CARNEGIE MELLON UNIVERISTY

DATA ANALYSIS PROJECT

Detecting friendship with smartphone co-location data

Author:
Momin M. MALIK

Committee:
Dr. Anind K. DEY, Human-Computer Interaction Institute
Dr. Afsaneh DORYAB, Human-Computer Interaction Institute
Dr. Nynke M. D. NIEZINK, Department of Statistics & Data Science

A proposal for a thesis submitted in fulfillment of the requirements for the degree of Master's of Science in the

Secondary Master's in Machine Learning program
Machine Learning Department
School of Computer Science

May 15, 2018

1 INTRODUCTION

Friendship is a "voluntary, personal relationship typically providing intimacy and assistance", associated with characteristics of trust, loyalty, and self-disclosure [39]. It is one of the most important aspects of human existence, lending meaning to life, and providing for material, cognitive, and social-emotional needs in ways that lead to greater health and well-being [39].

Understanding the friendship relationship between people can be helpful for creating technology that serves people better. If an individual's friendships are known, these can be leveraged for applications supporting help-seeking behavior such as requests for recommendations or for favors, or for automatically establishing trust between users' devices. Friendships may also be leveraged for carrying out social interventions around diet and exercise [70] or for preventing disease transmission [68], such as in mobile applications that facilitate friends adopting and holding each other accountable to healthy behaviors. Conversely, it may be of interest to create applications to make recommendations about friendships in order to help bring people together, for example at conferences [18, 19] or for information dissemination in workplaces [62]. Longitudinal information about changes in friendships could help detect the onset of isolation and help design interventions to strengthen friendships.

But in order to incorporate information about one's friendship network in personal informatics and mobile applications, we need ways of detecting friendship.[1] One easy way to get 'ground truth' is to rely on ties from online social networking platforms; but such ties are not necessarily good proxies for the underlying construct of friendship [26, 76, 107]. Survey instruments have been the standard network data collection method in social network analysis for decades, but involve a high burden for users that make them impractical as a basis for mobile applications.

Previous social science work has established the strong link between individuals being physically close and being friends [39, 41, 60]. There is a two-way causal connection in that people who spend time together are more likely to become friends [41], and that people who are friends spend time together [39]. This suggests that there should be a robust signal of friendship in measurements of physical proximity that can then be leveraged for analysis, services, or interventions.

Our work is the first effort to detect friendships using feature extraction from smartphone location data. Previous work either linked location data and friendship without cross validation [38], looked at ties on location-based online social network services [27], or used mobile phone call and SMS logs [105, 106]. We believe that detecting friendship from mobile phone co-location data is a realistic approach for future mobile applications and interventions that seek to leverage friendship for other tasks.

This paper presents the results of a 3-month study of a cohort of 53 participants, with final analysis performed on 9 weeks of data from 48 participants. We combine mobile phone sensor data collection with established social network survey instruments, and use rich feature extraction from co-location data to detect both friendship and *changes* in friendship.

Our contributions are as follows:

- Presenting results of an original collection of high-quality friendship data alongside rich mobile phone sensor data;
- Carrying out an extraction of pairwise co-location features, and investigating their importance in predicting the evolution of friendship. To our knowledge, this is the first work that studies the impact of co-location features extracted from smartphone location coordinates on a friendship detection task.
- A new baseline, and a rigorous standard for model testing. We demonstrate that our model, using only smartphone co-location features, we build a classifier that is correlated (Matthews correlation coefficient) with the true values 30% better than random. We measure model performance using multiple cross-validation

[1]Detecting friendship is a link prediction problem [66], although we use the term 'detection' to emphasize that our predictions are of concurrent, not future, values. We avoid 'infer' as we do not have standard errors.

schema that control for different dependencies, giving a more rigorous appraisal of likely out-of-sample performance.

- We discuss implications for developing mobile applications and interventions, and for basic research. We discuss how our findings can support the design of Ubicomp applications such as detecting friendship and changes in friendship from co-location and for interventions based on such detected friendships. We also discuss how smartphone-based characterization of pairwise proximities can provide support for basic research looking at the co-evolution of friendship and co-location.

Below, first we review background work in social network analysis and in the ubiquitous and pervasive computing literature around existing approaches to extracting interactions and/or friendships from mobile sensor data. We describe our study design, and our feature extraction process. With extracted features, we build a predictive model and use feature importance as a step towards characterizing the aspects of co-location most useful for detecting friendship.

2 BACKGROUND AND RELATED WORK

2.1 Friendship ties

Friendship is an example of a *relational phenomenon*. Relational phenomena are two-unit or *dyadic* relations [12]. Instead of the n individuals of a dataset being the observations, they have as observations $\binom{n}{2}$ undirected (symmetric) relations (e.g., co-location), or $2 \times \binom{n}{2}$ directed (potentially asymmetric) relations (e.g., self-reported friendships) between individuals.

Research throughout the 20th century has provided examples of friendship and other ties having explanatory and predictive power, such as in explaining why girls ran away from a school for delinquent teenage girls and predicting future runaways [77], explaining the breakup of a monastery [91], of the split of a karate club into two separate clubs [108], or using structures of informal networks to predict the success or failure of institutional reorganization [58]. Insights from these approaches have proved robust as they have been successfully integrated into search engines, recommender systems, and the structure of social media platforms.

We represent a collection of friendship ties between n people as an $n \times n$ adjacency matrix, A, where

$$A_{ij} = \begin{cases} 1 \text{ if there is a tie } i \longrightarrow j \\ 0 \text{ otherwise.} \end{cases}$$

and where $A_{ii} = 0$ (no self-loops). A is not necessarily symmetric, as friendship ties collected from sources like surveys can and do yield cases of $A_{ij} \neq A_{ji}$. As an (asymmetric) adjacency matrix defines a (directed) graph, which in this case is a *friendship network*, we refer to the n people as nodes. We refer to an (unordered) pair of individuals (i, j) as a *dyad*, and a value where $A_{ij} = 1$ as an *edge*, and also interchangeably as a *tie* or *link*.

While A_{ij} represents a 'dependency' between units i and j, for example $A_{ij} = 1$ will frequently be associated with similar outcomes from i and j, such ties are themselves dependent [94]. For example, while friendship ties are not necessarily mutual or *reciprocated*, we still have $A_{ij} = A_{ji}$ more often than we would expect at random. In symbolic terms, $P(A_{ij}) \not\perp P(A_{ji})$

Other such dependencies include *transitivity*, which are friend-of-a-friend connections, $P(A_{ij}) \not\perp P(\{A_{ik}, A_{kj} : k \neq i, j\})$, and preferential attachment [85], $P(A_{ij} \not\perp P(\{A_{kj} : k \neq j\})$.

Not accounting for such dependencies can, in explanatory models, lead to too-small standard errors and omitted variable bias [32–34]. In predictive models [14, 93], the effect of dyadic dependencies is on the validity of cross-validation estimates of model performance. This was discussed for pairwise similarity in activity recognition by Hammerla and Plötz [47]; in our case, a pairwise entity itself is the modeling target. In any link prediction task, having a cross-validation scheme that does not reflect an application setting (for example, removing ties at random from a network to form a test set, when in a real setting, unlabeled ties would be arriving in time) can potentially give a misleading picture of what true out-of-sample performance would be.

Unfortunately, there is no general solution to cross-validation in the presence of dyadic dependencies; even approaches that make strong assumptions about the data-generating process are challenging [28]. In order to best simulate an application setting, where we would not have a picture of the entire network or of previous values of ties, we explicitly avoid the use of both network and temporal features.

This does not solve the problem completely: for example, if for dyad (i, j) we have $A_{ij} = A_{ji} = 1$, and if A_{ij} is in the training set and A_{ji} in the test set (where A_{ij} and A_{ji} would be associated with the same undirected/symmetric proximity relations), we would be training and testing on the exact same feature levels, with the same label.

To somewhat address this problem, in addition to using *naïve k*-fold cross-validation that ignores dependencies, we employ two alternative cross-validation schemes, described below, to control for some dependencies.

2.2 Friendship and co-location

The connection of friendship and proximity has long been a topic of study in social science. In a foundational work of social psychology carried out in 1946, the 'Westgate study', Festinger et al. [41] carried out a field experiment around soldiers returning from WWII and attending graduate school at MIT on the GI bill. They and their families were randomly placed into the units of the relatively isolated, newly built Westgate housing complex. The study authors were able to quantify the extent to which people living close by, or passing one another on the way to their residences, were more likely to become friends than with others with whom they did not have opportunities for interaction (although the study looking only at men may have neglected an important causal process in the role of women, [17]). Then, in 1954 and 1955, the 'Newcomb-Nordlie fraternity' study [80, 81] recruited two waves of 17 male students who did not know each other, gave them free 'fraternity-style' housing, and studied how their personal characteristics, political positions, and interests affected their eventual friendship formation. These two studies established that proximity plays an important role in friendship, but also that proximity is not sufficient, and that other characteristics matter.

Later work [60] further quantified the relationship between distance and friendship, finding "the inverse square of the distance separating two persons" to be a good fit to measures of social impact. A retreat for all incoming sociology majors at the University of Groningen provided another opportunity for studying the emergence of friendship, with van Duijn et al. [102] finding that friendships developed due to one of four main effects: physical proximity, visible similarity, invisible similarity, and network opportunity.

There are also a number of studies using data from online social networks or other online platforms to study the connection between friendship and geography, although many of these are at the global scale [4, 20, 64, 67, 87] and do not have resolution at the scales at which interactions can occur. Furthermore, users of location-based social networks, or those who share locations on general online social networks like Twitter and Facebook, are a fraction of total users and form an unrepresentative sample [72] for a general smartphone-using population. Other work has taken call logs and used calls as the ties to study with respect to geography [82, 104], although call logs and SMS have a surprisingly poor relationship with friendship as measured by self-report [105, 106].

A landmark series of papers by Bernard, Killworth, and Sailer [9–11, 51, 52] showed that people are generally bad at recalling objective patterns of interaction. However, subsequent work [43, 57] argued that *psychological perceptions of the network*, rather than objectively measurable ties of interaction, were what were causal for individuals' behavior: subjective data may in some cases be more valuable for predicting and explaining, and thus the psychological perceptions captured by survey data may be more valuable for certain tasks than objective measurements.

This is related to adams' [sic] [1] critique of the model in Eagle et al. [38] predicting friendship from co-location; adams notes that there are 'close strangers and distant friends', both of which a method based on co-location would misclassify. In response, Eagle et al. [35] argue that some causal network processes might happen unconsciously, and sensor measurements might be able to detect these. The correlation of friendship and proximity can make it difficult to sort out which may be causal for a given process [25]. But in the future, if we can decorrelate friendships and proximity by controlling for one or the other, it could help us understand whether social influence or environmental factors are more likely to be the causal factor.

