Inferring Physical Interactions and Social Ties Using WiFi Signals

Piotr Sapiezynski
Technical University of Denmark
pisa@dtu.dk

Arkadiusz Stopczynski
Google Inc.
astopczynski@google.com

David Kofoed Wind
Technical University of Denmark
dawi@dtu.dk

Jure Leskovec
Stanford University
jure@cs.stanford.edu

Sune Lehmann
Technical University of Denmark,
Niels Bohr Institute
sljo@dtu.dk

ABSTRACT

Today's societies are enveloped in an ever-growing telecommunication infrastructure. This infrastructure has, in recent years, offered important opportunities for sensing and recording a multitude of human behaviors. Human mobility patterns constitute a prominent example that has been studied based on cell phone towers, Bluetooth beacons, or WiFi networks as proxies for location. However, while mobility is an important aspect of human behavior, understanding complex social systems requires studying not only the movement of individuals but also their interactions. Sensing social interactions is a technical challenge and many commonly used approaches—including RFID badges or Bluetooth scanning—offer limited scalability and resolution. Here we show that it is possible, in a scalable and robust way, to accurately infer person to person physical proximity based only on the list of WiFi access points measured by smartphones carried by the two people. We then demonstrate how tracking these interactions over time allows us to infer social ties in a larger population, revealing the internal structure of a complex social system. In both inference tasks our models achieve AUC scores of 0.9 and higher. Our results demonstrate both the value of WiFi signals in social sensing as well as potential threats to privacy that they carry.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous

Keywords

mobility; wifi; face-to-face; proximity; interactions; social networks

1. INTRODUCTION

In today's world we are surrounded by an ever-increasing number of telecommunication infrastructures, such as mobile phone networks, WiFi access points, or Bluetooth beacons. In addition to their intended function of providing connectivity, these infrastructures offer an unprecedented opportunity for sensing, modeling, and subsequent analyzing of a wide range of human behaviors [32]. For example, our interactions with other people and social ties we form are fundamental building blocks of human societies, and—as we show here based on WiFi access points—these can be inferred in a reliable and scalable way.

Investigating person-to-person proximity interactions between individuals is crucial for modeling a number of phenomena, such as epidemic spreading [56] or formation of social ties [15]. The knowledge of the social ties, in turn, allows us to study dissemination of information [65], adoption of habits [2], or performance of teams [13].

Despite the importance of understanding networks of close proximity interactions, there is a scarcity of scalable and efficient ways to obtain large scale datasets. This is due to the fact that technology has only recently developed to the point, where collection of such high resolution data has become technically feasible. The data sources used for investigating mobility of individuals, such as call detail records (CDRs) from mobile operators [19], are too coarse in terms of temporal and spatial resolution to allow inference of person-to-person proximity. On the other hand, the most popular methods for measurement of physical proximity require using specialized hardware (e.g. sociometric badges) [42, 47] or smartphones sensing each other through Bluetooth [14, 2, 59]. Specialized hardware adds cost and complexity to experimental deployments, effectively limiting their scale. Bluetooth scanning realized on participants' mobile phones increases power consumption [17] (limiting temporal resolution that can be achieved) and requires the devices to be in Bluetooth discoverable mode. This requirement raises privacy [69] and security concerns [50]. When a phone is in "discoverable mode" the location of its owner can be tracked by third parties (a fact commonly used both by researchers [31, 44], and advertisers [12]). Moreover, whenever a phone is discoverable, a malicious actor can attempt to pair to it in order to steal contact lists, content of short messages, etc. For these reasons phone manufacturers make it difficult (or impossible) for a handset to remain discoverable. iOS devices disable discoverability whenever the user leaves the Blutooth settings screen. Android devices let the user set the discoverability timeout to, at maximum, five min-

utes. In our study we relied on the fact that in Android versions 4.1 - 6.0 it is still possible to set unlimited discoverability timeout, but this might change at any point in the future. Apart from the privacy and security issues of using Bluetooth for sensing, another shortcoming is that Bluetooth data lacks location context. When co-presence of individuals is inferred through devices sensing each other, an additional step is usually required to estimate the location of the meeting, for example by associating Bluetooth scans with GPS signal [53], by using fixed infrastructure of RFID transmitters [55], or Bluetooth beacons [31]. In the light of these problems, it is clear that alternative methods for tracking person-to-person interactions are needed. We explore the possibility of using WiFi for this purpose.

Present work. Here we study the problem of inferring physical proximity between people from a list of WiFi signals sensed by their phones. We use a longitudinal dataset containing WiFi and Bluetooth scan results from hundreds of participants, collected as part of the Copenhagen Network Study [59]. Using Bluetooth as ground-truth for physical proximity, we train a model for comparing the results of WiFi scans from two devices to determine whether two individuals were in close physical proximity. We employ a number of interpretable metrics to compare the lists of visible WiFi access points, such as Jaccard similarity or correlation of received signal strengths. By investigating the total number of near-by access points and the time of the event we place each interaction in context: more populated areas tend to have more routers available; people are more likely to meet during work hours, or on a Friday afternoon than on a Sunday night. Importantly, our algorithm for using WiFi signals to infer proximity does not rely on positioning the routers in physical space. Co-location is not inferred by thresholding the distance between the estimated location of two individuals. Instead, their WiFi environments are compared and then we estimate the similarity directly. As a final step we are able to combine these insights using machine learning models to achieve the area under receiving operator curve (AUC ROC) scores of up to 0.9. We show that our model works in a range of environments, does not depend on particular access points, and its performance does not deteriorate over time. Our experiments demonstrate that we are able to track close-proximity interactions over time and in different social and spatio-temporal contexts. Overall, our approach performs better than previously suggested solutions.

Moreover, our work shows how physical proximity information obtained through WiFi signals can be used to infer social links in real-world networks. When identifying social ties we find that the size of the groups people meet in is an important factor in accurately inferring social ties, with friends spending more time in smaller groups. We also show that interactions between friends tend not to follow a particular schedule, as opposed to interactions between non-friends. We are able to identify distinct observable behaviors among pairs of students who are friends on Facebook, call each other, or exchange text messages, and we can distinguish these ties from pairs who are not connected. Using interpretable features pertaining to observable behaviors and simple machine learning approaches, we achieve AUC ROC scores over 0.9 in the link prediction task.

Contribution. Our contribution is twofold. Firstly, we present a novel approach for tracking close-proximity person-to-person interactions based on existing infrastructure of WiFi networks and off-the-shelf consumer smartphones (Sections 3 and 4). Secondly, we show that social networks can be accurately inferred from the behaviors measured using this approach (Section 5). Both findings are verified using a real-world longitudinal dataset comprised of high-resolution data of hundreds of people. Our findings can be applied to aid research and empower social applications, but also raise an important questions regarding privacy of millions of smartphone users.

2. EXPERIMENTAL DESIGN

2.1 The Copenhagen Networks Study

The dataset used in this work was collected as part of the Copenhagen Networks Study [59]. In this study, we tracked lives of approximately 1 000 students at Technical Universisty of Denmark. The population is densely connected, including a majority of freshman students beginning their studies in the same year. The data collection was not limited to the university context, thus the data contains a rich variety of interactions, both work-related and recreational [53]. Following students' lives 24/7, we collected two full years (2013-2015) of the rich social dynamics in a large population.

The data was collected primarily using a custom-built application running on Android smartphones handed out the study participants (LGE Nexus 4). The collector software was based on Funf Open Sensing framework [2] and ran in the background, collecting a variety of sensor readings. The data, compressed and encrypted, was periodically uploaded to a server located at the university.

The data was collected with high temporal resolution and includes:

- Bluetooth scans (every 5 minutes): each scan contains a list of discoverable devices[1], their unique identifiers, user defined names, and received signal strength (RSSI). Because we know which anonymized participant identifier corresponds to which Bluetooth unique identifier, we can monitor proximity between the participants.
- WiFi scans (every 5 minutes): each scan contains a list of WiFi access points (both traditional routers and mobile hotspots), their unique identifiers (BSSIDs or MAC addresses), network names they transmit (SSIDs), and RSSI.
- Location (every 5 minutes): both measured by the GPS and estimated using Google's location API based on available beacons such as WiFi access points and GSM towers.
- Connected cell tower (every 10 minutes).
- SMS log (snapshot every 12 hours): meta information (one-way hashed phone number, direction of communication, time) and a one-way hash of the content.

[1]smartphones in the study were specifically configured to be in Bluetooth discoverable mode

- Call log (snapshot every 12 hours): meta information (one-way hashed phone number, direction of communication, time, duration) without any content; because we know the hashes of phone numbers of our participants, we can see how they exchange messages and call each other.
- Screen on/off events (opportunistically).
- Device time offset (every 6 hours): to compensate for errors in how the users set time on their devices.
- Aside from the data from smartphones, we also collected the Facebook friendship graph of the participants, as well as their public interactions in the online social network.

It is important to note that while the reported data intervals were requested by the collector application, but collection also happened in an opportunistic fashion. Every time any application on the participant's phone requested any data of interest (for example location), our collector received and collected this datapoint. This way, for some data types, the effective temporal resolution is significantly higher than the reported intervals, resulting for example in multiple WiFi scans per minute or location update every couple of minutes [49]. In addition to data collected directly from the participants' phones, we also collected snapshots of participants' Facebook data every 24 hours (server-side collection), including lists of their friends, as well as likes, tags, and posts.

All data in the Copenhagen Networks Study was collected with participants' informed consent, with emphasis on ensuring awareness of the complexity and sensitivity of the collected data [57]. The study setup, including security, privacy, and informed consent has been approved by Danish Data Protection Agency. Further details of the study can be found in Ref. [59].

