Inferring friendship network structure by using mobile phone data

Nathan Eagle[a,b,1], Alex (Sandy) Pentland[b], and David Lazer[c]

[a]Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501; [b]MIT Media Laboratory, Massachusetts Institute of Technology, E15–383, 20 Ames Street, Cambridge, MA 02139; and [c]Departments of Political Science and Computer Science, Northeastern University, Boston, MA 02115

Edited by Susan Hanson, Clark University, Worcester, MA, and approved July 1, 2009 (received for review January 14, 2009)

Data collected from mobile phones have the potential to provide insight into the relational dynamics of individuals. This paper compares observational data from mobile phones with standard self-report survey data. We find that the information from these two data sources is overlapping but distinct. For example, self-reports of physical proximity deviate from mobile phone records depending on the recency and salience of the interactions. We also demonstrate that it is possible to accurately infer 95% of friendships based on the observational data alone, where friend dyads demonstrate distinctive temporal and spatial patterns in their physical proximity and calling patterns. These behavioral patterns, in turn, allow the prediction of individual-level outcomes such as job satisfaction.

engineering-social systems | relational inference | social network analysis | reality mining | relational scripts

The field devoted to the study of the system of human interactions—social network analysis—has been constrained in accuracy, breadth, and depth because of its reliance on self-report data. Social network studies relying on self-report relational data typically involve both limited numbers of people and a limited number of time points (usually one). As a result, social network analysis has generally been limited to examining small, well-bounded populations, involving a small number of snapshots of interaction patterns (1). Although important work has been done over the last 30 years to analyze the relationship between self-reported and observed behavior, much of the social network literature is written as if self-report data are behavioral data.

There is, however, a small but emerging thread of research examining social communication patterns based on directly observable data such as e-mail (2, 3) and call logs (4, 5). Here, we demonstrate the power of collecting not only communication information but also location and proximity data from mobile phones over an extended period, and compare the resulting behavioral social network to self-reported relationships from the same group. We show that pairs of individuals that report themselves as friends demonstrate distinctive behavioral signatures as measured only by the mobile phone data. Further, these purely objective measures of behavior show powerful relationships with key outcomes of interest at the individual level—notably, satisfaction.

The Reality Mining study followed 94 subjects using mobile phones preinstalled with several pieces of software that recorded and sent the researcher data about call logs, Bluetooth devices in proximity of approximately five meters, cell tower IDs, application usage, and phone status (6, 7). Subjects were observed using these measurements over the course of nine months and included students and faculty from two programs within a major research institution. We also collected self-report relational data from each individual, where subjects were asked about their proximity to, and friendship with, others. Subjects were also asked about their satisfaction with their work group Full details on data collection and variable construction are available in the *SI Text*. We will hereafter refer to data collected purely from mobile phones as "behavioral" data as opposed to "self-report" data.

We conducted three analyses of these data. First, we examined the relationship between self-report and behavioral data. Second, we analyzed whether there were behaviors identified in the mobile phone data that were characteristic of friendship. Third, we studied the relationship between behavioral data and individual satisfaction.

Results

Behavioral Versus Self-Report Data. The reliability of existing measures for relationships has been the subject of sharp debate over the last 30 years, starting with a series of landmark studies in which it was found that behavioral observations were surprisingly weakly related to reported interactions (8–10). There are multiple layers of cognitive filters that influence whether a subject reports a behavior (11). Existing research suggests that people are good at recalling long-term, but not short-term, social structures (12). We examine whether there are systematic biases in recall that have been observed in other areas with respect to human memory (13), specifically, whether there are recency and salience biases in recall of physical proximity. A recency bias is one where memories are biased toward recent events. A salience bias is one where memories are biased toward more vivid events. Here, we capture recency by the quantity of interactions in a fixed period preceding the survey, and salience by whether the individual in question is a friend or nonfriend.

We test for recency and salience biases by comparing self-reported proximity to observed proximity, examining whether self-reports were biased toward recent and salient proximity. Specifically, subjects were asked about their typical proximity to the other individuals in the study. These self-reports were compared with average daily proximity based on the Bluetooth scans. Although most (69%) observed proximity >0 was reported as nonproximity, when proximity was reported, it was typically overestimated: The average reported proximity was 87 min per day whereas the average observed proximity was only 33 min per day. We also found that for proximity >0, friends were much more accurate at reporting proximity than nonfriends. In this case, there was a statistically significant but small correlation between observed and reported proximity among individuals who worked together but did not consider each other friends ($r = .155$, $P < 0.001$), whereas there was a substantially stronger relationship for friends ($r = .412$, $P < 0.001$). Additionally, we found that when subjects were asked about long-term proximity patterns, recent proximity had a large and significant ($P < 0.001$)

Author contributions: N.E., A.S.P., and D.L. designed research; N.E. and A.S.P. performed research; N.E. and D.L. contributed new reagents/analytic tools; N.E. and D.L. analyzed data; and N.E. and D.L. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

See Commentary on page 15099.