2.3 Sensors for social networks

Since the first work with the 'sociometer' (later, 'sociometric badge') in 2002 [21], research groups have been using sensors to collect data about social systems. Studies often used sensor 'nodes' or other custom devices [3, 42, 44, 48, 59, 84], and with the exception of RFID tags [5], studies have largely used mobile devices [29, 36, 53, 54, 62, 65, 88, 90, 92, 92, 98, 99] because of their wide range of existing sensors, and their use results in lower participant burden than when participants are required to wear or carry an additional device. By and large, the data used is either co-location data (from either GPS, self-reported check-ins, or mutual detection of fixed sensors or WiFi hotspots) or proximity (through Bluetooth or RFID). There are also studies that use video to detect interactions [16, 50, 56, 103], and sociometric badges [21] or headsets [71] that gather audio data that allows for detecting conversations between specific individuals, but, in our work, we focus on the co-location and proximity sensing made possible with mobile phones.

These existing studies largely fall into three categories. Most common are studies that describe infrastructure, technical details, and study design, followed sometimes by descriptive modeling of network characteristics and some bivariate relationships [2, 3, 5–7, 15, 22–24, 42, 44, 49, 53, 63, 65, 75, 88, 92, 92, 98, 99]. This has been important work that has contributed to present-day mobile sensing tools – tools that are capable of going beyond exploration and overall descriptives into measuring specific processes, and applications designed on top of those processes like network-based health interventions [68].

The second category is of those that try to build systems other than the specific sensing platform. They may build or lay the groundwork for recommender systems [18, 19, 45, 62], or present models or algorithms for mining information about interactions from sensor data [29, 30, 36, 86].

The third category is of those that employ statistical models or techniques to make conclusions using sensor data. Stehlé et al. [97] look at the connection between 'spatial behavior', measured by RFID badges, and gender similarity. The 'Friends and Family' or 'SocialfMRI' study and dataset [2] has been used to look at the connection between interaction and financial status [83] and interaction and sleep and mood [78]. Madan et al. [69] used sensor measurements to look at the relationship between social interactions and changes in political opinion. The 'Reality Mining' dataset [36] has similarly been used to look at obesity and exercise in the presence of contact between people [70]. Another approach is that of Eagle & Pentland [37], which presents a spectral clustering system for extracting daily patterns from time series. Staiano et al. [96] use ego networks (induced subgraphs of single nodes and all their respective neighbors) in call logs, Bluetooth-based proximity networks, and surveys to predict Big-5 personality traits. Dong et al. [31] modeled the co-evolution of behavior and social relationships from mobile phone sensor data.

Also in this category, and most similar to our work is a paper based on the Reality Mining dataset: Eagle and Pentland [38] used mobile phone data proximity to infer a network of self-reported friendship ties. They first calculated a 'probability of proximity' score over the range of a week, calculated as an average frequency of proximity over nine months of data, which they showed was systematically different for each of reciprocated self-reported ties $(A_{ij}, A_{ji}) = (1, 1)$, non-reciprocated ties $(A_{ij}, A_{ji}) \in \{(0, 1), (1, 0)\}$, and no ties $(A_{ij}, A_{ji}) = (0, 0)$. Their model was an explanatory one, meaning that cross-validation was not necessary, and it was appropriate to group nine months of data to model even a wave of friendship self-report in the first month. However, such an approach does not match a mobile use case, which would involve detecting friendships only from recently gathered batches of location data. Neither the 'probability of proximity' score, nor the friendship inferences, were subjected to cross-validation, such that the reported 95% accuracy for inferring friendships is effectively training accuracy. In this paper, we attempt to replicate their approach; while we also found that, when aggregated over the entire time period of data collection, there was a major difference between mutual friendship ties and both non-ties and non-reciprocated ties (fig. 3), this proved ineffective for building a classifier. Aggregation like this, over time, dyads, and splits of training and test sets, obscures the variance that ruins test performance.

We now describe the questions that we are trying to answer and the study and analysis we conducted to answer them.

3 STUDY DESIGN AND PROCEDURE

The goal of our study is to understand the feasibility of inferring social relationships (friendship in particular) from (only) passive smartphone data. We are especially interested in the following questions:

(1) How well can we detect friendships from co-location features? In other words, if all we know about two people is their location patterns, how accurately can we say if they are friends?
(2) If we know that friendships exist, how well can we detect if these friendships are *close* friendships?
(3) How accurately can we detect whether a friendship is likely to change? Will co-location patterns over time provide information about the creation or dissolution of friendships?

To answer these questions, we carried out a 3-month study among members of a fraternity to use smartphone data to try and capture interactions and relationships as they were formed and evolved during that period. The following section describes the study setup and data collection process.

3.1 Participants and recruitment

We recruited members of an undergraduate fraternity in a research university in the northeastern United States. The fraternity had 60 members at the start of the study, with an additional 21 prospective members going through the 'pledging' process during the study duration, of which 19 completed the process. Of this cohort of 79 men, we recruited 66 participants, of which 53 ultimately participated in sensor data collection, and of which 48 responded to at least one survey wave. Having this sort of well-defined *boundary specification* [61] let us ask each study participant about their friendships with each member of the fraternity, giving negative examples that are explicit, unlike open-ended solicitation for friendships (such as from 'name generator' instruments) in which individuals are only implicitly not friends by not being mentioned.

The fraternity was relatively loose-knit; about 20 fraternity members live in a fraternity house, with the rest living elsewhere and required to be in the fraternity only one day a week (for a fraternity chapter-wide meeting).

The study lasted approximately three months. Participants were compensated $20 a week for having the passive and automated sensor data collection software, AWARE [40], installed on their smartphones, with additional $5 incentives for each survey wave they completed.

3.2 Data collection

Our task was to use mobile phone sensor data relating to location and proximity in a model that could recover self-reported friendship ties, and changes in such ties. Consequently, we collected survey data about friendships in three waves, and used a tracking app to collect Wifi, Bluetooth, and location data from mobile phones.

3.2.1 Survey data. During the study, participants were asked to fill out a survey asking about their social connections, based off of existing instruments [55, 101]. There was a public listing of fraternity members, and consequently we were able to ask about respondents' ties to all fraternity members (i.e., ask about ties to everybody in the specified boundary), not just those participating in the study; while we were not able to relate friendships with non-participant fraternity members to sensor data, since non-participation meant we do not have sensor data, it does give a sense of the importance of non-participants in the social system.[2]

The surveys were collected three times over 9 weeks: shortly after the beginning of the study, then four weeks after, and lastly at the end of the study five weeks later (we made the second period longer, as one of these five weeks was spring break, when many study participants were away from campus). Participants were asked about five different quantities: their recollections about who they interacted with frequently; who they considered to be a friend; who they considered to be a *close* friend; who they went to for advice on personal matters; and who they went to for advice on professional/academic matters. The correlation between these collected networks, between each other and over time, is given below in figure (4). Friendship can change at shorter intervals than six weeks; but since friendship is an internal and subjective psychological construct, currently the only way of getting data on friendship is surveys with high respondent burden that makes it infeasible to collect at more frequent intervals.

3.2.2 Passive smartphone data. We equipped each participant with the AWARE mobile phone framework[3] [40] on their iOS devices (≈90% of participants) or Android devices (the remaining ≈10%; there were no users of Windows or other mobile operating systems). We used AWARE to record Bluetooth and WiFi detections, each at 10 minute sampling intervals. We also had continuous monitoring of battery and screen status (on/off), and complete records of call and message metadata (with hashed values for phone numbers). For location, the Android AWARE client uses the Google fused location plugin, which has several options for trading off accuracy and battery usage, and for which we selected the low power option. The iOS AWARE client uses the iOS location services, in which we similarly selected an option with low battery usage.

We also performed WiFi fingerprinting in the fraternity house to help us determine when participants were co-located in rooms in the house.

4 DATA PROCESSING

4.1 Data completeness

4.1.1 Survey data. The completeness of the survey data is shown in fig. (1a). The response rate dropped in each survey round; compared to survey 1, survey 2 had a response rate of 59%, and survey 3 had a response rate of 51%. In total, there were 48 participants providing network data, 34 of which responded to 2 surveys giving us longitudinal network data (the minimum requirement for use in our chosen statistical social network model), including 20 participants that responded to all 3 surveys. In total, out of $\binom{48}{2} = 1128$ potential pairs, we were able to train and/or test on 830 pairs.

4.1.2 Sensor data. The completeness of the sensor data is shown in fig. (1b). Some logistical problems prevented all participants from starting smartphone data collection on the first day, and some participants discontinued the use of the app because of technical issues (battery life, sporadic interference with certain external Bluetooth devices, etc.).

There were two sources of missing values in the calculated features: either artifacts relating to no observations fulfilling a certain criteria (e.g., no co-locations within 50m on mornings), or else actual missing data (one or both mobile devices were not providing a certain sensor's data during a given period, e.g., mornings of a given week). For the former (artifacts), we replaced missing values with appropriate substitutes, such as 0s or the maximum possible value. For logarithmic features, some of which could be less than 1, we replaced $-\infty$ with

[2]For example, if non-participants were all seldom nominated by respondents (corresponding to low indegree in the collected networks), it would mean that study non-participation is related to being unimportant in the social system, which would be encouraging, although this did not turn out to be the case (see below).

[3]http://www.awareframework.com

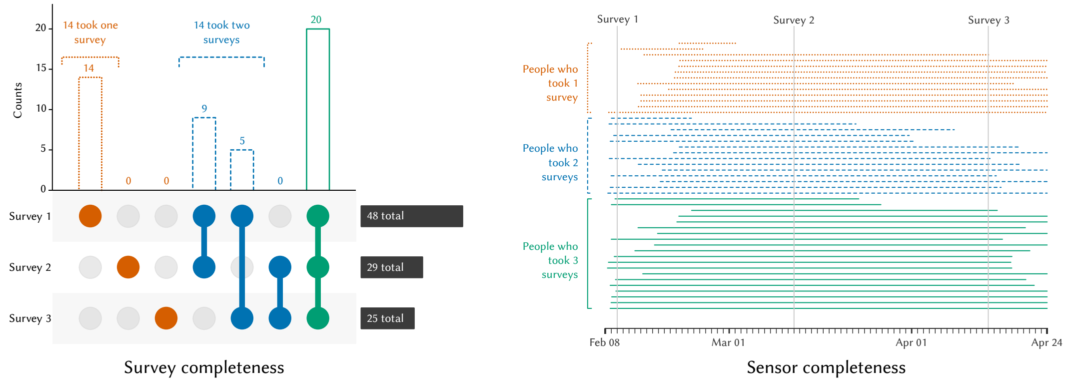

Fig. 1. (Left) Looking at longitudinal completeness, 14 people completed survey 1 only, and none completed surveys 2 or 3 only. 9 people completed surveys 1 and 2, 5 people completed surveys 1 and 3, none completed 2 and 3 only, and 20 people completed all three waves. This is shown in the vertical bars at the top. This comes out to 48 respondents for survey 1, 29 respondents for survey 2, and 25 respondents for survey 3, shown in the solid horizontal bars on the right side. (Right) We show the time periods in which sensor data was collected for people who answered one survey (dotted lines), two surveys (dashed lines), or all three surveys (solid lines). The times of the three surveys are marked with vertical lines.

zeros. For inverse-squared features, we replaced ∞ with a value, 200, slightly larger than the largest observed inverse-squared value. For the missing values resulting from an actual lack of data, we kept the missing values in the cells of the feature matrix. This necessitated using classifiers that can handle missing values among the features, like the R random forest implementation `rpart` [100] which has procedures for handling NAs when constructing decision trees, and other packages built on top of `rpart`.