2.2 Inferring physical proximity and social ties

The high-resolution data collected in the Copenhagen Networks study offers an opportunity for an unprecedented insight into dynamics of a complex social system seen across multiple channels. In this study we use Bluetooth scans as a ground truth indicator for measurements of close physical proximity. Due to the small range of Bluetooth transmission, two devices sensing each other over Bluetooth can be assumed to be within distance of up to 10 meters (30 ft) from each other [15, 2, 52, 7]. Moreover, the location data allows for high-resolution fixed rate tracking of the participants' mobility, both indoors and outside. Calls, texts, and Facebook interactions reveal social structure of the population. Communication on these channels is rarely incidental, thus considering the number of interactions on these channels provides a proxy for quantifying the strength of social ties.

We should note that discovering internal social structure of a densely-connected population—such as student population considered here— based on physical proximity is not an easy task. Non-acquainted participants in the study meet on a regular basis during lectures, in cafeterias, or at gym. Here we attempt to extract social signal above and beyond the already-dense hairball of interactions [53] to discover particular close-proximity events and to reconstruct the entire social network of participants.

	training	test
total observations	.5M	115.5M
% positive	31%	31%
unique users	812	820
median number of access points per observation	7.0	7.0
mean number of access points per observation	11.3	11.3

Table 1: Summary statistics of the dataset used to infer proximity events.

3. INFERRING CLOSE-PROXIMITY INTERACTIONS

3.1 Problem statement

In brief, our task is to compare the lists of WiFi routers seen by users A and B approximately at the same time (with at most $\Delta t = 300$ seconds difference) and determe whether the two users were in close physical proximity. We use Bluetooth data to train and verify our models.

3.2 Data preparation

WiFi. We found that in our dataset there are multiple WiFi routers that share the same MAC address, a phenomenon which might confound our task. We use a simple heuristic to remove these "ambiguous" routers since finding the optimal way of identifying them would warrant a publication on its own. Here we simply rely on the network name they broadcast. Because the routers at campus broadcast up to four network names (SSID) per MAC address, we remove the scans of routers which broadcast five or more network names throughout the observation. **X PS X**Put some numbers here: % of mac addresses have more than 4 ssids but they are seen in % of all scans. % of scans become empty afer removing the offending routers.

Next, we identify the home routers for each participant. We assume one home location per month per participant and that the home location corresponds to the routers being scanned most during that month.

Bluetooth. Due to imperfect firmware and software running on the phones, Bluetooth data is not always available—not all users are scanning and discoverable at all times. Therefore, for each hour of the observation period we only consider WiFi scans from users who were seen and who saw at least one other person through Bluetooth during that hour.

To train our model we also need to provide negative examples. For dyads in this category we choose potential interactions between two people who did not see each other on Bluetooth, but whose lists of scan results share at least one overlapping router. Compared to selecting negative samples by randomly sampling dyads this definition brings the task closer to a real-life scenario of discovering very close physical proximity (up to approximately 10 meters).

3.3 Dataset statistics

Table 1 shows the details about the dataset. Through a year of data we found 116M potential interactions. We randomly select .5M of them to train the models.

category	features
AP presence	overlap, union, jaccard
RSSI	spearman, pearson, manhattan, euclidean
AP presence + RSSI	top AP, top AP $+/-6dB$
timing	hour of week
popularity	min popularity, max popularity, Adamic-Adar

Table 2: Features used to infer close-proximity interactions.

We also observe that in 99% of cases of Bluetooth sightings the corresponding WiFi scans overlap by at least one access point. This indicates that there is a potential in using WiFi scan results to infer the co-presence with high recall. Conversely, in more than 31% of cases where there is at least one overlapping access point, the two devices are also close according to Bluetooth. This indicates that WiFi signals can be applied to the task resulting in a high precision solution. The majority (53%) of meetings happen during working hours (from 8am to 7pm) on campus.

3.4 Methods of comparison

We use a number of metrics to compare two lists of WiFi scan results and use these metrics as features in a supervised machine learning approach. We divide the features into the following categories: availability of access points, received signal strength, presence + RSSI, timing, and popularity. Table 2 lists the features we apply, and Figure 1 shows how the probability of an interaction changes as a function of each feature's value. In this section we describe each feature in detail. Citations refer to the first articles using the features for the purpose of face to face contact detection.

Availability of access points (AP presence). First, we compare the list of routers seen by the two phones, regardless of their received signal strength. We introduce the following measures: **overlap**: the raw count of overlapping routers [28]; **union**: size of the union of the two lists; **jaccard**: ratio between the size of the intersection and the size of the union of the two lists [27]. **non-overlap**: the raw count of non-overlapping routers (union-overlap) [28]; Figure 1a-c presents the interplay between the values of the three parameters and the probability of an interaction. Intuitively, the more common routers two phones see in a scan, the higher the probability of them being in close proximity. Perhaps surprisingly, this probability also depends on the size of the union: the larger the union of the two lists the lower the probability of an interaction. This can be explained by the fact that the number of available access points is positively correlated with the population density. Hence, popular places are likely to attract people who do not necessarily interact with one another. Conversely, two people in a relatively unpopular location are more likely to be there together. The visible dip in the union plot, corresponding to lower probability of meeting with around 30 routers present might correspond to a particular location where many non-interactions happen (for example the dining hall), but we

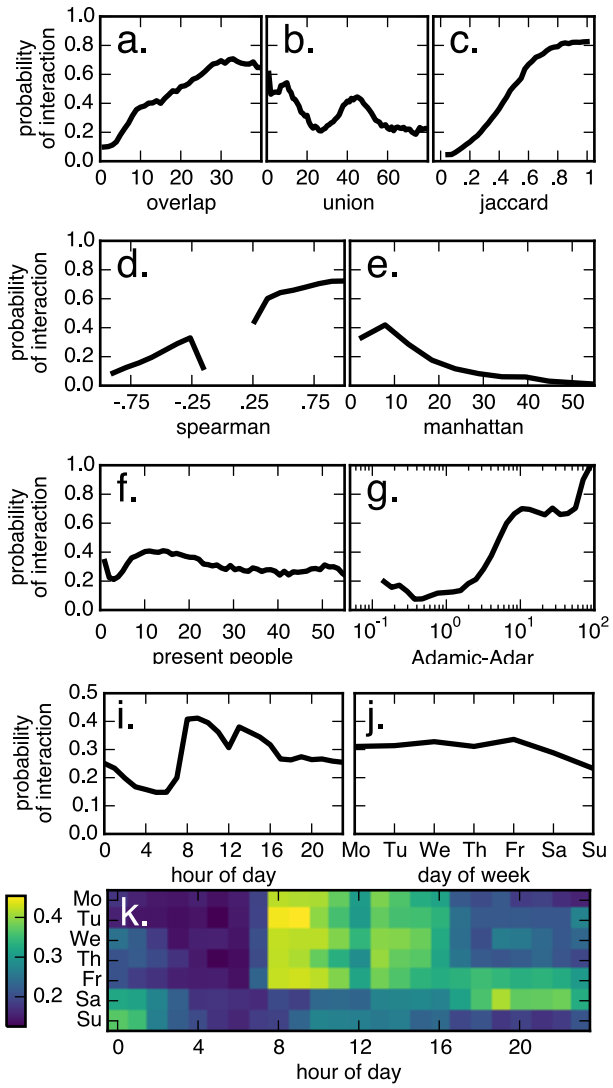

Figure 1: The more common routers two phones see, the higher the probability of close proximity, but the more routers they see in total, the lower the probability of an interaction - densely populated areas have more routers and more people who are not necessarily interacting. Jaccard similarity allows us to recognize interactions regardless of the number of visible access points.

expect that the general tendency holds outside of the context of our experiment. Using Jaccard similarity between the two lists allows to recognize interactions regardless of the number of visible access points.

Received Signal Strength Indicator (RSSI). Next, we focus on comparing the received signal strength of the overlapping routers. While received signal strength (RSSI) is not a good proxy for distance from the router in general terms [48], two colocated people can be expected to have similar RSSI readings for the overlapping routers. Therefore we investigate the **spearman** and **pearson** correlation

coefficients of received signal strengths of the overlapping routers and present the results in Figure 1d. Note that because there are instances where the correlation is undefined (*not a number*) or not statistically significant (with $p_{val} > 0.05$), we replace such values of the coefficients with the mean values of valid correlations (see section 3.5 for details of the imputation). This implies that there are no examples of small correlations (which given only a few values to compare are not statistically significant) and there is a dip in probability of interactions corresponding to the mean value of correlation coefficients.

Furthermore, we also calculate the difference between RSSI of overlapping routers by measuring the ℓ_1 and ℓ_2 distances and dividing the results by the number of overlapping routers. For simplicity we call these features **manhattan** and **euclidean** and define them in Equations 1 [38] and 2 [27] respectively.

$$m = \frac{\sum_i |RSSI_{A,i} - RSSI_{B,i}|}{N} \quad (1)$$

$$e = \frac{\sqrt{\sum_i (RSSI_{A,i} - RSSI_{B,i})^2}}{N} \quad (2)$$

where $RSSI_{A,i}$ is the received signal strength or access point i as measured by user A and N is the total number of overlapping routers. Figure 1e shows that with growing distance, the probability of an interaction falls.

AP presence + RSSI. It has been previously shown that a good heuristic for determining whether a user is in the same location during two measurements is to verify whether they measure a common strongest router [18]. Here, we verify whether this approach can be used for inferring co-location: if two users measure the same router as the strongest one, we assume they are in close proximity. We investigate the strict case, **top AP**. Additionally, we allow for some variability in the measured strength: feature **top AP** $+/-6dB$ takes a positive value if there is at least one overlapping access point in the lists of routers of A and B within $6dB$ from the top router.

