Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks

Sinan Aral[a,b,1], Lev Muchnik[a], and Arun Sundararajan[a]

[a]Information, Operations and Management Sciences Department, Stern School of Business, New York University, Kaufmann Management Center, 44 West 4th Street, New York, NY 10012; and [b]Center for Digital Business, Sloan School of Management, Massachusetts Institute of Technology, 5 Cambridge Center–NE25, Cambridge, MA 02142

Edited by Matthew O. Jackson, Stanford University, Stanford, CA, and accepted by the Editorial Board October 6, 2009 (received for review August 4, 2009)

Node characteristics and behaviors are often correlated with the structure of social networks over time. While evidence of this type of assortative mixing and temporal clustering of behaviors among linked nodes is used to support claims of peer influence and social contagion in networks, homophily may also explain such evidence. Here we develop a dynamic matched sample estimation framework to distinguish influence and homophily effects in dynamic networks, and we apply this framework to a global instant messaging network of 27.4 million users, using data on the day-by-day adoption of a mobile service application and users' longitudinal behavioral, demographic, and geographic data. We find that previous methods overestimate peer influence in product adoption decisions in this network by 300–700%, and that homophily explains >50% of the perceived behavioral contagion. These findings and methods are essential to both our understanding of the mechanisms that drive contagions in networks and our knowledge of how to propagate or combat them in domains as diverse as epidemiology, marketing, development economics, and public health.

dynamic matching estimation | peer influence | social networks | identification

The recent availability of massive networked data sets has enabled studies of population-level human interaction at unprecedented scale (1–3). Such studies document the persistent structural properties of networks (4), how they form, evolve, and dissolve (5), and how their structure is correlated with social interaction (1, 6, 7), individual and collaborative team performance (8–11), health outcomes (12–14), and global product demand patterns (15). Networks of interactions among individuals also provide the primary pathways along which viral contagions spread in social, biological, technological, and economic systems (16–18), which may explain why network structure is correlated with such a variety of outcomes. Yet although many studies model the dynamics of viral spreading by using assumptions about susceptibility rates, transition probabilities, and their relationships to network structure, few large-scale empirical observations of networked contagions exist to validate these assumptions (16–18).

We analyze a new, large scale dataset which comprehensively captures the diffusion of a mobile service product over a social network for 5 months after its launch date. A key challenge in identifying true contagions in such data is to distinguish peer-to-peer influence, in which a node influences or causes outcomes in its neighbors, from homophily, in which dyadic similarities between nodes create correlated outcome patterns among neighbors that merely mimic viral contagions without direct causal influence (19). Although the diffusion patterns created by peer influence-driven contagions and homophilous diffusion are similar, they are likely to result in significantly different dynamics. Influence-driven contagions are self-reinforcing and display rapid, exponential, and less predictable diffusion as they evolve (18, 20), whereas homophily-driven diffusion processes are governed by the distributions of characteristics over nodes. These distinctions make distinguishing true contagions from homophilous diffusions at early stages important for the success or failure of contagion management efforts. As more of a perceived contagion is explained by homophily rather than peer influence, intervention strategies should shift from peer-to-peer methods based on network structure to outreach based on population segmentation across individuals' characteristics. Formal procedures for separating influence and homophily are therefore essential to support policies that encourage or discourage the spread of behaviors in networks, from health interventions to viral marketing campaigns.

Contagions and homophilous diffusion are both typified by correlations between network structure and individual outcomes over time (1–3, 5–11, 17, 21, 22). Two empirical patterns have been used to substantiate claims of peer influence and contagion in networks (*i*) *assortative mixing*—correlations of behaviors among linked nodes (23, 24)—and (*ii*) *temporal clustering*—temporal interdependence of behaviors among linked nodes (12–14, 25–27). Because peer influence is likely to lead to assortative mixing, some studies claim assortative mixing is evidence of peer influence (12–14, 25–27). Evidence of temporal clustering is used to corroborate these claims because as Anagnostopoulos et al. (25) argue "if influence does not play a role, even though an agent's probability of activation could depend on her friends, the timing of such activation should be independent of the timing of other agents." Yet, while evidence of assortative mixing and temporal clustering in outcomes may indicate peer influence, social contagion, and viral spreading, such outcomes may also be explained by homophily—the demographic, technological, behavioral, and biological similarities of linked nodes (28). If ties are more likely between similar nodes, their outcomes could be correlated because of inherent similarities in their characteristics rather than as a consequence of their interactions. On one hand, linked nodes may directly influence one another to exhibit similar outcomes, creating viral contagions. On the other hand, linked nodes may simply have greater likelihoods of displaying correlated outcomes, in time and in network space, as a consequence of their similarities.

Here we develop a matched sample estimation framework to distinguish influence and homophily effects in dynamic networks, and we apply this framework to a unique dataset documenting product adoption in a large network. We find that previous methods significantly overestimate peer influence in this network, mistakenly identifying homophilous diffusion as influence-driven contagion.

Data

We apply our statistical framework to a longitudinal dataset that combines: (*i*) the global network of daily instant messaging (IM) traffic among 27.4 million users of Yahoo.com (Fig. 1) with (*ii*) data on the day-by-day adoption of a mobile service application

Author contributions: S.A., L.M., and A.S. designed research, performed research, contributed new reagents/analytic tools, analyzed data, and wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission. M.O.J. is a guest editor invited by the Editorial Board.

[1]To whom correspondence should be addressed. E-mail: sinan@stern.nyu.edu.

This article contains supporting information online at www.pnas.org/cgi/content/full/0908800106/DCSupplemental.