4.2 Data handling

4.2.1 Missing data. Based on the distribution of lengths of time where data was missing (fig. 2), we would have needed to interpolate up to eight hour intervals to have any real impact on the proportion of time that has missing data. Consequently, we chose to not use partial interpolation. We did try out last value carried forward interpolation, but this did not change results.

4.2.2 Spring break. Spring break may be extremely informative, for example if two people are proximate to each other but far from everybody else it may be that they are more likely to be friends. However, spring break is systematically different from every other week, such that if we train on spring break, we have no meaningful test set. Thus, we removed spring break from the data set. This is also why the two periods have an unequal number of weeks, with 4 weeks between survey waves 1 and 2, and 5 weeks between survey waves 2 and 3; spring break fell between survey waves 2 and 3, such that removing it leaves 4 weeks in each period.

4.2.3 Alternative approaches. In addition to the main approach described here, we also looked at interpolation alternatives. Time series interpolation, performed prior to feature extraction, did not improve performance. Interpolation on the extracted features also did not improve performance. Trying thresholds of equal width (specifically, trying thresholds from 50m to 500m in 50m increments, or 200m increments from 600m to 2000m as in fig. 3) also did not improve performance. Lastly and most significantly, treating the classification task as being between mutual dyads and other dyads, as per the Eagle et al. work [38], resulted in worse test performance.

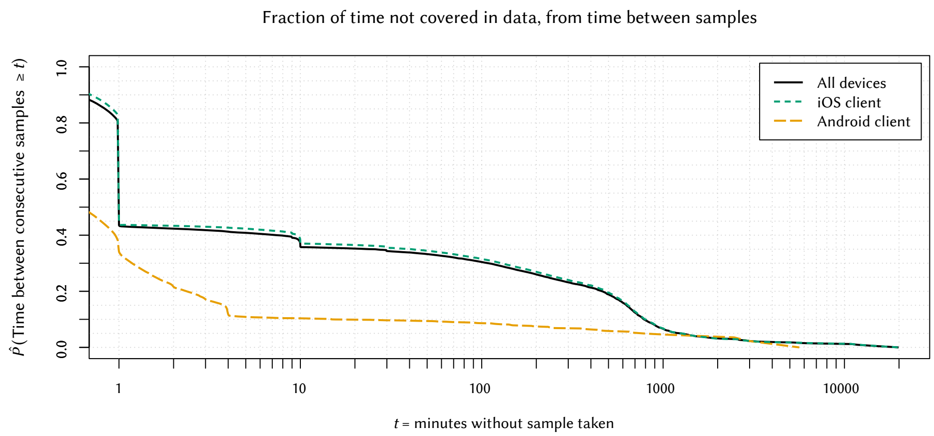

Fig. 2. A plot of the survival function (the empirical complementary cumulative distribution) for time not covered by sensor data. 40% of the study time across all participants is covered by sensor readings less frequent than 10 minutes, meaning that 60% of the data has no gaps in coverage.

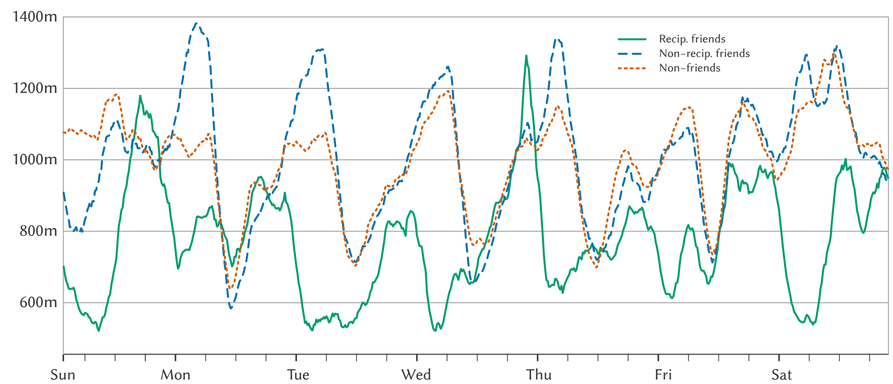

Fig. 3. The median weekly pairwise distances between reciprocated (mutual) friendships, $A_{ij} = A_{ji} = 1$, non-reciprocated friendships, $A_{ij} = 1 \neq A_{ji} = 0$ or $A_{ij} = 0 \neq A_{ji} = 1$, and non-friendships, $A_{ij} = A_{ji} = 0$, for times when pairs are within the area of the university, and aggregated over the entire period of data (i.e., no training/test split). This is analogous to the approach of [38], and this figure reproduces their figure 2 (except with median distance, rather than mean frequency of proximity). While it appears there is a strong pattern, it is a result of an aggregation that obscures the variance between weeks and in data splitting, such that this seeming pattern proved ineffective as a basis of classification in testing.

4.3 Collected surveys and sensors

4.3.1 Network survey instrument. We asked participants about five different types of ties in each of the three surveys: following previous social science literature, we asked about advice-seeking relationships (both personal advice seeking, and academic/professional advice-seeking), in addition to asking about friendships and, for each reported friendship, asking if it was also a close friendship. For comparison with work on recall [9–11, 51, 52] and memorability of social interactions [60], we also asked about frequency of interaction.

The similarities between these collected networks, both for the same network across the three waves and between the different networks, is given in fig. (4). The similarity metric used is the Jaccard index, a common method for comparing networks (as it looks only at ties shared across the two networks, not shared non-ties), potentially of overlapping but unequal sets of nodes, which in our case happens because of non-response. For two networks N_A and N_B, with $n_{A \cap B}$ overlapping nodes and adjacency matrices A and B restricted to these nodes, the Jaccard index is

$$J(A, B) = \frac{|\{A_{ij} = B_{ij} = 1\}|}{2 \times \binom{n_{A \cap B}}{2}}$$

As we can see, there is a much higher correlation between self-reported frequent interaction and friendship than there is between friendship and close friendship. There is also a high correlation between close friendship and the two types of advice ties; while we did not use advice ties in the current analysis, this similarity

giving insight into what types of relationships the prompt about ‘close friendships’ elicit (like the prompt about friendship, we explicitly do not define what we mean by ‘close’, letting participants interpret the term).

Looking at the changes in the networks from survey to survey, we see that close friendships and both types of advice-seeking relationships are much less variable over time than are friendships or self-reports of frequent interaction.

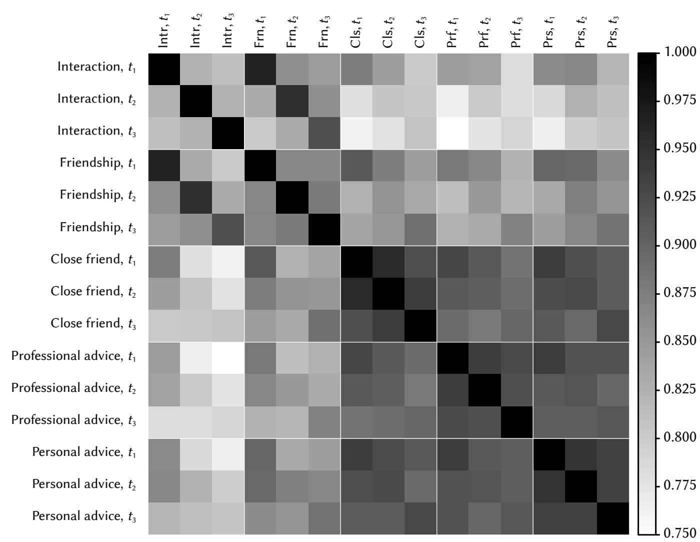

Fig. 4. The similarity between networks (5 types of ties, each collected 3 times), measured via the Jaccard index. The self-reported frequent interaction and friendship networks are more similar than the other networks, and both also exhibit more variation across the three waves.

4.3.2 Bluetooth. Collected Bluetooth data turned out to be unusable. Both Android and iOS no longer make the 16 hex digit Bluetooth MAC addresses of detected devices available to app developers. Instead, detected devices are recorded in terms of a 32 hex digit universally unique identifier (UUID), which are assigned by the detecting device uniquely to each detected device and used to recognize those detected devices in the future.

4.3.3 Wifi. One candidate for characterizing proximity is when two devices detect or connect to the same Wifi device. Here, Wifi hotspot MAC addresses are unique (unlike the hotspot name/label, which for example with 'eduroam' is shared not only across multiple hotspots in the same university, but across multiple cities across the world!), and mutual detection of this picks up when two devices are proximate. Out of 830 potential pairs, 406 pairs of mobile devices detected at least one Wifi hotspot in common (although not necessarily at the same time).

As mentioned above, we also conduced Wifi fingerprinting in the fraternity house, including collecting all Wifi devices detected in each room along with the received signal strength indication (RSSI) of the respective signals. In order to perform Wifi fingerprinting (match a set of hotspots that are detected by a mobile phone in a given scan and with respective RSSIs to previously collected profiles from specific rooms), we needed multiple detected hotspots per scan (every 10 minutes). However, we found that only about 6.7% of scans for Wifi hotspots recorded more than one detected hotspot; in the frat house as well, we could tell when a device was connected to one of the frat house's Wifi hotspots, but not which other hotspots were detected in order to determine a specific room. Thus, we only use as the basis for features whether at least one Wifi hotspot was detected in common at the scan of a specific 10 minute interval from two devices, ignoring the tiny fraction of detections that include multiple devices, and also ignoring RSSI. The frat house has 5 main Wifi devices for about 30 rooms over 3 floors. Based on the size of rooms in the fraternity house and the relative coverage of its Wifi devices, we estimate that at least within the fraternity house, our Wifi localization approach is accurate to within a bit of a smaller radius than its general 32m accuracy, perhaps 20m or so; however, we do not have similar measurements for the rest of campus.

4.3.4 Location. Since we have, from previous theory, that co-location is causally related to friendship through interaction, we would ideally want to extract features from pairwise distance measurements that will be effective as a proxy for interaction. However, in the Google Fused Location plugin that AWARE uses to collect location data, we used the PRIORITY_LOW_POWER option, which prioritizes low power usage, as previous testing with AWARE had showed battery drain was a major cause of participant dropout. This low power option does not actively use GPS, instead using a combination of cell phone towers and detected Wifi hotspot with known geolocations, and is advertised as being accurate to within about 10km.[4] In the iOS client, the accuracy setting corresponding to low power use was to set `desiredAccuracy` option to 1km, with a threshhold for recording new movements of 1000m.[5] In practice, the reported accuracy was usually much better, with a significant portion of readings reporting an accuracy of within 10m. Still, even if the location is not accurate to within the range of potential interactions, similarity in location (as measured by co-location) is a form of dyadic similarity that can be directly or indirectly related causally to friendship.