Popularity. Additionally, we inspect how many different participants of the study scanned the overlapping routers within five minutes of the meeting - intuitively if only a few persons were in a given location they were more likely to be there together, rather than by chance. We find the least and the most popular among the overlapping routers and report **min_popularity** and **max_popularity**. As we show in Figure 1f., this intuition is not be confirmed by the data. There is no correlation between the number of individuals present and the probability that any two of them are interacting. Note that popularity and the size of union are correlated (Spearman's $\rho = 0.48$, $p_{val} < 0.001$) - more routers are located in popular places, so the more routers there are around, the more people see each of them. However, to achieve a good estimation of popularity, we need data from the entire population, while the number of routers around can be obtained just from data of just the two individuals. Additionally, we use a score inspired by a measure introduced by Adamic and Adar [1], defined as:

$$score(u_1, u_2) = \sum_i \frac{1}{\log(popularity(AP_i))}. \quad (3)$$

Here, each overlapping router is weighted more the fewer people scanned it. In this case, the higher the value, the higher the probability of a meeting between two people.

Timing. In contrast to the other features we described, timing does not rely on comparing the list of scan results. Instead, we use the timestamp of each potential face to face meeting to exploit the temporal characteristics of human interactions. As a reminder, we only consider a potential interaction if both parties have WiFi scans within 300 seconds from one another. For simplicity, we assume here that the timestamp of the potential interaction is the lower of the two scan timestamps. We notice that the prior probability of two people being proximate depends on the time of day and the day of week, as shown in Figure 1i-k. While there is only a small variability between the days of the week (Figure 1j.), the probability of the interaction during a day (Figure 1i.) appears to be driven both by the class schedule—the probability is the highest during classes, and drops during lunchtime—and by after-school social activities. Only by combining the two factors (Figure 1k.), we get the full picture: the probability of interactions from Monday to Tuesday is driven by the school schedule; Friday is a mixture of scheduled and social interactions, with the probability remaining high far into the night hours; Saturday is characterized by interactions starting in the late afternoon and into the night; and on Sunday our participants interact mostly during daytime, with no visible lunch breaks. We add a feature to capture these patterns: **hour of week**: from 0 to 167.

3.5 Imputing missing values

Two of our features are Pearson and Spearman correlations. There are two cases in which it is not possible to calculate the correlation: (1) if there are fewer than three routers available for comparison, (2) if at least one person reads all the signal strenghs at the same level. In such cases we assume a NaN (not-a-number) value of ρ to be imputed later on. Additionally, we assume a NaN value of ρ if the correlation is not significant with the $p_{val} < 0.05$. This results in multiple missing values for the two features. The simplest approach is to skip such observations, but that would imply not training the model in cases with few routers available. We therefore impute the values by assigning the mean value of the feature (averaged over all the non-NaN training examples) when we encouter NaN values. This average from training is preserved and used to impute missing values in the test set. Other approaches, such as using the median value of the feature or using k nearest neighbors to impute the missing value [62] yield similar, but not better results.

4. RESULTS OF INFERRING PHYSICAL INTERACTIONS

4.1 Performance of single features

We first show how well one can infer close-proximity interactions by simply thresholding a single feature. We determine the threshold resulting in the highest area under Receiver Operating Curve (AUC ROC) and then apply the

found threshold on previously unseen data. Additionally, for each category we train a random forest model based on corresponding features. Finally, we combine all the features to train a random forests classifier.

The results are presented in Table 3. We find that the single best performing feature is Jaccard similarity between the two lists of routers. As expected, thresholding on time information is not meaningful (it is equivalent to assuming that all interactions after a certain hour of a certain day of week are close proximity interactions). It is important to note that the performance in test does not drop compared to training, which means that the models and thresholds are not overfitted to the training data.

4.2 Performance of feature sets

Combining the various features using machine learning approaches such as random forests yields a higher performance, with AUC ROC of 0.89, compared to the best single feature with AUC ROC of 0.84.

Furthermore, we compare the model based on the features proposed by Krumm et al. [28] to models based on richer sets of features, see Table 4. The NearMe model introduced by Krumm is based on four features and its performance is not higher than using a single feature: Jaccard coefficient. Our Simple model is based on features that do not require long term data collection and are not specific to our deployment. It performs better than any single feature and than the NearMe model. Enhancing the model with the information on popularity (the General model) further improves the performance. Finally, using all features, including timing and location (which might be specific to this experiment as they depend on our campus as location and the time schedule typical for students), allows us to build the best performing model.

		AUC ROC	
category	feature	training	test
AP presence	overlap	0.77	0.77
	jaccard	0.84	0.84
	union	0.53	0.53
	non-overlap	0.74	0.74
	combined	0.85	0.85
RSSI	spearman	0.66	0.68
	pearson	0.67	0.68
	manhattan	0.60	0.60
	euclidean	0.59	0.59
	combined	0.72	0.76
Location	at DTU	0.61	0.61
	at home	0.64	0.64
	combined	0.65	0.65
Presence + RSSI	top AP	0.60	0.60
	top±6	0.75	0.74
	combined	0.75	0.75
Timing	time of week	0.51	0.51
Popularity	min	0.54	0.54
	max	0.59	0.59
	adamic_adar	0.77	0.77
	combined	0.73	0.73

Table 3: Performance of single features and feature categories in the task of inferring close proximity interactions. We learn the optimal thresholds for each feature (set) based on randomly selected .5 million samples out of the total of 116 million. While training the classifiers we employ a 5-fold cross validation. Note that the performance, measured using area under Receiver Operating Curve (AUC ROC), remains equally high in the test period as in the training period. Jaccard similarity between lists of routers seen by the two devices is the best performing single feature.

5. INFERRING SOCIAL TIES

Next we focus on inferring infer social network ties. We create a set of features describing the properties and the dynamics of the inferred physical interactions between participants to infer the links of the underlying social network. We exclusively use the interactions inferred from the WiFi data, omitting the 0.5 million samples used for training. We aggregate interactions across one month windows, thus creating static friendship networks for each month of the data.

5.1 Data availability

Figure 2 summarizes the properties of the dataset showing the number of users (a), dyads (b), and the density of all five networks (c): face to face, facebook friendship graph, facebook interaction graph, sms, and call. Note in Figure 2a. that while there are behavioral (WiFi) data for significantly more users than the ground truth data. The decline in number of Facebook data is caused by expiring Facebook authorsations, which the users did not renew. The relatively low number of users with calls and short messages is not caused by the users failing to collect this data: it shows that these forms of communication are not popular with all the participants of the experiment. Figure 2b emphasizes that the sets of interactions in different communication channels are highly imbalanced: there are an order of magnitude more links in the facebook friendship network compared to call/sms and facebook interaction network, and an order of magnitude more interactions face to face compared to facebook friend links. As shown in Figure 2c contacts through communication networks are not driven by the academic schedule as much as face to face interactions: the fraction of active dyads does not decrease during the summer time, as opposed to the face to face network.

5.2 Method

We use different social networks as the ground truth for social ties: the Facebook graph of friendships, the Facebook graph of comments (there is a link between user A and user B if they comment on each other's wall content), the network of calls, and the network of text messages. For simplicity we treat all networks as non-directional and assume that the links are reciprocated (although we acknowledge that in real-world networks the perceived friendships are not always reciprocal [3]). This way we show that inference of social ties from co-occurrences is robust with respect to the specific definition of a social tie and that it is, in fact, possible to capture subtle differences between Facebook, text, and call friends. We contextualize each inferred meeting by describing its social makeup, timing, and location. We construct features to capture long term characteristics of dyadic relations. For each dyad A, B we define 16 features grouped into the following categories (see Table 5 for an overview):

		AUC ROC	
featureset	features	training	test
NearMe	overlap, non-overlap, spearman, euclidean	0.84	0.83
Simple	AP presence, RSSI, Presence + RSSI	0.86	0.86
General	all but at DTU and Timing	0.88	0.88
Full	all	0.89	0.89

Table 4: Performance of feature sets in the task of inferring close proximity interactions. We train a RandomForest classifier on selected subsets of features: NearMe [28], Simple (no features that are specific to this experiment or require longer term data collection), General (without features that could be specific to this experiment), and Full (all listed features). Even the Simple model performs better than NearMe. Using features which could be specific to the experiment marginally improves performance.

Time spent together in various contexts. After reducing the time resolution to one minute, we compute **total time together** in minutes. Using the presence or lack of routers with the network name (SSID) of `dtu`, which is the university network, we describe the location of each meeting and calculate the time **on campus** and **outside of campus**. Additionally, assuming that the access point observed the most is one's home router, we compute the time the two people spent together **at home** of one of the individuals. Following the intuition that friends meet in smaller groups [63], we weigh each meeting by the size of the union of people A and B met within 300 seconds of the meeting and then sum the co-occurrences **weighted by the number of people** present. In a similar way, we weigh each meeting by the number of access points as a proxy of a population density [49] and then sum the co-occurrences **weighted by the number of APs**.

Regularity of meetings. In a university setting, many people meet during classes without necessarily forming social ties. To distinguish between organic and on-schedule meetings, we calculate **entropy of hour of the day**, **entropy of the day of the week**, **entropy of the hour of the week** of meetings. This way dyads who meet only on particular days and during work hours have a low entropy, whereas dyads meeting irregularly have higher entropy scores. Additionally, we compute the **mean-** and **median time between meetings** of each dyad: we expect friends to meet more often, even if for a short time, and non-friends to meet rarely, but possibly for longer periods at a time (e.g. at classes lasting up to 4 hours each). Furhtermore, we report the **entropy of locations** in which the meetings take place; intuitively, if two people always meet in the same location they are less likely to be friends than if they are seen together in different places. Using entropy of time and location of meetings have been previously described as the Entropy Based Model [45].

