_* and compare them with the corresponding effect of explicit foci. As shown in figure 5 (solid diamonds), the overall probability of new ties forming decreases approximately exponentially with network distance and stabilizes for distances greater than 5 at a value 2,500 times less than that for individuals who share an acquaintance ($d_{ij} = 2$).[19] Given that the majority of individuals in our network do not share any explicit foci, the likelihood of a tie forming via triadic closure—that is, between two individuals with a mutual acquaintance—is on average roughly 30 times greater than when they are removed only one step further. Strikingly, however, when two students attend one or more classes together (i.e., they share an explicit focus), the maximum probability of tie formation (fig. 5, circles) increases by a factor of roughly 50, and it remains high even for longer distances. Sharing an explicit focus, in other words, clearly exerts a tremendous impact on the likelihood of a tie forming—with a magnitude comparable to that of sharing a friend.

Next, we observe that a qualitatively similar pattern holds for sharing an implicit focus, but only for sufficiently high cutoff values of g_*. Specifically, defining an implicit focus at the level of the 50th percentile of g_{ij} produces a curve (fig. 5, open squares) similar to that for all pairs. Yet for increasingly restrictive cutoff values—for example, the 95th and 99th percentiles (fig. 5, open triangles and inverted open triangles, respectively)—the probability of new tie formation for shared implicit foci increasingly resembles that for shared explicit foci. We further observe that pairs above the 99th percentile of shared bulk messages (equivalent to $g_* \approx 140$) are on average 55 times more likely to form a tie than pairs below the 99th percentile—a relationship that closely matches the effect of sharing a class for student pairs. For purposes of subsequent analysis, therefore, we say that two nodes share an implicit focus if they received at least 140 of the same bulk e-mails per semester ($q_{ij} = 1$; 0 otherwise), corresponding to the 99th percentile of the number of jointly received bulk messages. Aside from establishing a natural calibration of implicit foci, the equivalence relation between implicit and explicit foci is extremely helpful for our analysis, as implicit foci are not biased toward

[19] The average probability of a tie forming between two nodes at distance $d_{ij} = 2$ is about 30 times greater than that for a pair at $d_{ij} = 3$; the corresponding probability for a pair at $d_{ij} = 3$ is 10 times that for a pair at $d_{ij} = 4$; and so on.

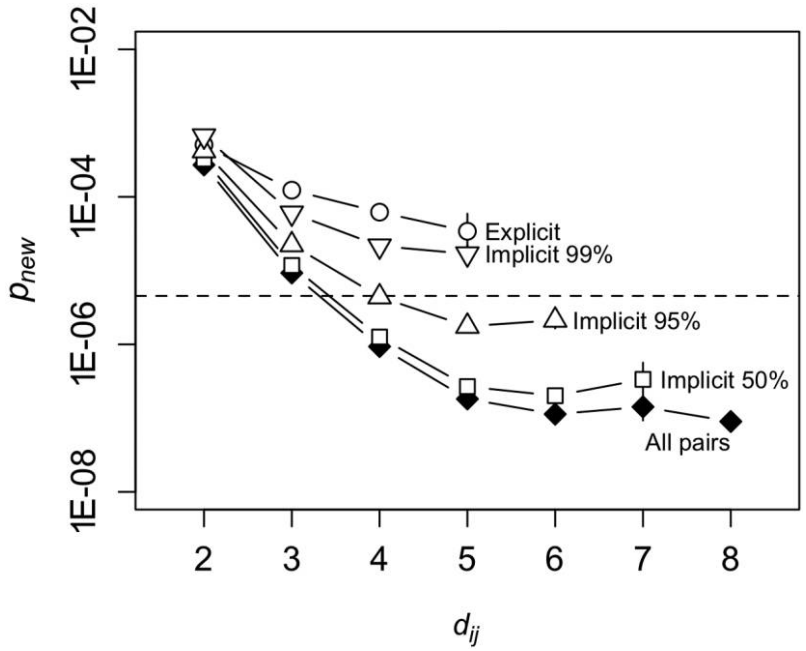

FIG. 5.—Estimated average daily probability of new tie as a function of network distance. Bootstrap 95% confidence intervals are shown unless smaller than symbol size. Where there are fewer than 50 observations for some values of $d_{ij} > 5$, data points are not displayed. The dashed line indicates the average probability of new tie formation.

students in the same way that classes are; thus, by using implicit rather than explicit foci as an indicator of structural proximity, we can extend our analysis to the entire population, rather than being constrained to focus on students.[20]

THE ORIGINS OF HOMOPHILY

As we have discussed, disentangling the individual and structural origins of some observed pattern of homophily requires dynamic data in which both network interactions as well as social and organizational foci are recorded for the same population over time (McPherson et al. 2001). Our data, comprising both social foci and networks over two semesters, are

[20] More generally, the use of implicit foci as indicators of structural proximity may be particularly helpful in the analysis of e-mail-based data sets, which frequently do not code what we have called explicit foci, but usually do retain multirecipient messages (see, e.g., Ebel, Mielsch, and Bornholdt 2002; Eckmann et al. 2004).

therefore well suited to this address this problem. In particular, a major advantage of this data set over those used in previous studies of homophily is that, whereas most studies have relied on egocentric samples, here we have the complete network. As a result, not only can we compare adjacent with nonadjacent pairs, but we are also able to study the full functional dependency of similarity with network distance (d_{ij}) between every pair of nodes at every point in time. Our data set also codes for several individual attributes that may represent different dimensions of homophily (see app. A), in particular *gender*, *age*, *status*, *field, year*, and *state* (this last we convert to a *from U.S./foreigner* dichotomy in order to obtain a more balanced distribution).[21] Because we find a positive homophily effect on tie formation with respect to each of these variables (i.e., ties are more likely to form between people of the same gender, similar age, etc.), we introduce an aggregate measure of pairwise similarity, S_{ij}, which is computed for each pair (i, j) as the number of matches over the six individual attributes named above.[22] Aggregate similarity therefore varies between 0 and 6, where $S_{ij} = 0$ corresponds to a pair with no common attributes and $S_{ij} = 6$ implies identical attributes.[23]

Observed homophily.—As described in the previous section, we prepared 210 daily, undirected network snapshots for days 60–270 (where two consecutive snapshots overlap by 59 days). With each snapshot, we computed the following quantities for all pairs of individuals in the network: (*a*) network distance (shortest path length), d_{ij}; (*b*) the number of shared bulk messages, g_{ij}, and the corresponding implicit focus indicator, q_{ij}; and (*c*) the number of jointly attended classes (for student pairs only) in the current semester, c_{ij}.[24] In figure 6, we plot aggregate similarity as a function of "structural proximity," which is represented by network

[21] We note, however, that our data do not code for some dimensions of homophily, such as race or economic status, that are clearly of interest to sociologists and that have been the focus of previous studies of homophily (McPherson and Smith-Lovin 1982; Marsden 1987, 1988; Ibarra 1995; Louch 2000).

[22] There are many ways to compute similarity between two sets of attributes (factor analysis, cosine similarity, etc.); we have chosen to employ a simple additive scale for ease of interpretation.

[23] When either or both individuals in a pair have a missing value for a particular variable, instead of throwing away the pair, we use the population mean for the corresponding similarity-scale component. For example, if ages in a particular pair *(i, j)* are, say, 24 and 25, then according to our definition *age match* $(i, j) = 1$. If j's value is "missing," we will assign *age match* $(i, j) = 0.175$ because 17.5% of all pairs are of the same age. Then we add up the results of matching other variables to obtain S_{ij}. This approach is arguably the simplest imputation technique that alleviates the problem of nonrandom missing values (e.g., nonstudents having more missing values than students).

[24] Precisely, this quantity is computed for all pairs of students who have taken a class in each given semester.

