

Inferring Person-to-person Proximity Using WiFi Signals

Piotr Sapiezynski
Technical University of Denmark
pisa@dtu.dk

Arkadiusz Stopczynski
Technical University of Denmark
MIT Media Lab
arks@dtu.dk

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 offers important opportunities for sensing and recording a multitude of human behaviors. Human mobility patterns are a prominent example of such a behavior which has been studied based on cell phone towers, Bluetooth beacons, and 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 on a large scale is a technical challenge and many commonly used approaches—including RFID badges or Bluetooth scanning—offer only limited scalability. Here we show that it is possible, in a scalable and robust way, to accurately infer person-to-person physical proximity from the lists of WiFi access points measured by smartphones carried by the two individuals. Based on a longitudinal dataset of approximately 800 participants with ground-truth interactions collected over a year, we show that our model performs better than the current state-of-the-art. Our results demonstrate the value of WiFi signals in social sensing as well as potential threats to privacy that they imply.

Categories and Subject Descriptors

H.4 [Information Systems Applications]: Miscellaneous

Keywords

social sensing; wifi; proximity; interactions; social networks;

1. INTRODUCTION

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 [26]. Here we show how our interactions with other people can be inferred in a reliable and scalable way, using signals from WiFi access points.

Being able to infer person-to-person proximity events with high spatio-temporal resolution enables modeling of phenomena such as spreading of diseases and information [21], formation of social ties [12], as well as group dynamics [42].

Commercial applications vary from distributed ad hoc networking [27] to romantic matchmaking [10].

Despite the importance of understanding networks of close proximity interactions, there is a scarcity of scalable and efficient ways to obtain data for large populations. This is due to the fact that technology has only recently developed to the point, where collection of such high resolution data has become technologically feasible. The data sources used for investigating mobility of individuals, such as call detail records (CDRs) from mobile operators [16], are too coarse in terms of temporal and spatial resolution to allow inference of person-to-person proximity. On the other hand, the current state-of-the-art methods for measurement of physical proximity require using specialized hardware (*e.g.*, sociometric badges) [32, 37] or smartphones sensing each other through Bluetooth [11, 3, 48]. Specialized hardware adds cost and complexity to experimental deployments, effectively limiting their scale. Bluetooth scanning realized on participants’ mobile phones increases power consumption [14]—limiting temporal resolution that can be achieved—and requires the devices to be in Bluetooth *discoverable mode*. This requirement raises privacy [52] and security concerns [40]. When a phone is in discoverable mode the location of its owner can be tracked by third parties, a fact commonly used by researchers [25, 34], and advertisers [9]. Moreover, whenever a phone is discoverable, a malicious actor can attempt to pair to it in order to steal contact lists or content of messages. For these reasons phone manufacturers make it difficult (or impossible) for a handset to remain discoverable indefinitely. iOS and Android 6.0+ devices disable discoverability whenever the user exits the Bluetooth settings screen. Older Android devices let the user set the discoverability timeout to, at maximum, five minutes. 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 programmatically, 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 comparing Bluetooth scans with GPS measurements [42], by using fixed infrastructure of RFID transmitters [44], or Bluetooth beacons [25]. In the light of these problems, it is clear that alternative methods for tracking person-to-person interactions are needed. There have been attempts at exploiting WiFi signals for social sensing (*e.g.*, [30, 24, 29,

23] further described in the related work section) but their general applicability is unclear. The previous methods relied on a single feature for comparing list of detected WiFi devices, they were only trained and tested in controlled environments, and they lack verification on longer timescales.

Present work. Here we study the problem of inferring physical proximity between pairs of individuals 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 over a year as part of the Copenhagen Network Study [48]. 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. Apart from comparing the lists directly, we can derive context from just the number of routers seen in the lists: more populated areas tend to have more routers available. Furthermore, we exploit the characteristics of interaction dynamics, for example that 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 0.89 in the proximity inference task. 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.

Contribution. 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 and compare its performance against existing methods.

2. EXPERIMENTAL DESIGN

The dataset used in this work was collected as part of the Copenhagen Networks Study [48]. It covers mobility and interaction records of approximately 1000 students at Technical University of Denmark, over a two year period. Each student was equipped with a LGE Nexus 4 Android smartphone as a data collecting device. On each phone, an application based on the Funf Open Sensing framework [3] gathered readings from multiple sensors including:

- Bluetooth scans (every 5 minutes): each scan contains a list of discoverable devices,*their unique identifiers, user defined names, and received signal strength (RSSI). Because we know which anonymized participant identifier corresponds to which Bluetooth unique

*smartphones in the study were specifically configured to be in Bluetooth discoverable mode

	training	test
total observations	0.5M	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.