Network similarity. It is commonly assumed that social relations tend to be transitive: if A is friends with C and C is friends with B, then A is likely to be friends with B.

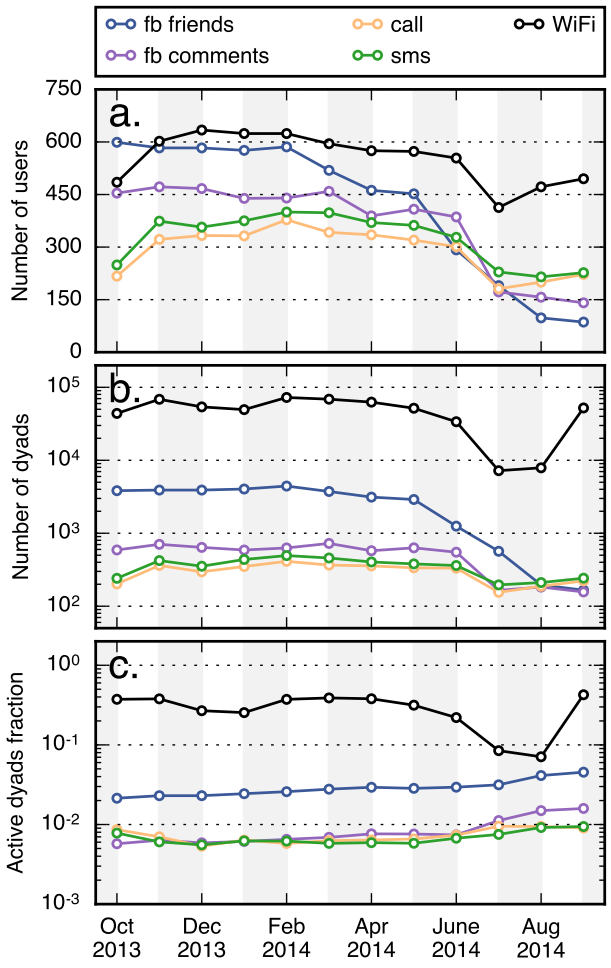

Figure 2: a. The number of participants with particular sources of data as a function of time. The fact that the users did not renew the authorization to access their Facebook data causes the decline of availability. The low but stable number of people with call and sms data shows that not everybody uses the traditional communication channels. b. The number of dyads with interactions in each of the data sources. This shows the severity of the class imbalance problem: in the peak month of February 2014 there are 72570 dyads interacting on WiFi but only 412 that exchanged phone calls. c. The fraction of dyads who interacted among all possible dyads. Because all participants in the study are students at the same university, as many as 37% of possible dyads actually interact in the offline world. At the same time only 0.6% of dyads between people who use the call functionality call each other. Note that while the fraction of active face-to-face dyads drops in the vacation period (July-August 2014), such behavior is not seen in the other communication channels.

We measure the Jaccard similarity of top contacts between A and B, assuming that the more similar their top contacts are, the more likely A and B are to be friends. We use the values of **overlap among top 5 contacts**, **over-**

category	features
Time spent together	total time together, on campus, outside of campus, at home, weighted by the number of people, weighted by the number of APs
Regularity	entropy of hour of the day, entropy of the day of the week, entropy of the hour of the week, mean time between meetings, median time between meetings, entropy of locations
Network similarity	overlap among top 5 contacts, overlap among top 15 contacts, overlap among top 25 contacts, overlap among top 50 contacts

Table 5: Features used to infer social ties. Each feature has three variants: total, in-role (considering only interactions during working hours on weekdays), and extra-role (considering only interactions outside of working hours and on weekends).

lap among top 15 contacts, **overlap among top 25 contacts**, and **overlap among top 50 contacts** (we note that extending the search beyond 50 top contacts does not increase the performance of the models). The network similarity has been previously exploited in the problem of link prediction for example in [11].

Each of the 16 features has three variants: total, in-role (considering only interactions during working hours on weekdays), and extra-role (considering only interactions outside of working hours and on weekends). Previous research showed that the distinction is crucial and that extra-role interactions are more indicative of friendship [15]. We train a random forests classifier for each month of the data using the four networks (Facebook friendship, facebook interactions, call, sms) as ground truth.

5.3 Results

We find that all four networks—Facebook friendships, facebook interactions, sms, and call—can be reliably inferred from the close proximity events collected over a year. Our data suggests that there are different levels of psychological propinquity necessary for an edge to exist in each of the four networks. The population we study is better connected on Facebook than via sms and calls, which indicates a lower threshold for becoming *friends* online than maintaining telecommunication exchanges.

Furthermore, comparing the performance on the Facebook and the telecommunication networks, we find that edges in telecommunication networks can be inferred more reliably. This means that our physical interactions are reflected better in our communication networks than to the Facebook friendship graph. For example in the month of March 2014:

- **Facebook friends.** 73% of Facebook friends met at least once during the month, including 43% who met outside of campus; 35% spent at least one hour together.
- **Facebook friends, who actively interact online.** 89% of Facebook friends who interact on Facebook met at least once during the month, including 64% who met outside of campus; 58% spent at least one hour together.
- **SMS friends.** 94% of sms contacts met at least once, including 80% who met outside of campus; 68% spent at least one hour together.
- **Call friends.** 93% of call contacts met at least once, including 80% who met outside of campus; 71% spent at least one hour together.

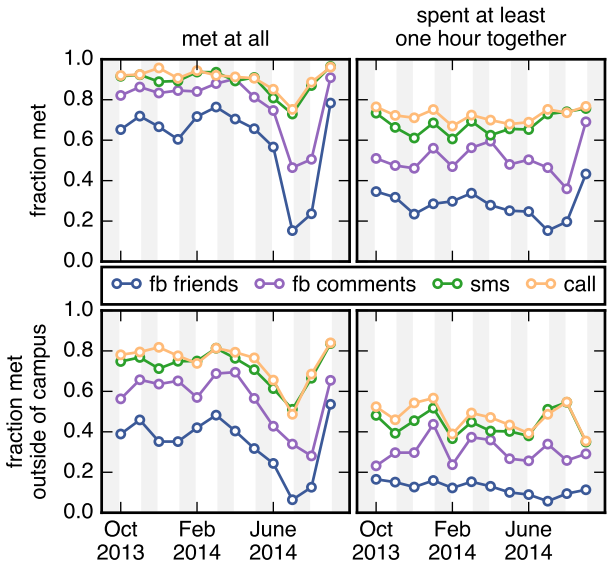

Figure 3: A vast majority of people who contact each other via phone/sms also meet in the real world, often outside of campus and for longer periods. People who interact on Facebook do too, but only a smaller subset of them.

The results for each month are presented in Figure 3, revealing that the tendency holds during other months as well, with higher fraction of telecommunication contacts meeting on and outside of campus, and spending longer time together.

Given the 16 features, we infer the four kinds of friendship among them. We perform a five-fold cross validation procedure with random forests classifier and report the mean area under ROC and MCC scores for each of the models in each prediction task in Figure 5a and c. Additionally, we perform an analog procedure only investigating links among people studying the same majors. The results reported in Figure 5b and d indicate that there is still a strong signal of friendship even among people who are "forced" to spend multiple hours per day together. Finally, we present the importance of each feature as estimated by the Random Forest Classifier in Figure 4. We evaluate the predictive performance of the models using the area under Receiver Operating Curve (AUC ROC). The value can be interpreted as the answer to the question: given a pair of friends and a pair of non-friends

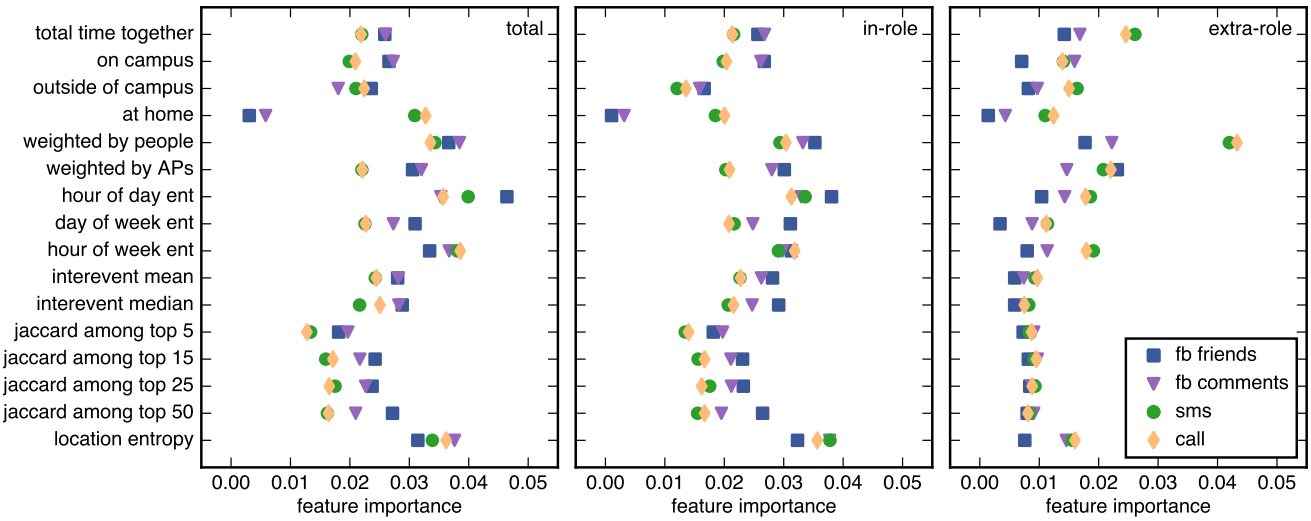

Figure 4: Relative importance of features in predicting the three kinds of links. The most important feature for predicting call/sms networks it is the extra-role time weighted by the number of people present. In-role interactions are more important for inferring facebook links than extra-role. Time at home is consistently the least important feature for inferring facebook ties. The presented values are median importances of 10 runs of five-fold cross validation training of a Random Forest Classifier.

what fraction of times is the model able to tell which are friends. Based on the reported AUC ROC scores it is worth noting that the difference between a pair Facebook friends and a pair of average non-friends is less pronounced in the behavioral data that it is the case with interaction (facebook comments, call, sms) networks, see Figure 5. This finding is consistent with previous research on strength of Facebook ties [67, 25].