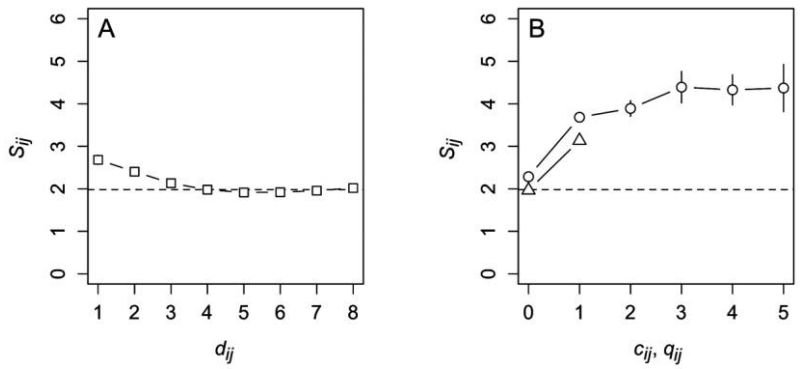

FIG. 6.—Average pairwise similarity (S_{ij}) as a function of (A) network distance and (B) number of shared classes (circles) or shared implicit foci (triangles). The dashed line shows the overall average.

distance, on the one hand (pt. A), and number of shared classes or, alternatively, shared implicit foci, on the other (pt. B). To begin with, we observe that adjacent pairs exhibit nearly 40% higher similarity than the population average, thus confirming many previous results showing that acquaintances are more similar than strangers and alleviating concerns that the particular population in question may be too homogeneous to show a measurable pattern of homophily. In fact, part A of figure 6 also shows that similarity is not only lower for nonfriends than for friends, but decreases monotonically with distance from $d_{ij} = 1$ to $d_{ij} = 4$, where it approaches the population average. In other words, the usual result that friends are more similar than strangers can be seen as a special case of a more general rule that individuals who are "close" are more similar than those who are "distant."

As part B of figure 6 shows, the same general rule also applies to our other measure of structural proximity: first, because individuals who share either explicit or implicit foci are significantly more similar than those who do not, and second, because individuals who share multiple explicit foci (i.e., classes) are increasingly similar. The striking impact of shared foci on similarity, however, also raises a potential concern with respect to the similarity measure itself—namely, that for students, a number of its components are highly correlated with choice of classes. One might therefore be concerned that our measure of individual similarity acts, in effect, as an indicator variable for sharing a class, and that controlling for shared classes (as we do later) would effectively eliminate the potential for similarity to have any additional impact on tie formation, thereby artificially increasing the apparent importance of induced homophily vis-à-vis choice

homophily.[25] To address this potential systematic bias in our data, we consider in figure 7 (top row) the distribution of similarity for student pairs who shared classes with that for student pairs who did not. As expected from figure 6, students who shared classes (fig. 7, pt. B) are, on average, much more similar than students who did not (pt. A). However, its higher average notwithstanding, the distribution in part B of figure 7 also exhibits higher variance (1.8) than that in part A (1.3); thus, the potential for differences in similarity to impact tie formation is not in fact diminished for pairs who share classes versus those who do not. As a further check we compare distributions of similarity for pairs who share implicit foci (fig. 7, pt. D) with those who do not (pt. C). As before, we find that the distribution of similarity for individuals sharing an implicit focus is higher than that for all individuals. We also find that the relative increase in variance for sharing implicit foci is even higher (1.8 in pt. D vs. 0.6 in pt. C) than that for explicit foci. Since we find that implicit foci, when properly calibrated, are qualitatively similar to explicit foci, and also because implicit foci apply to the entire population rather than just to students, we use implicit foci as our primary measure of shared affiliation, whenever appropriate.[26]

Effect of similarity on tie formation.—Having established the presence of homophily in our population, we turn now to our main interest of understanding its origins, which, as emphasized earlier, requires us to restrict our attention exclusively to the formation of new ties. Specifically, we identify all tie-formation events in our data by comparing the network on day $t - 1$ and day t, for $t = 61 \ldots 270$, and then estimate the impact of similarity on the probability, $p_{new}(i, j)$, that a new tie will form between nodes i and j, averaged over the entire time interval, fitting logistic regressions of the general form $\log[p/(1 - p)] = b_0 + b_1 S_{ij} + \varepsilon$, where b_1 is the coefficient of interest.[27] To control for the effects of other covariates of interest (in particular, distance and shared implicit foci) we fit the model separately to different subsets of our data—for example, to the subset of nodes that share an implicit focus and that are also at distance $d_{ij} = 2$. Fitting a model with a single covariate to multiple subsets allows us to interpret our results more easily than if we were to estimate multiple covariates simultaneously for a single model.[28]

[25] We are grateful to an *AJS* referee for pointing out this potential problem.

[26] We have checked our results for explicit foci as well and recorded similar findings.

[27] We note that in principle, because of our sampling procedure, a link can be considered new multiple times if it dissolves and forms again. Fortunately, however, we have chosen the parameter τ precisely to avoid such cases; hence, they occur only rarely.

[28] We have, however, fit a single model with all covariates and checked that we obtain comparable results to those we report.

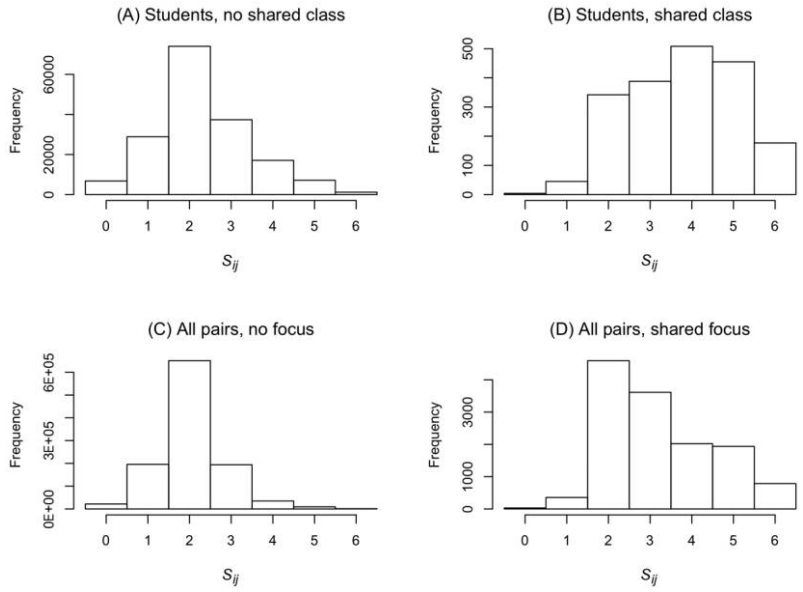

FIG. 7.—Distribution of pairwise similarity (S_{ij}) for pairs sharing explicit and implicit foci

In particular, for any given subset we can simply exponentiate the estimated value of b_1 to obtain the odds ratio

$$\frac{P[\text{new tie}|S_{ij} = 1]}{P[\text{no tie}|S_{ij} = 1]} \Big/ \frac{P[\text{new tie}|S_{ij} = 0]}{P[\text{no tie}|S_{ij} = 0]},$$

which can in turn be rearranged as

$$\frac{P[\text{new tie}|S_{ij} = 1]}{P[\text{new tie}|S_{ij} = 0]} \Big/ \frac{P[\text{no tie}|S_{ij} = 1]}{P[\text{no tie}|S_{ij} = 0]}.$$

Because new tie formation occurs so rarely (on average, just one in 200,000 pairs of nodes not connected at present will have been connected in the next snapshot), the term $P[\text{no tie}|S_{ij} = 1]/P[\text{no tie}|S_{ij} = 0] \approx 1$ can be factored out of the above expression; thus the odds ratio can interpreted roughly as the relative risk

$$P[\text{new tie}|S_{ij} = 1]/P[\text{new tie}|S_{ij} = 0],$$

which is just the relative change in the probability of an event corresponding to a unit change in S_{ij}. Thus, for any set of nodes, we can easily compute the impact on tie formation of increasing similarity in a way that is intuitive and also easily comparable across different conditions.

As noted above, tie formation is a rare event in our network; thus, we employ the case-control sampling methodology (King and Zeng 2001, 2002). For each type of transition that we study, and for each daily network snapshot, we determined the appropriate "risk sets" of pairs of nodes that could experience the transition (for tie formation, all pairs not currently connected; for tie dissolution, all current ties). We then sampled a total of about 100,000 *cases* (pairs that actually experienced the transition) over the 210-day period as well as twice that number of *controls* (pairs from the risk set that did not experience the transition of interest—e.g., for the dissolution example, ties that did not dissolve).[29] Using the case-control weights, we can compute transition probabilities and fit logistic regression models efficiently (King and Zeng 2002).[30] Table 3 shows the set of models that predict the daily probability of tie formation.