We sought to pick several choice thresholds that might characterize geographic similarity in simple way. First, we plot an empirical complementary cumulative distribution function (i.e., a survival function) in log-x scale (fig. 5a) to better see the overall distribution. There is an 'elbow' around 2000m, which is about the size of the university and surrounding area. Within 2000m, we cluster the weighted distances [104] to find thresholds, which we show over a kernel density estimate of the head of the distribution (fig. 5b).

Next, to compare self-reported ties and co-location, it is necessary to summarize four week spans of relational geolocation data into a set of features. t

After data processing, we have the following:

- Continuous-valued time series of *pairwise distances*
- Binary time series of whether both members of a given pair were within a geobox around the university campus

[4] https://developers.google.com/android/reference/com/google/android/gms/location/LocationRequest, and http://www.awareframework.com/plugin/?package=com.aware.plugin.google.fused_location

[5] https://developer.apple.com/library/content/documentation/UserExperience/Conceptual/LocationAwarenessPG/CoreLocation/CoreLocation.html

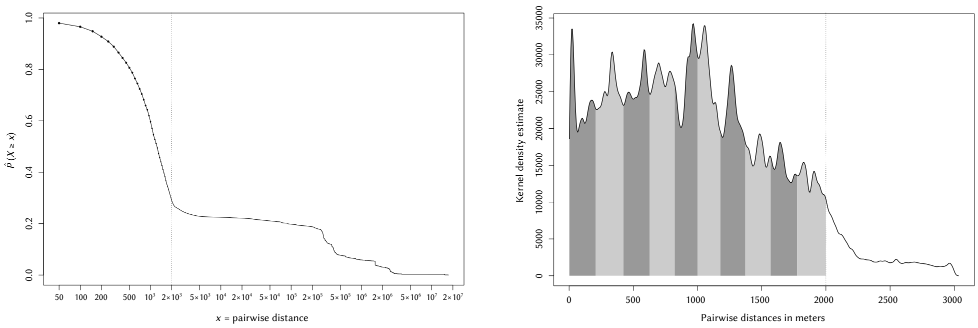

Fig. 5. A survival function (left), plotted in log-x scale, shows pairwise distances over time. Based on the 'elbow' around 2000m (approximately the size of the university and surrounding area), marked with a vertical dotted line, we only found clusters for pairwise distances below 2000m. Below 2000m, we clustered distances (again weighted by the time spent at that distance). The fitted clusters are shown on top of a kernel density estimate (right) that gives a detail of the head of the distribution. The cluster breaks are at 207m, 422m, 626m, 822m, 1001m, 1178m, 1373m, 1570m, 1776m, and then our cutoff of 2000m. These are also listed in table (1).

- Binary time series of whether both members of a given pair were within a geobox around the fraternity house
- Binary time series of whether both members of a given pair detected at least one Wifi hotspot in common
- Binary time series of whether both members of a given pair detected a Wifi hotspot visible from the fraternity house in common

All the above maintain missing values that result from one or both members of a pair not having recorded sensor data.

The set of features are summarized in table (1). The first set of features are calculated from the continuous-valued time series of pairwise distances.

Then, we set thresholds on these pairwise distances of 50m, 100m, ..., 500m, which give binary time series. The second set of features are calculated on each of these 10 binary time series, and are based off of summary statistics for Bernoulli variables.

Additionally, when calculating the continuous-valued time series of pairwise distances, we also generated binary time series for if the locations of both members of the pair fell within a geobox around the university's campus, and a geobox around the fraternity house. This gave 2 additional binary time series, and multiplying each of these with each of the 10 thresholded time series gave an additional 20 time series.

Lastly, we can consider the length of sequences of consecutive 1s (spans of co-location at the given threshold) and of consecutive 0s (gaps between co-location at the given threshold). These are integer-valued but we treat them as continuous, and calculate summary statistics accordingly.

Each of these feature types are crossed with time periods: weekdays only and weekends only, nights only (12am - 6am), mornings only (6am - 12pm), afternoons only (12pm - 6pm), and evenings only (6pm - 12am). We employed summaries of distributions after logarithmic transformation after observing that the original distributions were often heavily right-skewed.

In total, there are $9 + 12 \times (5 + 20) = 309$ features, each taken over seven settings, for $309 \times 7 = 2163$ candidate location features. Wifi features were $2 \times (5 + 20) = 50$, and $50 \times 7 = 350$ for an additional 350 features, for a total of 2,513 features. We extracted these over two 4-week periods, corresponding to the 4 weeks between surveys 1 and 2, and the 5 weeks between surveys 2 and 3 with the week of spring break subtracted out.

Distribution to summarize		Statistic		Timeframe
Pairwise distances (continuous-valued)	⊗	Mean Median Standard deviation Mean of logarithm Median of logarithm Standard deviation of logarithm Mean of inverse-squared distance Median of inverse-squared distance Standard deviation of inverse-squared distance	⊗	4-week period Weekdays only within 4-week period Weekends only within 4-week period Mornings [6am - 12pm) only within 4-week period Afternoons [12pm - 6pm) only within 4-week period Evenings [6pm - 12am) only within 4-week period Nights [12am - 6am) only within 4-week period
Pair is within 207m (binary) Pair is within 422m (binary) Pair is within 626m (binary) Pair is within 822m (binary) Pair is within 1001m (binary) Pair is within 1178m (binary) Pair is within 1373m (binary) Pair is within 1570m (binary) Pair is within 1776m (binary) Pair is within 2000m (binary) Pair is within geobox around campus (binary) Pair is within geobox around fraternity house (binary) Wifi device detected in common (binary) Wifi device in fraternity house detected in common (binary)	⊗	Count of times within threshold Mean of count Standard deviation of count Standard deviation squared of count Count(1 - Count) *Spans (distribution of lengths of consecutive 1s)* Mean of lengths Median of lengths Standard deviation of lengths Minimum length Maximum length Mean of logarithm of lengths Median of logarithm of lengths Standard deviation of logarithm of lengths Minimum of logarithm of lengths Maximum of logarithm of lengths *Gaps (distribution of lengths of consecutive 0s)* Mean of lengths Median of lengths Standard deviation of lengths Minimum length Maximum length Mean of logarithm of lengths Median of logarithm of lengths Standard deviation of logarithm of lengths Minimum of logarithm of lengths Maximum of logarithm of lengths	⊗	

Table 1. Extracted features. "⊗" indicates taking all pairwise combinations. The thresholds are irregularly spaced because they are empirically derived from 1-dimensional clustering; see fig. (5b) for these clusters.

5 DETECTION OF FRIENDSHIPS, STRENGTH, AND CHANGE

5.1 Modeling friendships

We use sensor data from weeks 1-4 to predict the self-reported friendships in survey wave 2 (given to study participants at the end of week 4), and sensor data from weeks 5-9 (excluding the week of spring break), to predict the self-reported friendships in survey wave 3 (given to participants at the end of week 9). This is a standard binary classification task. In this task, we do not make use of survey wave 1.

5.2 Modeling friendship evolution

We can treat changes in friendship (whether tie creation or dissolution) as another binary classification task, with targets

- $P(A_{ij}^{(t)} = A_{ij}^{(t+1)} \mid X^{[t,t+1)})$: No change in friendship (either no friendship, or maintained friendship)
- $P(A_{ij}^{(t)} \neq A_{ij}^{(t+1)} \mid X^{[t,t+1)})$: Change in friendship (either tie creation or tie dissolution)

While we ideally would be able to separately model tie creation and dissolution, as they are distinct processes [95], in our data only a small proportion of ties changed in either direction such that modeling became difficult. We will see below that our results for this task were poor, although treating it as a multiclass problem over the direction of change only led to worse performance.

5.3 Cross-validation schema

To try and avoid network and temporal dependencies giving us inflated model performance, we use three different cross-validation schemes:

- Naïve CV, which is 'naïve' in the same way as is Naïve Bayes. This is k-fold cross validation, with $k = 10$, randomly assigning each A_{ij} to a fold.
- Dyad-based CV. This k-fold cross validation, grouped by the values of a pair of participants (a dyad) in both directions and across all surveys: group $(A_{ij}^{(1)}, A_{ji}^{(1)}, A_{ij}^{(2)}, A_{ji}^{(2)}, A_{ij}^{(3)}, A_{ji}^{(3)})$ and assign the entire 6-tuple to a single fold. Not all of these values in the tuple will be non-missing, meaning that the folds will not be of equal size. But since assignment to fold is not dependent of the number of missing values, sizes will be the same in expectation.
- Temporal cross-validation. Train on $(A^{(2)}, X^{[1,2)})$, test on $(A^{(3)}, X^{[2,3)})$ for friendship detection, and train on $(A^{(1)}, A^{(2)}, X^{[1,2)})$, test on $(A^{(2)}, A^{(3)}, X^{[2,3)})$ for changes in friendship.

5.3.1 Dyad-based cross-validation. When carrying out cross-validation normally, dependencies may impact the validity of CV estimates of true out-of-sample performance in the sense that performance in an application setting will turn out to be far worse than was performance on held-out test data. To see how dependencies could cause this to happen, consider the naive approach of randomly split edges into training and test sets. We will sometimes get A_{ij} in the training set and A_{ji} in the test set; but in our data—as is typical in friendship network data—81% of friendship edges self-reported in a given period are reciprocated (either both are 0 or both are 1), the exact same feature values on which the model was trained would also appear in the test set with the same response. Clearly, reciprocity is one way that information can leak across training and test sets. But even if we make sure that both A_{ij} and A_{ji} are on the same side of the data split, there could be another person k who had similar patterns of co-location with both i and j, and who is friends with both, creating the same problem; cross-validation schemes have been proposed to deal with this [see 28, for a discussion], but they involve either partitioning the adjacency matrix into a training block and test block, which involves throwing out all the data of the cross-blocks, or partitioning by cells of the adjacency matrix (an edge coloring problem), but neither scheme will break arbitrary network dependencies. The practical consequences of this are that cross-validation estimates of model performance will potentially be inflated as compared to performance in real application settings.

5.3.2 Temporal block cross-validation. Another form of leakage is temporal; especially in models that use running means or other such window-based metrics as features, randomly assigning observations to folds amounts to 'time traveling', using future knowledge to predict the past. This clearly will make test estimates of out-of-sample performance massively inflated. One way to address this, discussed in [89] and [8], is to perform training and testing in temporal blocks, with the test set always chronologically after the training set, and preferably with a 'buffer' block between the two to break some autoregressive dependence.

In our case, we do not use previous friendship values as features, because we are assuming a case where ground truth is unknown or known from another setting, rather than only currently unknown but known at a previous value in the same setting.