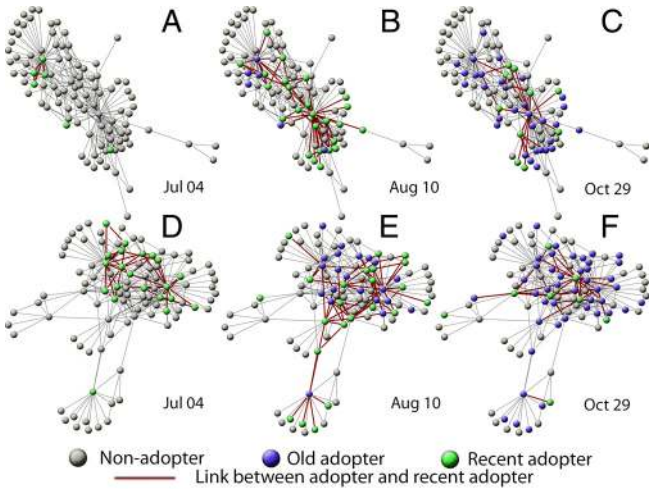

Fig. 1. Diffusion of Yahoo! Go over time. (*A–C* and *D–F*) Two subgraphs of the Yahoo! IM network colored by adoption states on July 4 (the Go launch date), August 10, and October 29, 2007. For animations of the diffusion of Yahoo! Go over time see Movies S1 and S2.

launched in July 2007 (Yahoo! Go) (Fig. 2*A*), and (*iii*) precise attribute and dynamic behavioral data on users' demographics, geographic location, mobile device type and usage, and per-day page views of different types of content (e.g., sports, weather, news, finance, and photo sharing) from desktop, mobile, and Go platforms. Much of these data, such as mobile device usage and page views of different types of content, provide fine-grained proxies for individuals' tastes and preferences. The complete set of covariates includes 40 time-varying and 6 time-invariant individual and network characteristics. Taken together, the sampled users of the IM network registered >14 billion page views and sent 3.9 billion messages over 89.3 million distinct relationships. For details about the service, the data, and descriptive statistics see the *Data* section of the *SI*.

Evidence of Assortative Mixing and Temporal Clustering

We observe strong evidence of both assortative mixing and temporal clustering in Go adoption. At the end of the 5-month period, adopters have a 5-fold higher percentage of adopters in their local networks ($t - stat = 100.12, p < 0.001$; $k.s. - stat = 0.06, p < 0.001$) and receive a 5-fold higher percentage of messages from adopters than nonadopters ($t - stat = 88.30, p < 0.001$; $k.s. - stat = 0.17$, $p < 0.001$). Both the number and percentage of one's local network who have adopted are highly predictive of one's propensity to adopt (Logistic: $\beta_{(\#)} = 0.153, p < 0.001$; $\beta_{(\%)} = 1.268, p < 0.001$), and to adopt earlier (Hazard Rate: $\beta_{(\#)} = 0.10, p < 0.001$; $\beta_{(\%)} = 0.003$, $p < 0.001$). The likelihood of adoption increases dramatically with the number of adopter friends (Fig. 2*C*), and correspondingly, adopters are more likely to have more adopter friends (Fig. 2*B*), mirroring prior evidence on product adoption in networks (29).

Adoption decisions among friends also cluster in time. We randomly reassigned all Go adoption times (while maintaining the adoption frequency distribution over time) and compared observed dyadic differences in adoption times among friends to differences among friends with randomly reassigned adoption times, a procedure known as the "shuffle test" of social influence (25). Compared with these randomly reassigned adoption times, friends are between 100% and 500% more likely to adopt within 2 days of each other, after which the temporal interdependence of adoption among friends disappears (Fig. 1*D*).

Evidence of assortative mixing and temporal clustering may suggest peer influence in Go adoption, but is by no means conclusive. Demographic, behavioral, and preference similarities could simultaneously drive friendship and adoption, creating assortative mixing. Such homophily could also explain the temporal clustering

