Supporting Information

Eagle et al. 10.1073/pnas.0900282106

SI Text

Supporting Online Material

Below we provide all relevant details on data collection and analysis, including an explanation of the subject pool, data-collection protocols, a description of variable construction, and a summary of data analyses.

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*. Of these 94 subjects, 68 were colleagues working in the same building on campus (90% graduate students, 10% staff) whereas 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)†. 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 by 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 were stored on each phone's internal 32 MB 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 at http://reality.media.mit.edu/download.php.

Data Description. Phone log.

(TIME) 20060720T211505 (DESCRIPTION) Voice call (DIRECTION) Outgoing (DURATION seconds) 23 (NUMBER) 6175559821

Bluetooth. (TIME) 20060721T111222 devices: 000e6d2a3564 [Amy's Phone] 000e6d2b06ea [Jon's PalmPilot]

Location. (TIME) 20060721T111222 (CELL AREA) 24127, (CELL TOWER) 111, (SERVICE PROVIDER) AT&T Wirel (USER DEFINED LOCATION NAME) My Office

Observational Accuracy. Although 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, although 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 and powered-off devices. On average we have logs accounting for ≈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. 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), it has been shown that a phone's position can be localized to within 100–200 m in urban areas (3). Fig. S1 shows a representation of subjects' locations, as inferred by the cellular tower transition data.

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 5–10 m 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 m of each other.

Human Subjects Approval. Continuously recording a subject's daily behavior over an extended period 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 explaining 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. Although we found that subjects were initially concerned about the privacy implications, <5% of the subjects ever disabled the logging software throughout the nine-month study.

Constructing the Dyadic Observational Variables. Conducting periodic Bluetooth scans at 5-min intervals generated ≈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

*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.

†ContextLog is freely available software under the GNU General Public License. It can be downloaded from the University of Helsinki at http://www.cs.helsinki.fi/group/context/.

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 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 min.

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 number of minutes of daily phone communication between each pair of subjects.

Self-Report Survey Data. At the midterm of the nine-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 (Within 10 feet) with Each Person.
- 5 - at least 4–8 h per day. 4 -at least 2–4 h per day. 3 - at least 2 h - 30 min per day . 2 - at least 10 - 30 min per day. 1 - at least 5 min. 0 – 0–5 min (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

Dyadic Data Analysis: Multiple Regression Quadratic Assignment Procedure (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 MRQAP, a standard technique by which 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. Fig. S2 is the sociomatrix corresponding to the reported proximity and the observed proximity. Although most proximity is not reported (69%), when a proximity event is reported, it is typically overestimated as evidenced by the darker values in the reported proximity sociomatrix in Fig. S2. The average reported amount of nonzero proximity is 86.5 min, whereas the average amount of nonzero observed proximity is 32.8 min.

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. By using MRQAP, we show 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. Fig. S3 contrasts the observed and reported behavior with the travel and socializing behavior of the same subject.

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. Fig. S4 presents scatter plots of responses and observed proximity values for (*i*) friend dyads (both reciprocated and nonreciprocated); and (*ii*) reciprocated nonfriend dyads[‡]. It is striking that although there seems to be little correlation between individuals who work together but do not consider each other friends ($r = .155, P < 0.001$), there is clearly a relationship between self-report proximity and the observations for friends ($r = .414, P < 0.001$).

Recency. Recent behavior is a powerful predictor of reported behavior. Fig. 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. We chose one week as the time window because, as Fig. 1 shows, this amount of time yields the largest correlation between the survey response and the subjects' prior behavior. In Table S1, the MRQAP regression shows that both observed average proximity ($b = .303, P < 0.001$) and recent proximity ($b = .225, P < 0.001$) are significant predictors of reported proximity ($r = 0.478, P < 0.001$) for the 32 colleagues who were at work during the week leading up to the survey (number of dyads = 992).

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. By 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), nonreciprocal friends (dyads where only one subject identifies the other as a friend), and reciprocal nonfriends (dyads who work together, but neither consider the other a friend).

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. 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).[¶] There are

[‡]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.