To better understand the driving forces behind each kind of friendship we investigated, we plot the cumulative distribution of values of each feature for the different relationships in Figure 7. We find that across all aspects, people who call or text each other (yellow and green lines) express a stronger tie in physical space than other types of friends, and non-friends (black line). Figure 7 shows that they tend to spend most time together (e), both at campus (f) and outside (g), and at each others' homes (h). We also notice that they tend to meet with fewer people present (i) and in places with a lower population density (j). Their interevent time is the lowest (k, l), which means that they not only spend more time together, but also meet more often. Higher entropy values (m-o) indicate that the timing of their meetings has less of a scheduled character than it is the case with other relationships. We also observe that call friends also have most similar friends. Behavioral signatures of Facebook friends (blue line) are somewhere between those of calling friends and people who are not friends on Facebook.

Finally, we investigate whether extending the observation time can help in the inference task. We build a matrix in which each row is a dyad and each column is a month. Each cell of this matrix contains the probability that the corresponding dyad exchanged phone calls in the corresponding month as estimated by the month's random forest classifier. Then, for each month, we build a Logistic Regression model which uses these probabilities for the month and N months before (Figure 6a) or after (Figure 6b). Figure 6 reports the

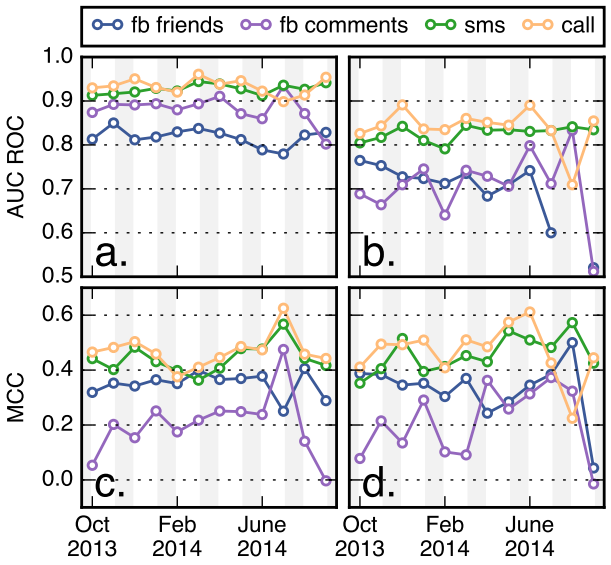

Figure 5: Results of inferring friendship networks among all students (a. and c.) and students who study the same major (b. and d.). Evidently it is possible to infer friendships even among people who, beacuse of the class schedule, spend multiple hours per day together.

increase in area under receiver operator curve introduced by exploiting more than one month of data. Using past or future data increases the performance of the inference for all months.

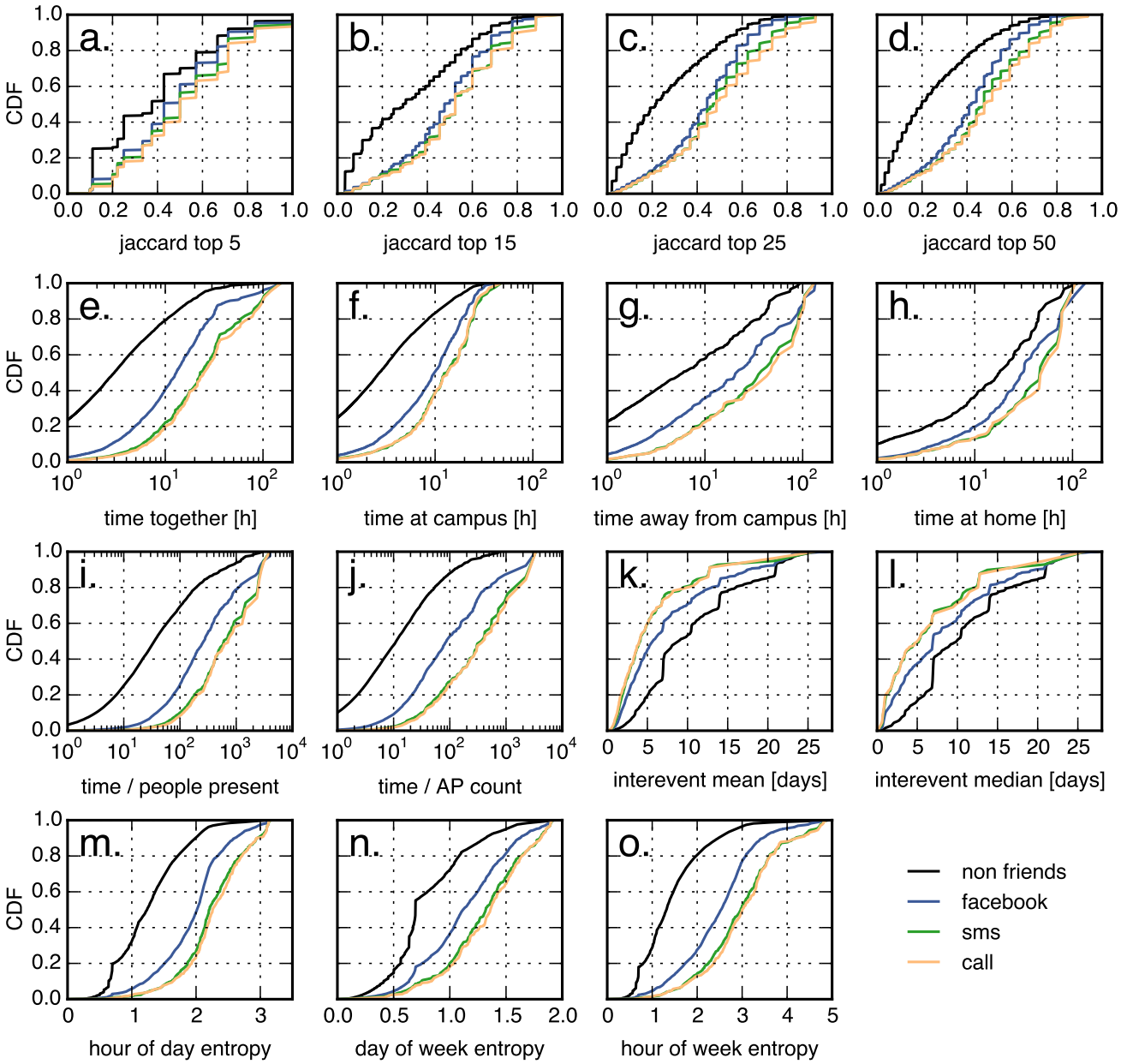

Figure 7: Compared to Facebook friends (blue line) and Facebook non-friends (black line) people who call (yellow line) or message (green line) each other spend more time together (e-h) especially with only few others around (i, j), meet more often (k, l), and irregularly (m-o). Their share closest contacts (a-d). Behavioral signatures of Facebook friends lie between those of calling friends and people who are not friends on Facebook. The ties of people who call each other appear slightly more pronounced in face to face meetings than those who message each other.

6. RELATED WORK

In this section we discuss related work that explores the application of mobile data to deepen our understanding of aspects relevant to this paper.

Smartphones as social sensors. Until recently, social scientists did not have access to methods for observing behavioral patterns and interactions of large populations. Instead, they relied on information reported by subjects through surveys. This method of collecting data enables the researchers to ask in-depth, personal questions possibly leading to thorough understanding of some aspects of the investigated individuals' lives. However, questionnaires suffer from a number of shortcomings, some of which cannot be resolved without a complete paradigm shift. Questionnaire data can be biased by low temporal resolution of the

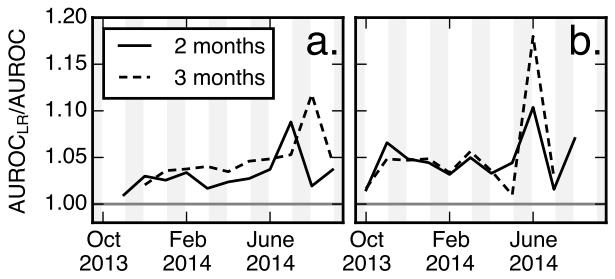

Figure 6: Adding information from previous (a.) and future (b.) months increases the performance of predicting links (here: calls) throughout the year.

responses as well as the limited capacity of subjects to objectively recall and report past events. For example, it has been shown that people are typically unable to order their social contacts by the frequency of interactions [5] and are biased towards more recent events [15]. Moreover, because of the cost and organizational complexity, surveys are not typically administered to large-scale cohorts.

On the other side of the spectrum big data approaches, as for example using call detail records (CDR) or public online activity data, allieviate some of these problems. They enable investigating populations orders of magnitude larger, with studies reporting between 2.5 million [30] and 25 million [70] participants in mobile phone experiments and as many as 41 million in a study of Twitter [29]. Furthermore, these data contain 'objective' events, not their subjective perceptions, and its time resolution is not, in principle, limited. However, such datasets only offer access to only a thin slice of human activity and, in practice, still suffer from non-contiguity, with samples only created when a subject performs a certain activity.