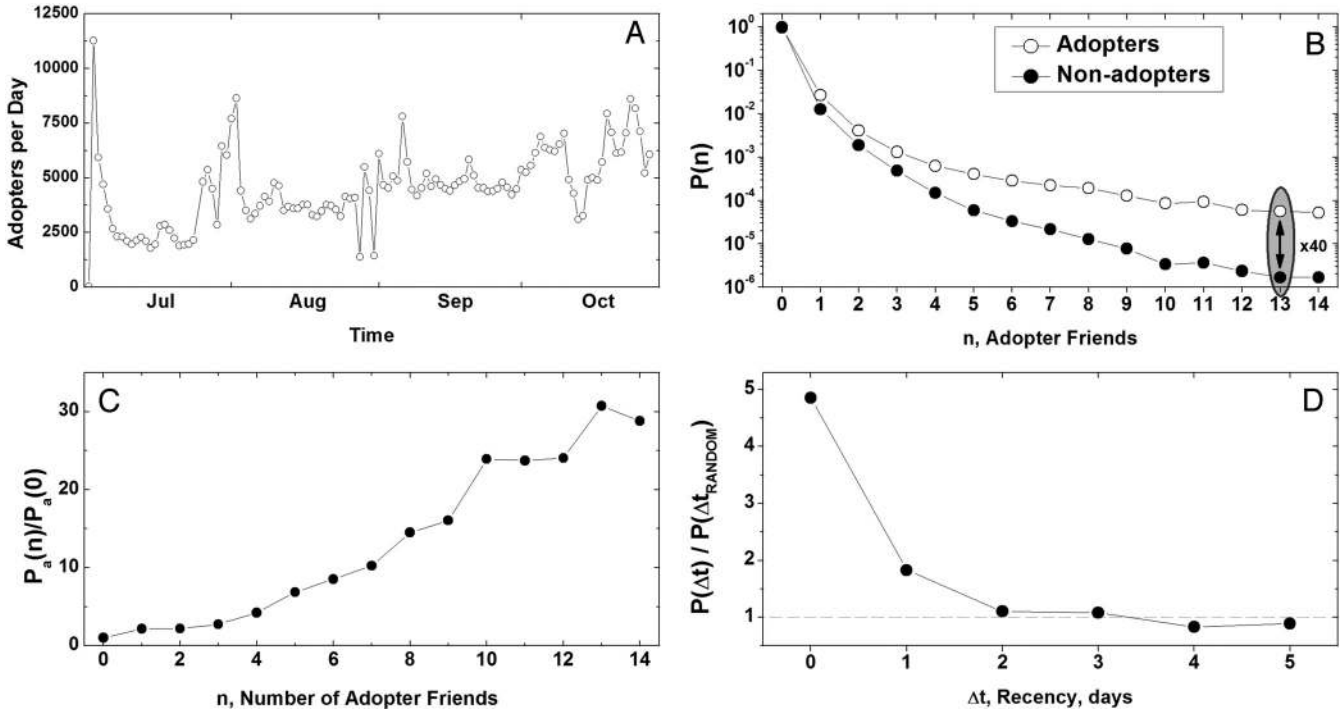

Fig. 2. Assortative mixing and temporal clustering. (*A*) The number of Go adopters per day from July 1 to October 29, 2007. (*B*) The fraction of adopters and nonadopters with a given number of adopter friends. (*C*) The ratio of the likelihood of adoption given n adopter friends $P_a(n)$ and the likelihood of adoption given 0 adopter friends $P_a(0)$ where the number of adopter friends is assessed at the time of adoption. (*D*) Frequency of observed dyadic differences in adoption times between friends compared with differences in adoption times between friends with randomly reassigned adoption times. $\Delta t = t_i - t_j$, where t_i represents the time of i's adoption.

of adoption decisions. If friends are more similar, they are more likely to have similar strengths of preference for Go and similar desires to be "early adopters" of mobile technology services, making them more likely to adopt contemporaneously even if they do not influence one another. These alternative explanations frame a foundational puzzle: Do social choices and behaviors exhibit assortative mixing and temporal clustering in networks because of influence (friends induce friends to adopt), or homophily (friends have similar backgrounds and tastes), and when is one explanation more likely than the other? Robust answers to this question require a statistical framework that estimates influence by taking into account how individual characteristics and similarities among linked nodes may drive assortative mixing and temporal clustering. Some work on the identification of peer effects in networks [e.g., Oestreicher-Singer and Sundararajan (15), Brock and Durlauf (30), and Bramoulle et al. (31)] has developed following seminal work by Manski (32) and Frank and Strauss (33), or models of the co-evolution of networks and behaviors by Snijders (34), methods based on exogenous shocks to peers [e.g., Tucker (35)], or examination of random assignments [see the Dartmouth Roommate studies, e.g., Sacerdote (36)]. However, identification conditions are strict, methods are not typically scalable to large networks, observation of random assignment is rare, and shocks to peers used as instruments are rarely truly exogenous because social relationships typically signal unobserved reasons why these shocks should be correlated among peers. We therefore attempt to describe a scalable and widely applicable alternative method to distinguish homophily and influence, one of which complements existing research on the identification of peer effects.

Methods

In the context of product adoption, peer influence is associated with the presence of adopters in one's local network (the treatment). However, identification of *causal* peer influence effects (37) is complicated by the unobservability problem (38). Each user either has adopter friends or not, making it impossible to observe whether those with adopter friends (the treated) would have adopted had they not had adopter friends. Homophily in this case creates a selection bias because treatments are not randomly assigned: adopters are more likely to be treated because of similarity with their neighbors. Thus, frequently used methods such as regression analysis, which can only establish correlation, are insufficient. Causal treatment effects can, on the other hand, be estimated by matched sampling, which controls for confounding factors and overcomes selection bias by comparing observations that have the same likelihood of treatment.

Toward this end we adapt matched sample estimation (2, 38) for use in dynamic networked settings. Conditioning matches on a vector of observable characteristics, behaviors, and attributes yields influence estimates that account for the homophily that may make product adoption decisions cluster in the network even if no influence exists. This procedure establishes upper bounds on the degree to which influence (rather than homophily) explains assortative mixing and temporal clustering in networks. Because influence can vary over time, our framework provides estimates of its evolution. We can also assess the marginal influence of having any number of friends.

We created a dynamic matched sample of treated and untreated nodes over time, where receiving various degrees of the treatment is defined as having 1, 2, 3, or 4 or more friends who adopted the product. We matched treated nodes with untreated nodes that were as likely to have the same number of adopter friends, conditional on a vector of observable characteristics and behaviors (X), but who did *not* have as many adopter friends. For every period, we estimated p_{it}, the propensity to have been treated at time t, using a logistic regression of the likelihood of having a friend who adopted as a function of users' attributes and dynamic behaviors up to and on day t, as follows:

$$p_{it} = P(T_{it} = 1|X_{it}) = \frac{\exp[\alpha_{it} + \beta_{it}X_{it} + \varepsilon_{it}]}{1 + \exp[\alpha_{it} + \beta_{it}X_{it} + \varepsilon_{it}]},$$

where T_{it} is the treatment status of i on day t and X_{it} represents the vector of demographic and behavioral covariates of i. As treatment status (the number of friends who have adopted), adoption outcome (whether the focal node has adopted), and the vector of observable characteristics X_{it} all vary over time, we performed daily, weekly, and biweekly matched sample tests over the 4-month period. We dropped matched pairs for which the distance of propensity scores exceeded two standard deviations of the observed distribution of propensity score differences. For all treated nodes i, ($\forall i$, $T_{it} = 1$) we chose an untreated match j such that $\|p_{it} - p_{jt}\|$ is minimized subject to $\min\|p_{it} - p_{jt}\| < 2\sigma_d$ where $d = p_{it} - p_{jt}$. This process yielded matched pairs who are equally likely to have a certain number of adopter friends because of observed and correlated latent homophily, contrasting them on the sole dimension of their neighbors' actual adoption status—treated nodes had more adopter neighbors than their untreated matches. We then compared fractions of treated (n_+) and untreated (n_-) adopters over time.