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.

The collector app additionally collected the data requested by other applications on the phone. Therefore, the temporal resolution of the data for some of the users can be even higher than one sample every 5 minutes.

All data in the Copenhagen Networks Study was collected with the participants’ informed consent, with an emphasis on ensuring awareness of the complexity and sensitivity of the collected data [46]. 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. [48].

3. METHODS

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 determine whether the two users were in close physical proximity. We use Bluetooth data as ground truth for physical proximity to train and verify our models.

3.1 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 rely on the network name they broadcast. Because the routers at the DTU 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. We found 3950 offending MAC addresses, which corresponds to only 0.04% of all unique MAC addresses in the data. However, scans of these routers constitute 1.4% of all scan results.

Next, we identify one home router for each participant per month. We employ the following heuristic for each participant:

1. Bin the time information of WiFi scan history. The size of the bin does not influence the results significantly, here we use 10 minutes.

- Sort the list of routers by the number of timebins in which they appear, in descending order.
- The router that appears in the biggest number of timebins is assumed to be the home router.

The details of the procedure are described in Ref. [39].

Bluetooth. Due to the imperfect firmware and software running on the phones, Bluetooth data is not always available—not all users are scanning and discoverable at all times. This can introduce a situation in which two persons are proximate, but Bluetooth does not capture that event. We divide the dataset into one hour subsets and select only the WiFi and Bluetooth data from people who were seen and who saw at least one other person through Bluetooth. This strict approach makes the task more difficult, as it removes long periods where individuals are alone, for example night-time samples of students who do not live with other participants.

Negative samples. 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). As a result, the dataset has 31% positive and 69% negative samples.

3.2 Dataset statistics

Table 1 shows the details about the dataset. Through a year of data we found 116M potential interactions. We randomly select 0.5M of them to train the models.

We note that in our dataset people are near to access points more than 95% of time, and the average count of routers in a single scan is 12, see Figure 1A. 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. In general, pairs of people who are in Bluetooth proximity scan more routers in common than those who are not, see Figure 1B. The majority (53%) of meetings happen during working hours (from 8am to 7pm) on campus.

3.3 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, popularity, and location. Table 2 lists the features we apply, and Figure 2 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 person to person contact detection.

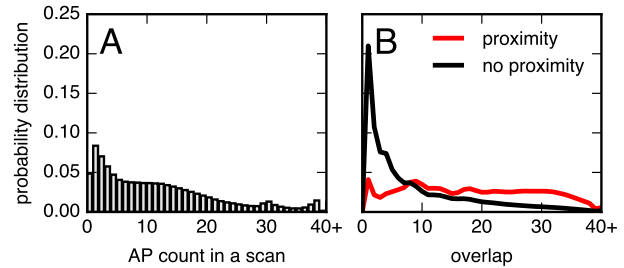


Figure 1: a. More than 95% of scans report at least one access point, and 12 APs on average. b. People in Bluetooth proximity scan more overlapping routers than those who are not proximate.

category	features
AP presence	overlap, non-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
location	at home, at DTU

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

Availability of access points (AP presence). First, we compare the list of routers seen by the two phones, regardless of their received signal strengths. We introduce the following measures: **overlap**: the raw count of overlapping routers [24]; **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 [23]. **non-overlap**: the raw count of non-overlapping routers (size of union minus size of overlap) [24]; Figure 2A-C presents the interplay between the values of the three parameters and the probability of an interaction. Intuitively, the greater the number of 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 [39]. 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 a dining hall). Nevertheless, we expect that, in general, the probability of interaction is negatively correlated with the size of union. Using Jaccard similarity between the two lists allows to recognize interactions regard-

