Inferring the Social-Connectedness of Locations from Mobility Data

Tristan Brugman, Mitra Baratchi, Geert Heijenk, Maarten van Steen

University of Twente, Enschede, The Netherlands
{t.w.r.m.brugman, m.baratchi, geert.heijenk, m.r.vansteen}@utwente.nl

Abstract. An often discriminating feature of a location is its social character or how well its visitors know each other. In this paper, we address the question of how we can infer the social contentedness of a location by observing the presence of mobile entities in it. We study a large number of mobility features that can be extracted from visits to a location. We use these features for predicting the social tie strengths of the device owners present in the location at a given moment in time, and output an aggregate score of social connectedness for that location. We evaluate this method by testing it on a real-world dataset. Using a synthetically modified version of this dataset, we further evaluate its robustness against factors that normally degrade the quality of such ubiquitously collected data (e.g. noise, sampling frequency). In each case, we found that the accuracy of the proposed method highly outperforms that of a state-of-the-art baseline methodology.

Keywords: Spatial profiling, link prediction, mobility data mining, Wi-Fi scanning, mobility modeling

1 Introduction

Just like people, locations have a social profile. This profile reflects the social connectedness of people who visit those locations and it dynamically changes over time as people with different social ties enter and leave a location. Having knowledge about the social profile of a location before visiting it is a useful addition to location-based recommender systems. If we want to meet new people and make new friends, we may want to visit the most social pub among all the pubs in town. However, if we want to go to a quiet library for studying, a less social one may be more appropriate. Likewise, having knowledge about the social profile of a location helps to improve the services offered there. This has been shown to be important in, for example, elderly care facilitates [1, 2, 3].

To create social profiles for locations, in this paper, we design a method that can infer the social connectedness of that location from ubiquitously generated mobility data (such as GPS coordinates, Wi-Fi scans, or check-in records in location-based social networks). While research in spatial profiling [4] from mobility data has previously addressed characterizing locations from such data, creating a social profile for locations has not been addressed before. In order to know how socially connected a location is, we investigate to what extent we can extract the social tie strength of people based on their visit to a single location. Previous research has mainly examined methods to infer the

strength of social ties between pairs of individuals. This is achieved by using the global trajectory of mobile entities over *many* locations. We, however, consider characterizing *locations* rather than *individuals*. This means that, in our case the available input data is limited to that acquired from a single location.

Extracting the social context from visiting patterns of people in only a single location is a challenging problem. Mobility data is of limited social interaction content. For example, working in the same building does not guarantee that two people have strong ties or even know each other. Due to the inherent differences in the functionality of locations, the social context can be reflected in different features of visits. It is not yet clear which features of a visit can be used for this purpose. Furthermore, oftentimes additional data with strong social interaction content does not accompany mobility datasets. In most cases ground truth on social ties can only be collected from a small sample of visitors. To address these problems, in this paper, we propose a data-driven technique for extracting an aggregate measure for the social connectedness of a single location by detecting only the presence of people in that location. More specifically:

- We study a large number of features that can be extracted from mobility data acquired from presence of people in a single location.
- We propose a supervised method for selecting among this list of features and consequently learning social ties from them.
- We validate the performance of our method in predicting social tie strengths using a dataset of Wi-Fi mobility data. As ground truth, we use an estimate score of social tie strengths derived from the similarity of devices' SSID (Service Set IDentifiers) sets and show how the method performs by learning from a sample of these social tie indicators.

2 Related Work

There have been a number of previous studies that describe methods for inferring social ties between individuals from either Wi-Fi protocol-specific information, or more general mobility data.

Wi-Fi protocol-specific data: when Wi-Fi-enabled devices try to connect to nearby access points, they often broadcast *probe request* messages, which can be used to infer the social ties between the owners of devices. Detections of probe requests can be used to create a mobility trace representing timestamped presence of mobile devices near Wi-Fi access points (or scanners) [5]. Furthermore, probe requests contain the names of access points that the mobile device has been connected to before (SSIDs). Previous research has shown that it is possible to extract information about the social ties between device-owners from their similarity in SSID lists. The authors of [6] have examined different similarity metrics between the SSID lists of pairs of devices observing a high correlation between SSID lists and social tie strengths acquired by surveying device owners. The authors of [7] use SSID list similarity to extract a social network to confirm the sociological theory of homophily [8]. The study in [9] proposes a framework to use the social information acquired from the SSID list similarity and location visitation frequencies for calculating a venue reputation score. The method in [10] employs different techniques to infer social ties between device owners.

