Inferring Social Network Structure using Mobile Phone Data

Nathan Eagle[1,3*], Alex (Sandy) Pentland[3], David Lazer[2]

[1]MIT Design Laboratory, Massachusetts Institute of Technology, Cambridge, MA
[2]John F. Kennedy School of Government, Harvard University, Cambridge, MA
[3]MIT Media Laboratory, Massachusetts Institute of Technology, Cambridge, MA

*****To whom correspondence should be addressed: nathan@media.mit.edu

We analyze 330,000 hours of continuous behavioral data logged by the mobile phones of 94 subjects, and compare these observations with self-report relational data. The information from these two data sources is overlapping but distinct, and the accuracy of self-report data is considerably affected by such factors as the recency and salience of particular interactions. We present a new method for precise measurements of large-scale human behavior based on contextualized proximity and communication data alone, and identify characteristic behavioral signatures of relationships that allowed us to accurately predict 95% of the reciprocated friendships in the study. Using these behavioral signatures we can predict, in turn, individual-level outcomes such as job satisfaction.

In a classic piece of ethnography from the 1940s, William Whyte carefully watched the interactions among Italian immigrants on a street corner in Boston's North End (*1*). Technology today has made the world like the street corner in the 1940s—it is now possible to make detailed observations on the behavior and interactions of massive numbers of people. These observations come from the increasing number of digital traces left in the wake of our actions and interpersonal communications. These digital traces have the potential to revolutionize the study of collective human behavior. This study examines the potential of a particular device that has become ubiquitous over the last decade—the mobile phone—to collect data about human behavior and interactions, in particular from face-to-face interactions, over an extended period of time.

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. Self-reports are potentially mediated by confounding factors such as beliefs about what constitutes a relationship, ability to recall interactions, and the willingness of individuals to supply accurate information about their relationships. Whole network studies relying on self-report relational data typically involve both limited numbers of people (usually less than 100) and a limited number of time points (usually 1). 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 (*2*). While important work has been done over the last 30 years to parse the relationship between self-reported and observed behavior, much of social network research is written as if self-report data *are* behavioral data.

There is, however, a small but emerging thread of literature examining interaction data, e.g., based on e-mail (*3*, *4*) and call logs (*5*). In this paper we use behavioral data collected from mobile phones (*6*) to quantify the characteristic behaviors underlying relational ties and cognitive constructs reported through surveys. We focus our study on three types of information that can be captured from mobile phones: communication (via call logs), location (via cell towers), and proximity to others (via repeated Bluetooth scans). The resulting data provide a multi-dimensional and temporally fine grained record of human interactions on an unprecedented scale. We have collected 330,000 hours of these behavioral observations from 94 subjects. Further, in principle, the methods we discuss here could be applied to hundreds of millions of mobile phone users.

Measuring Relationships

The core construct of social network analysis is the relationship. 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 (*7, 8, 9*). These studies, in turn, were subject to three critiques: First, that people are far more accurate in reporting long term interactions than short term interactions (*10*). Second, that it is possible to reduce the noise in network data because every dyad (potentially) represents two observations, allowing an evaluation (and elimination) of biases in the reports (*11*). Third, that in many cases the construct of theoretical interest was the cognitive network, not a set of behavioral relations (*12*). Here, behavior is defined as some set of activities that is at least theoretically observable by a third party, whereas a cognitive tie reflects some belief an individual holds about the relationship between two individuals (*13*).

There are multiple layers of conscious and subconscious cognitive filters that influence whether a subject reports a behavior (*10*, *14*). Cognitive sub-processes are engaged in the encoding and retrieval of a behavior instance from a subject's memory; the subject must understand the self-report request (i.e., survey question) to refer to the particular behavior; and the subject still gets to decide whether to report a particular behavior as a tie or not – a particular issue in the study of sexual or illicit relationships, for example (*15*). These filtering processes contribute to a problematic gap between actual behaviors and self-report data.

Divergences between behavior and self-reports may be viewed as noise to be expunged from the data (*11*), or as reflecting intrinsically important information. For example, if one is interested in status, divergences between the two self-reports of a given relationship between two people, or between reported and observed behavior, may be of critical interest (*18*). In contrast, if one is focused on the transmission of a disease, then the actual behaviors underlying those reports will be of central interest, and those divergences reflective of undesirable measurement error (*15*).

None of the above research examines the relationship between behavior and cognition for relationships that are intrinsically cognitive. Observing friendship or love is a fundamentally different challenge than observing whether two people talk to each other; e.g., two individuals can be friends without any observable interactions between them for a given period.

In this paper we demonstrate the power of collecting behavioral social network data from mobile phones. We first revisit the earlier studies on the inter-relationship between relational behavior and reports of relational behavior, but focusing in particular on some of the biases that the literature on memory suggest should arise. We then turn to the inter-relationship between behavior and reported friendships, finding that pairs of individuals that are friends demonstrate quite distinctive relational behavioral signatures. Finally, we show that these purely behavioral measures show powerful relationships with key outcomes of interest at the individual level—notably, satisfaction.

Research Design

This study follows ninety-four subjects using mobile phones pre-installed with several pieces of software that record and send the researcher data on call logs, Bluetooth devices in proximity, cell tower IDs, application usage, and phone status (*19*). These subjects were observed via mobile phones over the course of nine months, representing over 330,000 person-hours of data (about 35 years worth of observations). Subjects included students and faculty from a major research institution; the resulting dataset is available for download. We also collected self-report relational data, where subjects were asked about their proximity to and friendship with others. Subjects were also asked about their satisfaction with their work group (*20*).

We conduct three analyses of these data. First, we examine the relationship between the behavioral and self-report interaction data. Second, we analyze whether there are behaviors characteristic of friendship. Third, we study the relationship between behavioral data and individual satisfaction.

Relationship between Behavioral and Self-Report Data

Subjects were asked how often they were proximate to other individuals at work. The boxplot shown in Figure 1 illustrates the remarkably noisy, if mildly positive, relationship between these self-report data and the observational data from Bluetooth scans. The literature on memory suggests a number of potential biases in the encoding into and retrieval from long term memory. We focus on two potential biases: recency and salience. Recency is simply the tendency for more recent events to be recalled (*21*). Salience is the principle that prominent events are more likely to be recalled (*22*). We therefore incorporate into our data analysis a measure of recent interactions (the week before the survey was answered), and a variety of measures of salience. The key question is whether recent and salient interactions significantly affect the subject's ability to accurately report average behaviors.

Using a multiple regression quadratic assignment procedure, common to the analysis of the adjacency matrices representing social networks, we can assess the significance of the predictive value of variables (*18*, *23*). While proximity at work was significantly related to self-reports, remarkably, proximity *outside* work was the single most powerful predictor of reported proximity at work. Other relational behavior, including proximity that was recent, on Saturday night, and between friends, were independently and significantly predictive of whether an individual reported proximity to someone else during work (p<.0001). These systematic biases limit the effectiveness of strategies designed to reduce noise in self-report data through modeling the biases of particular individuals (*10*), since these biases will affect both members of a dyad in the same direction (e.g., recency).

Behavioral Characteristics of Friendship

What does a friendship “look like”? Certainly, we would anticipate relatively more phone calls and proximity between a pair of people who view one another as friends. More generally we anticipate that there are culturally embedded relational routines that friends tend to follow—for example, getting together outside of workplace hours and location, especially Saturday nights. We constructed seven dyadic behavioral variables: volume of phone communication and six contextualized variants of proximity. Figure 2 confirms that for all the dyadic behavioral variables, reciprocal friends score far higher than reciprocal non-friends (subjects who work together but neither considers the other a friend). A multivariate analysis confirms that the seven behavioral variables are significantly and independently related to reciprocated friendship/nonfriendship (p < .001). Further, in all but one case, non-reciprocal friends have intermediate scores. That one case is proximity at work, which suggests that there is a cultural/cognitive ambiguity as to whether this particular set of behaviors constitutes “friendship.”