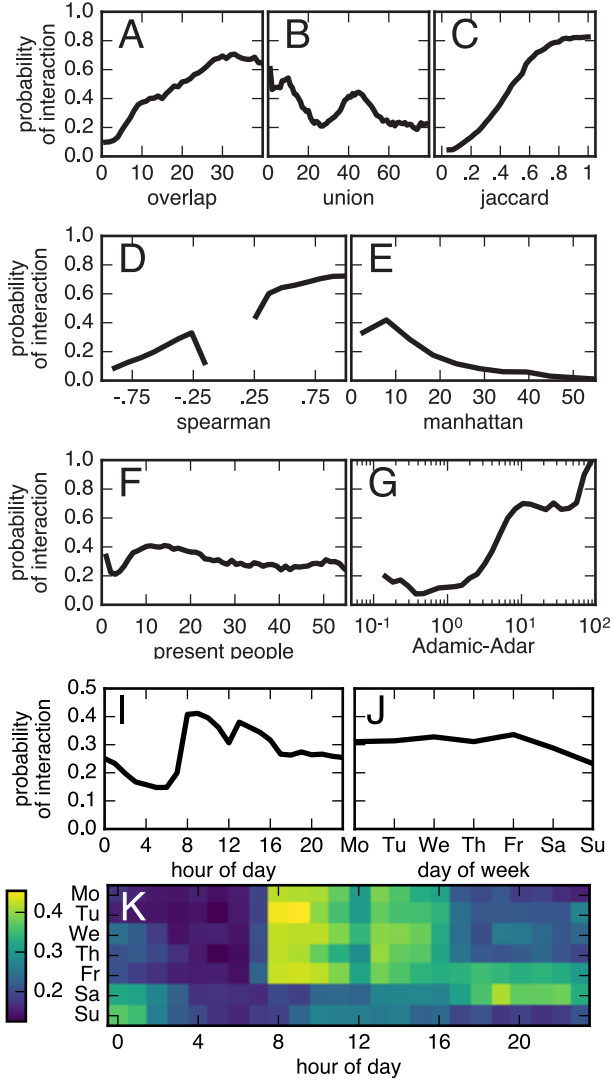


Figure 2: The larger the number of common routers two phones see, the higher the probability of close proximity. At the same time, the more routers they see in total, the lower the probability of an interaction — densely populated areas have more routers and more people who are not necessarily interacting. Jaccard similarity allows us to recognize interactions regardless of the number of visible access points.

less 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 generally a reliable proxy for distance [38], two co-located people can be expected to have similar RSSI readings for the overlapping routers. We investigate the **spearman** and **pearson** correlation coefficients of received signal strengths of the overlapping routers. For brevity we only present the results for the spearman metric Figure 2D — the values of the two metrics are highly correlated (Spearman’s $\rho = 0.89$,

$p_{val} < 0.001$). 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 3.4 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 **eucclidean** and define them in Equations 1 [30] and 2 [23] 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 2E 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 [15]. 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 $\pm 6dB$** assumes 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 2F, this intuition is not entirely confirmed by the data. The correlation between the number of individuals present and the probability that any two of them are interacting is low (Spearman’s $\rho = 0.15$, $p_{val} < 0.001$). 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 [2], defined as:

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

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 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 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 2I-K. While there is only a small variability between the days of the week (Figure 2J), the probability of the interaction during a day (Figure 2I) 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 2K), 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.

Location. The last category, location, contains two binary features. A meeting is considered **at home** if at least one of the routers in the union corresponds to the home router of one of the users (the heuristic for home location detection is explained in 3). A meeting is assumed to take place **at DTU** if at least one of the routers in the union broadcasts a WiFi network name of `dtu`, as all access points on the campus do.

3.4 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 strengths 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 encounter NaN values. This average from training is preserved and used to impute missing values in the test set. We verified in our data that other approaches, such as using the median value of the feature or using k nearest neighbors to impute the missing value [50], do not improve the consecutive predictive performance.

4. RESULTS

In this section we evaluate the performance of each feature and each featureset in the task of proximity inference. Then,

we examine the robustness of our best model to short training as well as the various types of environments in which the interactions happen.

4.1 Performance of single features

We first show how well one can infer close-proximity interactions using single features. We report the area under Receiver Operating Characteristic curve (AUC ROC) as the first metric of performance in Table 3. Then, we select the threshold at which the F_1 score (the harmonic mean between precision and recall) is maximized in the training set. We also report the F_1 score at the threshold optimal for the training set along with the AUC ROC for the test data (111.5 million previously unseen samples).

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 thresholds are not just specific to the training data.