General mobility data: data acquired from location-based social networks, GPS, and cellular networks have also been used for extracting social information. Research presented in [11] describes a method to improve social tie prediction, focusing on user check-ins in location-based social networks. Two different mobility features are extracted for each pair of users that have visited the same location: the minimum place entropy across all venues they have both visited, and the sum of the inverse of each place entropy value. In [12] mobility data from cellular networks is used along with phone call communications to infer a reciprocal friendship social network. Authors of [13] use two information-theoretic indicators to infer social link types of people relying on similarities of their visits extracted from their GPS mobility data.

While the first group of research show that it is possible to infer social relationships from the SSID list, they have not investigated extracting such information purely from the mobility data acquired from probe requests. While SSID lists are specific to the Wi-Fi protocol, datasets made by the detection of probe requests can represent a more general class of mobility datasets such as those acquired from GPS, Bluetooth, cellular networks, etc. Our approach in this paper is to investigate to what extent mobility data acquired from probe requests can be used for extracting social information. The second group of studies, on the other hand, successfully makes use of mobility features to predict the existence of social relationships. By having access to the global trajectories, these methods are successful while only employing a limited number of mobility features. We, however, consider extracting the social context from mobility data collected from a single location. This is implied by our goal, which is characterizing *locations* rather that *individuals*.

3 Preliminaries

The goal of this study is to derive an aggregate social connectedness score for a location using mobility information collected from visitor's devices. In a specific location, we consider having a system that detects presence of visitor devices in it. Wi-Fi scanning [5] near a Wi-Fi scanner allows collection of such a dataset. Before explaining the problem, we define a number of terminologies used in the rest of the paper:

Definition 1. *A **location** is a defined spatial area where presence of devices can be detected. An example of a location is the area covered by a Wi-Fi scanner.*

Definition 2. *A **detection** in a location is a tuple $\langle \mathsf{d}, t \rangle$, in which d represents the identifier of the visiting device and t represents the moment in time that the device is detected.*

Definition 3. *A **mobility trace**, denoted by $\langle \mathbf{MT}, t_{start}, t_{end} \rangle$, is a collection of detections acquired in a timespan ranging from t_{start} to t_{end}.*

Definition 4. *The **pairwise social tie strength** between two device owners, denoted by $\{s_{ij} | s_{ij} \in [0,1]\}$, is the strength of the social tie between the owners of devices i and j.*

Definition 5. *The **social connectedness score**, denoted by $\overline{s}$, is the normalized social tie strength between a group of users. It is a score $\overline{s} = \frac{1}{|\mathbf{D}|} \sum_{i,j \in \mathbf{D}} s_{ij}$ for each set of devices $\mathbf{D} = \{d_1, d_2, ..., d_n\}$, in which each d_k represents a different device.*

Problem: Given a mobility trace $\langle \mathbf{MT}, t_{start}, t_{end} \rangle$ and a timestamp $t \in [t_{start}, t_{end}]$, we are interested to infer the aggregate social connectedness score $\bar{s}_t$ for the group of devices that are present in the location at time t.

4 Approach

In this section, we present our approach in extracting the social connectedness score of a location based on visits of people to only that location. Our approach is based on learning the relationship between mobility features and social tie strengths in a supervised manner. For this purpose, we train a model that identifies the relationship between mobility features and samples of social tie strengths. We explore a variety of possible mobility features and use a feature selection algorithm in order to identify a subset of important mobility features related to social ties. In order to determine an indicator for ground truth on social tie strengths, we consider the similarity between devices' Wi-Fi SSID lists. Based on the mobility dataset, other type of ground truth can also be used for this purpose.