The dyad-based scheme lacks any ability to control for triadic effects, but better controls for any temporal autocorrelation, and controls completely for reciprocity. The temporal scheme better controls for triadic effects that may be at work within a single time period and for reciprocity within a single time period, but if people have set habits of location and co-location, there will still be temporal information leakage, and if a tie is reciprocated with a delay, there will be information leakage from this.

The application settings that each cross-validation scheme represent are, respectively, if we trained on known dyads and used it to predict unknown dyads, and if we trained on known friendships in one period and then used sensor data to predict changes or future values of those friendships.

The ideal application setting would not need any self-reported survey data to detect friendships or evolution, which would mean taking a model trained on one social system and applying it in another system; for this, the temporal cross-validation scheme is likely a better approximation, given how much the system of the fraternity potentially changed with the introduction of the new pledges (and how the new pledges themselves may have changed in developing new friendships, or weakening previous ones).

5.4 Modeling and feature selection

For our classifier, we used an implementation of AdaBoost [79] built on the R random forest implementation of `rpart`, as this implementation has explicit procedures for handle missing values. We wanted to preserve missing values for substantive reasons, although we also did try interpolation on both the time series and in the feature matrix, neither of which improved results.

Ensembles of decision tree also provide a measure of feature importance, the mean decrease in GINI coefficient for variables across trees. However, we found the most important features varying wildly over different folds, and different CV schema. Methods like stability selection [74] that might address this, like most implementations of ensembles of logistic regression, did not have packages that could handle missing values. We explored implementing our own heuristic version of stability selection for an ensemble of trees, but ultimately this was not stable either. Our final approach was to separately fit random forests; while their performances were not as good as those of AdaBoost, the features selected into the random forest had much more consistency across folds, seeds, and CV schemes.

6 RESULTS

6.1 Friendship detection

Results for the three cross-validation schemes are given in table (2). In each case, the no information rate corresponds to the proportion of the majority class, 0, and would be the accuracy we would get if we always predicted no tie. Subtracting this from 1 also gives the *network density*, which is the number of edges divided by the maximum number of possible edges. The friendship networks have a density of 0.23.

As anticipated, the naïve CV scheme gives better results than either of the other two CV schema. We use a one-sided binomial test of the accuracy against the No Information Rate (NIR), equal to the frequency of the majority class, and find that both naïve and dyadic CV are significant at the usual $p < 0.05$ level, but under temporal block CV, the classifier is only significantly better than the NIR at the $p < 0.1$ level.

To summarize classifier performance, we use Matthews correlation coefficient (MCC). This is the same as Pearson's ϕ or mean square contingency coefficient, an analog for a pair of binary variables of Pearson's product-moment correlation coefficient, but was rediscovered by Matthews [73] for use as a classification metric. For the count of true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN), the MCC is

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(FP + TP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}}.$$

The MCC has several desirable properties. It has an interpretable range (0 for random predictions, -1 for perfect misclassification, and 1 for perfect classification), has a connection to the χ^2 statistic used to test independence in a 2×2 contingency table in that $|\text{MCC}| = \sqrt{\chi^2/n}$ and, most helpfully, is a good summary of performance in cases of class imbalance [13].

In our classifications, the MCC ranges from .30 under the most optimistic test performance, that of the naïve CV, to .21 under temporal block CV. We also tried class balancing, which improved recall but at the expense of MCC and of overall accuracy.

	Naive CV	**Dyad-based CV**	**Temporal block CV**
Accuracy	0.8006	0.7920	0.7913
Accuracy, 95% CI	(0.7882, 0.8125)	(0.7794, 0.8042)	(0.7726, 0.8091)
(No Information Rate / Majority class)	*(0.7740)*	*(0.7740)*	*(0.7785)*
Binomial test, Accuracy vs. NIR, *p*-value	p=1.5e-05	p=0.0025	p=0.0901
Precision (Positive predictive value)	0.6918	0.6508	0.6812
Recall/Sensitivity (True positive rate)	0.2122	0.1723	0.1088
Specificity (True negative rate)	0.9724	0.9730	0.9855
F1 score	0.3248	0.2724	0.2964
AUC	0.7148	0.7039	0.1876
Matthews correlation coefficient	0.3039	0.2562	0.2120

Table 2. Friendship detection, test performance across the three CV schema. The no information rate corresponds to the baseline accuracy given by predicting no ties; in the case of networks, this

6.2 Detecting close friendships

We repeat the assessment of the above models, conditioning on the presence of a friendship, and making our detection target whether or not a friendship is reported to be *close*. In this case, the network of close friendships has a network density of .41, making the no information rate .59.

	Naive CV	**Dyad-based CV**	**Temporal block CV**
Accuracy	0.6817	0.6670	0.5741
Accuracy, 95% CI	(0.6511, 0.7112)	(0.6361, 0.6969)	(0.5259, 0.6212)
(No Information Rate / Majority class)	*(0.5861)*	*(0.5861)*	*(0.5185)*
Binomial test, Accuracy vs. NIR, *p*-value	p=7.6e-10	p=1.8e-07	p=0.0117
Precision (Positive predictive value)	0.6904	0.6711	0.7069
Recall/Sensitivity (True positive rate)	0.4188	0.3832	0.1971
Specificity (True negative rate)	0.8674	0.8674	0.9241
F1 score	0.5213	0.4879	0.3083
AUC	0.6997	0.6695	0.5889
Matthews correlation coefficient	0.3250	0.2906	0.1777

Table 3. *Close* friendship detection, conditioned on the presence of a friendship, test performance across the three CV schema.

We see a similar pattern of performance, with temporal block CV being the most conservative/pessimistic, and naïve CV being far more optimistic.

6.3 Detecting changes in friendship

Detecting loss in friendships could be particularly important for social interventions, such as preventing the onset of isolation. However, the rarity of changes in friendship (only 13% of ties change, either being created or dissolving) complicates modeling.

our feature extraction and modeling approach proved unable to meaningfully detect changes in friendship. AdaBoost failed to predict any positive test cases for any CV schema; a random forest performed better, with table (4) showing these results. Although with a Matthews correlation coefficient of .07 for the naïve CV and .03 for the dyadic-based CV, this is minimal quality, and the classifier output does not pass a statistical test for being significantly better than the No Information Rate.

	Naive CV	**Dyad-based CV**
Accuracy	0.6842	0.8645
Accuracy, 95% CI	(0.6692, 0.6989)	(0.8532, 0.8752)
(No Information Rate / Majority class)	*(0.8710)*	*(0.8710)*
Binomial test, Accuracy vs. NIR, p-value	p=1	p=0.8902
Precision (Positive predictive value)	0.1676	0.2093
Recall/Sensitivity (True positive rate)	0.3651	0.0183
Specificity (True negative rate)	0.7315	0.9898
F1 score	0.2297	0.0336
AUC	0.5483	0.5167
Matthews correlation coefficient	0.0720	0.0256

Table 4. *Change* detection, random forest test performance. AdaBoost made only negative test classifications, but random forests (performance shown here) did make some positive classifications under naïve and dyad-based CV, although under temporal block CV again there were no positive classifications.

6.4 Feature importance and selection

Feature selection can often improve classifier performance, but it is also useful for diagnostic and exploratory analysis. In our case, we use Correlation-based Feature Selection (CFS) [46]. This tries to pick out a set of features that are both correlated with the class label, and uncorrelated with one another.

However, for both this and other approaches (e.g., feature importance measures from random forests), we found that important features across the different folds and different CV schema were highly variable. Stability selection [74], like many techniques built on logistic regression, does not have available implementations that handle missing data; our own heuristic implementation of this with random forests, rather than with ℓ_1 regularized logistic regression, did not work well.

We focus feature selection for the friendship detection task, as for friendship change in particular, feature selection did not improve classifier performance, and the poor overall performance suggests that these my not be the 'right' extracted features to characterize changes in friendships.

The results here are from applying CFS to the training set for the temporal block CV, as temporal block CV is the most conservative cross-validation method. Specifically, we split the temporal block training set into 10 folds, and perform CFS within each. Again taking the idea from stability selection of looking only if whether a given feature was selected inside a given fold, and not looking at its specific importance, we choose those features that appeared in CFS runs on at least 9 of the 10 folds. This selects in 19 features, the correlations between which are shown in figure (6).

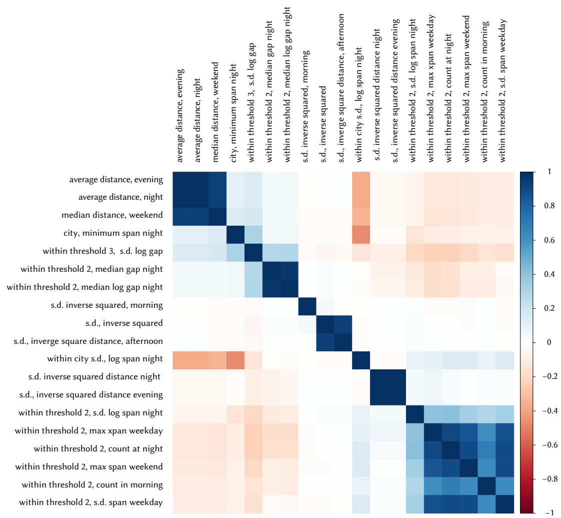

Fig. 6. Correlations between the features selected via CFS on the training set of a temporal block cross-validation scheme.

Performance with these selected features did improve slightly over the temporal block CV, raising the MCC from .21 to .26, although the improvement in accuracy was not significant change.

There are seemingly three mutually uncorrected groups of features. Metrics based on inverse squared distances are helpful for classification and are also relatively uncorrelated with other selected features. There are several characterizations of average distance, a number of features connected to the second threshold at 422m, and many features characterizing variance.

These are:

The Pearson correlations between these are shown in fig. (6). While there are groups of highly linearly correlated features, many of the features are not correlated, giving an independent signal.

This feature selection suggests two things. First, that Latané et al.'s [60] finding that inverse-squared distance fits well to reports of memorable social interactions may be a helpful metric around which to extract features. Second, the second threshold, at 422m (fig. 5b), seems particularly relevant versus others. Otherwise, being in the fraternity house in the *mornings*, in the city in the *afternoons*, and within 422m of each other in the *evenings* seems to capture some relevant patterns of friendship. However, these results are tentative, first because predictive models can be misleading [14, 93] and should only be substantively interpreted in exploratory, conditional ways, and second as the better-performing models built with AdaBoost had a great deal of variance in which features had nonzero coefficients across folds. Conversely, an ensemble made with AdaBoost using only these features has significantly worse performance. The random forest performance is barely effected using only these 30 features instead of all 2,163, but its performance is overall below that given by AdaBoost.

7 DISCUSSION AND CONCLUSIONS

We had three goals: to detect friendships, to detect close friendships given that a friendship exists, and to detect a change in friendship. We were generally successful in the first two goals. While these results may be improved upon, the fact that proximity is not related perfectly to friendship means there will always be a limit to how well we can use co-location can detect friendship ties.