Model 1 shows that similar individuals are far more likely to become acquainted than dissimilar individuals; specifically, the odds ratio of 1.9, which is highly significant, implies that the average tie-formation rate for a highly similar pair ($S_{ij} = 6$) is $1.9^6 \approx 50$ times that for a highly dissimilar pair ($S_{ij} = 0$) and about 13 times that for a pair with average similarity ($S_{ij} = 2$). We emphasize that this effect is not merely highly significant in a statistical sense, but also extremely large—more than 1,000% for highly similar pairs of individuals, compared with average pairs. Model 1 therefore appears to provide strong support for the "individualistic" argument, made in the introduction, that acquaintances are more similar than strangers because individuals preferentially select similar others when forming new ties. As we have also shown, however, almost all new tie formation

[29] Because we sampled independently from each daily network snapshot, it follows that the same pair could be drawn multiple times as long as it stayed "at risk." Some nodes were more likely to be part of the dyads that experienced transitions or were in the risk set day after day and hence might be overrepresented in our case-control samples. Furthermore, because the number of dyads at a given distance d increases approximately exponentially with d up to $d = 5$ (and then tapers off due to the finite network size), the risk set for the first and second transition types included practically all node pairs, and thus every node had a roughly equal chance of being selected as part of a pair, given the selection probabilities for cases and controls. But when sampling from the tie-dissolution process, the higher-degree nodes were more likely to be present in the risk set of existing dyads and therefore more likely to be selected as part of a case or control dyad. To correct for the unequal node-selection probabilities, we tried using robust standard errors and additionally weighting every pair inversely proportionally to the product of the respective nodes' frequencies in the sample. Neither of these modifications affected the results.

[30] We do not apply the additional adjustment for case-control sampling ratio (King and Zeng 2001), in part because we are more interested in the effects of similarity on the probability of events (measured by the corresponding odds ratios) than in estimating event probabilities per se, and in part because the adjustment—as implemented in the Zelig package (Imai, King, and Lau 2007)—is computationally intensive and takes a very long time with our data.

TABLE 3
DAILY PROBABILITY OF TIE FORMATION AS A FUNCTION OF PAIRWISE SIMILARITY

Model	Subset	N pairs	β (SE)	P	Odds Ratio
1	All pairs	312,440	.638 (.004)	.00	1.89**
2	$q_{ij} = 1$	44,017	.028 (.014)	.05	1.03^{+}
3	$d_{ij} = 2$	68,931	.234 (.018)	.00	1.26**
4	$d_{ij} = 3$	50,325	.295 (.008)	.00	1.34**
5	$d_{ij} = 4$	83,559	.495 (.010)	.00	1.64**
6	$d_{ij} \geq 5$	109,625	.704 (.017)	.00	2.02*
7	$d_{ij} = 2$; $q_{ij} = 1$	32,805	.059 (.034)	.08	1.06^{+}
8	$d_{ij} = 3$; $q_{ij} = 1$	8,381	.031 (.021)	.14	1.03
9	$d_{ij}=4$; $q_{ij} = 1$	2,009	.100 (.031)	.00	1.11*
10 ...	$d_{ij}\geq 5$; $q_{ij} = 1$	822	.232 (.055)	.00	1.26*

NOTE.—Average probability of tie formation is 5×10^{-6}. Pairs were sampled independently from each daily network snapshot (i.e., a pair could be drawn multiple times if still "at risk").

$^{+}$ $P < .10$.

* $P < .05$.

** $P < .01$.

takes place between individuals who are structurally proximate (fig. 5), and individuals who are structurally proximate tend to be similar (fig. 6). Clearly, therefore, one might suspect that at least some of the effect observed in model 1 can be attributed simply to the effects of structural proximity. We test this hypothesis in two ways, corresponding to our two measures of structural proximity.

First, model 2 controls for shared implicit foci (which, recall, correspond to pairs of individuals who are corecipients of at least 140 bulk e-mails per semester). As indicated in table 3, we find that essentially all of the effect in model 1 can be accounted for in terms of shared foci; that is, when only individuals who share an implicit focus are considered, increasing similarity has no impact on tie formation (i.e., the odds ratio drops from 1.9 to roughly 1 and has low significance). We emphasize, moreover, that this result cannot be dismissed simply with the reasoning that individuals sharing foci show no tendency to connect to similar others simply because they have no opportunity to—as described above, and shown in figure 7, the variance in similarity within foci is on par with the population variance. Rather, what model 2 suggests is that the higher average similarity of easily available options is the primary determinant of who forms a tie with whom—not a strong preference for similar others among the alternatives available.

Next, in model 3, we examine the effect of similarity on new tie formation when individuals share a mutual acquaintance ($d_{ij} = 2$). Unlike for shared foci, here we find that the effect of similarity remains positive

and highly significant; however, we note that it is weakened considerably. In quantitative terms, the odds ratio drops from 1.9 to 1.3, meaning, in effect, that highly similar pairs are now about only four times as likely to connect to highly dissimilar pairs, compared with a factor of 50 in model 1, and only about 2.5 times as likely as average pairs, compared with a factor of 13 in model 1. Thus, although similarity continues to play an important role in new tie formation even when it is brokered by a mutual acquaintance, once again the restricted opportunities afforded by structural proximity appear to account for much of the process of selecting alters. Consistent with this interpretation, we also find (in models 4–6) that as network distance increases, the importance of similarity reemerges—in particular, for "distant" pairs ($d_{ij} \geq 5$), the effect of similarity is roughly the same as in model 1—a result that we also find when both network distance and shared foci are accounted for simultaneously (models 7–10).

The overall message of table 3, in other words, is that although similar pairs are more likely to connect to each other, much of this effect is accounted for by the tendency of similar pairs to be structurally proximate either in terms of shared foci—a result that is in agreement with some previous findings (McPherson and Smith-Lovin 1987)—or in terms of shared acquaintances. The strong effect of similarity in model 1, in fact, appears to be driven by the large number of distant individuals (more than half of all pairs are at distance $d_{ij} \geq 4$ and do not share a focus), for whom choice homophily does appear to play a role—for example, at $d_{ij} = 4$, highly similar pairs are roughly 20 times more likely to form ties than completely dissimilar pairs.[31] Individuals, therefore, are not indifferent to similarity—in the absence of readily available opportunities to interact, they do indeed seem to flock to similar others—but their actual choices appear to be strongly determined by structural factors, and under those constraints, their preference for similar others does not appear to strongly affect their choices.

Effect of similarity on selection to risk sets.—It appears that much of the homophily observed in our population can be attributed to the homogeneity of opportunities within the limited population of alters with whom any given individual A is "at risk" (i.e., has a high probability) of forming a tie—in particular, individuals with whom A shares either a social focus or a mutual acquaintance. However, as we have discussed, these "risk sets" are themselves the products of previous choices that A

[31] We note, however, that the actual rate of tie formation between these pairs is so low (roughly one in 10^7; see fig. 5) that relatively few ties are ever formed in this way.

has made.[32] If, as seems plausible, A has made these choices precisely in order to enter a risk set that is similar to himself—for example, joining the "right" groups, or spending time with an acquaintance whom he knows to be friends with the right people—then one might conclude that what seems like induced homophily may in fact reflect individual preferences with respect to potential alters after all. Conversely, if A's selection of risk sets is itself determined mostly by the opportunities available to him at some earlier point in time—in effect, his preexisting contacts and memberships—then one would conclude that indeed his choices of similar others are induced by the structure of which he is a part.

We now investigate this question by examining the origins not of tie formation itself—our focus in table 3—but of the formation of two types of risk set: first, the set of individuals at $d_{ij} = 2$, and second, the set of individuals sharing social foci. Although we could consider other risk sets as well (e.g., the set of nodes at $d_{ij} = 3$), we observe that roughly 60% of all new ties are formed via triadic closure, and 30% are formed via focal closure (noting that these mechanisms are not mutually exclusive); thus, any bias in the creation of these sets will in turn exert considerable influence over any subsequent tie formation. First, we examine the effect of similarity on pairs of individuals transitioning from distances greater than 2 (where they are at low risk of forming a tie with each other) to the "friend-of-a-friend" (FoaF) risk set ($d_{ij} = 2$), in which their probability of forming a tie rises dramatically; and second, we examine the corresponding effect on pairs joining the "shared focus" (SF) risk set.[33] In both cases, as before, we fit logistic regression models of the form $\log[p/(1 - p)] = b_0 + b_1 S_{ij} + \varepsilon$ and then exponentiate the estimated coefficient b_1 to obtain approximate relative risk associated with a unit increase in S_{ij}.

Considering first the FoaF risk set, table 4 indicates that the probability, $p_{FoaF}(i, j)$, of a pair transitioning from $d_{ij} > 2$ to $d_{ij} = 2$ varies with S_{ij} in the same fashion as p_{ij} in table 3. On the one hand, model 11 shows that p_{FoaF} increases with each unit change in S_{ij} roughly by a factor of 1.6 on average, implying that not only are similar individuals more likely to

[32] The term "risk set" reflects its origin in the epidemiological literature. When discussing social tie formation, it may be more appropriate to speak of "opportunity sets," and correspondingly, when discussing tie dissolution, of "risk sets." We have chosen to use the term "risk set" throughout for consistency.