4.1 Ground Truth Metric

In order to train a model, we need to obtain ground truth for the social tie strengths between pairs of devices. We take the approach of using the anonymized Wi-Fi SSID lists of each pair of devices, and computing a value that measures their overlap [6, 10, 7, 9, 14]. The general intuition is that elements of this list represent presence in places such as device owner's homes which are only probable to be shared when people have strong social ties. The more two lists share such kind of rare SSIDs, the probability that they have stronger ties increases. The research in [6] has compared a variety of similarity metrics, among which a modified version of Adamic-Adar metric known as Psim-3 performs best in determining social ties between individuals. This metric is calculated as $\sum_{z \in X \cap Y} \frac{1}{f_z^3}$, in which X and Y are the two SSID sets, and f_z is the number of times that identifier z occurs in the dataset. This measure can be normalized between zero and one based on the maximum strength found in the whole dataset.

4.2 Method

Our method performs in two phases of learning and inference. During the initialization phase, the model is trained and its features are selected based on a mobility trace and knowledge about the pairwise social tie strengths. This is followed by the utilization phase, in which the mobility trace is supplied to the model, inferring pairwise tie strengths. By combining these strengths with the devices present at a given timestamp, an aggregate social connectedness score is then calculated. Figure 1 provides an overview of the proposed method.

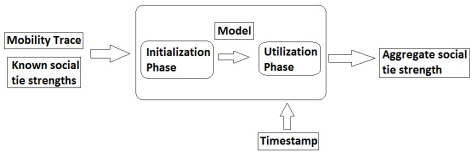

Fig. 1: High-level overview of the proposed method

4.3 Features

To train the model, we have selected 5 general feature classes leading to 124 mobility features which are extracted from each pair of devices. Table 2 provides an overview of these features. This table can be interpreted using the notations introduced in Table 1. The feature classes group the features by their source of information; overlapping visits, devices themselves, and the environment. There is no class related to the location only, because its features would be independent of the pair of devices, and have the same values for each sample. These features are:

- **Overlap Only** (51 features) These features characterize the mutual overlapping visits of the pair of devices to the location. Examples are the number and total length of the overlapping visits, and the average amount of time that one device waits before and after the other arrives.
- **Individual Only** (16 features) These features characterize the individual visits of each of the two devices to the location. They include the total number and length of visits by the devices, and their average and median visit lengths.
- **Overlap and Individual** (8 features) These features relate the overlapping visits to the individual devices' overall visiting pattern. For example, this group contains the ratio between overlapping visits and the total number of visits made by each device, or the ratio between overlap length and total visit length of each device.
- **Overlap and Location** (7 features) These features characterize the state of the location when the devices had overlapping visits to it. Specifically, it considers how busy the location was when the visit took place. Example features are the average and maximum number of devices present during overlapping visits.
- **Individual and Location** (42 features) These features characterize the state of the location during the individual visits of devices. Examples are the average and maximum popularity during individual visits.

The features are defined based on the mobility of any pair of devices i and j, such that $\{(i,j) \in \mathbf{D}\}$. For each of these devices, we consider the set of all of its n visits to the considered location, $\mathbf{V}_{\mathsf{d}} = \{v_1, v_2, ..., v_n\}$ with d being the id of the detected device. Each v_{p} in this set is defined by tuples of the form $\langle s,e \rangle_{\mathsf{p},\mathsf{d}}$, with s and e being the timestamp of when the visit started and ended, respectively. Additionally, we consider the set of overlapping visits $\mathbf{O}_{i,j}$ of these two devices, by determining which visits in $\mathbf{V}_i$ took place during a visit in $\mathbf{V}_j$. The set $\mathbf{O}_{i,j} = \{ov_1, ov_2, ..., ov_m\}$ is composed of m number of overlapping visits these two devices had. Therefore, each ov_q is composed of tuples of the form $\langle s_i, s_j, e_i, e_j, s_o, e_o \rangle_q$, in which each element is a timestamp. The elements s_o and e_o are defined as follows: $s_o = max(s_i, s_j)$ and $e_o = min(e_i, e_j)$.