[¶]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 noninteraction 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 nonfriend dyads have far more 0's than friends, driving up the "accuracy" of their self-reports. A more rigorous test thus looks at only dyads where there were was nonzero observed interactions. (Friends reported 0's 35% accurately, and nonfriends reported 0's 99.5% accurately.) We also ran this analysis including noncolleague dyads ($N = 2,555$), and produced substantively identical results. We present only colleague dyads in this analysis because distinguishing friends from nonfriend colleagues is a tougher test than distinguishing friends from nonfriend noncolleagues. 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 noncolleagues are available on request.

three types of dyads that occur: reciprocal friends, nonreciprocal friends, and reciprocal nonfriends. Reciprocal friends occur when both subjects mark the other as a friend ($n = 22$). Nonreciprocal friends occur when only one of the two subjects marks the other as a friend ($n = 28$). Reciprocal nonfriends occur when neither subject marks the other as a friend are colleagues ($n = 935$).

Table S2 lists the variables we selected for the factor analysis based on the trends in the data highlighting that proximity is generally much higher for friends, but time and location are important predictors as well. We therefore divided proximity into variables corresponding to on campus/off campus and daytime/nights (separated at 8 a.m. and 8 p.m.), and conducted a factor analysis. The analysis demonstrated there are two common factors ($P < 0.005$), explaining 59% of the variance in these seven variables. Communication seems to break down into two factors: in-role communication/proximity and extra-role communication/proximity. In-role communication is simply the amount of work-associated communication that takes place, which is dominated by proximity at work during the weekdays. Extra-role communication is driven off-campus proximity and quantity of phone calls.

Both in-role and extra-role communication are strongly predictive of friendship. After a promax rotation on the factor scores, a threshold of 2.5 on extra-role communication correctly classifies 21/22 (95%) reciprocal friends and 901/935 (96%) reciprocal nonfriends. By using a threshold of 1.1 on in-role communication, we correctly classify 21/22 (95%) reciprocal friends and 844/935 (90%) reciprocal nonfriends. Although there were no thresholds that could identify the nonreciprocal friend dyads with these levels of accuracy, we show below that nonreciprocal dyads do form a group that behaviorally falls between reciprocated friend and nonfriend 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) by using the Bonferroni adjustment to confirm that the behavior from reciprocal friends and nonfriends do indeed come from different distributions [$F(1, 3) = 192.49, P < 0.0001$]. We also found that nonreciprocal friends were significantly different from both the dyads labeled as reciprocal friends [$F(1, 2) = 9.23$, $P < 0.005$] and the dyads labeled as reciprocal nonfriends [$F(2, 3) = 77.80$, $P < 0.0001$], as shown in Fig. S6[‖].

Behavior Correlations for Different Time Scales. 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) and a high of 0.82 (for proximity at work during the day). Fig. S7 shows the correlations of the aggregate proximity data by using different time windows ranging from seven to fifty days.

[‖]Conducting the pairwise ANOVA by using 25 randomly sampled reciprocal nonfriend dyads to maintain a similar sample size generated *F*-statistics that were not qualitatively different. Results are available on request.

1. Eagle N, Pentland A (2006) Reality mining: Sensing complex social systems. *Personal Ubiquitous Comput* 10:255–268.
2. Raento M, Oulasvirta A, Petit R, Toivonen H (2005) ContextPhone: A prototyping platform for context-aware mobile applications. *IEEE Pervasive Computing* 4:51–59.
3. A. LaMarca, et al. (2005) Place lab: Device positioning using radio beacons in the wild. In *Proceedings of the Third International Conference on Pervasive Computing* (Springer Munich), pp 116–133.
4. Haythornthwaite C (2005) Social networks and Internet connectivity effects. *Info Comm Soc* 8:125–147.
5. Baker FB, Hubert LJ (1981) The analysis of social interaction data. *Sociol Methods Res* 9:339–361.
6. Krackhardt D (1988) Predicting with networks—Nonparametric multiple-regression analysis of dyadic data. *Soc Networks* 10:359–381.

Fig. S1. Visualization of the Reality Mining data. Individual subjects can be positioned on the map based on their phones' reported cellular tower data. A yellow line connects the subjects during an ongoing call. A dynamic animation of this behavior is available at http://reality.media.mit.edu.