4.2 Performance of feature sets

We train a Gradient Boosting Classifier for each category of features and present the results in Table 4. The parameters of the classifier are tuned each time through a grid search of the parameter space with 5-fold cross validation. Furthermore, we compare the model based on the features proposed by Krumm *et al.* [24] to models based on richer sets of features, see Table 4. In the original work, Krumm *et al.* did not find any performance improvements of using a combined model over using single features. Here, we show that combining the features they proposed does improve the performance. 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 or group of features, and it outperforms the model based on the features introduced by Krumm. 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), does not improve the performance of the classifier.

4.3 WiFi similarity and physical proximity

Here, we verify whether there is a correlation between how close people are in physical space (approximated by the received Bluetooth signal strength measured on their phones) and the probability that our models misclassify the sample as “non-interaction”. As we show in Figure 3, the shorter the distance over which an interaction happens (high Bluetooth RSSI), the lower the probability of missing that interaction. This shows that the similarity measure between WiFi lists introduced by our models has a physical interpretation: a more similar WiFi environment indicates proximity in a more granular way than just the Bluetooth 10 meter range.

4.4 Training period and performance in test

Figure 4 shows how the number of samples used for train-

category	feature	AUC ROC		F_1	
		train	test	train	test
AP presence	overlap	0.77	0.77	0.61	0.61
	jaccard	0.84	0.84	0.69	0.68
	union	0.53	0.53	0.48	0.48
	non-overlap	0.74	0.74	0.58	0.57
RSSI	spearman	0.70	0.70	0.57	0.58
	pearson	0.71	0.71	0.59	0.59
	manhattan	0.60	0.60	0.51	0.51
	euclidean	0.59	0.59	0.51	0.51
Presence + RSSI	top AP	0.60	0.60	0.48	0.48
	top AP±6dB	0.75	0.74	0.65	0.65
Popularity	min_popularity	0.54	0.54	0.48	0.48
	max_popularity	0.59	0.59	0.49	0.50
	adamic_adar	0.77	0.77	0.62	0.62
Timing	hour of week	0.51	0.51	0.48	0.48
Location	at DTU	0.61	0.61	0.51	0.51
	at home	0.64	0.64	0.55	0.55

Table 3: Performance of single features and feature categories in the task of inferring close proximity interactions. Jaccard similarity between lists of routers seen by the two devices is the best performing single feature. F_1 are given for a threshold that maximizes F_1 in the training set.

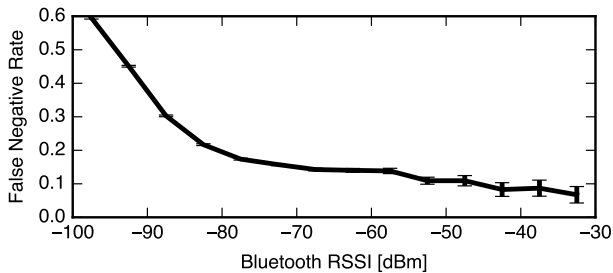


Figure 3: The distance over which an interaction happens can be approximated using Bluetooth received signal strength (RSSI). Very close proximity contacts are unlikely to be misclassified as non-interactions. The lower the RSSI (the more distant the two potentially interacting people), the higher the probability, that our models miss the interaction.

ing influences the performance of the full model in test. We compare the performance of a random forest classifier and a gradient boosted classifier and find that the latter has a slightly higher performance for training sets larger than 1000 samples. On the other hand, training of the random forest classifier can be parallelized, thus making the process faster.

4.5 Importance of features

Here we show how important each feature is for the machine learning model. In the implementation we use [35] the feature importance is defined as the total decrease in node impurity weighted by the probability of reaching that node, averaged over all trees of the ensemble [1]. Figure 5 shows the accumulated results from 30 training rounds of the gradient boosted classifier on randomly selected subsets

featureset	AUC ROC		F_1	
	train	test	train	test
AP presence: overlap, non-overlap, jaccard, union	0.85	0.85	0.69	0.69
RSSI: spearman, pearson, manhattan, euclidean	0.78	0.79	0.62	0.62
Presence+RSSI: top AP, top AP±6dB	0.75	0.75	0.65	0.65
Popularity: min, max, adamic_adar	0.79	0.79	0.62	0.62
Location: at DTU, at home	0.65	0.65	0.55	0.55
NearMe: overlap, non-overlap, spearman, euclidean	0.87	0.87	0.71	0.71
Simple: AP presence, RSSI, Presence + RSSI	0.88	0.88	0.72	0.72
General: AP presence, RSSI, Presence + RSSI	0.89	0.89	0.73	0.73
Popularity, at home	0.89	0.89	0.73	0.73
Full: all features	0.89	0.89	0.73	0.73

Table 4: Performance of feature sets in the task of inferring close proximity interactions. We train a Gradient Boosted Classifier on selected subsets of features: each feature category listed in Table 3, NearMe [24], 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). Using features which could be specific to the experiment does not improve performance further.