Table 1: Function definitions

Function	Output	Definition
$length(v)$	$e-s$	Duration of the visit v
$present(t)$	$\{\mathsf{d} \in \mathbf{D} \mid (\exists v_{\mathsf{p}} = \langle s,e \rangle \in \mathbf{V}_{\mathsf{d}} : (s \leq t \leq e))\}$	Id of devices present at time t
$max\_present(\mathbf{MT})$	$max(\lvert present(t) \rvert)_{t \in \mathbf{MT}}$	Maximum number of devices present in the location at any time
$f(x)$	$min(x), max(x), std(x), sum(x),$ $mean(x), median(x), max(x)-min(x)$	Statistics applied on the set x

Finally, $\mathbf{Od}_{i,j}$ and $\mathbf{Od}_{j,i}$ are similar to $\mathbf{O}_{i,j}$, but each overlap is calculated from the point of view of a single device. For example, if a visit from device i starts during one visit of device j and ends during another, it counts as 1 overlapping visit for $\mathbf{Od}_{i,j}$ and as 2 for $\mathbf{Od}_{j,i}$. Table 1 defines a number of auxiliary functions that are required to compute the features. The function denoted by f generates seven statistical values from a given sequence of values, so it defines seven features.

4.4 Initialization Phase

We use the above-mentioned features in the initialization phase. The first part of this phase is creating the input samples from available data and the second part performs feature selection and trains a model.

Creating samples: Input samples consist of $n = 124$ number of mobility features which are labeled with social tie strengths. In this paper, both the tie strength and mobility features are computed based on a data set of Wi-Fi access probe requests, in which each device is identified by its MAC address. The social tie strength is derived from SSID list similarity by calculating Psim-3, as described in Section 4.1. The mobility features are computed based on the timestamped Wi-Fi probe requests, which represent the presence of a device nearby a Wi-Fi scanner. The timestamps are recorded in discrete and irregular time intervals. In order to compute the mobility features, the start and end time of visits need to be detected from such timestamps. When the distance between two consecutive timestamps is shorter than a specific threshold tr they are grouped as a single visit. Otherwise, these timestamps are considered to be part of separate visits. We chose the gap length threshold tr such that it was higher than 95% of all gaps in the dataset (2 and 4 minutes). Algorithm 1 in the appendix provides algorithmic details on this procedure.

Feature selection and learning: In this phase, a number of features are selected and a model is trained to relate the input mobility features to the social tie strength. As we are interested in learning the social tie strengths as numerical values from numerical features, we are dealing with a regression problem. We initially select the important features in a greedy manner and continue to do so until no feature improves the quality of the regression [15]. Next we train the regressor using the selected set of features. Algorithm 2 in the appendix provides more detail on this procedure.

Table 2: Features extracted for each pair of devices i, j

Feature class	Indices	Feature definition
Overlap Only	1	$\lvert\mathbf{O}_{i,j}\rvert$
	2	$\sum_i length(\mathbf{O}_{i,j})$
	3 - 16	$f(\{\langle s_o - s_k\rangle_{q,k} \lvert ov_q \in \mathbf{O}_{i,j}, k \in i,j\})$ [1]
	17 - 29	$f(\{\langle e_k - e_o\rangle_{q,k} \lvert ov_q \in \mathbf{O}_{i,j}, k \in i,j\})$ [1]
	30 - 36	$f(\{\langle s_o - min(e_i,e_j)\rangle_q \lvert ov_q \in \mathbf{O}_{i,j}\})$
	37 - 43	$f(\{\langle max(e_i,e_j) - e_o\rangle_q \lvert ov_q \in \mathbf{O}_{i,j}\})$
	44 - 51	$f(\{\langle (s_o - min(s_i,s_j)) + (max(e_i,e_j) - e_o)\rangle_q \lvert ov_q \in \mathbf{O}_{i,j}\})$
Individual Only	1 - 2	$\{\lvert\mathbf{V_k}\rvert \lvert k \in i,j\}$ [1]
	3 - 16	$f(\{length(v_p) \lvert v_p \in \mathbf{V_k}, k \in i,j\})$ [1]
Overlap and Individual	1 - 2	$\{\lvert\mathbf{Od}_k\rvert, k \in \langle i,j\rangle, \langle j,i\rangle\}$ [1]
	3	$\lvert\mathbf{O}_{i,j}\rvert / max(\lvert\mathbf{V_i}\rvert, \lvert\mathbf{V_j}\rvert)$
	4 - 5	$\lvert\mathbf{O}_{i,j}\rvert / \lvert\mathbf{V_k}\rvert, k \in i,j$ [1]
	6	$(\sum length(\mathbf{O}_{i,j})) / \sqrt{(\sum length(\mathbf{V_i}) * \sum length(\mathbf{V_j}))}$
	7 - 8	$(\sum length(\mathbf{O}_{i,j})) / (\sum length(\mathbf{V_k})), k \in i,j$ [1]
Overlap and Location	1 - 7	$f(\{\langle \lvert present(t)\rvert / max\_present(\mathbf{MT})\rangle_q \lvert t \in \{range(s_o,e_o) \lvert ov_q \in \mathbf{O}_{i,j}\}\})$
Individual and Location	1 - 14	$f(\{\langle \lvert present(s)\rvert / max\_present(\mathbf{MT})\rangle_p \lvert v_p \in \mathbf{V_{k,k\in i,j}}\})$ [1]
	15 - 28	$f(\{\langle \lvert present((s+e)/2)\rvert / max\_present(\mathbf{MT})\rangle_p \lvert v_p \in \mathbf{V_{k,k\in i,j}}\})$ [1]
	29 - 42	$f(\{\langle \lvert present(e)\rvert / max\_present(\mathbf{MT})\rangle_p \lvert v_p \in \mathbf{V_{k,k\in i,j}}\})$ [1]