To apply this framework to explain temporal clustering we defined treated users as those with friends who had adopted within certain time intervals of one another (1 day, 2 days, 3 days, etc). For a given recency (R), we considered a user as treated if one of his friends had adopted Go within the specified time interval ($\Delta t \equiv t_i^a - t_j^a = R$) where t_i^a is the adoption time of the adopter i, and t_j^a is the adoption time of adopter j, a friend of i. Multinomial logistic regression was again used to compute estimates of the propensity of a user to be treated, i.e., the likelihood to have had a friend who had adopted R days earlier. Once propensity scores were computed, treated users were matched with untreated users having the closest likelihood of being treated. Untreated users, as before, were those who have no adopter friends within the time window. We again dropped pairs for which the distance of propensity scores exceeded two standard deviations of the observed distribution of propensity score differences. Influence estimates are thus bounded from above by the ratio of the number of treated adopters (n_+) to the number of untreated adopters (n_-). This procedure was repeated for a range of time intervals from 0 to 6 days ($\Delta t \in [0, 6]$) (Fig. 3*D*) where 0 corresponds to friends adopting Go on the same day. Full details regarding propensity score matching methods are provided in *SI*, *Propensity Score Matching*).

Results

To assess the upward bias in influence estimates created by homophily we compare our method (Fig. 3*B*) to random matching (Fig. 3*A*) which matches each treated node to a randomly selected node without conditioning the match on the vector of observable characteristics X_{it} and is analogous to methods commonly used to assess influence in networks: comparison to randomized or shuffled networks (12–14, 25, 26). Because friendship is not random, the selection of a random control group does not control for homophily, which may lead to a greater assessed likelihood of adoption among those with adopter friends. Indeed, in the first biweekly comparison, the fraction of treated adopters is 9 times greater than the fraction of randomly matched untreated adopters when treatment is defined as having 1 or more friends who adopt the product (Fig. 3*A*, open circles), implying that those with 1 or more adopter friends are 9 times more likely to adopt than a randomly selected "control" group. The implied marginal increases in adoption likelihoods for having 2, 3, and 4 or more adopter friends (for which the results imply a 15-fold increase in the average adoption likelihood) are also shown.

When we compare these results to those produced by dynamic matched sampling, which accounts for homophily and individual characteristics, estimates of influence are substantially reduced (Fig. 3*B*). In the first biweekly comparison the fraction of treated adopters is only ≈3 times greater than the fraction of matched untreated adopters when the treatment is defined as having 1 or more adopter friends (filled circles). The random matching estimates are 7 times greater than our matched sample estimates for the effect of having four or more adopter friends, implying that random matching overestimates influence by up to 700%.

Random matching overestimates influence to a greater degree earlier in the product lifecycle, whereas matched sample estimates are consistent over time (Fig. 3 *A* and *B*). We speculated that exaggerated homophily among early adopters leads to greater upward bias in random matching influence estimates in earlier periods. Cosine distances of attribute vectors between adopters and their adopter and nonadopter friends over time confirm that early adopters are indeed more similar to each other and less similar to their nonadopter friends than later adopters are to their respective adopter and nonadopter friends (Fig. 3*C*). Intuitively, estimates of influence that do not account for homophily display greater upward bias in contexts where greater homophily exists, as is the case with