A factor analysis reveals that two factors capture most of the variance in these variables, where the first factor seems to capture in-role communication and proximity (those interactions likely to be associated with work, e.g. proximity at work), and the second factor extra-role communication and proximity (those interactions that are unlikely to be associated with work, such as Saturday night proximity). As depicted in Figure 3, a key finding of this study is that using just the extra-role communication factor from this analysis, it is possible to accurately predict 96% of symmetric non-friends and 95% of symmetric friends; in-role communication produces a similar accuracy. Thus we can accurately predict self reported friendships based only on objective measurements of behavior. These findings imply that the strong cultural norms associated with social constructs such as friendship produce differentiated and recognizable patterns of behavior. Leveraging these behavioral signatures to accurately characterize relationships in the absence of survey data has the potential to enable the quantification and prediction of social network structures on a much larger scale than is currently possible.

Unsurprisingly, non-reciprocal friendships fall systematically between these two categories. This probably reflects the fact that friendships are not categorical in nature, and that non-reciprocal 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, non-reciprocal friendships, and reciprocated non-friend relationships do indeed come from three distinct distributions (F>9, p<.005).

Predicting Satisfaction Based on Behavioral Data

The preceding analysis highlights the potential to use the digital traces of previous behavior to infer cognitive constructs such as friendship. Do those inferences, in turn, predict meaningful individual level outcomes? One of the longest standing findings in the study of social support is the positive impact of social integration on the individual (*24*). We examine here whether one can predict, in particular, satisfaction of the individual with their work group based solely on relational behavior. We begin with a standard analysis of the relationship between satisfaction and number of friends, which demonstrates a moderately positive and significant ($p < .05$), relationship. However, the model is significantly strengthened when we add two variables, combining self-report and behavioral data: average daily proximity to friends (a positive and significant relationship, $p< .001$), and average phone communication with friends (a negative and significant relationship, $p < .005$). In the final two analyses we reran these regressions, replacing the self-report data with the inferred friendship relationships, using first a binary network based on a cut off for the extra-role factor, and second, a weighted network using each dyad's factor score. These analyses produced a set of parameter estimates that are substantively identical to those based on self-reported friendships, with an improvement of model fit. That is, it is possible to accurately infer social integration and thus satisfaction based solely on behavioral data without apparent deterioration in the model.

Conclusions

This paper contains the results from a large scale study of physical proximity among individuals, encompassing 35 years worth of observations at five second increments, and combining them with phone log, locational, and self-report data. We anticipate that the methods outlined here will have a major impact in the social sciences, providing insight into the underlying relational dynamics of organizations, communities and, potentially, societies. At the micro level these methods, for example, provide a new approach to studying collaboration and communication within organizations—allowing the examination of the evolution of relationships over time. More dramatically, these methods allow for an inspection of the dynamics of macro networks that were heretofore unobservable. There is no technical reason why data cannot be collected from hundreds of millions of people throughout the course of their lives. Further, while the collection of such data raises serious privacy issues that need to be considered, the potential for achieving important societal goals is considerable. The implications for epidemiology alone are foundational, as they are for the study of sociology, politics, and organizations, among other social sciences.

This paper thus offers a necessary first step in this revolution, 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 provide insight not just into observable behavior, but also into purely cognitive constructs, such as friendship and individual satisfaction.

While 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.

[1] W. Whyte. 1993. *Street Corner Society: The Social Structure of an Italian Slum*. 4th edition. Chicago, IL: University of Chicago Press.

[2] For example, in what is the standard reference in social network analysis, none of the data sets referenced include as many as 100 subjects. S. Wasserman and K. Faust. 1994. *Social Network Analysis, Methods and Applications*. Cambridge, UK: Cambridge University Press.

[3] G. Kossinets and D. J. Watts. 2006. **“**Empirical Analysis of an Evolving Social Network,” *Science* 311: 88-90.

[4] H. Ebel, L-I. Mielsch, and S. Bornholdt. 2002. “Scale-free topology of e-mail networks,” *Phys Rev E* 66: 35103.

[5] W. Aiello, F. Chung, L. Lu. 2000. A random graph model for massive graphs, Annual ACM Symposium on Theory of Computing, Proceedings of the thirty-second annual ACM symposium on Theory of computing: 171–180.

[6] N. Eagle, A. Pentland. 2006. “Reality Mining: Sensing Complex Social Systems”, *Personal and Ubiquitous Computing*, 10(4): 255-268.

[7] W. H. Bernard, P. Killworth, and L. Sailer. 1979. “Informant accuracy in social networks. Part IV: A comparison of clique-level structure in behavioral and cognitive network data.” *Social Networks 2*: 191-218.

[8] P. D. Killworth and H. R. Bernard. 1976. “Informant accuracy in social network data,” *Human Organization 35(8)*: 269-286.

[9] P.V. Marsden. 1990. “Network data and measurement,” *Annual Review of Sociology* 16: 435-463.

[10] L.C. Freeman, A.K. Romney, S.C. Freeman. 1989. “Cognitive Structure and Informant Accuracy,” *American Anthropologist* 89.

[11] C. Butts. 2003. “Network inference, error and informant (in)accuracy: a Bayesian approach,” *Social Networks 25:2*: 103-140.

[12] W.D. Richards. 1985. “Data, models, and assumptions in network analysis. In *Organizational Communication: Traditional Themes and New Directions*, ed R. D. McPhee , P. K. Tompkins. Sage, Beverly Hills: 108-128.

[13] D. Krackhardt. 1990. “Assessing the Political Landscape: Structure, Power, and Cognition in Organizations.” *Administrative Science Quarterly*, 35: 342–369.

[14] L.C. Freeman. 1992. Filling in the blanks: a theory of cognitive categories and the structure of social affiliation. *Social Psychology Quarterly* 55(2): 118-127.

[15] Brewer, D. D., Potterat, J. J., Muth, S. Q., Malone, P. Z., Montoya, P. A., Green, D. A., Rogers, H. L., & Cox, P. A. 2005. “Randomized trial of supplementary interviewing techniques to enhance recall of sexual partners in contact interviews. *Sexually Transmitted Diseases*, 32, 189-193.

[18] K.M. Carley and D. Krackhardt. 1996. “Cognitive inconsistencies and non-symmetric friendships.” *Social Networks 15*: 377-398.

[19] M. Raento, A. Oulasvirta, R. Petit,H. Toivonen. 2005. “ContextPhone – A prototyping plat-form for context-aware mobile applications”. *IEEE Pervasive Computing*, 4 (2), 51-59.

[20] Full details on data collection and variable construction are available in the supporting online materials.

[21] Waugh, N. C., & Norman, D. A. 1965. Primary memory. *Psychological Review* 72: 89-104.

[22] Higgins, E. T. 1996. Knowledge activation: Accessibility, applicability, and salience. In E. T. Higgins & A. Kruglanski (Eds.), *Social psychology: Handbook of basic principles* (pp.133–168). New York: Guilford Press.

[23] Krackhardt, D. 1988. Predicting with Networks - Nonparametric Multiple-Regression Analysis of Dyadic Data. *Social Networks* 10 (4): 359-81.

[24] Durkheim, Emile. 1951. *Suicide: A Study in Sociology* translated by George Simpson and John A. Spaulding. New York: The Free Press.

Figure 1. Self-Report vs. Observational Data. Boxplots highlighting the relationship between self-report and observational proximity behavior for undirected friendship and reciprocal non-friend dyads. Self-report proximity responses, on the x-axis, are scored from 0 to 5 (see legend). The y-axis shows observed proximity in minutes per day. The height of the box corresponds to the lower and upper quartile values of the distribution and the horizontal line corresponds to the distribution's median. The 'whiskers' extend from the box to values that are within 1.5 times the quartile range while outliers are plotted as distinct points. Three outlier dyads with an observed proximity greater than 400 min/day have been excluded from the plot.