[1] This feature is generated for both devices, leading to a pair of values. In order to make the order of those values independent from the order of the devices, the actual features are their maximum and minimum values.

4.5 Utilization Phase

After the regressor has been trained, it can be used to predict the social tie strengths between each pair of devices from their mobility features, and those tie strengths can be used to determine the aggregate social connectedness score of the location. In this phase only mobility features are used and the actual ground truth indicator of social ties are not. While such ground truth is available in a special case for a dataset collected using Wi-Fi scanning, in other mobility datasets (e.g. GPS, Cellular networks) it is not and can only be collected in a small scale (e.g. through surveying visitors). Therefore, a utilization phase without such ground truth indicator is a valid approach. Algorithm 3 provided in the appendix shows the utilization phase in detail.

5 Evaluation

In this section, we present the result of two experiments to validate our method. Firstly, we evaluate the accuracy in prediction of social tie strengths using a dataset generated by Wi-Fi scanners in our university campus since the start of 2016 (both MAC addresses and SSIDs are anonymized through secure hashing and visitors are provided with an opt-out list). Secondly, we synthetically modified this dataset to analyze sensitivity of our method to various factors that degrade the quality of such datasets.

For each of the experiments we applied 10 fold cross-validation, by training both the proposed method and a baseline method and generating pairwise social ties as output. The generated score by the algorithms is then compared to the ground truth indicator acquired from SSIDs. The indicator of accuracy in these experiments is the coefficient of determination, which is defined as $R^2 = 1 - \frac{\sum_i (f_i - y_i)^2}{\sum_i (y_i - \bar{y})^2}$, in which y_i is the ith predicted result and f_i is the ith observed result. This metric describes which portion of the variation in the actual social tie strengths is explained by the predicted social tie strengths. The reason for choosing the metric was that we found that a high proportion of social ties measured are weak and relatively a much smaller proportion are strong. Compared to other alternative metrics (e.g. Mean squared error), this metric is not biased towards such unbalance in proportion of weak and strong social ties. The reasoning behind weak ties is better depicted by Figure 2 (a). This figure shows how the MAC addresses are distributed among the top 25 SSIDs. As seen, a large set of devices share one SSID. Plugging this number to Psim-3 indicator mentioned in Section 4.1 results in a week tie between all of these devices. Whereas, a much smaller number of devices share rare SSIDs leading to stronger ties.

Choosing a baseline: As mentioned before, none of the previous research has considered extracting social tie information from mobility data acquired of a single location. However, among the features used in previous research considering visits to multiple locations, the overlap feature calculated through measuring *co-occurrence probability* is the one that can also be calculated from a single location [10, 11]. Therefore, as our baseline, we trained the regressor using this feature. For a pair of devices i and j this feature is calculated as $(\sum length(\mathbf{O}_{i,j}))/\sqrt{(\sum length(\mathbf{V_i}) * \sum length(\mathbf{V_j}))}$.