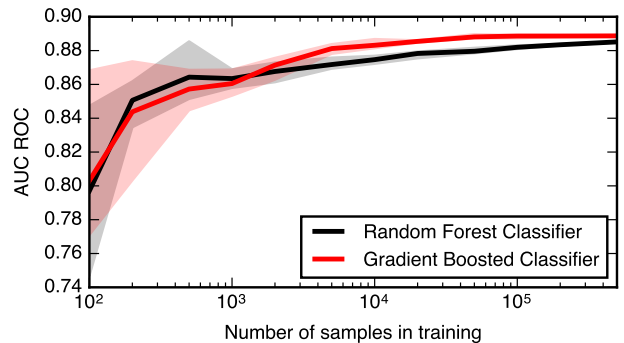


Figure 4: The more samples we use for training the interaction detection models, the better they perform in test, but after a certain threshold, the gains are negligible. The performance of the Gradient Boosted Classifier saturates at a higher level, but the time it takes to train the classifier is longer than it is the case with the Random Forest Classifier. Each of the model is trained 20 times for each number of samples, the shaded areas correspond to 25-75 percentiles and the solid lines to medians of the results for each training set size.

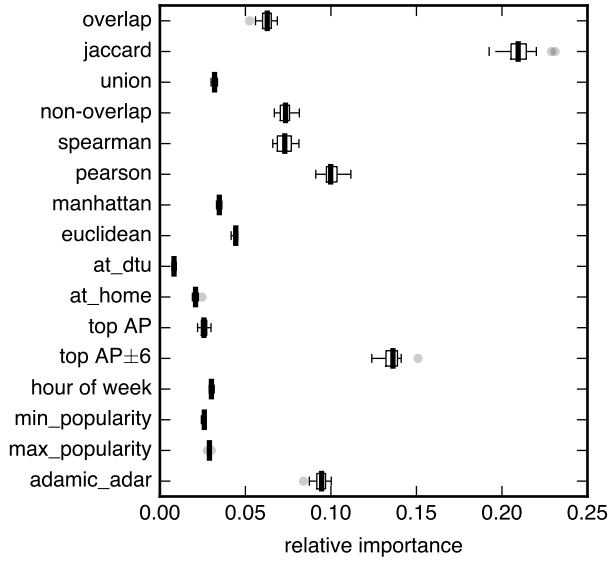


Figure 5: Gradient Boosted Classifier reports the relative importance of each feature (the decrease in node impurity it provides). After 30 training rounds we see that Jaccard is the most important feature, followed by overlap among the strongest routers (top AP±6dB), Adamic-Adar, and Pearson correlation between the signal strengths.

of the training data, each with 100 000 samples. We find that Jaccard similarity is the most important, followed by the overlap among the strongest routers, Pearson’s correlation of signal strengths, and Adamic-Adar (which exploits the overlap and the popularity of routers).

4.6 Validity of the model in different scenarios

Figure 6 shows the performance of the gradient boosting classifier in different contexts and across time.

Number of routers. As described before, the number of routers in an environment is positively correlated with the population density. We divide the test data in three equally-sized subsets, depending on the size of the union of routers seen by two people. Figure 6A shows that the performance of the model is best in the low and mid sets ($AUC > 0.9$) and observably lower ($AUC \approx 0.85$) for environments with the highest number of routers. Thus, we show that the model performs well in typical environments.

Location. Because our the data was collected by students of one university, with the majority of interactions happening on campus, there is a risk that the model would overfit towards such situation. This is, in fact not the case. Figure 6B shows that while the performance of the model is high on campus, it becomes even better for the meetings outside.

Timing. As shown in Figure 6C the performance of the model does not drop significantly during special periods, such as Christmas or summer vacation (gray areas in the plot correspond to periods with no university classes).