Figure 2. Normalized Dyadic Variables. The seven behavioral variables, normalized with respect to the reciprocal friendship data, are represented in the bar chart. The vertical dotted line at x=1 represents the values for reciprocal friend dyads. Reciprocal friends score higher than the other two groups for all dyadic variables with the exception of proximity at work. All three groups of dyads work together as colleagues.

Figure 3. 'In-Role' Communication vs. 'Extra-Role' Communication. Each point represents a pair of colleagues' 'in-role' and 'extra-role' communication factor scores. 95% (19/20) of the reciprocal friendships have extra-role scores above 2.3, while 96% (901/935) of reciprocal non-friends have extra-role scores below 2.3.

Figure 4a/b. Inferred, Weighted Friendship Network (a.) vs. Reported, Discrete Friendship Network (b.). The network on the left is the inferred friendship network with edge weights corresponding to the factor scores for factor 2, 'extra-role' communication. The network on the right is 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).

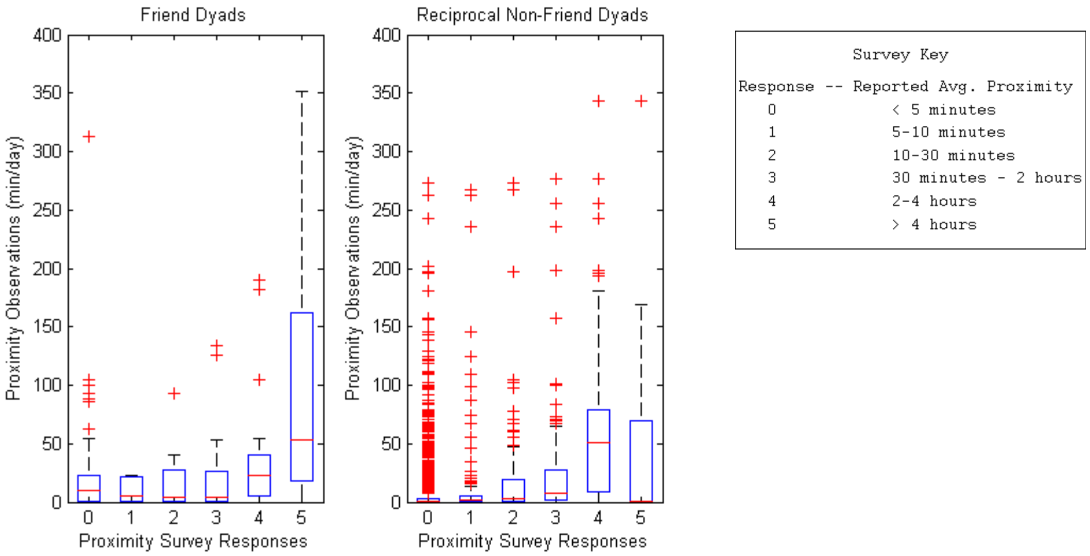


Figure 1.

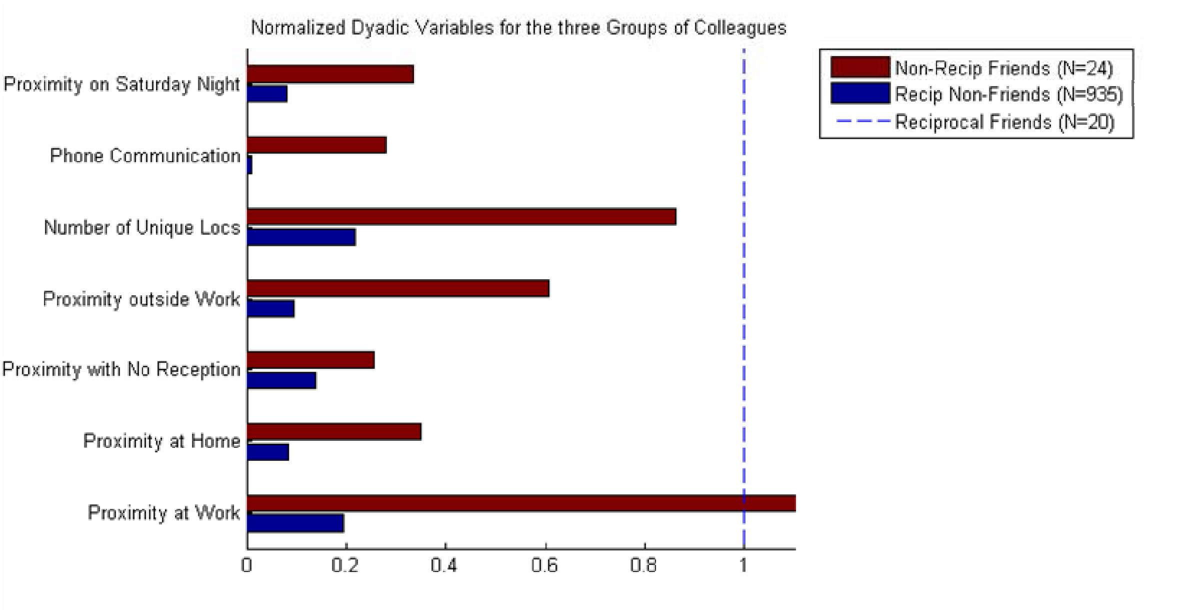


Figure 2.

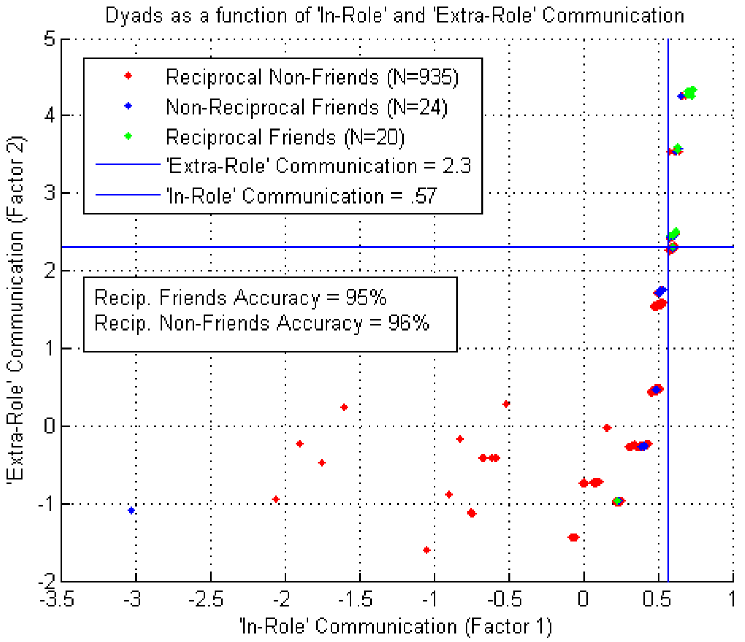


Figure 3.

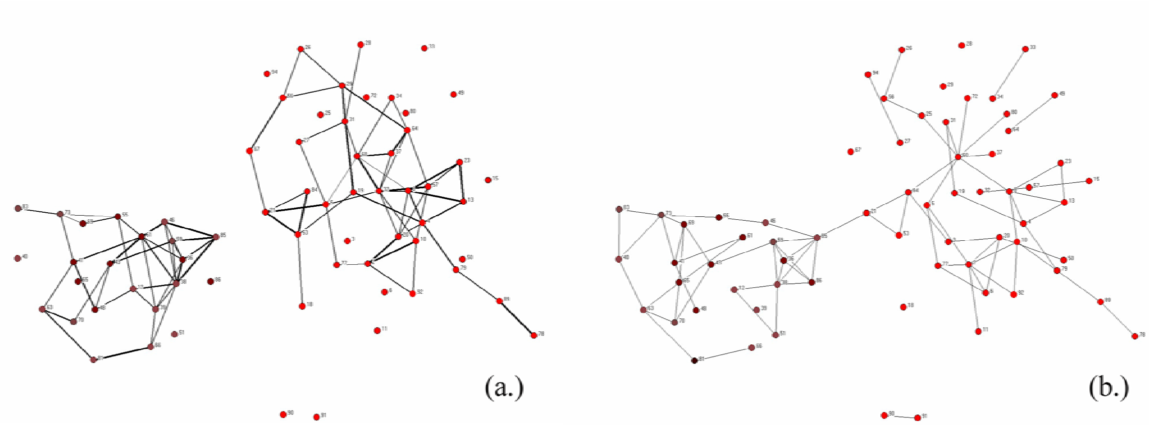


Figure 4a/b.