[33] Although we are interested in essentially the same question with respect to both risk sets, the structure of our data requires us to analyze them in slightly different ways. Because we have, in effect, continuously updated data on the network, the log-odds ratios that we give for the FoaF risk set are averaged over all time. The nature of university life, however, constrains much of the joining and leaving of social foci (especially classes) to coincide with the beginning of semesters. Thus, for the SF risk set, we compute coefficients for the entire spring semester, conditioned on the status of the pair in the fall semester.

TABLE 4
DAILY PROBABILITY OF A PAIR TRANSITIONING TO DISTANCE 2 FROM A LONGER DISTANCE

Model	Subset	N pairs	β (SE)	P	Odds Ratio
11 ...	All pairs	314,590	.438 (.004)	.00	1.55**
12 ...	$q_{ij} = 1$	18,882	−.012 (.016)	.45	.99
13 ...	$d_{ij} = 3$	109,870	.200 (.007)	.00	1.22**
14 ...	$d_{ij} = 4$	90,701	.358 (.009)	.00	1.43**
15 ...	$d_{ij} \geq 5$	110,650	.527 (.020)	.00	1.69**
16 ...	$d_{ij}=3$; $q_{ij} = 1$	15,114	.004 (.021)	.83	1.00
17 ...	$d_{ij} = 4$; $q_{ij} = 1$	1,753	.035 (.032)	.27	1.04
18 ...	$d_{ij} \geq 5$; $q_{ij} = 1$	522	−.001 (.060)	.98	1.00

NOTE.—Average probability of transition is 2×10^{-4}.
$^{+}$ $P < .10$.
* $P < .05$.
** $P < .01$.

meet, but they are also more likely to enter the FoaF risk set ($d_{ij} = 2$), in which they become likely to meet. Following the same logic as before, highly similar pairs ($S_{ij} = 6$) are 14 times more likely to acquire a mutual acquaintance than highly dissimilar pairs ($S_{ij} = 0$) and six times more likely than average pairs ($S_{ij} = 2$). As with the tie-formation process, however, this strong dependency on S_{ij} once again disappears when the focal proximity is present (model 12). Also as before, the effect of similarity on p_{FoaF} is attenuated at shorter network distances when no implicit focus is shared—for pairs that are at distance 3 (model 13), just outside the FoaF risk set, the corresponding multiplier effects for high similarity are reduced to 3.3 (compared with low similarity) and 2.2 (compared with average similarity). Finally, we see once again that as network distance increases (models 14–15), the effect of similarity observed in model 11 is recovered.

As expected, therefore, table 4 suggests that FoaF "neighborhoods"—subsets of nodes connected by a single intermediary—put similar people at risk of meeting one another in part because similar people are preferentially likely to join the FoaF neighborhood in the first place. Some of this effect, moreover, appears to reflect individuals choosing similar partners over dissimilar ones from among those available, thus suggesting that choice homophily is indeed present. Nevertheless, much of the effect once again seems attributable to the biased set of opportunities available to choose from—partly because people who select into a particular circle of friends previously belonged to the same social foci, and partly because they are already part of a larger, more inclusive friendship network defined by $d_{ij} = 3$. Friendship circles, in other words, like friendships themselves, are also the products of structural constraints; thus, their observed ho-

TABLE 5
PROBABILITY OF SHARING AN IMPLICIT FOCUS IN SPRING AS A FUNCTION OF PAIRWISE SIMILARITY

Model	Subset	N pairs	β (SE)	P	Odds Ratio
19 ...	F1 = 0	399,750	.957 (.028)	.00	2.60**
20 ...	F1 = 1	3,932	.230 (.036)	.00	1.26**
21 ...	$d_{ij} = 1$; F1 = 0	100	.049 (.377)	.90	1.05
22 ...	$d_{ij} = 2$; F1 = 0	3,627	.565 (.100)	.00	1.76**
23 ...	$d_{ij} = 3$; F1 = 0	47,933	.671 (.040)	.00	1.96**
24 ...	$d_{ij} = 4$; F1 = 0	148,975	.980 (.050)	.00	2.66**
25 ...	$d_{ij} \geq 5$; F1 = 0	199,115	1.186 (.094)	.00	3.28**
26 ...	$d_{ij} = 1$; F1 = 1	84	.304 (.375)	.42	1.36
27 ...	$d_{ij} = 2$; F1 = 1	765	.036 (.085)	.67	1.04
28 ...	$d_{ij} = 3$; F1 = 1	1,924	.278 (.054)	.00	1.32**
29 ...	$d_{ij} = 4$; F1 = 1	816	.283 (.069)	.00	1.33**
30 ...	$d_{ij} \geq 5$; F1 = 1	343	.226 (.109)	.04	1.25*

NOTE.—F1 = 1 denotes a shared focus in semester 1 (fall); F1 = 0 denotes no shared focus in fall.
$^{+}$ $P < .10$.
* $P < .05$.
** $P < .01$.

mogeneity is in part a consequence of the relative homogeneity of their larger surroundings.

Next, we consider how similarity affects the probability, $p_{SF}(i, j)$, that two individuals, who either do or do not share a focus in the fall semester, will have a shared implicit focus (defined as before as being at the 99th percentile of jointly received bulk e-mails) in the spring semester.[34] As we would expect from our previous results, model 19 (table 5) shows that pairs who were similar in the fall semester are far more likely to share a focus in the spring semester—the odds ratio is approximately 2.6, meaning that highly similar pairs are about 300 times as likely to share an implicit focus as nonsimilar pairs and 45 times as likely as pairs with average similarity. When we account for previous structural proximity, however, we find that the picture is slightly different from our previous results. Specifically, model 20 shows that the preference for similarity, although diminished in magnitude (the odds ratio is now 1.3), remains highly significant even when the individuals in question shared an implicit focus in the fall, and models 21–30 show that the effect gradually increases

[34] Our presentation is based on implicit foci for the sake of consistency; however, we obtained very similar results using class registration data. As discussed earlier, implicit foci by definition may represent multiple shared groups or quasi groups and do not have natural start and end points. Ideally, therefore, future studies should include nonclass foci—e.g., student organizations, sporting teams, and other interest groups—that are recorded explicitly.

TABLE 6
DAILY PROBABILITY OF TIE DISSOLUTION AS A FUNCTION OF PAIRWISE SIMILARITY

Model	Subset	N pairs	β (SE)	P	Odds Ratio
31 ...	All pairs	315,787	−.033 (.003)	.00	.97**
32 ...	$q_{ij} = 1$	133,519	.008 (.004)	.06	1.01^{+}

NOTE.—The average probability of dissolution is .01. It differs from the tie-formation probability because the network is roughly in equilibrium and there are many more disconnected pairs at risk of forming a tie than there are existing ties at risk of dissolution.

$^{+}$ $P < .10$.
* $P < .05$.
** $P < .01$.

with distance. In other words, for pairs who were friends in the fall, similarity had no effect on their likelihood to select the same groups in the spring semester, while for pairs who previously shared a friend or a group, the corresponding effect decreased in magnitude yet remained large and significant, a difference from models 2 and 12. Precisely why we find this difference is not entirely clear, but it may reflect the nature of group selection in a university setting, where factors like status, academic major, and seniority may be more relevant than whom one knows to which groups one chooses.

Tie dissolution.—Finally, we consider a third process—tie dissolution—that has been hypothesized to account for observed homophily (Felmlee et al. 1990; Leenders 1996). Specifically, we consider the effect of similarity on the probability, $p_{diss}(i, j)$, that a tie present in the current network snapshot will have dissolved by the next daily snapshot, defining a tie as dissolved when the dyad has not interacted for $\tau = 60$ days.[35] As indicated in table 6, we find that p_{diss} declines slightly with pairwise similarity—the odds ratio for model 31 is 0.97, which means that the odds of dissolution for highly dissimilar pairs are about 1.2 times higher than those for highly similar pairs; correspondingly, pairs of average similarity are about 1.14 times as likely to dissolve as highly similar pairs. Although these effects are weaker than those associated with tie formation (which were one to two orders of magnitude larger), model 31 does suggest that the observed

[35] Although this definition may count some ties as dissolved when in fact they are simply "dormant," it captures about 90% of all ties that remain inactive for the duration of our data-collection period; thus, to the extent that e-mail interactions reflect ongoing relations, a lapse in communication of 60 days is a reasonable measure of termination. There is no reason to suspect, moreover, that this criterion of tie dissolution either favors or disfavors similar over dissimilar dyads; thus, even to the extent that we are overcounting the number of ties dissolving, it should not affect our conclusions regarding differences in dissolution rates.

homogeneity of network neighbors might, to some degree, be a consequence of individuals' selectively preserving their relationships with similar others (or disassociating themselves from dissimilar others). As before, however, this effect all but vanishes once we control for shared foci; that is, model 32 shows that the odds ratio is not significantly different from 1 for pairs sharing an implicit focus. Shared social foci, in other words, not only increase the likelihood of tie formation, but also decrease the rate of dissolution (the dissolution rate goes down from 0.01 to 0.0086 for pairs with a shared implicit focus and to 0.0082 for student pairs with shared classes). In both cases, moreover, the impact of sharing a focus eliminates the impact of similarity.