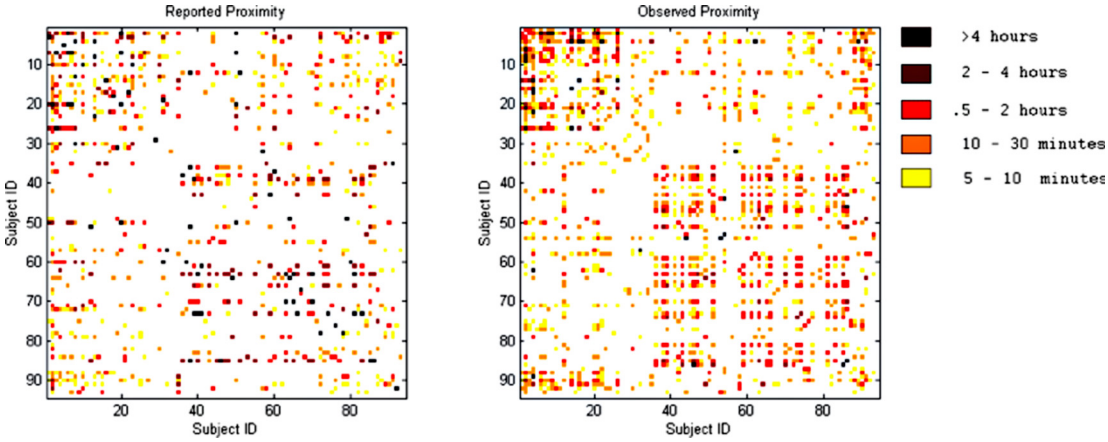

Fig. S2. Reported and observed proximity binned into five values for the 94 subjects. The empty (white) space indicates an average proximity of <5 min per day. Although a large fraction of dyads fail to report the observed proximity (69%), those dyads that do report proximity tend to overestimate it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

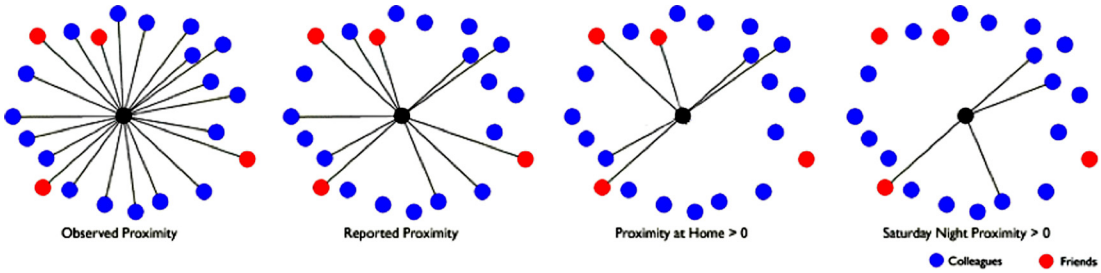

Fig. S3. Characteristic egocentric networks for an individual demonstrating the effects of saliency on reported behavior. The two networks on the right represent the subjects with whom the individual has had salient behavior: proximity at home and proximity on Saturday night (where Saturday night is defined as the times between 11 p.m. on Saturday and 3 a.m. on Sunday). Six of the eleven subjects reported as proximate were those to whom the individual had been proximate at home. Three of the four subjects to whom the individual was proximate on Saturday nights were also reported by the individual.