SUPPORTING ONLINE MATERIAL

Below we briefly discuss the behavior-cognition-report framework, and provide all relevant details on data collection and analysis, including: explanation of subject pool; data collection protocols; description of variable construction; and summary of data analyses.

BEHAVIOR-COGNITION-REPORT FRAMEWORK

Figure S1 presents a graphical portrayal of the behavioral and cognitive dimensions to relationships. Currently the large majority of social network research relies on self report data to measure both cognitive and behavioral relationships. We define behavioral relationships as relational information that is, in principle, observable to third parties; cognitive as relational information that private to the individual; and reports as those relationships that an individual indicates exist to the researcher. Relying solely on self report data thus makes it difficult to distinguish between cognitive and behavioral relationships—the measure of the belief that one has had lunch with Jane can not easily be distinguished from the actual behavior of having lunch with Jane. It is plausible that one might not report an apparent (i.e., observable) lunch with Jane because: one did not notice Jane at the table; you did not categorize this behavior as "lunch" but as something else (e.g., late morning snack); this behavior was categorized as "lunch" but not transferred to long term memory; this behavior was transferred to long term memory, but not retrieved when the survey was conducted; or that this memory was retrieved but the respondent was not willing to report it.

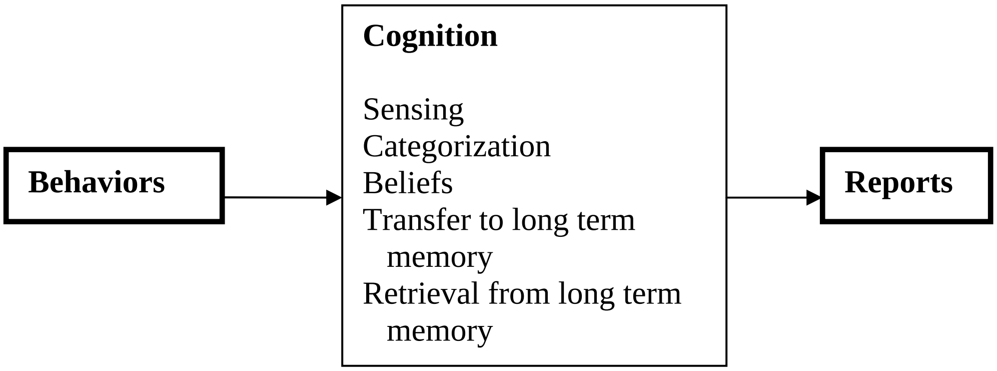


Fig. S1. This chart plots the behavioral and cognitive dimensions to relationships. The box on the left represents various relational behaviors, the box in the middle relational cognitive processes, and the third box observed reports regarding relationships. The first and third boxes are bolded to reflect that these are points where data collection is possible.

DATA COLLECTION AND ANALYSIS

Subject pool

The subjects from this study consisted of students and staff at a major university during the months between September 2004 and June 2005. For this paper's analyses, we used a subset of the data collected for the Reality Mining study (*1*), incorporating the 94 subjects that had completed the survey conducted in January 2005[1]. Of these 94 subjects, 68 were colleagues working in the same building on campus (90% graduate students, 10% staff) while the remaining 26 subjects were incoming students at the university's business school. The subjects volunteered to become part of the experiment in exchange for the use of a high-end smartphone for the duration of the study.

Observational Data from the Mobile Phone

Mobile Phone Logging Software

The data for this paper came from Nokia 6600 phones programmed to automatically run the ContextLog application as a background process at all times (*2*)[2]. This application continuously logs passive behavior such as location (from cell tower ids) and other proximate subjects (from Bluetooth device discovery scans at five-minute intervals). The application also logs all of the phone's activity, including voice calls and text messages, active applications (such as the calendar or games), and the phone's charging status.

Data were collected from the phones using two methods. Approximately 30 of the subjects were provided data plans (GPRS) on their mobile phone. For this group we had the phones directly connect to our data server during the night and upload the new data logged during previous the day. For the remaining subjects in the study, data was stored on each phone's internal 32MB memory card. The cards can store approximately four months of behavioral data before they need to be collected by the researchers.

An anonymized version of this dataset is currently available for download:

http://reality.media.mit.edu/download.php

Observational Accuracy

While the custom logging application on the phone crashes occasionally (approximately once every week), due to automatic restarts these crashes do not result in significant data loss. However, while the logging application can be assumed to be running anytime the phone is on, the dataset generated is certainly not without noise. Because we know when each subject began the study, as well as the dates that have been logged, we know exactly when we are missing data. These missing data are due to two main errors: data corruption

[1] There were 106 subjects in the Realty Mining experiment, however 12 of these subjects did not take the survey conducted in January of 2005 and were thus excluded from the analysis in this paper.

[2] ContextLog is freely available software under the GNU General Public License (GPL). It can be downloaded from the University of Helsinki: http://www.cs.helsinki.fi/group/context/.

and powered-off devices. On average we have logs accounting for approximately 85.3% of the time that the phones have been deployed.

Inferring Location from Cellular Towers

A mobile phone has reception when it is within the range of a fixed cellular tower. While 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. Using the tower IDs and respective transition timings (timestamps when the phone is handed off between cellular towers), it has been shown that a phone's position can be localized to within 100-200m in urban areas (*3*).

Inferring Proximity from Repeated Bluetooth Scans

Bluetooth is becoming an increasingly popular short-range RF protocol used as a cable replacement to wirelessly connect proximate mobile electronic devices (such as phones and laptops) together. A key feature of a Bluetooth device is the ability to scan for other nearby Bluetooth devices. When a Bluetooth device conducts a discovery scan, other Bluetooth devices within a range of 5-10m respond with their user defined name (e.g.: Mark's 6680), the device type (Nokia Mobile Phone), and a unique 12-digit MAC hardware address (e.g.: 0012d186e409). A device's MAC address is fixed and can be used to differentiate one subject's phone from another, irrespective of the device name and type. When a subject's MAC address is discovered by a periodic Bluetooth scan performed by another subject, it is indicative of the fact that the two subjects' phones are within 5-10 meters of each other.

Human Subjects Approval

Continuously recording a subject's daily behavior over an extended period of time has significant privacy implications. For example, under some circumstances, these data might be as sensitive as medical information. For IRB approval, we provided each subject with detailed information about the type of information that would be captured and instructions how to temporarily disable the logging application. We also had strict protocols limiting access to the data. All personal data such as phone numbers were one-way hashed (MD5), generating unique ids used in the analysis. While we found that subjects were initially concerned about the privacy implications, less than 5% of the subjects ever disabled the logging software throughout the 9-month study.

Constructing the Dyadic Observational Variables

Conducting periodic Bluetooth scans at 5 minute intervals has generated approximately 4 million proximity events in the dataset. For each proximity event we have logged the two proximate MAC addresses, the current associated cellular tower for each of the phones, and the time and date of the event. The dyadic variables below come from these proximity events, as well as phone communication logs and the report survey data.