Using smarphones as sensors offers a combination of the best of the two worlds — the richness of survey information, and the scale of big data studies — to ultimately solve the problem of low time resolution, by producing data independently from user input. Smartphones can capture a multitude of channels of information, from activity and mobility of individuals, encounters in the physical space, to communication events across a number of communication channels. Moreover, smartphones are becoming ever more widespread with 60% market penetration in the developed world, ranging from 51% in Europe to 70% in North America at the end of 2014 [24].

The use of smartphones in social observation was pioneered by Alex 'Sandy' Pentland and his collaborators in the seminal Reality Mining experiment [14]. In that work, Bluetooth was investigated for the first time as a proxy for face-to-face interactions and used to measure encounters among 100 students. Other deployments based on smartphones for data collection include, for example, the:

- Lausanne data collection campaign [26], with 12 months and 170 participants.
- Friends and Family study [2], with 15 months of data about 130 individuals. The experiment was designed to investigate the decision making and social influence.
- Social Evolution [36], with eight months of data about 60 individuals. The experiment was designed to study the adoption of political opinions and habits, spread of diseases, and the formation of social ties.
- NetSense [60], with two years of data about 200 individuals.
- Phone-Lab [40], a continouosly developing deployment, initially with 288 devices, focused more on using smartphones for measuring telecommunication infrastructure than on social networks.
- StudentLife Study [64], with 10 weeks of data about 48 participants
- Copenhagen Networks Study [59], more than two years of data with 850 participants, with a focus interaction dynamics offline

To our best knowledge none of the deployments to date have followed as many individuals and for as long as the Copenhagen Networks Study. Furthermore, we note that only the NetSense and StudentLife projects features data rich enough to investigate the questions posted in this article. Reality Mining, Social Evolution, and Friends and Family datasets do not report WiFi scan results. The Lausanne data collection is limited in its social aspect, with the participants spending only a small fraction of time with each other, thus making the friendship inference much simpler than it is in the case of Copenhagen Networks Study. The Phone-Lab dataset does not feature ground truth information on the social ties between the participants.

Location and mobility. CDR data has been used as a proxy for human mobility at large, societal scale. It has been shown that our movements are regular [19], stable [35], and predictable [54]. It yet remains to be verified whether these findigs hold fully if the analysis were to be performed on data with higher spatial and temporal resolution (such as WiFi data). At smaller scales, the scientific community inverstigated the potential of WiFi routers in applications of indoor [4, 21, 46] and outdoor [8, 39, 16, 23] localization. Our recent work investigates how large companies can crowd source the creation of databases with router locations [48, 39, 16] and how people's mobility on societal scale can be described using only a small subset of available routers [49]. WiFi signals can also be analyzed to discover places of interest and stop locations in an unsupervised manner, i.e. without explicit location information as reference [41, 68].

It is important to stress that the work presented in this article does not rely on any kind of location estimation but instead on relative comparison between the environments sensed by two parties.

Interactions. Complementary to mobility, the question of social interactions has been recently considered in various contexts, with the results indicating that collection of high-resolution behavioral traces is instrumental for understanding of complex processes in society [14, 53, 58, 56]. However, from a technical point of view, collection of such data remains a challenge.

The most popular methods for quantitative and scalable collection of close-proximity interactions include using specialized hardware (e.g. sociometric badges) [42, 47] or Bluetooth enabled smartphones [14, 2, 59]. In case of badges, interactions are usually inferred using radio-frequency identi-

fication (RFID) transmissions or infrared. This way, badges worn around participants' necks can usually sense not just proximity but also whether individuals are facing each other, resulting in recordings of face-to-face interactions. Sensing performed Bluetooth-enabled mobile phones is less granular, as only proximity but not orientation of the individuals can be sensed. Moreover, it requires the subjects' devices to remain in Bluetooth-discoverable state, which raises a number of security and privacy concerns, as described in the Introduction. There has been some developments in substituting Bluetooth with WiFi, an approach in which one of the phones acts as a hotspot and is sensed by others [6]. In controlled test environments this approach appears to offer a distance estimation resolution of 0.5m [43], providing a better understanding of the nature of the contacts [22]. However, the claim has not been tested in the wild and the method potentially introduces even more privacy and security problems than Bluetooth.

An alternative way of sensing interactions between two persons with smartphones relies on comparing the two devices' radio frequency perceptions of the envorinment. If a similarity is above a certain threshold, the two devices are assumed to be in physical proximity. The idea of comparing WiFi signals to measure proximity was initially explored more than a decade ago. Initially, researchers relied on single-feature measures of similarity, such as Manhattan discance [38] or overlap [37]. On the other hand, in the NearMe project [28] four features are used in conjunction to compare the lists of seen access points and estimate the distance between two persons: (1) the number of overlapping APs, (2) the Spearman rank correlation coefficient between the list of APs sorted by signal strength, (3) sum of squared differences in signal strength (Euclidean distance) (4) the numer of non overlapping APs. The authors found that using more than one feature does not further increase the accuracy of distance estimation and that the estimation error grows if previously unseen data are used for testing. in addition to short-distance proximity where the two devices must share at least one access point, introduces the notion of long-distance proximity. The long-distance proximity can be sensed using a precomputed network of proximity between access points. For example, when AP_1 and AP_2 are often seen by users in a single scan, they are defined as proximate. Then, when user A senses AP_1 and user B senses AP_2, they are considered to be in long-distance proximity. The distance between the two users can be estimated using average travel times between two routers.

Kjærgaard and Nurmi offer a comprehensive overview of factors which make the proximity sensing using WiFi difficult [27]. Among the key challenges they name body attenuation, the differences in sensing hardware, and the multitude of environments where the sensing takes place. They show how two features — Euclidean distance in received signal strength space and Jaccard coefficient — depend on the environment. In our our study all the participants used the same hardware (LG Nexus 4). We also note that the differences in environments instead of being directly addressed technologically, can actually be used to increase the performance of the model by exploiting the characteristics of human interactions: from a technical standpoint, environments with a smaller number of routers offer lower accuracy of distance estimation; however, two people in an environment with fewer access points are more likely to be actually interacting (see Figure 1).

Inferring social ties from co-presence events. Eagle et. al in their seminal work [15] were the first to explore the relationship between self-reported social ties and behavioral data collected through smartphones. They found that, while people fail to estimate the time they spend with others accurately, there are certain behaviors more indicative of friendship than just the total time spent together. Their analysis revealed a stronger correlation between the reported friendships and *extra-role* (off campus, off hours) than *in-role* meetings ($\rho = 0.35$ and $\rho = 0.08$ respectively). While Eagle focused on the co-locations of dyads with a limited view on the spatio-temporal context, Crawnshaw *et al.* . extended the apporach to include location context beyond the simple on/off campus indication [11]. By tracking the social composition of each location of interaction, they showed that meetings at less popular locations are a stronger signal of online friendship (using Facebook friendship links as ground truth). They also found that the more unpredictable the meeting schedule of a pair is (in terms of temporal, spatial, and spatiotemporal entropy) the higher their probability of being friends. Furthermore, they exploited transitivity inherent in the social networks to show that the similarity of neighborhoods can further aid the inference. They found that a model incorporating the newly introduced features far outperforms the simple approach of Eagle *et al.*

Using smartphones for social sensing offers unparalleled resolution of the data but is currently limited with respect to the number of participants. However, behavioral data at lower resolutions are available for significantly higher populations. Wang *et al.* [63] have shown that even co-locations inferred from crude CDR data can be used to infer social ties. The behavioral features they proposed (including the co-location rate weighted by the number of people present) have only marginally lower predictive power than the features based on network similarity between two poeple. As presented by Li *et al.* using trajectories instead of separate co-location events might further increase the performance of link inference [33].

In parallel to these development, reseachers have also worked on the link prediction problem in settings where the continous behavioral data is lacking. There, instead of being tracked in the background the subjects explicitely *check-in* at different locations at will. Crandall *et al.* investigated the relationship between the number of unique locations visited by two people and the probability of them being friends in a photo-sharing service [10]. Instead of defining the co-location as a simultaneous presence of two individuals in a very confined location, they allowed for visiting the location at different times (from one day to one year) and the location could be of arbitrary size (ranging from 80x80 meters to 800 x 800 kilometers). They found a surprisingly sharp growth of the probability of the friendship link with the growing number of unique locations visited. Perhaps surprisingly, lowering of the spatial resolution does not inherently introduce noise. Since the authors count only the unique locations, the authors require that with a lower resolution the meetings take place further apart; multiple encounters in distant locations are less likely to be co-incidental and are thus a strong signal of friendship. The data Crandall investigated has very few observations per user and only a very small fraction of friends (1%) "meet" at three different

locations on the same day. Scellato et al. extended this approach by introducing additional, inferred properties of locations shared among two people, such as the social entropy. Other works showed that probability of friendship decreases with growing geographical distance [34] and that clusters of friends tend to live nearby [51].

Pham *et al.* [45] suggest that the above mentioned methods might suffer from a high influence of coincidental meetings. In order to remedy the problem, they introduce the Entropy-Based Model (EBM). The model places ephasis on the diversity of locations in which the users meet as well as meetings in less popular places. The authors claim to deal with data sparseness better than other models.

There have also been developments into coupling the social and the mobility data beyond the task of link prediction. Intuitively, since maintaining a bond requires physical proximity, some of people's mobility is driven by social factors. Several works point to the realization that many non-routine travels observed in real data can be attributed to individuals seeking interaction with their social contacts [20, 61, 9].

Communication networks as proxies for real-world relationships. In this article we rely on Facebook friendship networks, and the links in the telecommunication network as proxies for the existence of social ties. It is therefore important to report on other research describing the applicability of such data in this context.