The results of the three cross-validation schema, and particularly now the dyad-based and temporal block CV schema gave results worse than those from the naïve CV, show the importance of considering dependencies when evaluating model performance. Again, temporal autocorrelation and dyadic reciprocity are not the only forms of dependencies relevant for the study of dyadic ties in time, but in designing CV schemes that take these dependencies into account, we show a practical alternate to naïve CV in the absence of well-developed, general methods for cross-validation with dyadic data.

This work is also the first to take the proposal of Eagle et al. [38] but to extend it to, and test it for, an application setting. They showed compelling patterns of reciprocated friendships, non-reciprocated friendships, and non-friends each having distinct proximity patterns. We were able to replicate this pattern (fig. 3), but found it to not be an effective basis for classification. In particular, their reported 95% accuracy in classifying friendships from proximity data was only training performance, with a single feature (proximity of probability) that was aggregated over period far after when the surveys about friendship ties were given. While the basic insight that friends tend to be proximate is still useful for detecting friendships, our extensive feature extraction and careful testing establish a new baseline for the extent to which this is feasible.

7.1 Limitations

Our process of feature extraction was still done using feature engineering, rather than automated methods that might do better at discovering a signal. Our CV schema still did not control for all dependencies, such as highly active or popular nodes, or transitive triads. Our attempts to classify changes in friendship were not successful; it is possible that extracting features that characterize *changes* in metrics over the given 4-week period would be more effective in modeling change. Lastly, while we were able to obtain a stable set of features, it is for a different model than the best-performing one, and this set of features does degrade the performance when used as a specification in the best-performing model. Further work will be needed to identify stable features relevant for the best-performing model.

Another significant limitation is that this approach of using co-location to detect friendship currently has not been demonstrated to be precise enough, to have high enough recall, to justify the potentially extreme intrusiveness of trying to predict friendships, a very intimate part of people's lives. Such systems would need to first improve accuracy, and second establish opt-in use cases to prevent the connection between friendship and proximity being the basis of surveillance systems.

7.2 Future work

In addition to the immediate directions of improving on these new baseline results, being able to characterize co-location can help us look at how friendship and co-location behavior co-evolve; there is a causal feedback loop of friends spending time together, and of people who spend time together becoming friends; knowing what aspects of co-location to consider can help separate out these two processes.

Next, while the precision and recall of the classifiers are lower than is ideal, it is a strong basis on which to build further data collection, for example in delivering prompts of likely friends for people to select, a difficult task outside of well-defined boundary specifications.

7.3 Summary

In this paper, we:

- Describe the collection of subjective, self-reported friendship data alongside objective sensor data within a given boundary specification;
- Design two cross-validation approaches with our data that control for dependencies that inflate test performance, giving a more realistic picture of true, out-of-sample performance;

- Build models that both beat a random baseline and a majority class baseline, and whose performance can serve as benchmarks for future work;
- Perform exploratory feature extraction as a way to see which engineered features will help classifier performance.

REFERENCES

[1] jimi adams. 2010. Distant friends, close strangers? Inferring friendships from behavior. *Proceedings of the National Academy of Sciences* 107, 9 (2010), E29–E30. https://doi.org/10.1073/pnas.0911195107
[2] Nadav Aharony, Wei Pan, Cory Ip, Inas Khayal, and Alex Pentland. 2011. Social fMRI: Investigating and shaping social mechanisms in the real world. *Pervasive and Mobile Computing* 7, 6 (2011), 643–659. https://doi.org/10.1016/j.pmcj.2011.09.004
[3] Constantinos Marios Angelopoulos, Christofoulos Mouskos, and Sotiris Nikoletseas. 2011. Social signal processing: Detecting human interactions using wireless sensor networks. In *Proceedings of the 9th ACM International Symposium on Mobility Management and Wireless Access (MobiWac '11)*. 171–174. https://doi.org/10.1145/2069131.2069163
[4] Lars Backstrom, Eric Sun, and Cameron Marlow. 2010. Find me if you can: Improving geographical prediction with social and spatial proximity. In *Proceedings of the 19th International Conference on World Wide Web (WWW '10)*. 61–70. https://doi.org/10.1145/1772690.1772698
[5] Alain Barrat, Ciro Cattuto, Vittoria Colizza, Francesco Gesualdo, Lorenzo Isella, Elisabetta Pandolfi, Jean-François Pinton, Lucilla Ravà, Caterina Rizzo, Mariateresa Romano, Juliette Stehlé, Alberto Eugenio Tozzi, and Wouter van den Broeck. 2013. Empirical temporal networks of face-to-face human interactions. *The European Physical Journal Special Topics* 222, 6 (2013), 1295–1309. https://doi.org/10.1140/epjst/e2013-01927-7
[6] Alain Barrat, Ciro Cattuto, Vittoria Colizza, Lorenzo Isella, Caterina Rizzo, Alberto Eugenio Tozzi, and Wouter van den Broeck. 2012. Wearable sensor networks for measuring face-to-face contact patterns in healthcare settings. In *Revised Selected Papers from the Third International Conference on Electronic Healthcare (eHealth 2010)*, Vol. 69. 192–195. https://doi.org/10.1007/978-3-642-23635-8_24
[7] Alain Barrat, Ciro Cattuto, Vittoria Colizza, Jean-François Pinton, Wouter Van den Broeck, and Alessandro Vespignani. 2008. High resolution dynamical mapping of social interactions with active RFID. arXiv:0811.4170. (2008). arXiv:https://arxiv.org/abs/0811.4170
[8] Christoph Bergmeir and José M. Benítez. 2012. On the use of cross-validation for time series predictor evaluation. *Information Sciences* 191 (2012), 1920–213. https://doi.org/10.1016/j.ins.2011.12.028
[9] H. Russell Bernard and Peter D. Killworth. 1977. Information accuracy in social network data II. *Human Communication Research* 4, 1 (1977), 3–18. https://doi.org/10.1111/j.1468-2958.1977.tb00591.x
[10] H. Russell Bernard, Peter D. Killworth, and Lee Sailer. 1979. Informant accuracy in social network data IV: A comparison of clique-level structure in behavioral and cognitive network data. *Social Networks* 2, 3 (1979), 191–218. https://doi.org/10.1016/0378-8733(79)90014-5
[11] H. Russell Bernard, Peter D. Killworth, and Lee Sailer. 1982. Informant accuracy in social-network data V: An experimental attempt to predict actual communication from recall data. *Social Science Research* 11, 1 (1982), 30–66. https://doi.org/10.1016/0049-089X(82)90006-0
[12] Stephen P. Borgatti, Ajay Mehra, Daniel J. Brass, and Giuseppe Labianca. 2009. Network analysis in the social sciences. *Science* 323, 5916 (2009), 892–895. https://doi.org/10.1126/science.1165821
[13] Sabri Boughorbel, Fethi Jarray, and Mohammed El-Anbari. 2017. Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric. *PLOS ONE* 12, 6 (06 2017), 1–17. https://doi.org/10.1371/journal.pone.0177678
[14] Leo Breiman. 2001. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science* 16, 3 (08 2001), 199–231. https://doi.org/10.1214/ss/1009213726
[15] Ciro Cattuto, Wouter van den Broeck, Alain Barrat, Vittoria Colizza, Jean-François Pinton, and Alessandro Vespignani. 2010. Dynamics of person-to-person interactions from distributed RFID sensor networks. *PLOS ONE* 5, 7 (2010), e11596. https://doi.org/10.1371/journal.pone.0011596
[16] Datong Chen, Jie Yang, Robert Malkin, and Howard D. Wactlar. 2007. Detecting social interactions of the elderly in a nursing home environment. *ACM Transactions on Multimedia Computing, Communications, and Applications* 3, 1 (2007). https://doi.org/10.1145/1198302.1198308
[17] Frances Cherry. 1995. One man's social psychology is another woman's social history. In *The stubborn particulars of social psychology: Essays on the research process*. Routledge, London, 68–83.
[18] Alvin Chin, Bin Xu, Hao Wang, Lele Chang, Hao Wang, and Lijun Zhu. 2013. Connecting people through physical proximity and physical resources at a conference. *ACM Transactions on Intelligent System Technologies* 4, 3 (2013), 50:1–50:21. https://doi.org/10.1145/2483669.2483683
[19] Alvin Chin, Bin Xu, Hao Wang, and Xia Wang. 2012. Linking people through physical proximity in a conference. In *Proceedings of the 3rd International Workshop on Modeling Social Media (MSM '12)*. 13–20. https://doi.org/10.1145/2310057.2310061

[20] Eunjoon Cho, Seth A. Myers, and Jure Leskovec. 2011. Friendship and mobility: User movement in location-based social networks. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '11)*. 1082–1090. https://doi.org/10.1145/2020408.2020579

[21] Tanzeem Choudhury and Alex Pentland. 2002. The sociometer: A wearable device for understanding human networks. In *Proceedings of the Workshop on Ad hoc Communications and Collaboration in Ubiquitous Computing Environments, Computer Supported Cooperative Work*.

[22] Tanzeem Choudhury and Alex Pentland. 2003. Modeling face-to-face communication using the sociometer. In *Workshop Proceedings of Ubicomp (Workshop: Supporting Social Interaction and Face-to-face Communication in Public Spaces)*.

[23] Tanzeem Choudhury and Alex Pentland. 2003. Sensing and modeling human networks using the sociometer. In *Proceedings of the 7th IEEE International Symposium on Wearable Computers (ISWC '03)*. 216–222. https://doi.org/10.1109/ISWC.2003.1241414

[24] Iolanthe Chronis, Anmol Madan, and Alex Pentland. 2009. SocialCircuits: The art of using mobile phones for modeling personal interactions. In *Proceedings of the ICMI-MLMI '09 Workshop on Multimodal Sensor-Based Systems and Mobile Phones for Social Computing (ICMI-MLMI '09)*. Article 1, 1:1–1:4 pages. https://doi.org/10.1145/1641389.1641390

[25] Ethan Cohen-Cole and Jason M. Fletcher. 2008. Is obesity contagious? Social networks vs. environmental factors in the obesity epidemic. *Journal of Health Economics* 27, 5 (2008), 1382–1387. https://doi.org/10.1016/j.jhealeco.2008.04.005

[26] Rense Corten. 2012. Composition and structure of a large online social network in the Netherlands. *PLOS ONE* 7 (2012), e34760. Issue 4. https://doi.org/10.1371/journal.pone.0034760

[27] Justin Cranshaw, Eran Toch, Jason Hong, Aniket Kittur, and Norman Sadeh. 2010. Bridging the gap between physical location and online social networks. In *Proceedings of the 12th ACM International Conference on Ubiquitous Computing (Ubicomp '10)*. 119–128. https://doi.org/10.1145/1864349.1864380

[28] Beau Dabbs and Brian Junker. 2016. Comparison of cross-validation methods for stochastic block models. arXiv:1605.03000. (2016). arXiv:https://arxiv.org/abs/1612.04717

[29] Trinh Minh Tri Do and D. Gatica-Perez. 2011. GroupUs: Smartphone proximity data and human interaction type mining. In *Proceedings of the 15th Annual International Symposium on Wearable Computers (ISWC 2011)*. 21–28. https://doi.org/10.1109/ISWC.2011.28