Because all of the phones are scanning every five minutes, if two subjects were together for 100 minutes there would be a total of 40 recorded proximity events. We therefore approximate each proximity event to be representative of a 2.5 minute time interval. To estimate the amount of proximity at a particular location such as ‘Work’, we multiply this time interval by the number of proximity events that involved the cellular towers associated with that location. A ‘Proximity at Work’ value of ‘15.7’ for a particular pair of individuals would thus mean that during the times when their phones have logged the cellular towers associated with campus, the individuals have had an average estimated daily proximity of 15.7 minutes.

The ‘Home’ label is associated with the tower where the subject is located at 5am during at least 70% of the nights in the study.[3] We were unable to resolve the tower ID associated with the homes of 23% of the subjects. This was due primarily to either these subjects moving houses during the study or living on campus.

Counting the number of unique tower IDs in these proximity events provides an estimate of the number of locations the dyad had been together while they were in the study. This variable is particular sensitive to behavior involving ground travel. If two subjects drive in the same car for an hour together, the number of unique locations can increase by more than 20.

Data logged for each voice conversation on the mobile phone during the study included the time the conversation started, the duration and direction (incoming or outgoing) of the call, and the other phone number involved. If this other number was associated with another subject in the study, we incorporate the duration of the call into a statistic that estimates the average amount of daily phone communication between each pair of subjects.

Self report Survey Data

At the midterm of the 9-month study we conducted an online survey, which was completed by 94 of the 106 Reality Mining subjects. This survey included dyadic questions regarding the average reported proximity and friendship with the other subjects, as well as questions concerning the individual’s general satisfaction with his or her work group. The questions used for this analysis are written below.

Dyadic Questions

- Estimate Your Average Proximity with Each Person
  - 5 - at least 4-8 hours per day... 4 -at least 2-4 hours per day... 3 - at least 2 hrs - 30 minutes per day .... 2 - at least 10 - 30 minutes per day... 1 - at least 5 minutes .. 0 – 0-5 minutes (default)
- Is this Person a Part of Your Close Circle of Friends?
  - Yes / No (default)

Individual Questions

- I am satisfied with the quality of our group meetings
  - 1 – Strongly Agree 2, 3, 4, 5,6, 7 – Strongly Disagree

[3] If a subject logs two adjacent towers over 70% of the nights the subject is in the study, we assume home lies in between these two towers and label both as ‘home’.

Table S1: Correlations between All Dyadic Variables*

No.	Variable Name	Mean	S.D.	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
1	Reported Proximity at Work	7.96	40.69								
2	Observed Proximity at Work	7.15	36.64	0.42							
3	Phone Communication	0.006	0.103	0.24	0.17						
4	Number of Unique Locations	3.28	6.27	0.42	0.63	0.34					
5	Proximity on Saturday Nights**	0. 10	0.60	0.19	0.26	0.25	0.50				
6	Proximity Outside Work	0.72	3.61	0.50	0.57	0.36	0.52	0.45			
7	Proximity with no Reception	0.90	4.66	0.30	0.37	0.11	0.28	0.18	0.28		
8	Proximity at Home	0.18	1.28	0.25	0.35	0.29	0.36	0.29	0.33	0.26	
9	Reported Friendship	0.01	0.12	0.37	0.19	0.39	0.29	0.25	0.27	0.14	0.16

** p<.005 for all values, significance calculated using the nonparametric quadratic assignment procedure (QAP).*

*** Saturday night proximity is defined as proximity between 11pm on Saturday and 3am Sunday morning. Saturday night proximity is measured in minutes/week.*

Proximity and communication variables measured in minutes/day unless otherwise noted.

Table S1. Means, Standard Deviations, and Correlations for all Dyadic Variables. Proximity and phone communication is measured in minutes per day with the exception of Saturday nights, which is measure in minutes per week. Unique locations are approximated as the number of unique mobile phone towers.

Table S1 shows the relationships among the different dyadic variables. The dyadic variables associated with amounts of proximity have been normalized to minutes per day and are all weakly correlated. More generally, all communication variables are positively correlated, confirming Haythornthwaite's (*4*) observation that communication via different media tends to be positively correlated (in this case, phone communication and various contextualized proximities). As we will discuss in Analysis 1, one striking result that can be seen in this table is the surprisingly small correlation between the reported proximity and observed proximity of only R=.42. In Analysis 2, we will go into more depth on the relationship between reported friendship and several of the observational variables such as phone communication (R=.36) and proximity on Saturday night (R=.25).

Dyadic Data Analysis: MRQAP

The interdependencies in observations inherent in whole network data present a challenge because these data cannot satisfy the assumptions necessary for traditional statistical regression techniques. For much of the analysis of dyadic variables in this paper, we will be using the nonparametric multiple regression quadratic assignment procedure (MRQAP), a standard technique to analyze social network data (*5,6*). 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.

Analysis 1: Discrepancies between Self-Report and Actual Behavior

In our first analysis, we highlight the major discrepancies between the self report proximity responses with the Bluetooth and location data collected from the mobile phones. We show in this section that these discrepancies are influenced by reported relationships, recent behaviors and the salience of particular proximity.

The majority (69%) of the observed average proximity of over 5 minutes/day was not reported. However when proximity was reported, it was typically overestimated as evidenced by the darker values in the reported proximity socio-matrix in Figure S1. The average reported amount of non-zero proximity is 86.5 minutes, while the average amount of non-zero observed proximity is 32.8 minutes.

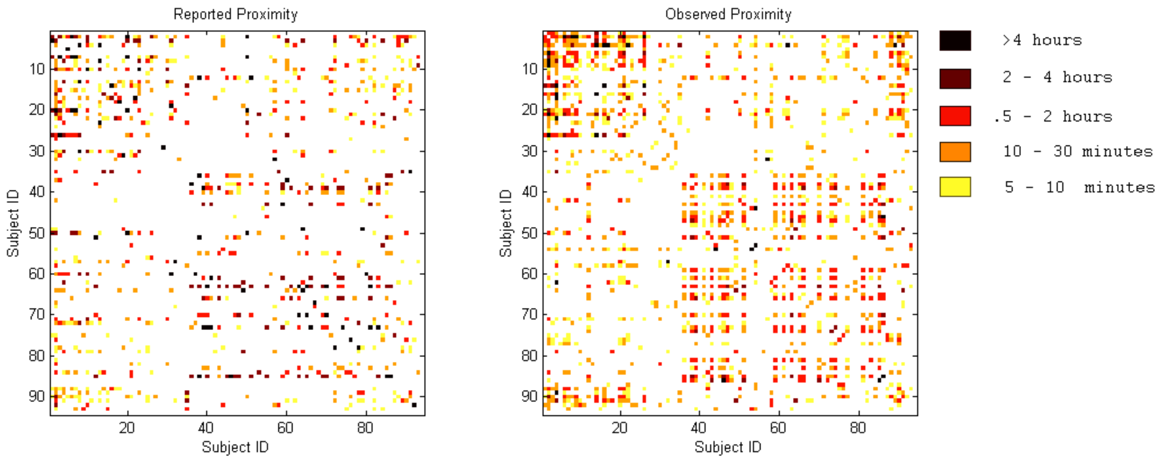


Fig. S2. Reported and observed proximity binned into 5 values for the 94 subjects. The empty (white) space indicates an average proximity of less than 5 minutes per day. While a large fraction of dyads fail to report the observed proximity (69%), those dyads that do report proximity tend to overestimate it (by a factor of 2.5 on average).

Salience. We hypothesize that prominent, or salient, events are more likely to be recalled. We consider salient proximity as proximity that occurs in locations and during times that are traditionally not associated with work, such as proximity at home or on Saturday night. Using MRQAP, we show in Table S2 that average proximity outside of work, at home, and on Saturday night all independently and powerfully predict reported proximity at work, controlling for observed proximity at work. Figure S3 contrasts the observed and reported behavior with the travel and socializing behavior of the same subject.