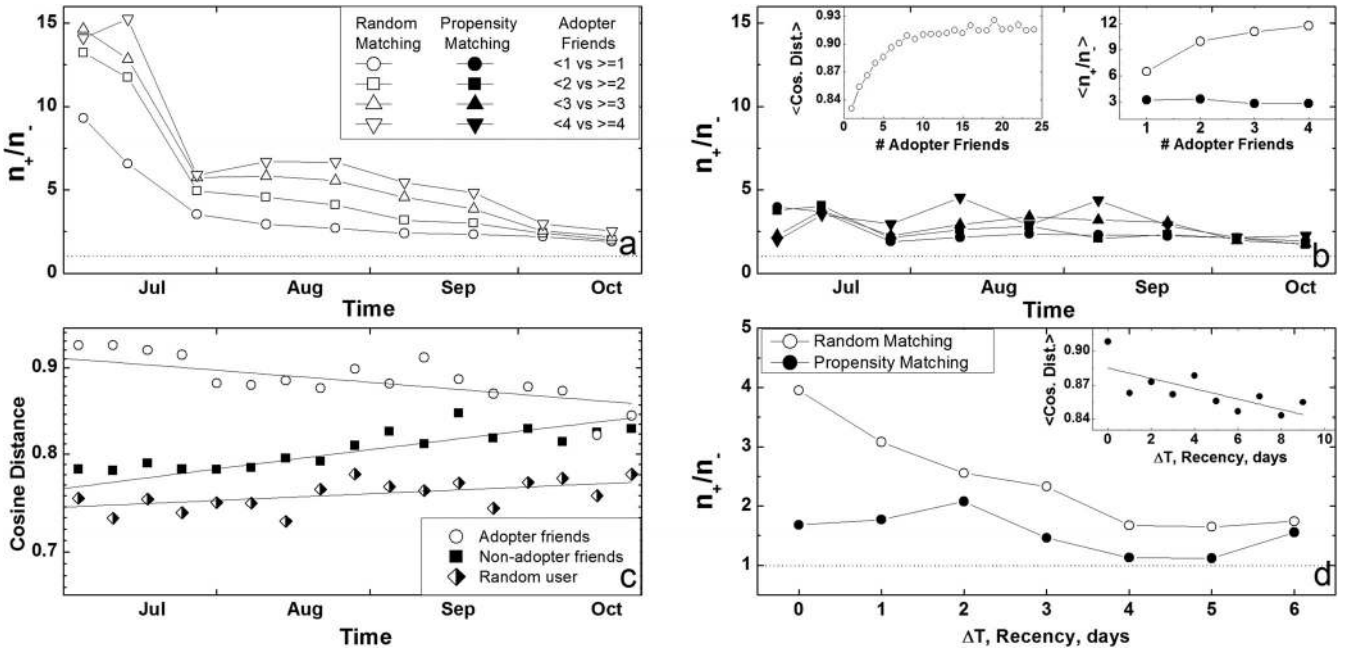

Fig. 3. Distinguishing homophily and influence. (*A* and *B*) The fraction of observed treated to untreated adopters (n_+/n_-) under random (*A*) and propensity score (*B*) matching over time. The dotted line shows a ratio of 1, when treatment has no effect. The *Right Inset* in *B* graphs the average marginal influence effects of having 1, 2, 3, or 4 adopter friends implied by random (open circles) and propensity score (filled circles) matching. The *Left Inset* graphs the average cosine distance of attribute and behavior vectors of adopters to adopter friends as the number of adopters in the local network increases ($\Sigma_{i,j}^{n} \cos(x_i^a, x_j^a)/n$). (*C*) Graphs the cosine distances of adopters to their adopter friends $\cos(x_{it}^a, x_{jt}^a)$, their nonadopter friends $\cos(x_{it}^a, x_{jt})$, and a random alter $\cos(x_{it}^a, x_{rt})$ over time with trend lines fitted by ordinary least squares. (*D*) The fraction of treated and untreated adopters, where treatment is defined as having a friend who adopted within a certain time period (or recency) ($\Delta t \equiv t_i^a - t_j^a = R$), under random matching (open circles) and propensity score matching (filled circles). The *Inset* graphs the cosine distances of dyads of adopters $\cos(x_{it}^a, x_{jt}^a)$ by the time interval between their adoption.

early Go adopters. Random matching also implies that the marginal influence of an additional adopter friend grows with the number of adopter friends, whereas propensity score results show linear to diminishing marginal influence effects of additional adopter friends (Fig. 3*B Right Inset*). This occurs in part because there is exaggerated homophily among larger clusters of adopter friends (Fig. 3*B Left Inset*). The more adopters there are in a group of friends the more likely they are to be more similar to one another. Comparisons to random therefore incorrectly imply that influence grows superlinearly with the number of adopter friends, whereas there is simply greater homophily in larger groups of adopters.

Homophily also accounts for temporal clustering. We redefined treatment to capture the effect of having a friend who adopted within a certain time period (or recency)($\Delta t \equiv t_i^a - t_j^a = R$) and reevaluated results under random and propensity score matching (Fig. 3*D*). Random matching overestimates the contribution of influence to the temporal clustering of adoption decisions by >200% for dyads that adopt on the same day ($\Delta t = 0$), >100% for dyads that adopt 1 day apart ($\Delta t = 1$), and so on. Friends who adopt contemporaneously are again more similar along observable demographic and behavioral dimensions [measured by $\cos(x_{it}^a, x_{jt}^a)$, Fig. 3*D Inset*], indicating that homophily explains a good deal of variance in the temporal clustering of Go adoption decisions.

Thus, homophily can, to a large extent, explain what seems at first to be a contagious process driven by peer influence. Over half of the cumulative adoption of treated users (those with at least one adopter friend) can be attributed to homophily effects (Fig. 4 *A* and *B*). The remaining adoption events (49.8%) represent the upper bound of influence effects established by our matched sample estimates. We also evaluated these influence effects under various environmental conditions (by holding out and varying one characteristic (x_i) at a time while matching on all other characteristics, *SI*, *Environmental Conditions*) and found the upper bounds of influence vary across different segments of the population. When ego's average strength of ties to adopter friends is above the median, the likelihood of adoption controlling for homophily is on average 2 times higher than when below the median (Fig. 4*C*). Those with cohesive, dense local networks (with more ties among their friends) adopt at a higher rate in the presence of an adopter friend controlling for observed homophily (Fig. 4*D*), reinforcing prior arguments that cohesive networks magnify information exchange and persuasion via redundancy and trust (39). Finally, greater consumption of news content makes ego more susceptible to potential influence. Because Yahoo! Go delivers personalized news, those with greater interest in such content are more susceptible to influence, demonstrating the importance of creating robust matches based on contextual behavioral variables (Fig. 4*E*). These estimates provide examples of the types of environmental conditions that affect the prevalence of influence in networks and demonstrate how to test them.

Discussion

We present a generalized statistical framework for distinguishing peer-to-peer influence from homophily in dynamic networks of any size. Application of this framework to a network of 27 million individuals connected by instant message traffic provides an estimate of the degree to which peer influence and homophily affect the diffusion of a new mobile service application across this network. Most critically, the results show that previous methods overestimate peer influence in this network by 300–700% and that homophily explains >50% of the perceived behavioral contagion in mobile service adoption. These findings demonstrate that homophily can account for a great deal of what appears at first to be a contagious process.