DISCUSSION

In concluding, we return to our opening observation that the presence of homophily in social networks has long been associated with other issues that are of interest to sociologists, such as segregation, inequality, and even the transmission of information between groups. To the extent that homophilous patterns of interactions are considered important outcome variables, therefore, related questions of social policy can only be answered on the basis of some understanding of how these patterns emerge. The lens through which we have viewed this question is that the formation of some particular relationship—and not others—is in part a consequence of individual preferences and in part the result of the opportunities available at the time. It is reasonable to suppose that most people have more opportunities to form social ties than they have the time, energy, or interest to pursue; thus, the particular individuals with whom they do choose to spend their time must presumably offer greater "benefit," broadly construed, than at least some of the available alternatives. Assuming that similarity between alters is associated with various benefits, one would expect to find that, all other things being equal, similar pairs of individuals would be more likely to form new ties—and less likely to terminate existing ties—than dissimilar pairs. All other things are rarely equal, however—in fact, it is also the case that at any point in time, the number of *unavailable* alters vastly outnumbers the available opportunities between which we are actively choosing. It may be that any number of these individuals would be at least as attractive to us as those with whom we have chosen to spend our time; yet the prohibitive cost of searching for and meeting these people renders the "benefit" part of the cost-benefit analysis irrelevant. To the extent, therefore, that choices are determined by who is readily available and that readily available alters tend to be more similar to the focal individual than the unavailable majority, one

would expect the effect of similarity on new tie formation to be mitigated once structural proximity is taken into account.

It is therefore the relative roles of similarity and proximity in determining observed homophily that we have investigated in this article, using a unique data set that combines dynamic network data with information regarding individual attributes and affiliations. To begin with, we have confirmed that our community indeed displays high levels of homophily: specifically, we showed (fig. 6) that individuals who are "proximate" in the network, because they either are connected by a short path length or share a social focus, are more similar than those who are "distant," with their similarity decreasing monotonically with network distance and increasing with number of shared classes. Next, we showed that similar individuals are more likely to form new ties with one another (model 1), consistent with the intuitive notion that friends are more similar than strangers because similar people prefer to become friends. However, we also showed (models 2 and 3) that this effect is strongly mitigated in instances where pairs are already socially proximate, either because they share mutual acquaintances or social foci, where once again the degree of mitigation increases with increasing proximity (measured in terms of either number of mutual friends or number of shared foci). These findings therefore support previous results of McPherson and Smith-Lovin (1987), who argued for the importance of induced homophily, although we emphasize that our analysis also shows that choice homophily continues to play an important, albeit diminished, role.

As we also argue, however, merely finding that structural proximity mitigates the observed tendency to connect preferentially to similar others does not, on its own, show that exogenously determined structural constraints necessarily undermine the ability of individuals to determine the composition of their friendship networks. That is, although, as we showed in figure 5, it is clearly true that structural proximity does in fact overwhelmingly determine new tie formation, it may well be that forward-looking individuals select into structural positions, such as classes, clubs, and even friendship circles, precisely in order to maximize their chances of meeting the people they want to meet. In other words, structural constraints that may initially appear exogenous are in fact generated endogenously and act effectively as proxies for unobserved individual preferences. To test this hypothesis, we have also considered (in addition to tie formation per se) the formation of "risk sets"—defined here as either friendship circles or shared foci—that subsequently act to constrain the set of potential alters with whom any individual can form ties. Here we find a slightly different picture depending on which risk set we consider. In the case of the friend-of-a-friend ($d_{ij} = 2$) risk set, the effect is greatly diminished for pairs that are transitioning to $d_{ij} = 2$ from $d_{ij} = 3$ and

also for those who have previously shared a social focus (table 4). In contrast, similar individuals are more likely to select into the same groups in the spring semester no matter what their relation was in the fall (table 5), though the effect is nonetheless mitigated considerably by previously shared social foci or friends.

In terms of our contrast between choice and induced homophily, therefore, we conclude that homophily observed in our population cannot be unequivocally attributed to either stylized mechanism but appears to depend on both in significant ways. Judging from the small number of ties that form between structurally distant individuals, preferences for attribute similarity will be expressed in the absence of other reasons to form new ties. As we observe, however, the vast majority of new ties form between individuals who already share a friend or a group, and once these conditions are met, similarity loses much, and in some cases all, of its effect. Friendship circles, moreover, also exist within relatively homogeneous networks and are subject to similar effects; thus, even when selecting into a new circle of friends, one is constrained by whom one already knows. When we consider selection to shared groups, we see a robust effect of similarity; but given the importance of groups to tie formation, this effect is an important one. We cannot of course know whether individuals select these groups because of whom they hope to meet or for some other reason, but the possibility that group choices are strategic is certainly present.

Moreover, we do not simply find that both choice and induced homophily matter—they appear to act as substitutes, each reinforcing the observed tendency of similar individuals to interact. Classes taken together beget future shared classes, and shared friends beget new shared friends, and so on. The selection of similar actors to similar foci and subsequent tie formation affects not only the individuals making the decisions—it also affects more distant pairs who are connected via those actors. These individuals now face new opportunities to form ties, and these opportunities are again skewed toward similar others. As the network and structure coevolve (McPherson and Ranger-Moore 1991), distant but similar individuals will be brought closer to each other in the network, creating a positive correlation between similarity and proximity. That correlation is then strengthened further by structural forces operating to facilitate connections between proximate individuals, which in turn bring the neighbors of those individuals closer than they already were, thus increasing the chances that they will also form ties, and so on.

Because such a large proportion of new ties form via this process, and because the process plays out over multiple "generations," we speculate that even a relatively weak preference for homophilous relationships will tend to be amplified over time, via a cumulative advantage–like process

(Simon 1955; Merton 1968; DiPrete and Eirich 2006), thereby producing striking patterns of observed homophily—analogous to so-called tipping models of residential segregation (Schelling 1978). This cumulative advantage view of homophily casts our original question—to what extent some observed pattern of homophily can be attributed to individual preferences versus structural constraints—in a new light. Although we do not doubt that most people do, in fact, have some preference for interacting with similar others (at least under some circumstances), our results raise the issue of how weak such a tendency would need to be in order for striking patterns of homophily *not* to arise. A thorough answer to this question would require the use of simulation models, in which choice homophily as well as focal and cyclic closure biases could be systematically varied. In this manner, one could establish, at least under certain simplifying assumptions, a lower bound on individual preferences for similar others, below which homophilous patterns of association would not emerge even with very strong structural constraints. Although such an extensive simulation exercise is beyond the scope of this article, it is certainly conceivable and would be an interesting direction for future research.

A second question suggested by our interpretation of homophily as a cumulative advantage process deals with the natural limits of such processes—that is, why do we not see more pronounced homophily than we do? In fact, why do successive rounds of induced homophily not lead to a balkanization of the network, possibly even into disconnected, homogeneous components? Once again, a satisfactory answer would require the aid of simulation models, but three possible mechanisms suggest themselves. First, any process that reduces distances preferentially between already proximate pairs is inherently self-limiting, simply because it is much more difficult for already closely separated pairs to become closer still than it is for distant pairs (Watts 1999). Second, while choices of new ties are dominated by structural proximity (in terms of either shared acquaintances or shared foci), a small fraction of "long-range" ties are always being formed as well. As is now well understood, even a small fraction of long-range ties is sufficient to ensure global connectivity of even a very large network (Watts 1999); moreover, in bridging previously distant, and presumably different, parts of a network (Granovetter 1973), these ties can also be expected to exert a natural brake on homophily. Thus, while even a weak preference for similar others may lead, over time, to striking patterns of homophilous relations, networks that are already highly homophilous will experience great difficulty in becoming more homophilous still. And finally, the use of e-mail as a means of professional communication suggests that people in the university community must to some extent interact with each other, regardless of similarity, simply in the

course of doing their jobs. Thus, the community could never become completely homophilous unless it so happened that actors altered their formal organizational positions so as to align their professional needs precisely along our recorded dimensions of homophily—an unlikely event, particularly over the relatively short time scale spanned by our data set.