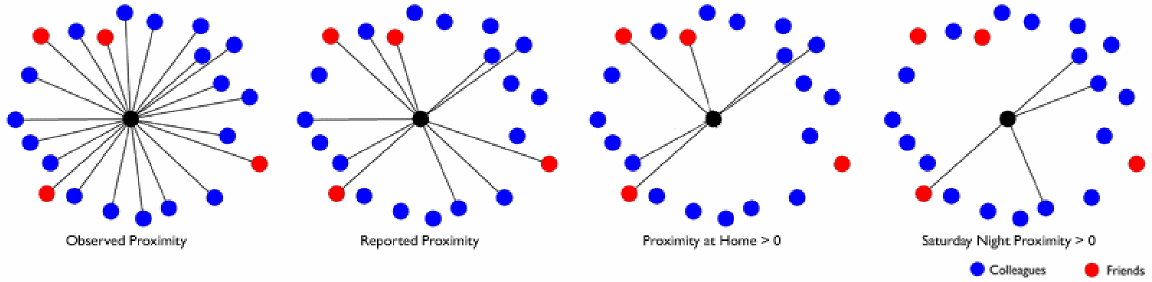


Fig. S3. Characteristic egocentric networks for an individual demonstrating the effects of saliency on reported behavior. The two networks on the left are identical to those in Figure S3, while the remaining two networks represent the subjects with whom the individual has had salient behavior: proximity at home and proximity on Saturday night (where Saturday night is defined as the times between 11pm on Saturday and 3am on Sunday). 6 of the 11 subjects reported as proximate were those whom the individual had been proximate to at home. 3 of the 4 subjects whom the individual was proximate to on Saturday nights were also reported by the individual.

Table S2: MRQAP Regression on Reported Proximity to Quantify Saliency Effects

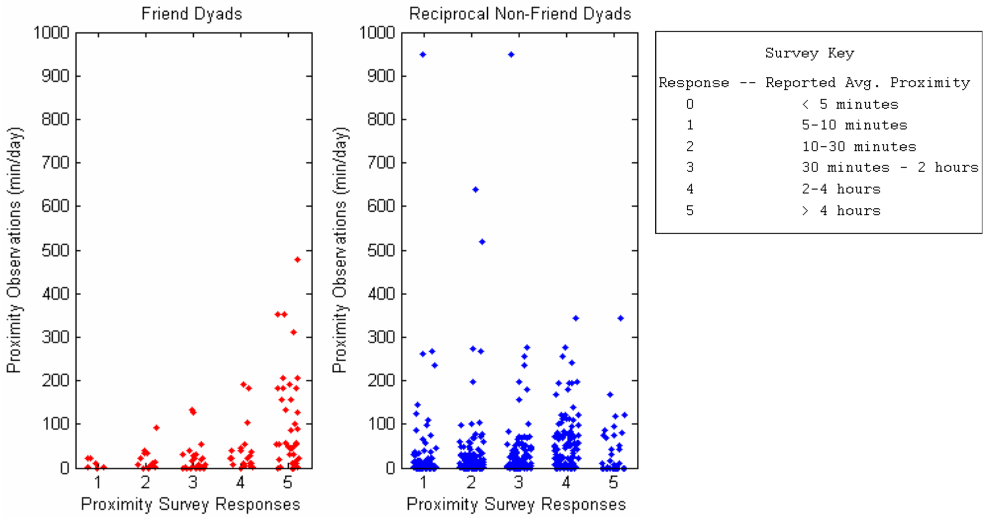
Variable Name	***Std. Coeff (b)***	***Sig. (p)***
Proximity at Work	.0932	0.000
Proximity at Home	.1243	0.000
Proximity Outside Work	.2396	0.000
Proximity on Saturday Nights	.0487	0.000
Adjusted R^2	.145 (p<.0001)	
# of Observations	8742	

Table S2. The Effects of Salient Proximity Events on Reported Average Proximity at Work. This table shows that while the "Proximity at Work" observational variable is correlated with the reported proximity at work, "Proximity Outside Work" is the reported data's single most powerful predictor.

As part of this analysis, we were also interested in quantifying how cognitive relationships affect the discrepancies between observed and reported behavior. Proximity to friends, for example, is likely more salient than proximity to people you may not even know. Figure S4 presents scatter plots of responses and observed proximity values for (a) friend dyads (both reciprocated and non-reciprocated); and (b) reciprocated non-friend dyads[4]. It is striking that while there seems to be very little correlation between individuals who work together but do not consider each other friends (R^2 =.024, p<.001), there is clearly a relationship between self report proximity and the observations for friends (R^2 =.171, p<.001).[5]

[4] It can be assumed throughout this paper that all dyads are colleagues. There were two groups of colleagues in this study; one group was made up of the 26 first-year business school students and the other group encompassed the 68 students and staff working together in the same building on campus.

[5] In this analysis we are only using dyads who have reported some proximity. The rationale for doing this is that it is, in certain ways, trivial to report non-interaction with people that you have never run across. In a dataset made up of many dyads with 0 interaction, achieving high accuracy is trivial. The non-friend dyads have far more 0's than friends, driving up the "accuracy" of their self reports. A more rigorous test thus



Response --	Reported Avg. Proximity
0	< 5 minutes
1	5-10 minutes
2	10-30 minutes
3	30 minutes - 2 hours
4	2-4 hours
5	> 4 hours

Fig. S4. Scatter plots highlighting the correlation between the self report and observation proximity behavior for (undirected) friendship and reciprocal non-friend dyads. Self-report proximity responses are listed on the x-axis with the observed proximity on the y-axis. A random number between +/- .5 is added to the survey responses for visual clarity.

Recency. Recent behavior is a powerful predictor of reported behavior. Figure S5 provides an illustrative example comparing the egocentric networks of a subject's observed proximity, reported proximity, and recent proximity. We define recent proximity as proximity that occurred during the seven days preceding the survey. It is clear from the figure that recent proximity has a strong influence over reported proximity for this particular subject. In Table S3, the MRQAP regression shows that both observed average proximity (b=.303, p<.001) and recent proximity (b=.225, p<.001) are significant predictors of reported proximity (R^2=.228, p<.001) for the 32 colleagues who were at work during the week leading up to the survey (N=992).

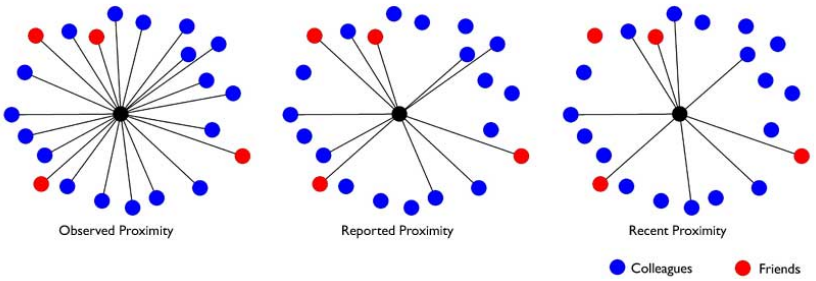


Fig. S5. Characteristic egocentric networks for an individual subject. The 22 surrounding nodes represent other subjects whom have been observed to be proximate at work to the individual for more than 5 minutes per day. Four of these subjects were labeled as a friend, while the remaining 18 are colleagues. The

looks at only dyads where there were was non zero observed interactions. (Friends reported 0's 35% accurately, and nonfriends reported 0's 99.5% accurately.)

individual correctly reported all 4 friends as proximate while only 7 of the 18 colleagues were reported. The network on the right shows that 9 of these 22 subjects were proximate to the individual for more than 5 minutes per day during the seven days leading up to taking the survey. 7 of the 11 reported subjects were recently proximate to the subject prior to the survey.

Table S3: MRQAP Regression on Reported Proximity to Quantify Recency Effects

Variable Name	***Std. Coeff (b)***	***Sig. (p)***
Proximity at Work	.303	0.000
Recent Proximity at Work	.225	0.000
Adjusted R^2	.227 (p<.0001)	
# of Observations	992	

Table S3. The Effects of Recent Proximity Events on Reported Average Proximity at Work. While the average proximity at work observational variable is strongly correlated with the self-report data, incorporating recent proximity provides substantial improvement to the model. Recent proximity is defined as the proximity events occurring during the week leading up to taking the survey.