[1]To whom correspondence should be addressed. E-mail: nathan@mit.edu.

This article contains supporting information online at www.pnas.org/cgi/content/full/0900282106/DCSupplemental.

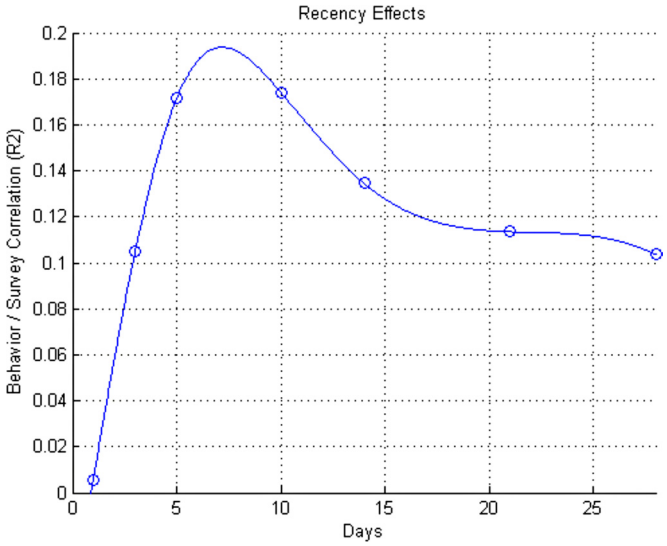

Fig. 1. The effect of recency on self-report data. When subjects were asked to report their general proximity patterns, the survey responses were biased in favor of recent behavior. Although the correlation is lowest when only using observed behavior during the day before the survey, as we expand the observational time window to seven days, the correlation between the self-report and observed behavior increases. However, expanding the time window beyond seven days results in a decreasing correlation, leading us to conclude that subjects recall of information about their interactions begins to degrade after approximately one week.

effect on self-reports, independent of the long-term observed proximity. Fig. 1 shows that this recency effect peaks when we use a seven-day window for recent interactions, suggesting that individuals recall of information about their interactions begins to degrade after approximately one week.

Relational Scripts. Observing intrinsically cognitive relationships, such as friendship or love, is a fundamentally different challenge than observing whether two people are near each other. It is clear, for example, that two individuals can be friends without any observable interactions between them for a given period. Context, however, especially spatial and temporal, is likely to be an important indicator of particular types of relationship, where spending a couple of hours in close proximity at a location away from work on a Saturday night is quite different from spending a couple of hours in close proximity at work on a Wednesday afternoon, for example. Here, we borrow from cognitive science the idea of scripts (14, 15). Specifically, we examine whether proximity, location, and time cluster together in a predictable fashion and whether these behavioral patterns, in turn, predict friendship.

Fig. 2 captures the average hour-by-hour levels of proximity for symmetric friend and nonfriend dyads, as well as asymmetric dyads. Proximity is generally much higher for friends, but time and location are important predictors as well, where the ratio of proximity off hours outside work is much higher for friends than nonfriends. We therefore divided proximity into variables corresponding to on campus/off campus, daytime/nights (separated at 8 a.m. and 8 p.m.), weekend proximity, and phone communication. A factor analysis (Table 1) revealed that two factors capture most of the variance in these variables. The first factor, which loads most heavily on proximity at work during the daytime is labeled "in-role," as it represents traditional behavior between colleagues. The second factor, which loads most heavily on off-campus proximity in the evening and on weekends, is labeled "extra-role" and is representative of behaviors outside the work environment. As depicted in Fig. 3, by using just the extra-role factor from this analysis, it is possible with a single parameter to accurately predict 96% of symmetric reports of nonfriendship and 95% reports of symmetric friendship. That is, we can accurately predict self-reported friendships based only on objective measurements of behavior because the strong cultural norms associated with social constructs such as friendship produce differentiated and recognizable patterns of behavior.