Wiese *et al.* [66] compared phone networks and self-reported friendships of 40 subjects. They found that while frequent communication indicates strong ties, lack of communication does not necessarily indicate a weak tie. Among other contributing factors they list the realization that people use multiple channels of communication (including face to face) and their phone networks do not fully describe their social networks. This might indicate that many of dyads who our models misclassifies as call friends, are friends indeed, but use different communication means. Furthermore, in the work we have already mentioned, Eagle *et al.* [15] found that the single feature most correlated with self-reported friendship is the phone communication between two people.

While studying users of Facebook Wilson *et al.* [67] found that only a fraction of links present in the social graph represent dyads which interact actively on Facebook. They recommend using the interaction graph instead of the declared friendships to better model the underlying social network. Wilson's work is further confirmed by Jones' research on inferring self-reported friendship ties from online interaction data [25]. They found that the strength of tie is correlated with the intensity of contact on Facebook, especially with commenting each other's wall content. Furthermore, they found that private messages, to which we do not have access in this study, do not constitute a better indicator of real word friendship than wall posts.

Given the research described in this section, we believe that the exitence of phone communication links can be treated as friendship signal among our population. We further confirm the findings from Wilson and Jones showing that Facebook interaction networks are more predictable from offline behavioral data than the Facebook friendship graph.

7. DISCUSSION

7.1 Privacy implications

WiFi scans routinely performed by modern mobile devices are not considered to be sensitive by a general public. As we have previously shown, WiFi can be efficiently used for high-resolution mobility tracking of entire populations [48, 49]. Here we go a step further and show that WiFi signals can also be used to infer who people interact with, not only where they are. Using machine learning models, both individual interactions and complex social networks can be inferred, even in densely-connected social systems. Thus, results of WiFi scans—collected by major manufacturers of mobile devices and available to majority of mobile application developers—constitute very sensitive datasets. For example, a vast majority of the applications available in Google Play Store has access to WiFi information, including all the scan results requested by the system as often as every minute [49]. While better permission models—requiring, for example, location permission to access WiFi scans results—are considered and implemented for the next mobile operating system releases, WiFi signals remain a major privacy risk for years to come. Such WiFi datasets deserve more attention from privacy advocates, mobile operating system and application developers, and general public at large.

7.2 Limitations of the WiFi-based social inference

While the presented approach to inference of social interactions and ties using WiFi scans offers an important new method in computational social science, here we want to recognize its limitations. WiFi-based inference of physical proximity requires data from both individuals for matching, as opposed to, for example, Bluetooth- or RFID-based approaches where data from just one device is generally sufficient (and additional data from both devices allows for redundancy and imputation of missing values). As a consequence dataset containing all the sensed access points for all the individuals can be significantly larger when compared to physical interactions data collected using Bluetooth- or RFID-based approaches.

The inference in the presented approach depends on the beacons being present in the environment—here WiFi access points. While today WiFi networks are omnipresent, especially in densely-populated areas [49], we find that in our longitudinal and diverse dataset approximately 5% of the WiFi scans did not report any nearby networks, preventing inference of physical proximity.

In this presented study, all phones collecting data were of the same make and model. When considering a broader application of the method, differences in WiFi hardware transmitters and firmware and software of the phones may result in less consistent scans data, making it more difficult to devise a robust model as the one presented here.

Finally, we should note that it is not our argument that the learnt model for discovering particular interactions and reconstructing the overall social network is generally applicable to different populations. Depending on the specific population and social context under consideration, the weights in the model might be different or even entirely new features might be useful. Our results however strongly indicate that physical proximity can be inferred in a feasible fashion using WiFi signals collected by smartphones, even if extremely densely-connected populations.

7.3 Conclusion

In this work we showed how seemingly innocuous WiFi scan results can reveal a great deal about our daily interactions with others and our social ties. By a clever use of behavioral traces, put in context through meta information and our basic understanding of the inner working of social systems, we can transform a noisy data source draw further insights about the social structure. Our findings have important privacy implications, especially given our previous demonstration about using WiFi signals for tracking mobility. On the other hand, they also constitute a great opportunity for companies with access to such data on a global, to contribute better epidemic and dissemination models built on proximity data of billions of people.

Acknowledgements

The authors would like to thank Andrea Cuttone for useful discussions as well as Urvashi Khandelwal and Jana Huisman for the important feedback.

8. REFERENCES

[1] L. A. Adamic and E. Adar. Friends and neighbors on the web. *Social networks*, 25(3):211–230, 2003.

[2] N. Aharony, W. Pan, C. Ip, I. Khayal, and A. Pentland. Social fmri: Investigating and shaping social mechanisms in the real world. *Pervasive and Mobile Computing*, 7(6):643–659, 2011.

[3] A. Almaatouq, L. Radaelli, A. Pentland, and E. Shmueli. Are you your friends? friend? poor perception of friendship ties limits the ability to promote behavioral change. *PLoS ONE*, 11(3):1–13, 03 2016.

[4] P. Bahl and V. N. Padmanabhan. Radar: An in-building rf-based user location and tracking system. In *INFOCOM 2000. Nineteenth Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings. IEEE*, volume 2, pages 775–784. Ieee, 2000.

[5] H. R. Bernard, P. D. Killworth, and L. Sailer. Informant accuracy in social network data IV: a comparison of clique-level structure in behavioral and cognitive network data. *Social Networks*, 2(3):191–218, Jan. 1979.

[6] I. Carreras, A. Matic, P. Saar, and V. Osmani. Comm2sense: Detecting proximity through smartphones. In *Pervasive Computing and Communications Workshops (PERCOM Workshops), 2012 IEEE International Conference on*, pages 253–258. IEEE, 2012.

[7] D. Chaffin, R. Heidl, J. R. Hollenbeck, M. Howe, A. Yu, C. Voorhees, and R. Calantone. The promise and perils of wearable sensors in organizational research. *Organizational Research Methods*, page 1094428115617004, 2015.

[8] Y.-C. Cheng, Y. Chawathe, A. LaMarca, and J. Krumm. Accuracy characterization for metropolitan-scale wi-fi localization. In *Proceedings of the 3rd International Conference on Mobile Systems, Applications, and Services*, MobiSys '05, pages 233–245, New York, NY, USA, 2005. ACM.

[9] E. Cho, S. A. Myers, and J. Leskovec. 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*, pages 1082–1090. ACM, 2011.

[10] D. J. Crandall, L. Backstrom, D. Cosley, S. Suri, D. Huttenlocher, and J. Kleinberg. Inferring social ties from geographic coincidences. *Proceedings of the National Academy of Sciences*, 107(52):22436–22441, 2010.

[11] J. Cranshaw, E. Toch, J. Hong, A. Kittur, and N. Sadeh. Bridging the gap between physical location and online social networks. In *Proceedings of the 12th ACM international conference on Ubiquitous computing*, pages 119–128. ACM, 2010.

[12] S. Datoo. High street shops are studying shopper behaviour by tracking their smartphones or movement. http://goo.gl/vGg8k8.

[13] Y.-A. de Montjoye, A. Stopczynski, E. Shmueli, A. Pentland, and S. Lehmann. The strength of the strongest ties in collaborative problem solving. *Scientific reports*, 4, 2014.

[14] N. Eagle and A. Pentland. Reality mining: sensing complex social systems. *Personal and ubiquitous computing*, 10(4):255–268, 2006.

[15] N. Eagle, A. S. Pentland, and D. Lazer. Inferring friendship network structure by using mobile phone data. *Proceedings of the National Academy of Sciences*, 106(36):15274–15278, 2009.

[16] A. Eustace. Wifi data collection: An update. http://goo.gl/VFJ9mM.

[17] R. Friedman, A. Kogan, and Y. Krivolapov. On power and throughput tradeoffs of wifi and bluetooth in smartphones. *Mobile Computing, IEEE Transactions on*, 12(7):1363–1376, July 2013.

[18] R. C. Gatej. An adaptive approach to mobile sampling. Master's thesis, Technical University of Denmark, Department of Applied Mathematics and Computer Science / DTU Co, Matematiktorvet, Building 303B, DK-2800 Kgs. Lyngby, Denmark, compute@compute.dtu.dk, 2013. DTU supervisors: Sune Lehmann, Jakob Eg Larsen, DTU Compute.

[19] M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabasi. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, 2008.

[20] P. A. Grabowicz, J. J. Ramasco, B. Gonçalves, and V. M. Eguíluz. Entangling mobility and interactions in social media. *PLoS One*, 9(3):e92196, 2014.

[21] A. Haeberlen, E. Flannery, A. M. Ladd, A. Rudys, D. S. Wallach, and L. E. Kavraki. Practical robust localization over large-scale 802.11 wireless networks. In *Proceedings of the 10th Annual International Conference on Mobile Computing and Networking*, MobiCom '04, pages 70–84, New York, NY, USA, 2004. ACM.

[22] E. T. Hall. The hidden dimension . 1966.

[23] D. Han, D. G. Andersen, M. Kaminsky, K. Papagiannaki, and S. Seshan. Access point localization using local signal strength gradient. In *Passive and Active Network Measurement*, pages 99–108. Springer, 2009.

[24] G. Intelligence. The mobile economy 2015. http://goo.gl/5MCrxK.

[25] J. J. Jones, J. E. Settle, R. M. Bond, C. J. Fariss, C. Marlow, and J. H. Fowler. Inferring tie strength from online directed behavior. *PloS one*, 8(1):e52168, 2013.

[26] N. Kiukkonen, J. Blom, O. Dousse, D. Gatica-Perez, and J. Laurila. Towards rich mobile phone datasets: Lausanne data collection campaign. *Proc. ICPS, Berlin*, 2010.