Analysis 2: What Does Friendship Look Like?

We hypothesize that certain behavioral regularities such as repeated proximity and communication on Saturday nights can be indicative of friendship. Using self-report data on each subject's friendships, we are able to examine the behavioral correlates of reciprocal friends (dyads that have both subjects identify the other as a friend), non-reciprocal friends (dyads where only one subject identifies the other as a friend), and reciprocal non-friends (dyads who work together, but neither consider the other a friend).

Table S4: MRQAP Regression on Friendship

Variable Name	***Std. Coeff (b)***	***Sig. (p)***
Proximity at Work	-0.0194	0.039
Proximity Outside Work	0.08317	0.005
Proximity with No Reception	0.05230	0.005
Proximity on Saturday Nights	0.07008	0.000
Number of Unique Locations	0.09964	0.001
Phone Communication	0.31822	0.000
Same Program	0.05115	0.000
Same Gender	0.00662	0.321
Adjusted R^2	.194 (p<.0001)	
# of Observations	8742	

Table S4. Self-Report Friendship Regression using the seven dyadic variables as well as shared program and gender. This table shows the correlations between friendship and the seven dyadic variables. While

phone communication is the best predictor of friendship, all of the observational variables have a significant correlation with friendship.

Table S4 shows that the number of unique locations, communication, proximity outside work and on Saturday night all add significant explanatory power to the model. It is also clear that neither program nor (especially) gender is a strong predictor of friendship independent of the behavior variables.

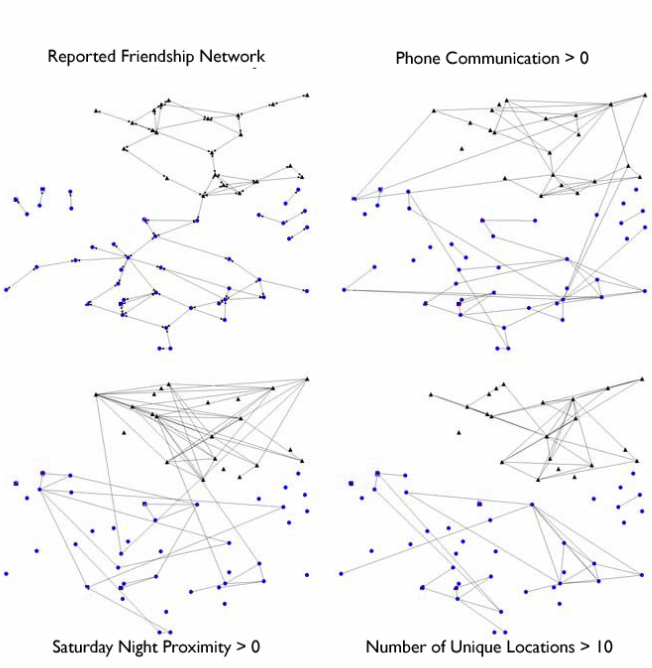

Reported Friendship Network
Phone Communication > 0
Saturday Night Proximity > 0
Number of Unique Locations > 10

Fig. S6. Networks representing reported friendship, phone communication, proximity on Saturday night and travel. Nodes color reflects the two groups of colleagues – the red nodes are first year business school students and the blue nodes work together in the same building on campus. For visual clarity, subjects are only included if they have at least one connection in the reported friendship network.

Friendship Inference via Factor Analysis

In this section we will construct a model to identify friendships based on the observational data. For an accurate comparison, we will only include colleague dyads, with no missing information.[6] A dyad qualifies as a "colleague dyad" only if the two members of the dyad work together (either as business school students or in the same building on campus).[7] There are three types of dyads that occur: reciprocal friends, non-reciprocal friends, and reciprocal non-friends. Reciprocal friends occur when both subjects mark the other as a friend (N=20). Non-reciprocal friends occur when only one

[6] The majority of the missing data involved the 'proximity at home' variable because we were unable to resolve the tower IDs associated with 23% of the subjects' homes. This resulted in the exclusion of 8 reciprocal friendships, 30 non-reciprocal friendships, and 851 reciprocal non-friend relationships. Interpolation of missing data did not qualitatively change the results—results available upon request.

[7] We also ran this analysis including non-colleague dyads (N=2555), and produced substantively identical results. We present only colleague dyads in this analysis because distinguishing friends from non-friend colleagues is a tougher test than distinguishing friends from non-friend non-colleagues. That is, the large majority of noncolleagues are not friends and almost never cross paths; thus inferring relationships in this group is fairly trivial. Results that include non-colleagues available upon request.

of the two subjects marks the other as a friend (N=24). Reciprocal non-friends occur when neither subject marks the other as a friend are colleagues (N=935).

A factor analysis was conducted for the seven dyadic variables after a log transformation. The analysis demonstrated there are two common factors (p<.005), explaining 51% of the variance in these seven variables. Communication seems to break down into two factors: in-role communication/proximity and extra-role communication/proximity, as shown in Table S5. In-role communication is simply the amount of work-associated communication that takes place, particularly proximity at work and number of unique locations[8]. Extra-role communication is driven by Saturday night proximity, proximity at home, and quantity of phone calls.

Table S5: Loadings from a Factor Analysis for Friendship

Variable Name	***Specific Variance***	***Factor1: 'In-Role' Communication***	***Factor 2: 'Extra-Role' Communication***
Proximity at Work	0.2064	0.9194	-0.0595
Proximity at Home	0.7694	0.0716	0.4401
Proximity with no Reception	0.4749	0.6697	0.0990
Proximity Outside Work	0.6288	0.3535	0.3491
Number of Unique Locations	0.1171	0.9927	-0.1162
Proximity on Saturday Nights	0.6689	-0.1584	0.6387
Phone Communication	0.6476	-0.1418	0.6523

Table S5. Factor Analysis Loadings. For relationship inference, 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 while extra-role communication corresponds to more personal behavior such as proximity on Saturday nights or at home. Either factor can be used to infer 95% of the reciprocal friendships.

Both in-role and extra-role communication are strongly predictive of friendship. After a promax rotation on the factor scores, a threshold of 2.3 on extra-role communication correctly classifies 19/20 (95%) reciprocal friends and 901/935 (96%) reciprocal non-friends. Using a threshold of .57 on in-role communication, we correctly classify 19/20 (95%) reciprocal friends and 868/935 (93%) reciprocal non-friends. While there were no thresholds that could identify the non-reciprocal friend dyads with these levels of accuracy, we show below that non-reciprocal dyads do form a group that behaviorally falls between reciprocated friend and non-friend dyads—perhaps reflecting that friendship is a continuous variable rather than bivariate. The behavioral data thus may be recapturing this underlying continuous variable.

