Mobile Phone Data for Inferring Social Network Structure

Nathan Eagle[†], Alex (Sandy) Pentland[Δ] and David Lazer[*]

[†] nathan@mit.edu, MIT Design Laboratory, Massachusetts Institute of Technology, Cambridge, MA
[Δ] sandy@media.mit.edu, MIT Media Laboratory, Massachusetts Institute of Technology, Cambridge, MA
[*] david_lazer@harvard.edu, John F. Kennedy School of Government, Harvard University, Cambridge, MA

Abstract 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.

1 Introduction

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.

2 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 [16]. 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.

3 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 [17]. 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 [18].

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.

4 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 [19]. Salience is the principle that prominent events are more likely to be recalled [20]. 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, 21]. 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).
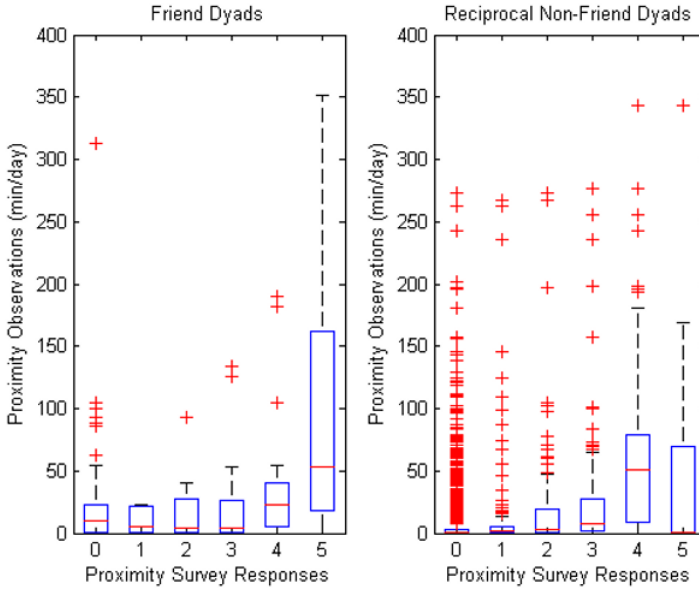
Figure 1. Self-Report vs. Observational Data. Boxplots highlighting the relationship between self-report and observational proximity behavior for undirected friendship and reciprocal non-friend dyads. Self-report proximity responses, on the x-axis, are scored from 0 to 5 (see legend). The y-axis shows 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.

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
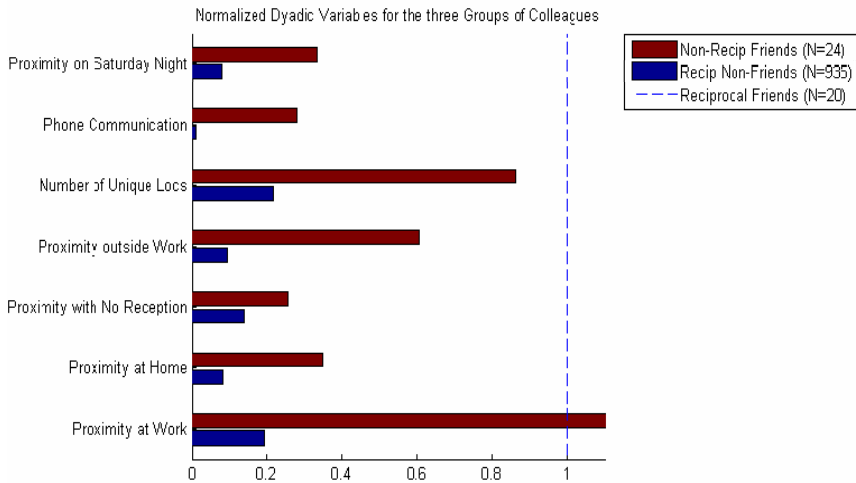

Figure 2. Normalized Dyadic Variables. The seven behavioral variables, normalized with respect to the reciprocal friendship data, are represented in the bar chart. The vertical dotted line at x=1 represents the values for reciprocal friend dyads. Reciprocal friends score higher than the other two groups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