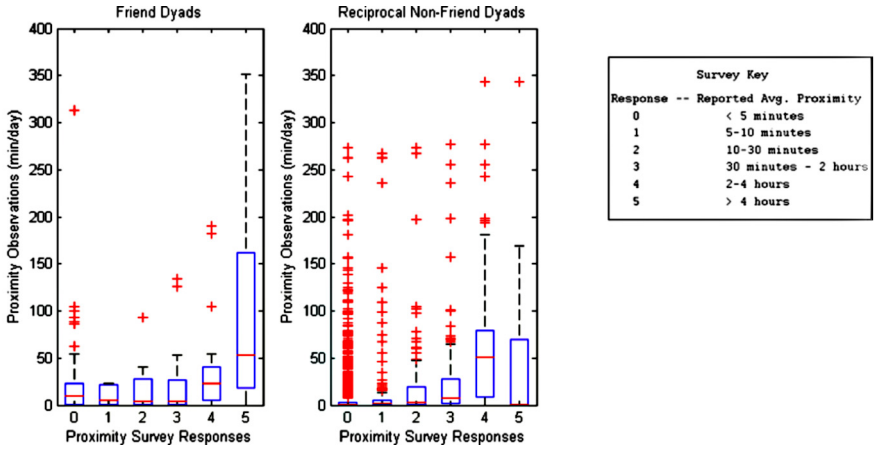

Fig. S4. Self-report vs. observational data. Box plots highlighting the relationship between self-report and observational proximity behavior for undirected friendship and reciprocal nonfriend dyads. Self-report proximity responses, on the *x* axis, are scored from 0 to 5 (see axis label). The *y* axis 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× the quartile range whereas outliers are plotted as distinct points. Three outlier dyads with an observed proximity >400 min/day have been excluded from the plot.

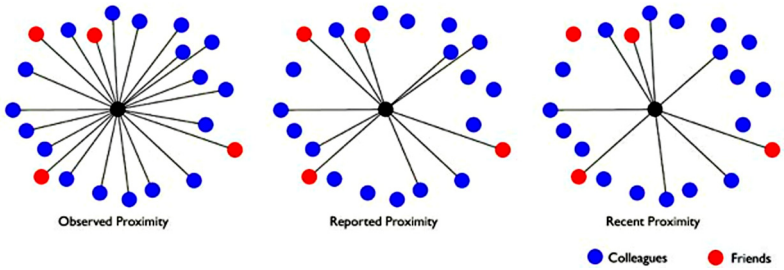

Fig. S5. Characteristic egocentric networks for an individual subject. The 22 surrounding nodes represent other subjects whom have been observed to be proximate at work to the individual for more than 5 min per day. Four of these subjects were labeled as a friend, and the remaining 18 are colleagues. The individual correctly reported all 4 friends as proximate whereas only seven of the 18 colleagues were reported. The network on the right shows that nine of these 22 subjects were proximate to the individual for more than 5 min per day during the seven days leading up to taking the survey. Seven of the eleven reported subjects were recently proximate to the subject before the survey.

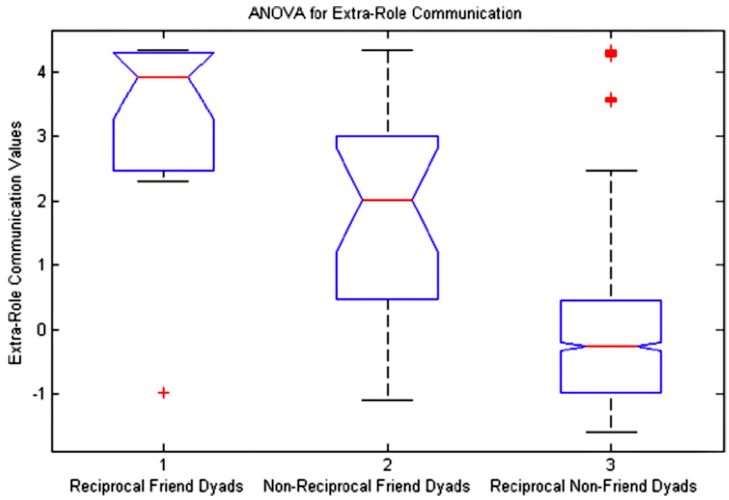

Fig. S6. Box-whisker plot generated on Factor 2, extra-role communication, for the three relationship types ($F > 9$, $P < 0.005$). Each box represents one of the dyad distributions. The height of the box corresponds to the lower and upper quartile values of the distribution and the horizontal line corresponds 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 × the quartile range whereas outlier dyads are plotted as distinct points.