Overestimates of influence are magnified at early stages of the diffusion process because those who are most susceptible are also

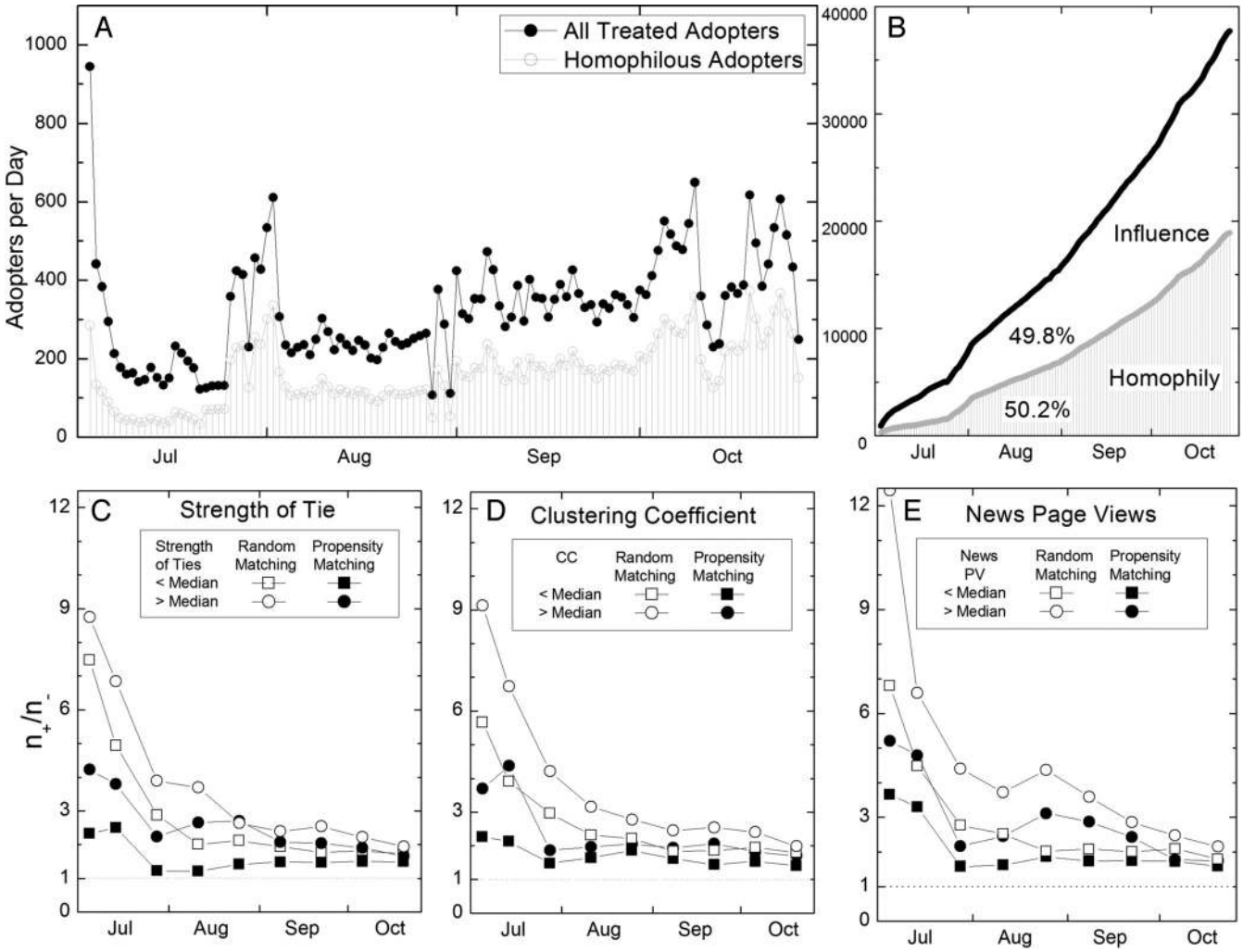

Fig. 4. Influence and homophily effects in Go adoption. (*A* and *B*) All treated adopters (filled circles) and the number of treated adopters that can be explained by homophily (open circles) per day (*A*) and cumulatively over time (*B*). (*C*–*E*) Treatment effects are then displayed when the average strength of ego's ties to adopter friends (measured by the volume of IM message traffic) is greater than and less than the median under random and propensity score matching (*C*); the clustering coefficient in the network around ego is greater than and less than the median (*D*); and ego's page views of news content are greater than and less than the median (*E*).

more similar to one another and more dissimilar vis-à-vis the rest of the population. Influence is also overestimated to a greater degree in large clusters of adopters because in these clusters the homophily effect is more pronounced. Large clusters of adopters tend to be more similar to one another, creating greater risk of overestimation of influence in the very cliques that seem to be the most susceptible to contagious spread. We also find that different subsets of the population, characterized by distributions of individual and relational characteristics such as the strength of ties and local clustering, display various susceptibilities to potential influence.

Our work is not without limitations. First, although we measure individuals' dynamic characteristics, preferences, and behaviors in great detail, the data are not necessarily comprehensive. Although the matching process accounts for homophily on all observed characteristics and those unobserved or latent characteristics that are correlated with what we observe, unobserved and uncorrelated latent homophily and unobserved confounding factors or contextual effects (such as correlated exposure to advertising among friends or information from common unobserved friends) may also contribute to assortative mixing and temporal clustering. The methods therefore establish upper bounds of influence estimates that account for homophily, and limitations in observability are likely to make our estimates of the homophily effect even more conservative. Second, a distinct but related body of literature examines selection and influence processes in the co-evolution of behaviors and network structure in cases where tie formation is likely to be a function of the behavior in question [see Snijders et al. (34)]. In our context (and in many important contexts) link formation is not likely to be driven by the behavior in question—Go adoption is unlikely to drive friendship. However, extending these methods to account for selection processes could prove useful in cases where selection effects are more prevalent. Third, Yahoo! Go 2.0 does not exhibit direct network externalities and its adoption is not likely to be driven by the desire to communicate with one's friends by using the application. We suspect that peer influence effects differ for products with direct network externalities and therefore encourage the application of these methods to influence estimation in the adoption of such products.