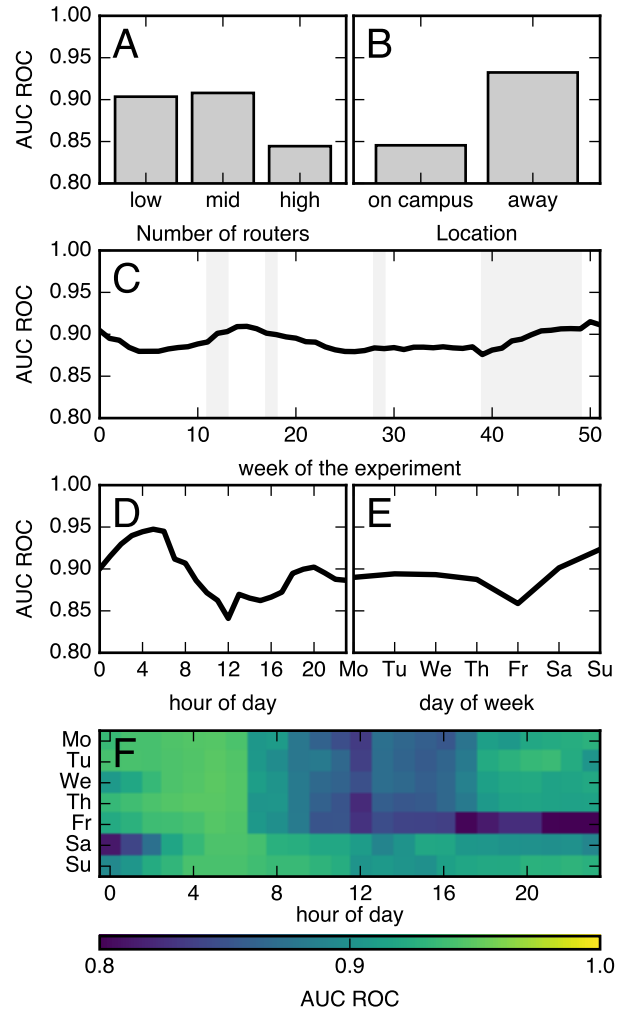


Figure 6: Our model for detecting person-to-person proximity events performs well regardless of the number of available routers (A) and location (B). Its performance does not drop during holidays (marked with gray areas in C). The situation in which the performance is the worst is the Friday evenings and nights (F), but even then, the AUC ROC is high.

Instead, it remains stable throughout the experiment.

The performance does vary with the hour of week, as shown in Figure 6D-F. When we compare it to Figure 2K, we see that the model performs better in situations where the prior probability of meeting is lower (for example during week nights). Nevertheless, it retains high performance of $AUC \geq 0.8$ throughout the week.

5. 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.

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 [16], stable [28],

and predictable [43]. Several works argue that many unpredictable travels observed in real data can be attributed to individuals seeking interaction with their social contacts [17, 49, 8]. It yet remains to be verified whether these findings 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 investigated the potential of WiFi routers in applications of indoor [4, 18, 36] and outdoor [7, 31, 13, 20] localization. Our recent work investigates how large companies can crowd source the creation of databases with router locations [38, 31, 13] and how people’s mobility on societal scale can be described using only a small subset of available routers [39]. 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 [22, 51].

It is important to stress that the work presented in this article does not rely on location estimation (in terms of geographical coordinates) 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 [11, 42, 47, 45]. 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) [32, 37] or Bluetooth enabled smartphones [11, 3, 48]. In case of badges, interactions are usually inferred using radio-frequency identification (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 using Bluetooth-enabled mobile phones is less granular. The proximity can be detected in a binary fashion or further refined using the received signal strength as a proxy for distance [41]. However, the orientation of the individuals can not be sensed. The subjects’ devices must 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 [33], providing a better understanding of the nature of the contacts [19]. 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 environment. 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 distance [30] or overlap [29]. NearMe project [24] introduced more features, such as rank correlation between the lists of

overlapping routers sorted by signal strength, Euclidean distance, and the number of non-overlapping APs. The authors explored combining the features into a regression model, but this approach did not outperform single features. Moreover, their model would overfit for the rooms where it was trained and thus under-perform in previously unseen environments. In fact, Kjærgaard and Nurmi name differences in environments where the sensing takes place among the most important obstacles in using WiFi for social sensing [23]. Carlotto *et al.* combine a number of previously suggested features using a Gaussian Mixture Model and claim that their model is not environment-dependent (performs equally well in both buildings where it was tested) [5].