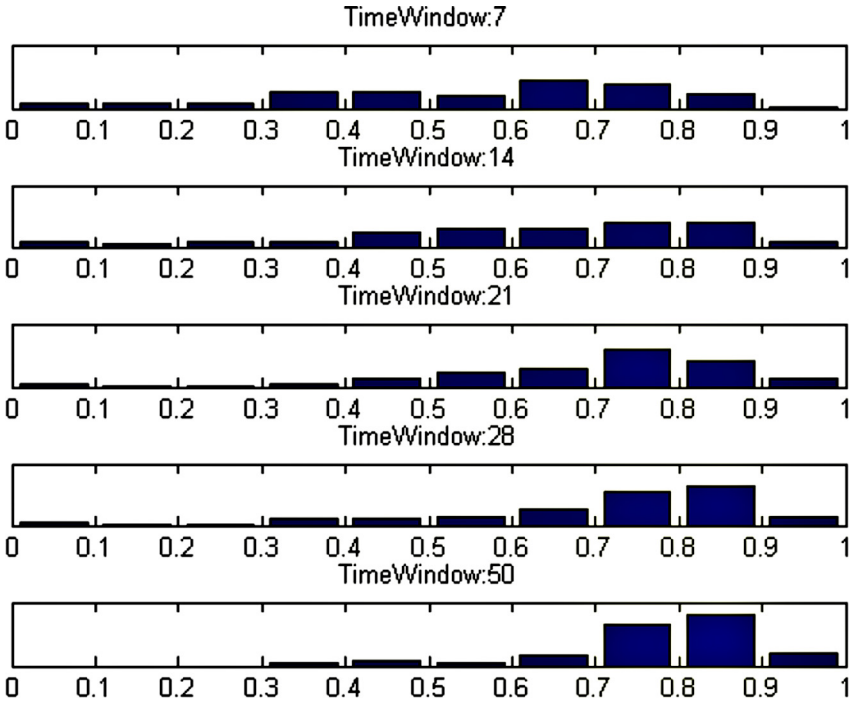

Fig. S7. Histogram of behavioral correlations for multiple time windows. The upper histogram shows the distribution of correlations between all of the proximity data and the proximity data from a randomly sampled seven-day window. As the time window expands, the sampled data becomes increasingly correlated to the nine-month dataset. However, due to the regularities inherent to human behavior, the lower histograms demonstrate the diminishing returns of an increased monitoring period; indeed, data collection over a two-week window will largely replicate the results we produced from nine months of observations.

Table S1. MRQAP regression on reported proximity to quantify recency effects

Variable name	Standard coefficient (*b*)	Signature (*P*)
Proximity at work	0.303	0.000
Recent proximity at work	0.225	0.000

The effects of recent proximity events on reported average proximity at work. Although 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. Adjusted R^2, 0.227 ($P < 0.0001$); number of observations, 992.

Table S2. Correlations between dyadic variables for the factor analysis*

No.	Variable name	Mean	SD	(1)	(2)	(3)	(4)	(5)	(6)	(7)
1	Work proximity, weekdays, 8 a.m.–8 p.m.	5.21	30.13							
2	Work proximity, weekdays, 8 p.m.–8 a.m.	0.28	1.60	0.48						
3	Work proximity, weekends	0.30	1.92	0.64	0.69					
4	Off-campus proximity, weekdays, 8 a.m.–8 p.m.	2.20	12.97	0.53	0.22	0.33				
5	Off-campus proximity, weekdays, 8 p.m.–8 a.m.	0.21	1.82	0.22	0.34	0.29	0.26			
6	Off-campus proximity, weekends	0.26	2.05	0.22	0.38	0.41	0.26	0.82		
7	Phone communication	2.44	38.21	0.10	0.09	0.12	0.11	0.53	0.52	
	Reported friendship	0.01	0.12	0.08	0.17	0.14	0.11	0.32	0.35	0.36

Means, standard deviations, and correlations for dyadic variables used in the friendship factor analysis. Proximity and phone communication is measured in minutes per day. Correlations with reported friendship are listed on the last row of the table. Proximity and communication variables measured in minutes/day unless otherwise noted.

*$P < 0.005$ for all values, significance calculated by using the nonparametric quadratic assignment procedure.