Unsurprisingly, the factor scores for nonreciprocal friendships fall systematically between the reciprocal friendship dyads and the nonfriend dyads. This probably reflects the fact that friend-

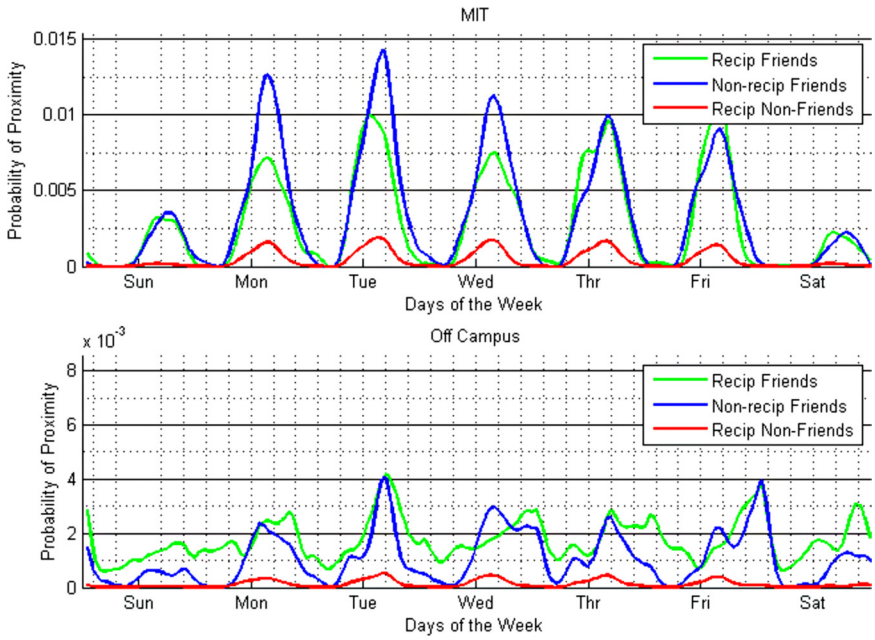

Fig. 2. Probability of proximity. Proximity probabilities at work and off campus for symmetric friend, asymmetric friend, and nonfriend dyads. Probability of proximity is calculated for each hour in the week and is generally much higher for friends than nonfriends. However, it is also apparent that asymmetric and symmetric friend dyads have different temporal and spatial patterns in proximity, with symmetric friends spending more time together off campus in the evenings.

Table 1. Factor analysis loadings

Variable name	Specific variance	Factor 1: Extra-role	Factor 2: In-role
Work proximity, weekdays, 8 a.m.–8 p.m.	0.005	−0.119	1.07
Work proximity, weekdays, 8 p.m.–8 a.m.	0.568	0.555	0.144
Work proximity, weekends	0.642	0.501	0.137
Off-campus proximity, weekdays, 8 a.m.–8 p.m.	0.310	0.691	0.195
Off-campus proximity, weekdays, 8 a.m.–8 p.m.	0.240	0.946	−0.123
Off-campus proximity, weekends	0.291	0.914	−0.119
Phone communication	0.806	0.469	−0.047

For relationship inference, based on a promax rotation, it is possible to divide the dyadic variables into the two factors above: in-role and extra-role communication. In-role communication consists of the behaviors typically associated with colleagues whereas extra-role communication corresponds to more personal behavior such as proximity on Saturday nights or at home.

ships are not categorical in nature, and that nonreciprocal friendships may be indicative of moderately valued friendship ties. Thus, inferred friendships may actually contain more information than is captured by surveys that are categorical in nature. A pairwise analysis of variance using the Bonferroni adjustment shows that data from friendships, nonreciprocal friendships, and reciprocated nonfriend relationships do indeed come from three distinct distributions ($F = 9.2$, $P < 0.005$).

It is clear that there is enormous redundancy in these data, where proximity of a pair one week is correlated with proximity the next week. We find that for most of the proximity variables, observation for two weeks will largely replicate the data we produced here from nine months of observations, where the median correlation of two weeks with the full nine months of data for each of the components of the factor analysis varied from a low of 0.38 (for phone communication) to a high of 0.82 (for proximity at work during the day).