Although it is interesting in its own right, we conclude by emphasizing that our question regarding the origins of homophily is just one particular instantiation of a more general class of "structure versus agency" (Emirbayer and Goodwin 1994) questions that have been debated over the years by advocates of "structuralism," on the one hand, and "individualism," on the other (Mayhew 1980). Given any empirical regularity of a sociological nature—whether patterns of homophily, choices of occupation, wealth distributions, or educational attainment—one can always ask to what extent the observed outcome reflects the preferences and intentions of the individuals themselves and to what extent it is a consequence of the social-organizational structure in which they are embedded. In this broader context, our finding that both structure and agency matter may not seem altogether surprising; however, the attention that individualistic explanations of social phenomena have received in the social sciences broadly—in particular, with respect to explanations associated with the rational choice theory tradition (Harsanyi 1969; Becker 1976; Coleman and Fararo 1992; Kiser and Hechter 1998), but also more generally within sociology (Boudon 1987)—suggests that our findings regarding the importance of structure are nonetheless worth emphasizing. We also note that in the absence of dynamic data, structure-versus-agency debates can be difficult to adjudicate, and such data have been prohibitively difficult to obtain until recently. Our results therefore imply that data derived from electronic communication should be useful in addressing a range of questions associated with individualistic versus structural explanations of empirically observed patterns in the social world.

As promising as we consider electronic communications data to be, however, the particular data set used in our study nevertheless exhibits some important limitations. As we have already indicated, our data clearly lack some attributes, like race and socioeconomic status, that might be more salient with respect to the choice of interaction partners than the available variables. Moreover, the university community in question may be relatively homogeneous on some dimensions, particularly educational background, in comparison to other kinds of communities. One might therefore suspect that individual preferences with regard to the similarity of potential interaction partners may in general be stronger, and structural proximity correspondingly weaker, than our results appear to suggest. Electronic communication, moreover, may differ in some systematic ways between formal and informal organizations and various communities,

depending both on the demographics of their constituents and on the purpose of the organizations themselves. A business firm, for example, attempting to coordinate the activity of many departments to achieve a single, coherent goal, may display quite different patterns of interactions than a university community. And finally, privacy considerations severely limit the prospects of obtaining message content (as users tend to reveal their identities by what they say and by using signatures; thus, encrypting labels is of little use), without which it is extremely difficult to interpret the meaning of any given pattern of relations.

Together, these various deficiencies present some serious challenges to the widespread and productive incorporation of electronic data into empirical social science. However, we propose that they are not insurmountable. It would certainly be possible, for example, to conduct comparative analyses of network evolution in other environments, such as business firms, government agencies, or voluntary organizations. It might also be possible to supplement the approach that we have developed in this article with text analysis and validation of inferred network ties by selective surveying of e-mail users. Alternatively, informed consent procedures can be imagined under which users would be willing to provide content in exchange for well-defined benefits as well as assurances on the use of the content. Although many important details remain to be worked out, we anticipate that a systematic program of comparative, dynamic network analyses will reveal much of interest about the evolutionary dynamics of network structure and its relation to other substantively interesting social processes such as the diffusion of information and influence.

APPENDIX A

Definition of Variables

Status.—Formal status is available for all individuals present in the database in each semester. This variable (table 1) is inferred from flags indicating several status categories (*undergraduate*, *graduate*, *nondegree*, and *professional* students; *faculty*; *administrator*; *staff*). Tenured research scientists are considered faculty. Postdoctoral researchers and visiting scholars are included in the *affiliate* category, which also includes exchange students and recent alumni. The *instructor* category was created to include those faculty members who have also registered for classes (as students) as well as affiliates and students who have been listed as primary course instructors in the course database.

Some individuals (about 15% of the individuals with valid status) have a combination of status flags; the most frequent combination being *AF* (administrator and faculty), which may reflect the fact that many faculty

members automatically receive certain administrative privileges. Also, a number of students work part-time on campus and therefore have the additional flag *S* (staff). For purposes of analysis, such as determining if two individuals have the same status, we compared actual flags: for example, we say that two individuals with flags *GS* and *SU* have the same status because they share the flag *S*, but they would be assigned different status for descriptive purposes (graduate and undergraduate, respectively). In order to simplify description in cases of multiple flags, we assigned primary status using the following heuristic rules, listed in order of precedence:

1. Assign status instructor if (*a*) flags include faculty and *courses taken* > 0 or (*b*) flags include affiliate or student and *courses taught* > 0;
2. assign undergraduate if *dormitory* is defined;
3. assign faculty if flags include faculty;
4. assign graduate if flags include graduate;
5. assign administrator if flags include administrator;
6. assign undergraduate if flags include undergraduate;
7. assign staff if flags include staff;
8. assign professional if flags include professional;
9. assign nondegree if flags include nondegree.

Thus, (for purposes of description only) an individual with flags *GS* (graduate student and staff member) would be categorized as a graduate student, but a person with flags *AP* (administrator and professional student) would be assigned the primary status of administrator, because many graduate students work part-time as staff members (library workers, computing support personnel, etc.) and many full-time administrators pursue professional degrees.

Age.—Age at the beginning of the spring semester (constructed from the year of birth).

Gender.—Gender of the individual.

Year.—Number of years since enrollment at the current school (the best cohort proxy available).

Department.—Primary academic department (encrypted).

School.—Encrypted school code (students only; e.g., graduate school, business school, etc.). The two smallest numbers (counts) associated with a school code appear to be 1 and 28, which suggests a database coding error. However, 17 out of 19 schools have counts of 147 and greater and account for 99.9% of the population for whom *school* is defined (20,256 individuals, mainly professional, graduate, and undergraduate students and administrative personnel).

Field.—Approximate academic field based on primary department; has 12 categories (containing 500–3,500 individuals):

A4—Fine and performing arts, architecture
A5—Literature, languages, education
A6—History, philosophy, religion
L2—Biological sciences
L3—Public health and environment
L4—Medical sciences
NC—Nonclassified, other
P2—Natural and mathematical sciences
P7—Engineering sciences
S2—Social sciences
S3—Economics and management
S4—Law, policy, political sciences

Campus.—Encrypted campus location. Tabulation of the data shows eight campus locations ; however, the smallest one appears to contain just one individual, which suggests a coding error in the database. The largest three locations account for 98.5% of the population.

Dormitory.—Encrypted dormitory building (undergraduate students only).

State.—Home state. Constructed from zip codes using U.S. census data. To protect privacy, zip codes with low counts were aggregated to zip-4 and zip-3 until the number of individuals in each aggregated area was greater than or equal to five.

From U.S.—This variable is coded 1 if an individual's home address is in the United States, 0 otherwise. Note that while *from U.S*. values largely agree with *state*, 15 individuals for whom *from U.S*. = 0 have a valid state code due to a database error.

Similarity.—A dyadic scale with a range between 0 and 6 constructed as the number of matching items between two individuals out of the following characteristics: *gender*, *status*, *field*, *age*, *year*, and *from U.S*. If any of the values is missing for either person in the pair, the result is replaced with a sample average. We have checked that interactions are homophilous with respect to each variable in the scale, independent of others, and used the scale as a summary measure of pairwise similarity.

Courses taken.—This count represents both courses taken for a grade and those audited. The maximum value is 15, which seems unusually high even assuming that some courses could be audited. Ninety-five percent of students registered for six or fewer courses. The value of 15 comes from 20 professional students who are all present in the same course record, which seems to be repeated 15 times in the course table under different course IDs. It is not clear whether those multiple course records should be treated as an administrative or database error or as 15 separate courses. Fortunately, our results are robust to data issues of this kind, as

we typically condition not on the exact number of shared classes but on the fact that a pair shared at least one focus.

Courses taught.—Count of the number of courses an individual taught in a given semester.

Days active.—Number of days on which an individual sent out one or more messages.

Messages sent.—Total number of messages sent by an individual in a given semester.

Messages received.—Total number of messages received by an individual in a given semester.

In-degree.—Number of people who sent messages to an individual.

Out-degree.—Number of people to whom an individual sent messages.

APPENDIX B

Note on Data Cleansing and Missing Values

The university databases from which the data were obtained contained some errors and missing values. We have made a special effort to correct those, making use of the longitudinal nature of the data set (this article reports analysis of only one academic year's worth of data; however, the full data set spans two calendar years, or six academic semesters). For example, a valid year of birth would be present in some semester snapshots within individual records but missing in others. In such cases, we replaced missing values with a valid number. In cases where conflicting values were present (e.g., gender coded as female in semester 1 but as male in semester 2), the best value was determined using a set of heuristics.