Figure 3, 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 na-

ture, 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).
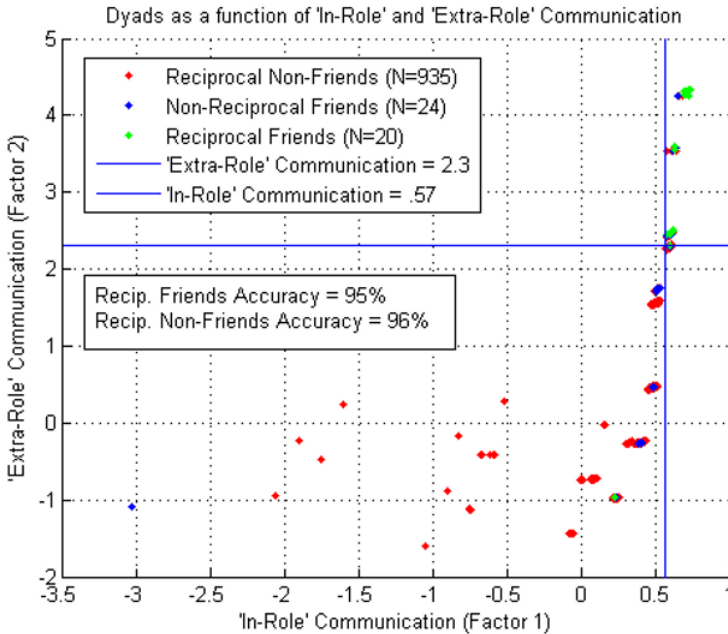

Figure 3. 'In-Role' Communication vs. 'Extra-Role' Communication. Each point represents a pair of colleagues' 'in-role' and 'extra-role' communication factor scores. 95% (19/20) of the reciprocal friendships have extra-role scores above 2.3, while 96% (901/935) of reciprocal non-friends have extra-role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

6 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 [22]. 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.
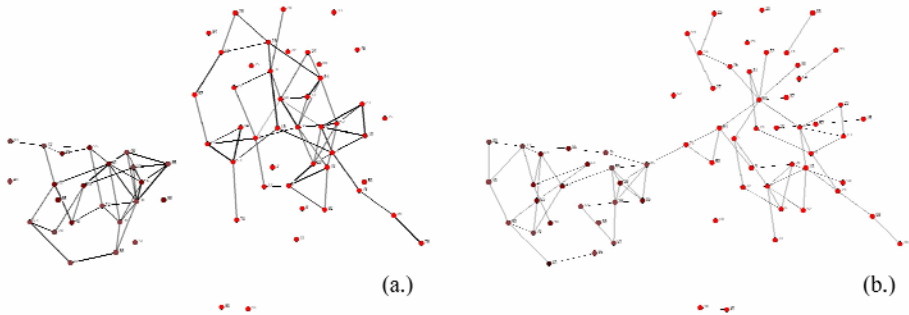

Figure 4a/b. Inferred, Weighted Friendship Network (a.) vs. Reported, Discrete Friendship Network (b.). The network on the left is the inferred friendship network with edge weights corresponding to the factor scores for factor 2, 'extra-role' communication. The network on the right is the reported friendship network. Node colors highlight the two groups of colleagues, first-year business school students (brown) and individuals working together in the same building (red).

7 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, provid-

ing 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.

References

1. W. Whyte. 1993.   Street Corner Society: The Social Structure of an Italian Slum.   4th edition. Chicago, IL: University of Chicago Press.
2. 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.
16. K.M. Carley and D. Krackhardt. 1996. "Cognitive inconsistencies and non-symmetric friendships." Social Networks 15: 377-398.
17. 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.
18. Full details on data collection and variable construction are available in the supporting online materials.
19. Waugh, N. C., & Norman, D. A. 1965. Primary memory. Psychological Review 72: 89-104.
20. 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.
21. Krackhardt, D. 1988. Predicting with Networks - Nonparametric Multiple-Regression Analysis of Dyadic Data. Social Networks 10 (4): 359-81.
22. Durkheim, Emile. 1951. Suicide: A Study in Sociology translated by George Simpson and John A. Spaulding. New York: The Free Press.