Understanding the dynamic mechanisms that govern contagion processes in networks is critical in numerous scientific disciplines and for the development of effective social policy, public health actions, and marketing strategies. A key challenge in identifying the existence and strength of true contagions is to distinguish peer

influence processes from alternative processes such as homophily that can lead to observed outcomes that mimic contagion, especially during early stages of diffusion. These findings, and the general statistical methods used to identify them, document the conditions under which peer influence exists and can help verify the implications of a broad class of social contagion models in a variety of contexts and disciplines (40, 41). The implications for research and policy are far reaching, because discovery of the mechanisms that drive contagions is critical for estimating viral marketing effectiveness, promoting health-related behavior change in large populations, and managing contagions in networks.

Materials and Methods

The data represent an anonymized sample of the Yahoo! Instant Messenger (IM) network where each node is an IM user for whom we collected detailed demographic, geographic, and behavioral information as well as daily IM message traffic. We first sampled all Yahoo! IM users who adopted Yahoo! Go between June 1, 2007, and October 29, 2007. This ''seed experimental sample'' consists of 532,365 users that we labeled ''service adopters.'' We then created a ''seed control sample'' by taking a random sample of 2% of the entire IM network. This seed control sample consists of 2,974,288 nodes that we labeled ''random control seeds.'' We executed a two-step snowball sampling procedure that traversed network links, defined by the existence of IM message traffic, two steps out from every control and experimental seed node, collecting the complete local network neighborhoods of all seed nodes. The first step of the snowball sampling procedure yielded 9.1 million new nodes (labeled ''first-step nodes'') that were IM contacts of the seed node populations. We then collected the local network neighborhoods of all first-step nodes by sampling all users who received at least one message from any of the first-step nodes. The second step of the snowball sampling procedure yielded an additional 14.9 million users, each of whom is two steps away from a seed node.

Behavior and network-related user characteristics such as numbers of page views, IM messages, and number of IM buddies are heavy tailed, a characteristic common to network data (*SI*). To normalize results to account for fat tails and the effects of outliers we use the logarithms of variable values. More specifically, each given value Y is normalized as $\log_{10}(Y + 1)$, where 1 is added to support cases in which $Y = 0$. We find that regression results are qualitatively similar in both cases, but that model fit is significantly better when the logarithm is used.

We test for assortative mixing by using t tests of mean differences, Kolmogorov–Smirnov tests of distributional differences, logistic regression and hazard rate models of the rate of Go adoption. We use logistic regression (33) to assess the effect of personal and local network characteristics on the probability of Go service adoption, defined as $y(X) = 1/(1 + \exp[-\alpha - \beta X])$, where X is a matrix of covariates for each user that may contain both categorical (such as gender or country of residence) and numerical user characteristics. We employ Cox proportional hazards regression (34) to assess the effect of individual user characteristics on the rate of adoption. The regression $h(t, X) = h_0(t)\exp[\alpha + \beta X + \varepsilon]$ estimates users' rate of Yahoo! Go adoption, where $h(t, X)$ represents the adoption rate, t is user time in the risk set, and $h_0(t)$ is the baseline adoption rate. The effects of independent variables are specified in the exponential power (*SI*, *Multivariate Survival Analysis*).

We estimate homophily over time by constructing a vector of 20 personal, behavioral, and local network attributes Table S7 and measure the cosine distance defined for vectors of characteristics, x_i and x_j for nodes i and j as follows:

$$\cos(x_i, x_j) = \frac{\sum_k x_{ik} \cdot x_{jk}}{|x_i| \cdot |x_j|}.$$

To assess the aggregate effect of peer influence on Go adoption across the entire population, we compute the fraction of treated to untreated adopters and use it to estimate the gross number of adopters who would have adopted had they not been treated (had they not had an adopter friend). We define n_+^T as the number of matched adopters treated with certain treatment T, n_-^T as the number of the matched untreated adopters, and N_+^T as the total number of treated adopters (matched or unmatched), and then estimate the number of adopters who would have adopted had they not been treated $\tilde{N}_-^T$ as follows: $\tilde{N}_-^T = \tilde{N}_+^T \cdot n_-^T / n_+^T$. And for all treatments, the estimated number of adopters $\tilde{N}_-$ is $\tilde{N}_- = \Sigma_t N_+^T \cdot n_-^T / n_+^T$, $t \in \{1, 2, 3, \leq 4\}$. Additional considerations related to this technique are provided in *SI*, *Aggregative Effect of Peer Influence*.

ACKNOWLEDGMENTS. We thank Yahoo! Inc. for their generosity in providing the data for this study. Financial support was received from the Institute for Innovation and Information Productivity, the Marketing Sciences Institute, and the New York University Stern School of Business. S.A. received financial support from National Science Foundation CAREER Award 0953832, IBM, and Oracle.