The longitudinal nature of the study also enabled us to track how the reported and observed relationships changed during the academic year. The differences in reported friendships over the course of the nine months can be explained as a combination of both reporting error (e.g., dyads failing to report a particular relationship when completing the survey) or the evolution of the relationships (e.g., dyads who become friends between January and May). Fig. 4 shows the extra-role factor distributions calculated by using only dyadic behavior from September until January. It is clear that when both sets of dyads report they were not friends in January, the autumn behaviors of the dyads that subsequently report a friendship in May are quite distinct from the autumn behaviors of the dyads who consistently reported they were not friends during both January and May surveys. This finding suggests that the observational data are capturing information about relationships that self-reports are missing.

Predicting Satisfaction Based on Behavioral Data. The preceding method still results in a high number of apparent false positives, because there are almost 50× as many mutual nonfriendships as there are mutual friendships. In particular, only 21 of the 67 predicted mutual friendships were reported in the surveys. However, this likely understates the accuracy of the behavior-based inference of friendship, because the self-report measure for friendship itself is probably not perfectly reliable. To compare the validity of these two measures of friendship, we examine the effectiveness of both measures in predicting social integration in work groups. We compare two predictive models, one based on self-reported friendship and one based on an inferred measure of friendship using the dyadic weights associated with the factor analysis described in *Relational Scripts* (see Fig. 1). In both models, the predictors are the number of friendships—with

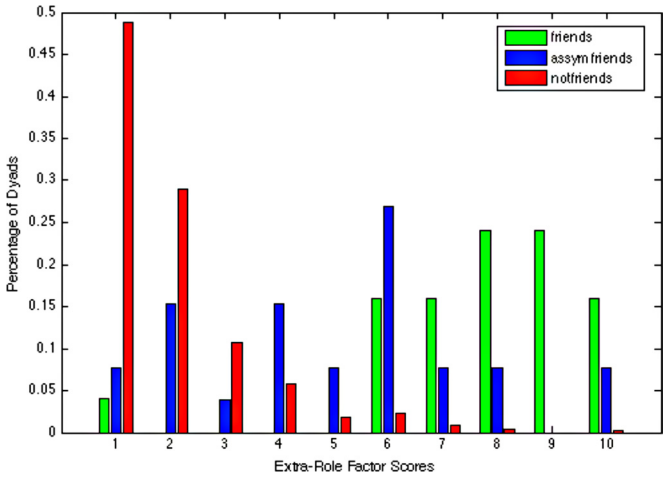

Fig. 3. Normalized extra-role histograms. The distributions of a pair of colleagues extra-role communication factor scores segmented by relationship. Ninety-five percent (21/22) of the symmetric friendships have extra-role scores above 5, whereas ninety-six percent (901/935) of symmetric nonfriends have extra-role scores below 5. The 28 asymmetric friends have more behavioral variance, drawing from behaviors characteristic of both nonfriends and friends.

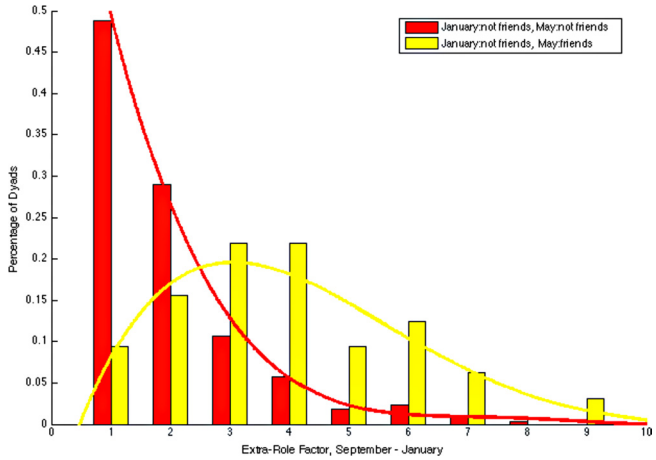

Fig. 4. A histogram of the extra-role distribution generated from behavioral data collected from September to January for two sets of dyads. The red bins represent the dyads that consistently confirmed they were not friends on both the January and May survey ($n = 2153$). The yellow bins represent the dyads that confirmed they were not friends on the January survey, but at least one individual named the other as a friend on the May survey ($n = 32$). Clearly these two sets of dyads come from distinct distributions; potential explanations for the yellow distribution could be survey error in January (i.e., the friendships existed, but were not reported in January), or that the dyads' behavior during the autumn was indicative of budding friendships that they only became aware of during the subsequent year.