[30] Trinh Minh Tri Do and Daniel Gatica-Perez. 2013. Human interaction discovery in smartphone proximity networks. *Personal and Ubiquitous Computing* 17, 3 (2013), 413–431. https://doi.org/10.1007/s00779-011-0489-7

[31] Wen Dong, Bruno Lepri, and Alex Pentland. 2011. Modeling the co-evolution of behaviors and social relationships using mobile phone data. In *Proceedings of the 10th International Conference on Mobile and Ubiquitous Multimedia (MUM '11)*. 134–143. https://doi.org/10.1145/2107596.2107613

[32] Malcolm M. Dow. 2007. Galton's Problem as multiple network autocorrelation effects: Cultural trait transmission and ecological constraint. *Cross-Cultural Research* 41, 4 (2007), 336–363. https://doi.org/10.1177/1069397107305452

[33] Malcolm M. Dow, Michael L. Burton, and Douglas R. White. 1982. Network autocorrelation: A simulation study of a foundational problem in regression and survey research. *Social Networks* 4, 2 (1982), 169–200. https://doi.org/10.1016/0378-8733(82)90031-4

[34] Malcolm M. Dow, Michael L. Burton, Douglas R. White, and Karl P. Reitz. 1984. Galton's Problem as network autocorrelation. *American Ethnologist* 11, 4 (1984), 754–770. https://doi.org/10.1525/ae.1984.11.4.02a00080

[35] Nathan Eagle, Aaron Clauset, Alex Pentland, and David Lazer. 2010. Reply to Adams: Multi-dimensional edge inference. *Proceedings of the National Academy of Sciences* 107, 9 (2010), E31. https://doi.org/10.1073/pnas.0913678107

[36] Nathan Eagle and Alex Pentland. 2006. Reality mining: Sensing complex social systems. *Personal Ubiquitous Computing* 10, 4 (03 2006), 255–268. https://doi.org/10.1007/s00779-005-0046-3

[37] Nathan Eagle and Alex Pentland. 2009. Eigenbehaviors: identifying structure in routine. *Behavioral Ecology and Sociobiology* 63, 7 (2009), 1057–1066. https://doi.org/10.1007/s00265-009-0739-0

[38] Nathan Eagle, Alex Pentland, and David Lazer. 2009. Inferring friendship network structure by using mobile phone data. *Proceedings of the National Academy of Sciences* 106, 36 (2009), 15274–15278. https://doi.org/10.1073/pnas.0900282106

[39] Beverley Fehr. 1996. *Friendship processes.* Sage Publications, Inc., Thousand Oaks, CA.

[40] Denzil Ferreira, Vassilis Kostakos, and Anind K. Dey. 2015. AWARE: Mobile context instrumentation framework. *Frontiers in ICT* 2, 6 (2015), 1–9. https://doi.org/10.3389/fict.2015.00006

[41] Leon Festinger, Kurt W. Back, and Stanley Schachter. 1950. *Social pressure in informal groups: A study of human factors in housing.* Stanford University Press, Stanford, CA.

[42] Anna Förster, Kamini Garg, Hoang Anh Nguyen, and Silvia Giordano. 2012. On context awareness and social distance in human mobility traces. In *Proceedings of the Third ACM International Workshop on Mobile Opportunistic Networks (MobiOpp '12)*. 5–12. https://doi.org/10.1145/2159576.2159581

[43] Linton C. Freeman, A. Kimball Romney, and Sue C. Freeman. 1987. Cognitive structure and informant accuracy. *American Anthropologist* 89, 2 (1987), 310–325. https://doi.org/10.1525/aa.1987.89.2.02a00020

[44] Adrien Friggeri, Guillaume Chelius, Eric Fleury, Antoine Fraboulet, France Mentré, and Jean-Christophe Lucet. 2011. Reconstructing social interactions using an unreliable wireless sensor network. *Computer Communications* 34, 5 (2011), 609–618. https://doi.org/10.

1016/j.comcom.2010.06.005

[45] Raghu K. Ganti, Yu-En Tsai, and Tarek F. Abdelzaher. 2008. SenseWorld: Towards cyber-physical social networks. In *Proceedings of the 2008 International Conference on Information Processing in Sensor Networks (IPSN '08)*. 563–564. https://doi.org/10.1109/IPSN.2008.48

[46] Mark A. Hall. 1999. *Correlation-based feature selection for machine learning*. Ph.D. Dissertation. Department of Computer Science, The University of Waikato.

[47] Nils Y. Hammerla and Thomas Plötz. 2015. Let's (not) stick together: Pairwise similarity biases cross-validation in activity recognition. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '15)*. 1041–1051. https://doi.org/10.1145/2750858.2807551

[48] Jeng-Cheng Hsieh, Chih-Ming Chen, and Hsiao-Fang Lin. 2010. Social interaction mining based on wireless sensor networks for promoting cooperative learning performance in classroom learning environment. In *Proceedings of the 6th IEEE International Conference on Wireless, Mobile and Ubiquitous Technologies in Education (WMUTE 2010)*. 219–221. https://doi.org/10.1109/WMUTE.2010.22

[49] Lorenzo Isella, Juliette Stehlé, Alain Barrat, Ciro Cattuto, Jean-François Pinton, and Wouter van den Broeck. 2011. What's in a crowd? Analysis of face-to-face behavioral networks. *Journal of Theoretical Biology* 271, 1 (2011), 166–180. https://doi.org/10.1016/j.jtbi.2010.11.033

[50] Shian-Ru Ke, Hoang Le Uyen Thuc, Yong-Jin Lee, Jenq-Neng Hwang, Jang-Hee Yoo, and Kyoung-Ho Choi. 2013. A review on video-based human activity recognition. *Computers* 2, 2 (2013), 88–131. https://doi.org/10.3390/computers2020088

[51] Peter D. Killworth and H. Russell Bernard. 1976. Informant accuracy in social network data. *Human Organization* 35, 3 (1976), 269–286. https://doi.org/10.17730/humo.35.3.10215j2m359266n2

[52] Peter D. Killworth and H. Russell Bernard. 1979. Informant accuracy in social network data III: A comparison of triadic structure in behavioral and cognitive data. *Social Networks* 2, 1 (1979), 19–46. https://doi.org/10.1016/0378-8733(79)90009-1

[53] Niko Kiukkonen, Jan Blom, Olivier Dousse, Daniel Gatica-Perez, and Juha Laurila. 2010. Towards rich mobile phone datasets: Lausanne data collection campaign. In *Proceedings of the 7th ACM International Conference on Pervasive Services (ICPS '10)*.

[54] Mikkel Baun Kjærgaard and Petteri Nurmi. 2012. Challenges for social sensing using WiFi signals. In *Proceedings of the 1st ACM Workshop on Mobile Systems for Computational Social Science (MCSS '12)*. 17–21. https://doi.org/10.1145/2307863.2307869

[55] Andrea Knecht, Tom A. B. Snijders, Chris Baerveldt, Christian E. G. Steglich, and Werner Raub. 2010. Friendship and delinquency: Selection and influence processes in early adolescence. *Social Development* 19, 3 (2010), 494–514. https://doi.org/10.1111/j.1467-9507.2009.00564.x

[56] Yu Kong and Yun Fu. 2016. Close human interaction recognition using patch-aware models. *IEEE Transactions on Image Processing* 25, 1 (2016), 167–178. https://doi.org/10.1109/TIP.2015.2498410

[57] David Krackhardt. 1987. Cognitive social structures. *Social Networks* 9, 2 (1987), 109–134. https://doi.org/10.1016/0378-8733(87)90009-8

[58] David Krackhardt. 1996. Social networks and the liability of newness for managers. In *Trends in Organizational Behavior*, Cary L. Cooper and Denise M. Rousseau (Eds.). Vol. 3. John Wiley & Sons, Inc., Chichester, NY, 159–173.

[59] Mathew Laibowitz, Jonathan Gips, Ryan Aylward, Alex Pentland, and Joseph A. Paradiso. 2006. A sensor network for social dynamics. In *Proceedings of the Fifth International Conference on Information Processing in Sensor Networks (IPSN 2006)*. 483–491. https://doi.org/10.1109/IPSN.2006.243937

[60] Bibb Latané, James H. Liu, Andrzej Nowak, Michael Bonevento, and Long Zheng. 1995. Distance matters: Physical space and social impact. *Personality and Social Psychology Bulletin* 21, 8 (1995), 795–805. https://doi.org/10.1177/0146167295218002

[61] Edward O. Laumann, Peter V. Marsden, and David Prensky. 1983. The boundary specification problem in network analysis. In *Applied network analysis: A methodological introduction*, Ron S. Burt and Michael J. Minor (Eds.). Vol. 61. Sage Publications, Beverly Hills, CA, 18–34.

[62] Jamie Lawrence, Terry R. Payne, and David De Roure. 2006. Co-presence communities: Using pervasive computing to support weak social networks. In *Proceedings of the 15th IEEE International Workshops on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE '06)*. 149–156. https://doi.org/10.1109/WETICE.2006.24

[63] Bruno Lepri, Jacopo Staiano, Giulio Rigato, Kyriaki Kalimeri, Ailbhe Finnerty, Fabio Pianesi, Nicu Sebe, and Alex Pentland. 2012. The SocioMetric badges corpus: A multilevel behavioral dataset for social behavior in complex organizations. In *Proceedings of the 2012 ASE/IEEE International Conference on Social Computing and 2012 ASE/IEEE International Conference on Privacy, Security, Risk and Trust (SOCIALCOM-PASSAT '12)*. 623–628. https://doi.org/10.1109/SocialCom-PASSAT.2012.71

[64] Jure Leskovec and Eric Horvitz. 2014. Geospatial structure of a planetary-scale social network. *IEEE Transactions on Computational Social Systems* 1, 3 (2014), 156–163. https://doi.org/10.1109/TCSS.2014.2377789

[65] Minshu Li, Haipeng Wang, Bin Guo, and Zhiwen Yu. 2012. Extraction of human social behavior from mobile phone sensing. In *Proceedings of the 8th International Conference on Active Media Technology (AMT 2012)*. 63–72. https://doi.org/10.1007/978-3-642-35236-2_7

[66] David Liben-Nowell and Jon Kleinberg. 2007. The link-prediction problem for social networks. *Journal of the American Society for Information Science and Technology* 58, 7 (2007), 1019–1031. https://doi.org/10.1002/asi.v58:7

[67] David Liben-Nowell, Jasmine Novak, Ravi Kumar, Prabhakar Raghavan, and Andrew Tomkins. 2005. Geographic routing in social networks. *Proceedings of the National Academy of Sciences* 102, 33 (2005), 11623–11628. https://doi.org/10.1073/pnas.0503018102

[68] Anmol Madan, Manuel Cebrian, David Lazer, and Alex Pentland. 2010. Social sensing for epidemiological behavior change. In *Proceedings of the 12th ACM International Conference on Ubiquitous Computing (UbiComp '10)*. 291–300. https://doi.org/10.1145/1864349.1864394