Because the three distributions from the 'extra-role' communication factor are approximately normally distributed, we were able to perform a pairwise one-way analysis of variance (ANOVA) using the Bonferroni adjustment to confirm that the behavior from

[8] "Number of Unique Locations" is strongly related to work, and thus in-role communication, because there are 18 cellular towers (managed by three mobile phone service providers) in the immediate vicinity of the work location

reciprocal friends and non-friends do indeed come from different distributions (F(1,3)=192.49, p<.0001). We also found that non-reciprocal friends were significantly different from both the dyads labeled as reciprocal friends (F(1,2)=9.23, p<.005) and the dyads labeled as reciprocal non-friends (F(2,3) =77.80, p<.0001).[9]

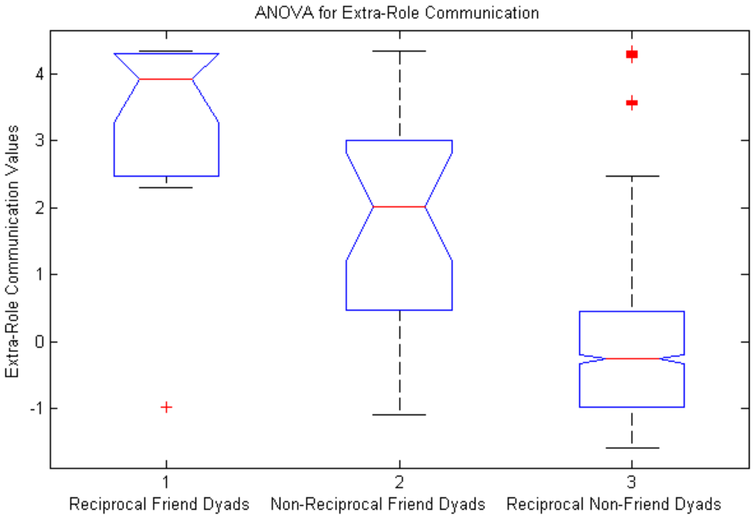


Fig. S7. Box-whisker plot generated on Factor 2, 'Extra-Role' Communication, for the three relationship types (F>9, p<.005). Each box represents one of the dyad distributions. The height of the box corresponds to the lower and upper quartile values of the distribution and the horizontal line corresponds to the distribution's median. The notches represent the length of the confidence interval for the median. Because the notches do not overlap, the true medians do differ with >95% confidence. The 'whiskers' extend from the box to values that are within 1.5 times the quartile range while outlier dyads are plotted as distinct points.

Analysis 3: Satisfaction vs. Proximity & Communication

In these analyses we examine whether social integration is related to satisfaction. We anticipate that individuals will be more satisfied the more friends they have (a standard measure of integration), as well as with the average amount of time they get to spend with those friends and the average phone communication with those friends. We also incorporate in these analyses a dummy variable set to zero if the subject has not reported any friends and one otherwise in order to capture any nonlinearity in the relationship between the jump from 0 friends to 1 friend.

The analysis below used the undirected self-reported friendship network to define friendship. While Table S6a shows that there is only a modest relationship between number of friends and satisfaction, Table S6b shows that incorporating the average amount of proximity to friends and average amount of phone communication with friends substantially improves the model fit (R^2=.161, p<.001). As the average amount of proximity to each friend increases, the more satisfied subjects are about their work group,

[9] Conducting the pairwise ANOVA using 25 randomly sampled reciprocal non-friend dyads to maintain a similar sample size generated F-statistics that were not qualitatively different. Results available upon request.

and (to our surprise), as the average amount of phone communication increases, satisfaction decreases. We include in the analyses summarized in Table S6c total proximity to and phone communication with all subjects as controls, which demonstrates that the findings in 6b do not reflect simply higher levels of satisfaction among those who just talk to everyone more.

Table S6a: Work Group Satisfaction Regression using Number of Friends (N=94)

Variable Name	***Corr. Coff. (b)***	***Stand. Error (SE)***	***t-stat***	***Sig. (p)***
Friendship Dummy Variable	-.606	.381	-1.59	.117
Number of Friends	.158	.064	2.46	.016
Adjusted R^2	.04 (p>.05)			

Table S6b: Work Group Satisfaction Regression using Number of Friends, Proximity to Friends, and Communication with Friends (N=94)

Variable Name	***Corr. Coff. (b)***	***Stand. Error (SE)***	***t-stat***	***Sig. (p)***
Friendship Dummy Variable	-.370	.175	-2.11	.038
Number of Friends	.377	.166	2.27	.026
Average Proximity to Friends	.719	.176	4.05	.000
Phone Communication with Friends	-.497	.171	-2.91	.005
Adjusted R^2	.161 (p<.01)			

Table S6c: Work Group Satisfaction Regression including Average Proximity and Average Communication (N=94)

Variable Name	***Corr. Coff. (b)***	***Stand. Error (SE)***	***t-stat***	***Sig. (p)***
Friendship Dummy Variable	-.402	.189	-2.13	.037
Number of Friends	.444	.180	2.46	.016
Average Proximity to Friends	.755	.194	3.90	.000
Average Phone Communication with Friends	-.530	.176	-3.01	.004
Average Proximity	-.123	.149	-.88	.378
Average Phone Communication	-.063	.`.37	-.46	.647
Adjusted R^2	.150 (p<.01)			

Table S6a/b/c. Satisfaction Regressed on Self-Report Friendships. Table S6a shows the weak, but positive relationship between job satisfaction and number of friends. Table S6b shows the model improves significant after adding average proximity to friends and phone communication with friends. Table S6c shows that adding average total proximity and phone communication does not help the model. A dummy variable has been incorporated into the regression to take into account the nonlinearity associated with subjects who did not list any other friends.

More interesting, from the perspective of this paper, is the comparison between regressions using self-report and inferred friendship networks. Substituting the self-report friendship matrix with an inferred friendship matrix using the extra-role communication factor, we produce substantively similar results, with somewhat improved overall fit of the model. In short, it is possible to produce a reasonable model of satisfaction based exclusively on behavioral observations of communication and proximity. In fact, the best fit of all is produced by using the continuous version of extra-role communication, suggesting that the (continuous) inferred friendship may be recapturing information about the underlying continuous construct.

Table S7a: Work Group Satisfaction Regression using *Discrete*, *Inferred* Friendship Network (N=94)

Variable Name	***Corr. Coff. (b)***	***Stand. Error (SE)***	***t-stat***	***Sig. (p)***
Friendship Dummy Variable	-.392	.170	-2.30	.024
Number (inferred) of Friends	.483	.164	2.94	.004
Average Proximity to (inferred) Friends	.698	.188	3.71	.000
Phone Communication with (inferred) friends	-.571	.182	-3.13	.003
Adjusted R^2	.180 (p<.001)			

Table S7b: Work Group Satisfaction Regression using *Weighted*, *Inferred* Friendship Network (N=94)

Variable Name	***Corr. Coff. (b)***	***Stand. Error (SE)***	***t-stat***	***Sig. (p)***
Friendship Dummy Variable	-.363	.167	-2.17	.034
Number (inferred) of Friends	.420	.162	2.60	.011
Average Proximity to (inferred) Friends	.799	.217	3.69	.000
Phone Communication with (inferred) friends	-.694	.182	-3.81	.000
Adjusted R^2	.200 (p<.001)			

Table S7a/b. Satisfaction Regressed on Inferred Friendships. A similar regression as Tables S6, however the (discrete) self-report friendship data has been replaced by the discrete (S7a) and weighted (S7b) *inferred* friendship network. The inferred networks are derived from the dyadic extra-role communication factor scores.

References

1. N. Eagle, A. Pentland. 2006. “Reality Mining: Sensing Complex Social Systems”, *Personal and Ubiquitous Computing*, 10 (4): 255-268.

2. M. Raento, A. Oulasvirta, R. Petit, H. Toivonen. 2005. “ContextPhone – A prototyping plat-form for context-aware mobile applications”, *IEEE Pervasive Computing*, 4 (2): 51-59.

3. A. LaMarca, Y. Chawathe, S. Consolvo, J. Hightower, I. Smith, J. Scott, T. Sohn, J. Howard, J. Hughes, F. Potter, J. Tabert, P. Powledge, G. Borriello, B. Schilit. 2005. “Place Lab: Device Positioning Using Radio Beacons in the Wild”. In Proceedings of Pervasive 2005, Munich, Germany.

4. C. Haythornthwaite. 2005. “Social networks and Internet connectivity effects”, *Information, Communication and Society*, 8 (2): 125-147.

5. F. B. Baker, L. J. Hubert. 1981. “The analysis of social interaction data”, *Sociol. Methods Res.* 9: 339–361.

6. D. Krackhardt. 1988. “Predicting With Networks - Nonparametric Multiple-Regression Analysis Of Dyadic Data”, *Social Networks*, 10 (4): 359-381.