Briefly, the following error-correction strategies were used: (*a*) modal value substitution for age, gender, and state (if there were several modes, then the most recent modal value was used, assuming that the more recent value was more likely to be correct); (*b*) backward interpolation for dormitory, field, department, campus, and school, and a combination of forward and backward interpolation with increment or decrement every three semesters for year (assuming that it typically takes three semesters including summer to advance to the next year in the program); and (*c*) a combination of merge and forward and backward interpolation for flags (e.g., when a person had a flag *undergraduate* for two semesters, then a gap, then a flag *staff*, we assumed that his or her status changed—the student likely graduated and was subsequently employed by the university—in which case we filled the gap by setting both *undergraduate* and *staff* flags).

Missing values and erroneous entries were interpolated only within the time bounds determined by an individual's presence in the community:

for example, if an individual first appeared in the database in semester 2 and was last recorded in semester 4, then error correction was applied only to the values pertaining to semesters 2–4. The procedure was most effective for age (22% of individual records augmented), gender (17%), and state (12%) and yielded marginal improvements for variables such as department, field, and campus (3% of all records). Details are available on request.

APPENDIX C

Descriptive Statistics of E-mail Exchange between Status Groups

TABLE C1
TOTAL NUMBER OF MESSAGES BY SENDER AND RECIPIENT CATEGORIES

	Recipient									Messages Sent	
Sender	Administrator	Affiliate	Faculty	Graduate	Instructor	Nondegree	Professional	Staff	Undergraduate	Total	Average
Administrator	912,793	120,454	227,804	86,333	21,588	1,665	45,967	89,540	93,174	1,599,318	539.8
Affiliate	136,543	615,633	151,686	107,579	10,445	763	121,313	17,799	146,356	1,308,117	175.3
Faculty	272,755	128,824	499,367	156,646	31,847	972	51,070	44,050	89,767	1,275,298	321.8
Graduate	94,609	104,019	180,626	360,948	15,712	604	32,963	19,718	91,376	900,575	198.4
Instructor	23,384	9,472	33,191	13,674	9,030	67	4,458	3,200	5,303	101,779	266.4
Nondegree	2,087	772	1,629	729	81	300	467	431	610	7,106	65.2
Professional	49,416	130,465	63,174	34,512	5,461	480	446,191	7,850	21,696	759,245	215.5
Staff	80,047	17,317	39,973	18,737	2,816	278	6,548	35,729	15,023	216,468	191.5
Undergraduate	105,070	153,213	103,742	110,458	6,558	549	22,360	16,643	469,663	988,256	156.3
Messages received:											
Total	1,676,704	1,280,169	1,301,192	889,616	103,538	5,678	731,337	234,960	932,968	7,156,162	235.43
Average	565.9	171.5	328.3	196.0	271.0	52.1	207.6	207.8	147.6	235.43	

TABLE C2

DISTRIBUTION OF OUTGOING MAIL, ADJUSTED FOR RECIPIENT'S GROUP SIZE (%)

	RECIPIENT								
SENDER	Administrator	Affiliate	Faculty	Graduate	Instructor	Nondegree	Professional	Staff	Undergraduate
Administrator	57.1	7.5	14.2	5.4	1.3	.1	2.9	5.6	5.8
Affiliate	10.4	47.1	11.6	8.2	.8	.1	9.3	1.4	11.2
Faculty	21.4	10.1	39.2	12.3	2.5	.1	4.0	3.5	7.0
Graduate	10.5	11.6	20.1	40.1	1.7	.1	3.7	2.2	10.1
Instructor	23.0	9.3	32.6	13.4	8.9	.1	4.4	3.1	5.2
Nondegree	29.4	10.9	22.9	10.3	1.1	4.2	6.6	6.1	8.6
Professional	6.5	17.2	8.3	4.5	.7	.1	58.8	1.0	2.9
Staff	37.0	8.0	18.5	8.7	1.3	.1	3.0	16.5	6.9
Undergraduate ...	10.6	15.5	10.5	11.2	.7	.1	2.3	1.7	47.5

NOTE.—The adjustment for recipient's group size means that each row is a mailing profile of an average sender if all groups were of equal size. All rows sum to 100%.

TABLE C3

DISTRIBUTION OF INCOMING MAIL, ADJUSTED FOR SENDER'S GROUP SIZE (%)

SENDER	RECIPIENT								
	Administrator	Affiliate	Faculty	Graduate	Instructor	Nondegree	Professional	Staff	Undergraduate
Administrator	54.4	9.4	17.5	9.7	20.9	29.3	6.3	38.1	10.0
Affiliate	8.1	48.1	11.7	12.1	10.1	13.4	16.6	7.6	15.7
Faculty	16.3	10.1	38.4	17.6	30.8	17.1	7.0	18.7	9.6
Graduate	5.6	8.1	13.9	40.6	15.2	10.6	4.5	8.4	9.8
Instructor	1.4	.7	2.6	1.5	8.7	1.2	.6	1.4	.6
Nondegree	.1	.1	.1	.1	.1	5.3	.1	.2	.1
Professional	2.9	10.2	4.9	3.9	5.3	8.5	61.0	3.3	2.3
Staff	4.8	1.4	3.1	2.1	2.7	4.9	.9	15.2	1.6
Undergraduate ...	6.3	12.0	8.0	12.4	6.3	9.7	3.1	7.1	50.3

NOTE.—Adjustment for sender's group size means that each column is a mailing profile of an average recipient if all groups were of equal size. All columns sum to 100%.

REFERENCES

Banks, David L., and Kathleen M. Carley. 1996. "Models for Network Evolution." *Journal of Mathematical Sociology* 21:173–96.

Barabási, Albert-Laszlo. 2005. "The Origin of Bursts and Heavy Tails in Human Dynamics." *Nature* 435:207–11.

Becker, Gary S. 1976. *The Economic Approach to Human Behavior*. Chicago: University of Chicago Press.

Blau, Peter M. 1977. *Inequality and Heterogeneity: A Primitive Theory of Social Structure*. New York: Free Press.

Blau, Peter M., and Joseph E. Schwartz. 1984. *Crosscutting Social Circles: Testing a Macrostructural Theory of Intergroup Relations*. Orlando, Fla.: Academic Press.

Boudon, Raymond. 1987. "The Individualistic Tradition in Sociology." Pp. 45–70 in *The Micro-Macro Link*, edited by Jeffrey C. Alexander, Bernhard Giesen, and Richard Munch. Berkeley and Los Angeles: University of California Press.

Burt, Ronald S. 1991. "Age as a Structural Concept." *Social Networks* 19:355–73.

Coleman, James S., and Thomas J. Fararo. 1992. *Rational Choice Theory: Advocacy and Critique*. Newbury Park, Calif.: Sage.

Cortes, Corinna, Daryl Pregibon, and Chris Volinsky. 2003. "Computational Methods for Dynamic Graphs." *Journal of Computational and Graphical Statistics* 12:950–70.

DiPrete, Thomas A., and Gregory M. Eirich. 2006. "Cumulative Advantage as a Mechanism for Inequality: A Review of Theoretical and Empirical Developments." *Annual Review of Sociology* 32:271–97.

Doreian, Patrick, and Frans N. Stokman, eds. 1997. *Evolution of Social Networks*. Amsterdam: Gordon & Breach.

Ebel, Holger, Lutz-Ingo Mielsch, and Stefan Bornholdt. 2002. "Scale-Free Topology of E-mail Networks." *Physical Review E* 66 (3): 035103.

Eckmann, Jean-Pierre, Elisha Moses, and Danilo Sergi. 2004. "Entropy of Dialogues Creates Coherent Structures in E-mail Traffic." *Proceedings of the National Academy of Sciences of the United States of America* 101:14333–37.

Emirbayer, Mustafa, and Jeff Goodwin. 1994. "Network Analysis, Culture, and the Problem of Agency." *American Journal of Sociology* 99 (6): 1411–54.

Feld, Scott L. 1981. "The Focused Organization of Social Ties." *American Journal of Sociology* 86:1015–35.

———. 1982. "Structural Determinants of Similarity among Associates." *American Sociological Review* 47:797–801.

Felmlee, Diane, Susan Sprecher, and Edward Bassin. 1990. "The Dissolution of Intimate Relationships: A Hazard Model." *Social Psychology Quarterly* 53:13–30.

Festinger, Leon. 1957. *A Theory of Cognitive Dissonance*. Stanford, Calif.: Stanford University Press.

Girvan, Michelle, and M. E. J. Newman. 2002. "Community Structure in Social and Biological Networks." *Proceedings of the National Academy of Sciences of the USA* 99:7821–26.

Goodreau, Steven M. 2007. "Advances in Exponential Random Graph (p^*) Models Applied to a Large Social Network." *Social Networks* 29 (2): 231–48.

Granovetter, Mark S. 1973. "The Strength of Weak Ties." *American Journal of Sociology* 78:1360–80.