[69] Anmol Madan, Katayoun Farrahi, Daniel Gatica-Perez, and Alex Pentland. 2011. Pervasive sensing to model political opinions in face-to-face networks. In *Proceedings of the 9th International Conference on Pervasive Computing (Pervasive 2011)*. 214–231. https://doi.org/10.1007/978-3-642-21726-5_14

[70] Anmol Madan, Sai T. Moturu, David Lazer, and Alex Pentland. 2010. Social sensing: Obesity, unhealthy eating and exercise in face-to-face networks. In *Proceedings of Wireless Health 2010 (WH '10)*. 104–110. https://doi.org/10.1145/1921081.1921094

[71] Anmol Madan and Alex Pentland. 2006. VibeFones: Socially aware mobile phones. In *Proceedings of the Tenth IEEE International Symposium on Wearable Computers (ISWC 2006)*. 109–112. https://doi.org/10.1109/ISWC.2006.286352

[72] Momin M. Malik, Hemank Lamba, Constantine Nakos, and Jürgen Pfeffer. 2015. Population bias in geotagged tweets. In *Papers from the 2015 ICWSM Workshop on Standards and Practices in Large-Scale Social Media Research (ICWSM-15 SPSM)*. 18–27. http://www.aaai.org/ocs/index.php/ICWSM/ICWSM15/paper/view/10662

[73] Brian W. Matthews. 1975. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta (BBA) - Protein Structure* 405, 2 (1975), 442–451. https://doi.org/10.1016/0005-2795(75)90109-9

[74] Nicolai Meinshausen and Peter Bühlmann. 2010. Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 72, 4 (2010), 417–473. https://doi.org/10.1111/j.1467-9868.2010.00740.x

[75] Andrew G. Miklas, Kiran K. Gollu, Kelvin K. W. Chan, Stefan Saroiu, Krishna P. Gummadi, and Eyal Lara. 2007. Exploiting social interactions in mobile systems. In *Proceedings of the 9th International Conference on Ubiquitous Computing (Ubicomp 2007)*. 409–428. https://doi.org/10.1007/978-3-540-74853-3_24

[76] Alan Mislove, Massimiliano Marcon, Krishna P. Gummadi, Peter Druschel, and Bobby Bhattacharjee. 2007. Measurement and analysis of online social networks. In *Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement (IMC '07)*. 29–42. https://doi.org/10.1145/1298306.1298311

[77] Jacob L. Moreno. 1934. *Who shall survive? A new approach to the problem of human interrelations.* Nervous and Mental Disease Publishing Co., Washington, D.C.

[78] Sai T. Moturu, Inas Khayal, Nadav Aharony, Wei Pan, and Alex Pentland. 2011. Using social sensing to understand the links between sleep, mood, and sociability. In *Proceedings of the 2011 IEEE International Conference on Privacy, Security, Risk and Trust and IEEE International Conference on Social Computing (PASSAT/SocialCom 2011)*. 208–214. https://doi.org/10.1109/PASSAT/SocialCom.2011.200

[79] Tofigh Naghibi. 2014. *maboost: Binary and multiclass boosting algorithms.* https://CRAN.R-project.org/package=maboost R package version 1.0-0.

[80] Theodore Mead Newcomb. 1961. *The acquaintance process.* Holt, Reinhard & Winston, New York, NY.

[81] Peter G. Nordlie. 1958. *A longitudinal study of interpersonal attraction in a natural group setting,.* Ph.D. Dissertation. University of Michigan.

[82] Jukka-Pekka Onnela, Samuel Arbesman, Marta C. González, Albert-László Barabási, and Nicholas A. Christakis. 2011. Geographic constraints on social network groups. *PLOS ONE* 6, 4 (2011), 1–7. https://doi.org/10.1371/journal.pone.0016939

[83] Wei Pan, Nadav Aharony, and Alex Pentland. 2011. Fortune monitor or fortune teller: Understanding the connection between interaction patterns and financial status. In *Proceedings of the 2011 IEEE International Conference on Privacy, Security, Risk and Trust and IEEE International Conference on Social Computing (PASSAT/SocialCom 2011)*. 200–207. https://doi.org/10.1109/PASSAT/SocialCom.2011.163

[84] Joseph A. Paradiso, Jonathan Gips, Mathew Laibowitz, Sajid Sadi, David Merrill, Ryan Aylward, Pattie Maes, and Alex Pentland. 2010. Identifying and facilitating social interaction with a wearable wireless sensor network. *Personal and Ubiquitous Computing* 14, 2 (2010), 137–152. https://doi.org/10.1007/s00779-009-0239-2

[85] Matjaž Perc. 2014. The Matthew effect in empirical data. *Journal of The Royal Society Interface* 11, 98 (2014). https://doi.org/10.1098/rsif.2014.0378

[86] Dinh Phung, Brett Adams, and Svetha Venkatesh. 2008. Computable social patterns from sparse sensor data. In *Proceedings of the First International Workshop on Location and the Web (LOCWEB '08)*. 69–72. https://doi.org/10.1145/1367798.1367810

[87] Daniele Quercia, Licia Capra, and Jon Crowcroft. 2012. The social world of Twitter: Topics, geography, and emotions. In *Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media (ICWSM-12)*. 298–305. http://www.aaai.org/ocs/index.php/ICWSM/ICWSM12/paper/view/4612

[88] Kiran K. Rachuri, Cecilia Mascolo, Mirco Musolesi, and Peter J. Rentfrow. 2011. SociableSense: Exploring the trade-offs of adaptive sampling and computation offloading for social sensing. In *Proceedings of the 17th Annual International Conference on Mobile Computing and Networking (MobiCom '11)*. 73–84. https://doi.org/10.1145/2030613.2030623

[89] Jeff Racine. 2000. Consistent cross-validatory model-selection for dependent data: hv-block cross-validation. *Journal of Econometrics* 99, 1 (2000), 39–61. https://doi.org/10.1016/S0304-4076(00)00030-0

[90] Mika Raento, Antti Oulasvirta, and Nathan Eagle. 2009. Smartphones: An emerging tool for social scientists. *Sociological Methods & Research* 37, 3 (2009), 426–454. https://doi.org/10.1177/0049124108330005

[91] Samuel Franklin Sampson. 1968. *A novitiate in a period of change: An experimental and case study of social relationships.* Ph.D. Dissertation. Cornell University.

[92] Vedran Sekara and Sune Lehmann. 2014. The strength of friendship ties in proximity sensor data. *PLOS ONE* 9, 7 (2014), 1–8. https://doi.org/10.1371/journal.pone.0100915

[93] Galit Shmueli. 2010. To explain or to predict? *Statistical Science* 25, 3 (2010), 289–310. https://doi.org/10.1214/10-STS330

[94] Tom A. B. Snijders. 2011. Statistical models for social networks. *Annual Review of Sociology* 37, 1 (2011), 131–153. https://doi.org/10.1146/annurev.soc.012809.102709

[95] Tom A. B. Snijders, Gerhard G. van de Bunt, and Christian E. G. Steglich. 2010. Introduction to stochastic actor-based models for network dynamics. *Social Networks* 32, 1 (2010), 44–60. https://doi.org/10.1016/j.socnet.2009.02.004

[96] Jacopo Staiano, Bruno Lepri, Nadav Aharony, Fabio Pianesi, Nicu Sebe, and Alex Pentland. 2012. Friends don't lie: Inferring personality traits from social network structure. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing (UbiComp '12)*. 321–330. https://doi.org/10.1145/2370216.2370266

[97] Juliette Stehlé, François Charbonnier, Tristan Picard, Ciro Cattuto, and Alain Barrat. 2013. Gender homophily from spatial behavior in a primary school: A sociometric study. *Social Networks* 35, 4 (2013), 604–613. https://doi.org/10.1016/j.socnet.2013.08.003

[98] Arkadiusz Stopczynski, Piotr Sapiezynski, Alex Pentland, and Sune Lehmann. 2015. Temporal fidelity in dynamic social networks. *The European Physical Journal B* 88, 10 (2015), 249. https://doi.org/10.1140/epjb/e2015-60549-7

[99] Arkadiusz Stopczynski, Vedran Sekara, Piotr Sapiezynski, Andrea Cuttone, Mette My Madsen, Jakob Eg Larsen, and Sune Lehmann. 2014. Measuring large-scale social networks with high resolution. *PLOS ONE* 9, 4 (04 2014), 1–24. https://doi.org/10.1371/journal.pone.0095978

[100] Terry Therneau and Beth Atkinson. 2018. `rpart`*: Recursive partitioning and regression trees.* https://CRAN.R-project.org/package=rpart R package version 4.1-13.

[101] Gerhard G. van de Bunt, Marijtje A. J. van Duijn, and Tom A. B. Snijders. 1999. Friendship networks through time: An actor-oriented dynamic statistical network model. *Computational & Mathematical Organization Theory* 5, 2 (1999), 167–192. https://doi.org/10.1023/A:1009683123448

[102] Marijtje A. J. van Duijn, Evelien P. H. Zeggelink, Mark Huisman, Frans N. Stokman, and Frans W. Wasseur. 2003. Evolution of sociology freshmen into a friendship network. *The Journal of Mathematical Sociology* 27, 2-3 (2003), 153–191. https://doi.org/10.1080/00222500305889

[103] Coert van Gemeren, Ronald Poppe, and Remco C. Veltkamp. 2016. Spatio-temporal detection of fine-grained dyadic human interactions. In *Proceedings of the 7th International Workshop on Human Behavior Understanding (HBU 2016)*. 116–133. https://doi.org/10.1007/978-3-319-46843-3_8

[104] Dashun Wang, Dino Pedreschi, Chaoming Song, Fosca Giannotti, and Albert-Laszlo Barabasi. 2011. Human mobility, social ties, and link prediction. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '11)*. 1100–1108. https://doi.org/10.1145/2020408.2020581

[105] Jason Wiese, Jun-Ki Min, Jason I. Hong, and John Zimmerman. 2014. *Assessing call and SMS logs as an indication of tie strength.* Technical Report CMU-HCII-14-101. Human-Computer Interaction Institute, School of Computer Science, Carnegie Mellon University.

[106] Jason Wiese, Jun-Ki Min, Jason I. Hong, and John Zimmerman. 2015. "You never call, you never write": Call and SMS logs do not always indicate tie strength. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW '15)*. 765–774. https://doi.org/10.1145/2675133.2675143

[107] Christo Wilson, Alessandra Sala, Krishna P. N. Puttaswamy, and Ben Y. Zhao. 2012. Beyond social graphs: User interactions in online social networks and their implications. *ACM Transactions on the Web* 6 (2012), 17:1–17:31. Issue 4. https://doi.org/10.1145/2382616.2382620

[108] Wayne W. Zachary. 1977. An information flow model for conflict and fission in small groups. *Journal of Anthropological Research* 33, 4 (1977), 452–473. http://www.jstor.org/stable/3629752