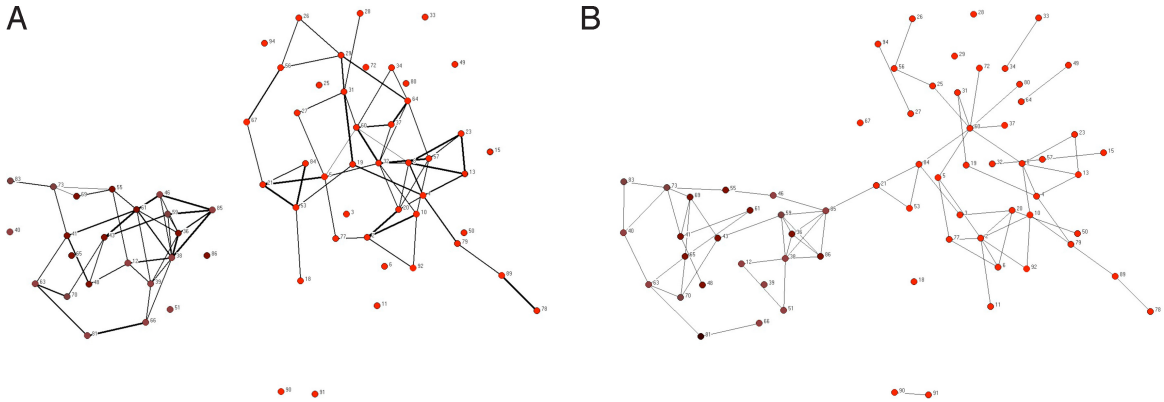

Fig. 5. Inferred, weighted friendship network vs. reported, discrete friendship network. Frame *A* shows the inferred friendship network with edge weights corresponding to the factor scores for factor 2, extra-role communication. Frame *B* shows the reported friendship network. Node colors highlight the two groups of colleagues, first-year business school students (brown) and individuals working together in the same building (red).

a dummy for zero friendships to capture nonlinearities (positive expected relationship), the average proximity to friends while at work (positive expected relationship), and phone communication with friends while at work (negative expected relationship).

The relationship between satisfaction and interactions patterns, shown in Table 2, was exactly as predicted, that is, having friends—especially ones to whom you were near at work—predicted satisfaction with the work group, and calling friends while at work was associated with lack of satisfaction with the work group. What is important, from the perspective of this paper, is that the inferred friendship network (see Fig. 5) produced substantively identical results to the self-report model, with a slightly improved fit. These nearly identical results suggest that it is possible to accurately infer subjective job satisfaction based solely on behavioral data, validating the inferred measure of friendship.

Discussion

Data collected from mobile phones have the potential to provide insight into the underlying relational dynamics of organizations, communities and, potentially, societies. At the microlevel, these methods provide, for example, a new approach to studying collaboration and communication within organizations—allowing the examination of the evolution of relationships over time. Leveraging these behavioral signatures to accurately characterize relationships in the absence of survey data also has the potential to enable the quantification and prediction of macro social network structures that were heretofore unobservable. There is no technical reason why data like these cannot be collected from millions of people throughout the course of their lives. Furthermore, although the collection of such data raises serious privacy issues that need to be considered (16–18), the potential for achieving important societal goals, from urban planning to public health, is considerable.

This paper thus offers a necessary first step, linking the predominant existing methodologies to collect social network data based on self-reports to data that can be collected automatically via mobile phones. Our results suggest that behavioral observations from mobile phones, as a complement to self-report surveys, provide insight not just into observable behavior but also into purely cognitive constructs, such as friendship and individual satisfaction. Although the specific results are surely embedded within the social milieu in which the study was grounded, the critical next question is how much these patterns vary from context to context.

Materials and Methods

Our extensive set of longitudinal behavioral data were collected by using 100 specially programmed Nokia 6600 smartphones. Details about the experimental design, the subject pool, data collection protocols, and a description of the dyadic variable construction are provided in the *SI Text*, Figs. S1–S7 and Tables S1 and S2.