We note that the differences in environments can actually be used to increase the performance of the model. We can exploit 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 2).

6. DISCUSSION

In this paper we evaluated the applicability of WiFi based social sensing. The idea of exploiting WiFi signals for this purpose is not new. However, to our best knowledge, researchers have not yet tested this approach in practice, over a long period, and in a large population that interacts in various environments. The growing popularity of WiFi access points and the phones’ inability to remain Bluetooth discoverable are two trends that make it feasible and important to begin using WiFi signals for social sensing.

6.1 Privacy implications

There are two main privacy implications of this work.

First, the ability to track face-to-face interactions using WiFi can help us move away from relying on Bluetooth. By not requiring the participants’ phones to remain Bluetooth discoverable we protect the privacy and security of the subjects. While currently most phones advertise their presence and identity by scanning for WiFi, this problem is being addressed. Both Android and iOS randomize the MAC address of the device every time it sends WiFi probe requests making it more difficult to identify the user.[†]

Second, our results indicate a potential erosion of privacy of Android users. As we have previously shown, WiFi can be efficiently used for high-resolution mobility tracking of entire populations [38, 39, 51]. Here we go a step further and infer who people interact with, not only where they are. 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 15 seconds [39]. This problem is addressed since Android 6.0—in the latest versions of the system an application has to hold a location permission to listen to WiFi scan results. However, the vast majority of handsets currently in use will not receive these crucial updates. Thus, WiFi signals remain

[†]The randomization can only happen when the device is not connected to any WiFi network. When it is, it announces its real MAC address in each probe request.

a major privacy risk for years to come.

6.2 Limitations of the WiFi-based social inference

While our approach to inference of social interactions using WiFi signals offers an important new method in computational social science, we want to recognize its limitations. The inference in the approach presented here depends on the WiFi routers being present in the environment. While today WiFi networks are omnipresent, especially in densely-populated areas [39], 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 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 scan data, making it more difficult to devise a robust model as the one presented here.

Furthermore, due to the lack of ground truth data, we cannot prove that our model accurately estimates the distance between users. We show, that our model is more likely to recognize interactions with a higher Bluetooth RSSI, but this property does not trivially translate to distance estimation.

Finally, we should note that it is not our argument that the values of all model features for discovering particular interactions and reconstructing the overall social network are 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 indicate, however, that physical proximity can be inferred in a feasible fashion using WiFi signals collected by smartphones, even in very densely-connected populations.

7. CONCLUSION

In this work we showed how WiFi scan results can reveal a great deal about our daily interactions with others and our social ties. By using behavioral traces, placed in context through meta information and our basic understanding of the inner working of social systems, we can transform a noisy data source to a strong social signal. Our findings have important privacy implications, especially given our previous work which shows that it is possible to use WiFi signals for tracking human mobility. On the other hand, WiFi scans also constitute a great opportunity for companies with access to such data on a global scale, to contribute *e.g.*, better epidemic models built on proximity data of billions of people. Finally, we hope that this method of social sensing will substitute Bluetooth sensing in future Computational Social Science deployments.

Acknowledgements

The authors would like to thank Andrea Cuttone for useful discussions as well as Urvashi Khandelwal and Jana Huisman for the important feedback. In this work we used the implementations of machine learning models from the scikit-learn [35] Python package.

8. REFERENCES

- [1] How are feature importances determined in Random Forest Classifier? <http://stackoverflow.com/a/15821880>. Accessed: 2015-10-17.
- [2] L. A. Adamic and E. Adar. Friends and neighbors on the web. *Social networks*, 25(3):211–230, 2003.
- [3] 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.
- [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] A. Carlotto, M. Parodi, C. Bonamico, F. Lavagetto, and M. Valla. Proximity classification for mobile devices using wi-fi environment similarity. In *Proceedings of the First ACM International Workshop on Mobile Entity Localization and Tracking in GPS-less Environments, MELT '08*, pages 43–48, New York, NY, USA, 2008. ACM.
- [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] 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.
- [8] 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.
- [9] S. Dato. High street shops are studying shopper behaviour by tracking their smartphones or movement. <http://goo.gl/vGg8k8>.
- [10] R. Dillet. Happn is a dating app powered by real life interactions. <http://goo.gl/0nHyIr>.
- [11] N. Eagle and A. Pentland. Reality mining: sensing complex social systems. *Personal and ubiquitous computing*, 10(4):255–268, 2006.
- [12] 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.
- [13] A. Eustace. Wifi data collection: An update. <http://goo.gl/VFJ9mM>.
- [14] 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.
- [15] R. C. Gatej. An adaptive approach to mobile sampling. Master’s thesis, Technical University of Denmark, 2013.
- [16] M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabasi.

- Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, 2008.
- [17] 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.
- [18] 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.
- [19] E. T. Hall. The hidden dimension . 1966.
- [20] 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.
- [21] L. Isella, J. Stehlé, A. Barrat, C. Cattuto, J.-F. Pinton, and W. Van den Broeck. What’s in a crowd? analysis of face-to-face behavioral networks. *Journal of theoretical biology*, 271(1):166–180, 2011.
- [22] J. H. Kang, W. Welbourne, B. Stewart, and G. Borriello. Extracting places from traces of locations. In *Proceedings of the 2nd ACM international workshop on Wireless mobile applications and services on WLAN hotspots*, pages 110–118. ACM, 2004.
- [23] 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.
- [24] J. Krumm and K. Hinckley. The nearest wireless proximity server. In *UbiComp 2004: Ubiquitous Computing*, pages 283–300. Springer, 2004.
- [25] 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.
- [26] 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.
- [27] J. Li, J. Jannotti, D. S. J. De Couto, D. R. Karger, and R. Morris. A scalable location service for geographic ad hoc routing. In *Proceedings of the 6th Annual International Conference on Mobile Computing and Networking, MobiCom '00*, pages 120–130, New York, NY, USA, 2000. ACM.
- [28] X. Lu, L. Bengtsson, and P. Holme. Predictability of population displacement after the 2010 haiti earthquake. *Proceedings of the National Academy of Sciences*, 2012.
- [29] M. McNett and G. M. Voelker. Access and mobility of wireless pda users. *SIGMOBILE Mob. Comput. Commun. Rev.*, 9(2):40–55, Apr. 2005.
- [30] J.-L. Meunier. Peer-to-peer determination of proximity using wireless network data. 2004.
- [31] B. Meyerson. Aol introduces location plug-in for instant messaging so users can see where buddies are. <http://goo.gl/2W1uYh>.
- [32] 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.
- [33] 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.
- [34] 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.
- [35] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.
- [36] 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.
- [37] 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.
- [38] P. Sapiezynski, R. Gatej, A. Mislove, and S. Lehmann. Opportunities and challenges in crowdsourced wardriving. In *Proceedings of the 15th ACM SIGCOMM conference on Internet measurement*. ACM, 2015.
- [39] P. Sapiezynski, A. Stopczynski, R. Gatej, and S. Lehmann. Tracking human mobility using wifi signals. *PLoS ONE*, 10(7):e0130824, 07 2015.
- [40] K. Scarfone and J. Padgett. Guide to bluetooth security. *NIST Special Publication*, 800:121, 2008.
- [41] V. Sekara and S. Lehmann. The strength of friendship ties in proximity sensor data. *PLoS one*, 9(7):e100915, 2014.
- [42] V. Sekara, A. Stopczynski, and S. Lehmann. Fundamental structures of dynamic social networks. *Proceedings of the National Academy of Sciences*, 113(36):9977–9982, 2016.
- [43] C. Song, Z. Qu, N. Blumm, and A.-L. Barabási. Limits of predictability in human mobility. *Science*, 327(5968):1018–1021, 2010.
- [44] 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.
- [45] A. Stopczynski, A. S. Pentland, and S. Lehmann. Physical proximity and spreading in dynamic social networks. *arXiv preprint arXiv:1509.06530*, 2015.

- [46] 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.
- [47] A. Stopczynski, P. Sapiezynski, S. Lehmann, et al. Temporal fidelity in dynamic social networks. *The European Physical Journal B*, 88(10):1–6, 2015.
- [48] 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.
- [49] J. L. Toole, C. Herrera-Yaqui, C. M. Schneider, and M. C. González. Coupling human mobility and social ties. *Journal of The Royal Society Interface*, 12(105):20141128, 2015.
- [50] 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.
- [51] D. K. Wind, P. Sapiezynski, M. A. Furman, and S. Lehmann. Inferring stop-locations from wifi. *PloS one*, 11(2):e0149105, 2016.
- [52] 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.