Hamm, Jill V. 2000. "Do Birds of a Feather Flock Together? Individual, Contextual, and Relationship Bases for African American, Asian American, and European American Adolescents' Selection of Similar Friends." *Developmental Psychology* 36 (2): 209–19.

Handcock, Mark S., Peter D. Hoff, and Adrian E. Raftery. 2002. "Latent Space Approaches to Social Network Analysis." *Journal of the American Statistical Association* 97:1090–98.

Harsanyi, John C. 1969. "Rational-Choice Models of Political Behavior vs. Functionalist and Conformist Theories." *World Politics* 21:513–38.
Holland, Paul W., and Samuel Leinhard. 1971. "Transitivity in Structural Models of Small Groups." *Comparative Group Studies* 2:107–24.
Ibarra, Hermina. 1993. "Personal Networks of Women and Minorities in Management: A Conceptual Framework." *Academy of Management Review* 18:56–87.
———. 1995. "Race, Opportunity and Diversity of Social Circles in Managerial Networks." *Academy of Management Review* 38:673–703.
Imai, Kosuke, Gary King, and Olivia Lau. 2007. "Toward A Common Framework for Statistical Analysis and Development." *Journal of Computational and Graphical Statistics* 17 (4): 892–913.
King, Gary, and Langche Zeng. 2001. "Logistic Regression in Rare Events Data." *Political Analysis* 9 (2): 137–63.
———. 2002. "Estimating Risk and Rate Levels, Ratios, and Differences in Case-Control Studies." *Statistics in Medicine* 21:1409–27.
Kiser, Edgar, and Michael Hechter. 1998. "The Debate on Historical Sociology: Rational Choice Theory and Its Critics." *American Journal of Sociology* 104:785–816.
Kleinberg, Jon, and Prabhakar Raghavan. 2005. "Query Incentive Networks." Pp.132–41 in *Proceedings of the 46th Annual IEEE Symposium on Foundations of Computer Science*. Los Alamitos, Calif.: IEEE Computer Society.
Kossinets, Gueorgi, and Duncan J. Watts. 2006. "Empirical Analysis of an Evolving Social Network." *Science* 311 (5757): 88–90.
Laumann, Edward O. 1966. *Prestige and Association in an Urban Community*. Indianapolis: Bobbs-Merrill.
Lazarsfeld, Paul F., and Robert K. Merton. 1954. "Friendship as a Social Process: A Substantive and Methodological Analysis." Pp. 18–66 in *Freedom and Control in Modern Society*, edited by Monroe Berger, Theodore Abel, and Charles H. Page. New York: Van Nostrand.
Leenders, Roger Th. A. J. 1996. "Evolution of Friendship and Best Friendship Choices." *Journal of Mathematical Sociology* 21:133–48.
Leskovec, Jure, and Eric Horvitz. 2008. "Planetary-Scale Views on a Large Instant-Messaging Network." p. 915–24 in *Proceedings of the 17th International Conference on the World Wide Web*. Beijing, China.
Liben-Nowell, David, Jasmine Novak, Ravi Kumar, Prabhakar Raghavan, and Andrew Tomkins. 2005. "Geographic Routing in Social Networks." *Proceedings of the National Academy of Sciences* 102:11623–28.
Lincoln, James R., Michael L. Gerlach, and Peggy Takahashi. 1992. "Keiretsu Networks in the Japanese Economy: A Dyad Analysis of Intercorporate Ties." *American Sociological Review* 57:561–85.
Louch, Hugh. 2000. "Personal Network Integration: Transitivity and Homophily in Strong-Tie Relations." *Social Networks* 22:45–64.
Malmgren, R. Dean, Daniel B. Stouffer, Adilson E. Motter, and Luís A. N. Amaral. 2008. "A Poissonian Explanation for Heavy Tails in E-mail Communication." *Proceedings of the National Academy of Sciences* 105:18153–58.
Marsden, Peter V. 1987. "Core Discussion Networks of Americans." *American Sociological Review* 52:122–313.
———. 1988. "Homogeneity in Confiding Relations." *Social Networks* 10:57–76.
Mayer, Adrian C. 1966. "The Significance of Quasi-Groups in the Study of Complex Societies." Pp. 97–122 in *The Social Anthropology of Complex Societies*, edited by M. Banton. London: Tavistock.
Mayhew, Bruce H. 1980. "Structuralism versus Individualism: Part 1, Shadowboxing in the Dark." *Social Forces* 59 (2): 335–75.
McPherson, J. Miller. 2004. "A Blau Space Primer: Prolegomenon to an Ecology of Affiliation." *Industrial and Corporate Change* 13:263–80.

McPherson, J. Miller, and J. R. Ranger-Moore. 1991. "Evolution on a Dancing Landscape: Organizations and Networks in Dynamic Blau Space." *Social Forces* 70: 19–42.

McPherson, J. Miller, and Lynn Smith-Lovin. 1982. "Women and Weak Ties: Sex Differences in the Size of Voluntary Associations." *American Journal of Sociology* 87:883–904.

———. 1987. "Homophily in Voluntary Organizations: Status Distance and the Composition of Face-to-Face Groups." *American Sociological Review* 52:370–79.

McPherson, J. Miller, Lynn Smith-Lovin, and James M. Cook. 2001. "Birds of a Feather: Homophily in Social Networks." *Annual Review of Sociology* 27:415–44.

Merton, Robert K. 1968. "The Matthew Effect in Science: The Reward and Communication Systems of Science Are Considered." *Science* 199 (3810): 55–63.

Mollica, Kelly A., Barbara Gary, and Linda K. Trevino. 2003. "Racial Homophily and Its Persistence in Newcomers' Social Networks." *Organization Science* 14:123–36.

Moody, James. 2001. "Race, School Integration, and Friendship Segregation in America." *American Journal of Sociology* 107:679–716.

Moody, James, Daniel McFarland, and Skye Bender-deMoll. 2005. "Dynamic Network Visualization." *American Journal of Sociology* 110:1206–41.

Moody, James, and Douglas R. White. 2003. "Social Cohesion and Embeddedness: A Hierarchical Concept of Social Groups." *American Sociological Review* 68:1–25.

Mouw, Ted. 2003. "Social Capital and Finding a Job: Do Contacts Matter?" *American Sociological Review* 68:868–98.

Newman, M. E. J., Duncan J. Watts, and Steven H. Strogatz. 2002. "Random Graph Models of Social Nectowkrs." *Proceedings of the National Academy of the Sciences* 99:2566–72.

Onnela, Jukka-Pekka, Jari Saramäki, Jörkki. Hyvönen, Gábor Szabó, David Lazer, Kimmo Kaski, János Kertész, and Albert-László Barabási. 2007. "Structure and Tie Strengths in Mobile Communication Networks." *Proceedings of the National Academy of Sciences* 104:7332–36.

Palla, Gergely, Albert-László Barabási, and Tamás Vicsek. 2007. "Quantifying Social Group Evolution." *Nature* 446:664–67.

Portes, Alejandro, and Julia Sensenbrenner. 1993. "Embeddedness and Immigration: Notes on the Social Determinants of Economic Action." *American Journal of Sociology* 98:1320–50.

Rapoport, Anatol. 1953. "Spread of Information through a Population with Sociostructural Bias: I. Assumption of Transitivity." *Bulletin of Mathematical Biophysics* 15:523–33.

Schelling, Thomas C. 1978. *Micromotives and Macrobehavior.* New York: W. W. Norton.

Simon, Herbert A. 1955. "On a Class of Skew Distribution Functions." *Biometrika* 42: 425–40.

Snijders, Tom A. B. 2001. "The Statistical Evaluation of Social Network Dynamics." Pp. 361–95 in *Sociological Methodology*, vol. 31. Edited by Michael E. Sobel and Mark P. Becker. Boston: Basil Blackwell.

Suitor, Jill, Barry Wellman, and David L. Morgan. 1997. "It's About Time: How, Why, and When Networks Change." *Social Networks* 19:1–7.

Verbrugge, Lois M. 1977. "The Structure of Adult Friendship Choices." *Social Forces* 56:576–97.

Wasserman, Stanley, and Katherine Faust. 1994. *Social Network Analysis: Methods and Applications.* New York: Cambridge University Press.

Watts, Duncan J. 1999. "Networks, Dynamics, and the Small-World Phenomenon." *American Journal of Sociology* 105:493–527.

Werner, Carol, and Pat Parmelee. 1979. "Similarity of Activity Preferences among

Friends: Those Who Play Together Stay Together." *Social Psychology Quarterly* 42 (1): 62–66.

White, Harrison C., Scott A. Boorman, and Ronald L. Breiger. 1976. "Social Structure from Multiple Networks: I. Blockmodels of Roles and Positions." *American Journal of Sociology* 81:730–80.