[27] M. B. Kjærgaard and P. Nurmi. Challenges for social sensing using wifi signals. In *Proceedings of the 1st ACM workshop on Mobile systems for computational social science*, pages 17–21. ACM, 2012.

[28] J. Krumm and K. Hinckley. The nearme wireless proximity server. In *UbiComp 2004: Ubiquitous Computing*, pages 283–300. Springer, 2004.

[29] H. Kwak, C. Lee, H. Park, and S. Moon. What is twitter, a social network or a news media? In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 591–600, New York, NY, USA, 2010. ACM.

[30] R. Lambiotte, V. D. Blondel, C. De Kerchove, E. Huens, C. Prieur, Z. Smoreda, and P. Van Dooren. Geographical dispersal of mobile communication networks. *Physica A: Statistical Mechanics and its Applications*, 387(21):5317–5325, 2008.

[31] J. E. Larsen, P. Sapiezynski, A. Stopczynski, M. Mørup, and R. Theodorsen. Crowds, bluetooth, and rock'n'roll: Understanding music festival participant behavior. In

Proceedings of the 1st ACM International Workshop on Personal Data Meets Distributed Multimedia, PDM '13, pages 11–18, New York, NY, USA, 2013. ACM.

[32] D. Lazer, A. S. Pentland, L. Adamic, S. Aral, A. L. Barabasi, D. Brewer, N. Christakis, N. Contractor, J. Fowler, M. Gutmann, et al. Life in the network: the coming age of computational social science. *Science (New York, NY)*, 323(5915):721, 2009.

[33] Q. Li, Y. Zheng, X. Xie, Y. Chen, W. Liu, and W.-Y. Ma. Mining user similarity based on location history. In *Proceedings of the 16th ACM SIGSPATIAL international conference on Advances in geographic information systems*, page 34. ACM, 2008.

[34] D. Liben-Nowell, J. Novak, R. Kumar, P. Raghavan, and A. Tomkins. Geographic routing in social networks. *Proceedings of the National Academy of Sciences of the United States of America*, 102(33):11623–11628, 2005.

[35] X. Lu, L. Bengtsson, and P. Holme. Predictability of population displacement after the 2010 haiti earthquake. *Proceedings of the National Academy of Sciences*, 2012.

[36] A. Madan, M. Cebrian, S. Moturu, K. Farrahi, et al. Sensing the" health state" of a community. *IEEE Pervasive Computing*, (4):36–45, 2012.

[37] M. McNett and G. M. Voelker. Access and mobility of wireless pda users. *SIGMOBILE Mob. Comput. Commun. Rev.*, 9(2):40–55, Apr. 2005.

[38] J.-L. Meunier. Peer-to-peer determination of proximity using wireless network data. 2004.

[39] B. Meyerson. Aol introduces location plug-in for instant messaging so users can see where buddies are. http://goo.gl/2W1uYh.

[40] A. Nandugudi, A. Maiti, T. Ki, F. Bulut, M. Demirbas, T. Kosar, C. Qiao, S. Y. Ko, and G. Challen. Phonelab: A large programmable smartphone testbed. In *Proceedings of First International Workshop on Sensing and Big Data Mining*, pages 1–6. ACM, 2013.

[41] T.-B. Nguyen, T. Nguyen, W. Luo, S. Venkatesh, and D. Phung. Unsupervised inference of significant locations from wifi data for understanding human dynamics. In *Proceedings of the 13th International Conference on Mobile and Ubiquitous Multimedia*, pages 232–235. ACM, 2014.

[42] D. O. Olguín, B. N. Waber, T. Kim, A. Mohan, K. Ara, and A. Pentland. Sensible organizations: Technology and methodology for automatically measuring organizational behavior. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, 39(1):43–55, 2009.

[43] V. Osmani, I. Carreras, A. Matic, and P. Saar. An analysis of distance estimation to detect proximity in social interactions. *Journal of Ambient Intelligence and Humanized Computing*, 5(3):297–306, 2014.

[44] E. O'Neill, V. Kostakos, T. Kindberg, A. Penn, D. S. Fraser, T. Jones, et al. Instrumenting the city: Developing methods for observing and understanding the digital cityscape. In *UbiComp 2006: Ubiquitous Computing*, pages 315–332. Springer, 2006.

[45] H. Pham, C. Shahabi, and Y. Liu. Ebm: An entropy-based model to infer social strength from spatiotemporal data. In *Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data*, SIGMOD '13, pages 265–276, New York, NY, USA, 2013. ACM.

[46] N. B. Priyantha, A. Chakraborty, and H. Balakrishnan. The cricket location-support system. In *Proceedings of the 6th annual international conference on Mobile computing and networking*, pages 32–43. ACM, 2000.

[47] M. Salathé, M. Kazandjieva, J. W. Lee, P. Levis, M. W. Feldman, and J. H. Jones. A high-resolution human contact network for infectious disease transmission. *Proceedings of the National Academy of Sciences*, 107(51):22020–22025, 2010.

[48] P. Sapiezynski, R. Gatej, A. Mislove, and S. Lehmann. Oportunities and challenges in crowdsourced wardriving. In *Proceedings of the 15th ACM SIGCOMM conference on Internet measurement*. ACM, 2015.

[49] P. Sapiezynski, A. Stopczynski, R. Gatej, and S. Lehmann. Tracking human mobility using wifi signals. *PLoS ONE*, 10(7):e0130824, 07 2015.

[50] K. Scarfone and J. Padgette. Guide to bluetooth security. *NIST Special Publication*, 800:121, 2008.

[51] S. Scellato, C. Mascolo, M. Musolesi, and V. Latora. Distance matters: Geo-social metrics for online social networks. In *WOSN*, 2010.

[52] V. Sekara and S. Lehmann. The strength of friendship ties in proximity sensor data. *PloS one*, 9(7):e100915, 2014.

[53] V. Sekara, A. Stopczynski, and S. Lehmann. The fundamental structures of dynamic social networks. *arXiv preprint arXiv:1506.04704*, 2015.

[54] C. Song, Z. Qu, N. Blumm, and A.-L. Barabási. Limits of predictability in human mobility. *Science*, 327(5968):1018–1021, 2010.

[55] J. Stehlé, N. Voirin, A. Barrat, C. Cattuto, L. Isella, J.-F. Pinton, M. Quaggiotto, W. Van den Broeck, C. Régis, B. Lina, et al. High-resolution measurements of face-to-face contact patterns in a primary school. *PloS one*, 6(8):e23176, 2011.

[56] A. Stopczynski, A. S. Pentland, and S. Lehmann. Physical proximity and spreading in dynamic social networks. *arXiv preprint arXiv:1509.06530*, 2015.

[57] A. Stopczynski, R. Pietri, A. Pentland, D. Lazer, and S. Lehmann. Privacy in sensor-driven human data collection: A guide for practitioners. *CoRR*, abs/1403.5299, 2014.

[58] A. Stopczynski, P. Sapiezynski, S. Lehmann, et al. Temporal fidelity in dynamic social networks. *The European Physical Journal B*, 88(10):1–6, 2015.

[59] A. Stopczynski, V. Sekara, P. Sapiezynski, A. Cuttone, M. M. Madsen, J. E. Larsen, and S. Lehmann. Measuring large-scale social networks with high resolution. *PLoS ONE*, 9(4):e95978, 04 2014.

[60] A. Striegel, S. Liu, L. Meng, C. Poellabauer, D. Hachen, and O. Lizardo. Lessons learned from the netsense smartphone study. In *Proceedings of the 5th ACM Workshop on HotPlanet*, HotPlanet '13, pages 51–56, New York, NY, USA, 2013. ACM.

[61] J. L. Toole, C. Herrera-Yaqüe, C. M. Schneider, and M. C. González. Coupling human mobility and social ties. *Journal of The Royal Society Interface*, 12(105):20141128, 2015.

[62] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R. B. Altman. Missing value estimation methods for dna microarrays. *Bioinformatics*, 17(6):520–525, 2001.

[63] D. Wang, D. Pedreschi, C. Song, F. Giannotti, and A.-L. Barabasi. Human mobility, social ties, and link prediction. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1100–1108. ACM, 2011.

[64] R. Wang, F. Chen, Z. Chen, T. Li, G. Harari, S. Tignor, X. Zhou, D. Ben-Zeev, and A. T. Campbell. Studentlife: assessing mental health, academic performance and behavioral trends of college students using smartphones. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 3–14. ACM, 2014.

[65] D. J. Watts and P. S. Dodds. Influentials, networks, and public opinion formation. *Journal of consumer research*, 34(4):441–458, 2007.

[66] J. Wiese, J.-K. Min, J. I. Hong, and J. Zimmerman. 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*, pages 765–774. ACM, 2015.

[67] C. Wilson, B. Boe, A. Sala, K. P. Puttaswamy, and B. Y. Zhao. User interactions in social networks and their implications. In *Proceedings of the 4th ACM European*

conference on Computer systems, pages 205–218. Acm, 2009.

[68] D. K. Wind, P. Sapiezynski, M. A. Furman, and S. Lehmann. Inferring stop-locations from wifi. *PloS one*, 11(2):e0149105, 2016.

[69] F.-L. Wong and F. Stajano. Location privacy in bluetooth. In R. Molva, G. Tsudik, and D. Westhoff, editors, *Security and Privacy in Ad-hoc and Sensor Networks*, volume 3813 of *Lecture Notes in Computer Science*, pages 176–188. Springer Berlin Heidelberg, 2005.

[70] H. Zang and J. Bolot. Anonymization of location data does not work: A large-scale measurement study. In *Proceedings of the 17th annual international conference on Mobile computing and networking*, pages 145–156. ACM, 2011.