1. Onnela J-P, et al. (2007) Structure and tie strengths in mobile communication networks. *Proc Natl Acad Sci USA* 104:7332–7336.
2. Hill S, Provost F, Volinsky C (2006) Network-based marketing: Identifying likely adopters via consumer networks. *Stat Sci* 21:256–276.
3. Lazer D, et al. (2009) Computational social science. *Science* 323:721–723.
4. Leskovec J, Horvitz E (2008) Planetary-scale views on a large instant-messaging network. *Proceedings of the 17th International World Wide Web Conference* (ACM, New York), pp 915–924.
5. Palla G, Barabási A-L, Vicsek T (2007) Quantifying social group evolution. *Nature* 446:664–667.
6. Kossinets G, Watts DJ (2006) Empirical analysis of an evolving social network. *Science* 311:88–90.
7. Adamic LA, Glance N (2005) The political blogosphere and the 2004 U.S. election: Divided they blog. *LinkKDD '05: Proceedings of the 3rd International Workshop on Link Discovery* (ACM, New York), pp 36–43.
8. Guimerà R, Uzzi B, Spiro J, Amaral LAN (2005) Team assembly mechanisms determine collaboration network structure and team performance. *Science* 308:697–702.
9. Aral S, Brynjolfsson E, Van Alstyne M (2009) Information, technology and information worker productivity. Available at http://ssrn.com/abstract=942310.
10. Aral S, Van Alstyne M (2009) Networks, information & brokerage: The diversity-bandwidth tradeoff. Available at http://ssrn.com/abstract=958158.
11. Aral S, Brynjolfsson E, Van Alstyne M (2008) Productivity effects of information diffusion in networks. Available at http://ssrn.com/abstract=987499.
12. Christakis NA, Fowler JH (2008) The collective dynamics of smoking in a large social network. *N Engl J Med* 358:2249–2258.
13. Christakis NA, Fowler JH (2007) The spread of obesity in a large social network over 32 years. *N Engl J Med* 357:370–379.
14. Fowler J, Christakis N (2008) Dynamic spread of happiness in a large social network: Longitudinal analysis over 20 years in the Framingham Heart Study. *BMJ* 337:2338.
15. Oestreicher-Singer G, Sundararajan A (2008) The visible hand of social networks in electronic markets. Available at http://ssrn.com/abstract=1268516.
16. Eubank S, et al. (2004) Modelling disease outbreaks in realistic urban social networks. *Nature* 429:180–184.
17. Balthrop J, Forrest S, Newman M, Williamson M (2004) Technological networks and the spread of computer viruses. *Science* 304:527–529.
18. Barthélemy M, Barrat A, Pastor-Satorras R, Vespignani (2004) A velocity and hierarchical spread of epidemic outbreaks in scale-free networks. *Phys Rev Lett* 92:178701.
19. McPherson M, Smith-Lovin L, Cook JM (2001) Birds of a feather: Homophily in social networks. *Annu Rev Sociol* 27:415–444.
20. Madar N, Kalisky T, Cohen R, ben-Avraham D, Havlin S (2004) Immunization and epidemic dynamics in complex networks. *Eur Phys J B* 38:269–276.
21. Hidalgo CA, Klinger B, Barabási A-L, Hausmann R (2007) The product space conditions the development of nations. *Science* 317:482–487.
22. Venkatesan K, et al. (2008) An empirical framework for binary interactome mapping. *Nat Methods* 6:83–90.
23. Newman MEJ (2002) Assortative mixing in networks. *Phys Rev Lett* 89:208701.
24. Park J, Barabási A-L (2007) Distribution of node characteristics in complex networks. *Proc Natl Acad Sci USA* 104:17916–17920.
25. Anagnostopoulos A, Kumar R, Mahdian M (2008) Influence and correlation in social networks. *14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM, New York), pp 7–15.
26. Crandall D, Cosley D, Huttenlocher D, Kleinberg J, Suri S (2008) Feedback effects between similarity and social influence in online communities. *14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM, New York), pp 160–168.
27. Java A, Kolari P, Finin T, Oates T (2006) Modeling the spread of influence on the blogosphere. *WWW Conference (Edinburgh, UK)*.
28. Jackson MO (2008) Average distance, diameter, and clustering in social networks with homophily. *Internet and Network Economics*, Lecture Notes in Computer Science, eds Papadimitriou C, Zhang S (Springer, Berlin), Vol 5385, pp 4–11.
29. Backstrom L, Huttenlocher D, Lan X, Kleinberg J (2006) Group formation in large social networks membership, growth, and evolution. *KDD '06: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM, New York), pp 44–54.
30. Brock WA, Durlauf SN (2001) Discrete choice with social interactions. *Rev Econ Stud* 68:235–260.
31. Bramoullé Y, Djebbari H, Fortin B (2009) Identification of peer effects through social networks. *J Econometrics* 150:41–55.
32. Manski CF (1993) Identification of endogenous social effects: The reflection problem. *Rev Econ Stud* 60:531–542.
33. Frank O, Strauss D (1986) Markov graphs. *J Am Stat Assoc* 81:832–842.
34. Snijders T, Steglich C, Schweinberger M (2006) Modeling the co-evolution of networks and behavior. *Longitudinal Models in the Behavioral and Related Sciences*, eds Montfort Kv, Oud H, Satorra A (Erlbaum, Philadelphia), pp 41–71.
35. Tucker C (2008) Identifying formal and informal influence in technology adoption with network externalities. *Management Sci* 54:2024–2038.
36. Sacerdote B (2001) Peer effects with random assignment: Results for Dartmouth roommates. *Q J Econ* 116:681–704.
37. Cartwright D (1965) Influence, leadership, control. *Handbook of Organizations*, ed March JG (Rand McNally, Chicago), pp 1–47.
38. Rosenbaum PR, Rubin DB (1983) The central role of the propensity score in observational studies for causal effects. *Biometrika* 70:41–55.
39. Coleman JS (1958) Relational analysis: The study of social organization with survey methods. *Human Organization* 17:28–36.
40. Centola D, Macy M (2007) Complex contagions and the weakness of long ties. *Am J Sociol* 113:702–734.
41. Dodds PS, Watts J (2004) Universal behavior in a generalized model of contagion. *Phys Rev Lett* 92:218701.