Data collected from these mobile phones consist of cellular tower transition, Bluetooth device discovery scans, and communication events. Although most cellular towers have ranges extending several square kilometers, in typical urban settings tower densities are significantly higher. Each tower has been assigned an ID that is logged by the mobile phones in our study. By using the tower IDs and respective transition timings (time stamps when the phone is handed off between cellular towers), we are able to estimate location and movement. Conducting periodic Bluetooth scans at 5-min intervals has generated ≈4 million proximity events in the dataset. For each proximity event, we have logged the two proximate Bluetooth phones, the current associated cellular tower for each of the phones, and the time and date of the event. Because all of the phones are scanning every five minutes, if two subjects were together for 100 min there would be a total of 40 recorded proximity events. We therefore approximate each proximity event to be representative of a 2.5-min time interval. To estimate the amount of proximity at a particular location such as ''work'', we multiply this

Table 2. Ordinary least squares regression of relationship between satisfaction and interaction patterns

Model	MC self-report	Model	CMC observational
Friendship dummy, equals 1 no friends (sr)	−0.370*(0.175)	Friendship dummy (inf)	−0.392*(0.170)
Number of reciprocatedfriendships (sr)	0.377*(0.166)	Number of friends (inf)	0.483** (0.164)
Average proximity tofriends (sr) while at work	0.719** (0.176)	Average proximity to friends (inf) while at work	0.698** (0.188)
Phone communication with friends (sr) while at work	−0.497** (0.171)	Phone communication with friends (inf) while at work	−0.571** (0.182)
Adjusted r^2	0.161	Adjusted r^2	0.180

Models are based on a combination of self-report or inferred relational and interaction data. $N = 94$ for all models, with $*P<0.05$, and $**P<0.01$, two-tailed test.

time interval by the number of proximity events that involved the cellular towers associated with that location.

Analysis of the dyadic variables was performed by using the nonparametric multiple regression quadratic assignment procedure (MRQAP), a standard technique to analyze social network data. The MRQAP technique treats square network matrices as distinct variables that can be incorporated into a regression by sampling from a repeated permutation to generate a random estimate of the relationship between multiple matrices. We performed additional tests on the dyadic variables by using the results from a factor analysis with a variety of rotations to assess the robustness of the friendship edge prediction.

ACKNOWLEDGMENTS. The authors would like to thank Mika Raento for technical support, Devon Brewer for useful comments, and Nokia for financial support and mobile phone donation. N. E. was supported by the Santa Fe Institute. N. E. and A. P. were sponsored by the Massachusetts Institute of Technology Media Laboratory.

1. Wasserman S, Faust K (1994) Social Network Analysis: Methods and Applications (Cambridge Univ Press, New York).
2. Kossinets G, Watts D (2006) Empirical analysis of an evolving social network. *Science* 311:88–90.
3. Ebel H, Mielsch L, Bornholdt S (2002) Scale-free topology of e-mail networks. *Phys Rev E* 66:35103.
4. Aiello W, Chung F, Lu L (2000) A random graph model for massive graphs. *Proceedings of the Thirty-Second Annual ACM Symposium on Theory of Computing* (Association for Computing Machinery, New York) pp. 171–180.
5. Onnela J, et al (2007) Structure and tie strengths in mobile communication networks. *Proc Natl Acad Sci* 104:7332–7336.
6. Eagle N, Pentland A (2006) Reality mining: Sensing complex social systems. *Personal Ubiquitous Comput* 10:255–268.
7. Raento M, Oulasvirta A, Petit R, Toivonen H (2005) ContextPhone: A prototyping platform for context-aware mobile applications. *IEEE Pervasive Computing* 4:51–59.
8. Bernard HR, Killworth PD (1977) Informant accuracy in social network data II, *Hum Comm Res* 4:3–18.
9. Killworth PD, Bernard HR (1976) Informant accuracy in social network data. *Hum Org* 35:269–286.
10. Marsden P (1990) Network data and measurement. *Ann Rev Sociol* 16:435–463.
11. Freeman L (1992) Filling in the blanks: A theory of cognitive categories and the structure of social affiliation. *Soc Psychol Q* 55:118–127.
12. Freeman L, Romney A, Freeman S (1987) Cognitive structure and informant accuracy. *Am Anthropol* 89:310–325.
13. Frensch P (1994) Composition during serial learning: A serial position effect. *J Exp Psychol Learn Mem Cognit* 20:423–443.
14. Abelson R (1981) Psychological status of the script concept. *Am Psychol* 36:715–729.
15. Krackhardt D (1990) Assessing the political landscape: Structure, cognition, and power in organizations. *Admin Sci Q* 35:342–369.
16. Butler D (2007) Data sharing threatens privacy.*Nature* 449:644–645.
17. Anonymous (2007) A matter of trust: Social scientists studying electronic interactions must take the lead on preserving data security. *Nature* 449:637–638.
18. Lazer D, et al. (2009) Computational social science. *Science* 323:721–723.