Table 3: Wi-Fi dataset statistics

Statistic	**Value**
Number of locations (scanner)	20
Data collection period	260 days
Number of unique MAC addresses	2,790,703
Number of non-random unique MAC addresses	281,562
Number of probes collected	130,279,931
Number of probes collected from non-random sources	126,807,946

5.1 Wi-Fi dataset

In this section, we evaluate the accuracy of our proposed method in predicting SSID-derived social tie strengths, and compare it to the baseline's accuracy. Table 3 describes various features of this dataset. For each location, the timestamps, scanner IDs and anonymized MAC addresses are used to create the mobility trace, after the Organizationally Unique Identifier (OUI) field has been used to filter out randomized MAC addresses[3] (by examining the OUI field of addresses [16]). The anonymized SSIDs are

[3]This anonymization approach is taken by recent mobile phone operating systems.

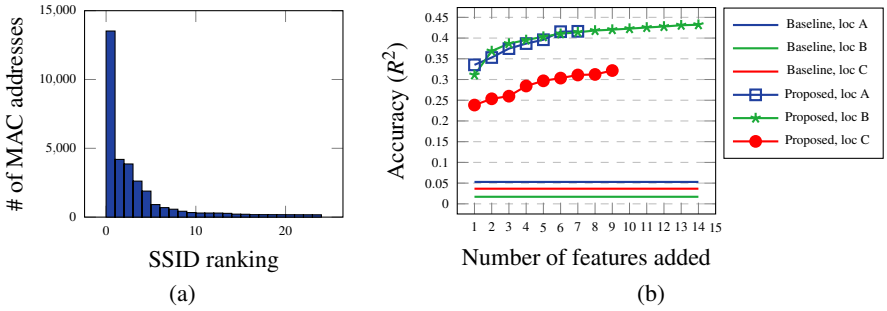

Fig. 2: (a) MAC addresses per SSID (b) Progression of performance by adding features

solely used to infer the ground truth indicator of social tie strength between devices. As previously described in Section 4.1, we use the Psim-3 metric for this purpose.

Progression of the algorithm: We initially present the results on three locations (denoted by A-C) to demonstrate how the algorithm works. Figure 2 (b) shows the performance of the methods during the progression of the feature selection algorithm (Algorithm 2). The algorithm keeps adding features and stops when no performance increase is observed. As seen, the proposed method reaches its optimal performance by using 7 - 14 features, reaching a coefficient of determination between approximately 0.3 and 0.45. Using a single feature the performance of the baseline is a constant value and significantly lower.

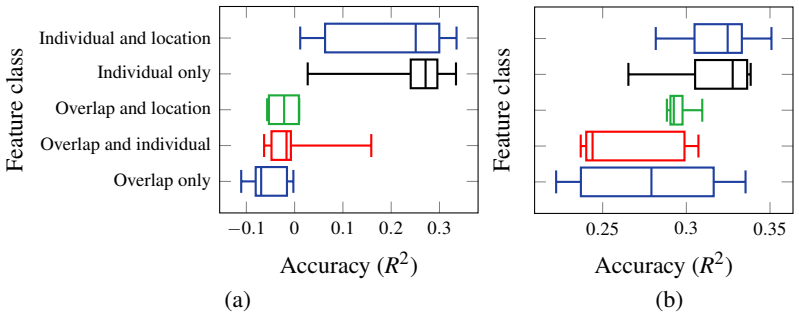

Fig. 3: Performance in location A after using (a) a single feature and (b) using two features

Figure 3 (a) shows the distribution of performances of the regressors generated during the first round of the feature selection for location A (within Algorithm 2). The figure shows that in the first round features based on individual mobility patterns outperform those based on overlapping visits.This result is interesting as it is not intuitively expected that individual mobility features represent social ties. One possible reason for this is that **Individual** features act very strong in determining the social tie indicator of devices that are always present in the location. Examples of these would be stationary Wi-Fi enabled devices such as access points, and printers that have a social tie strength

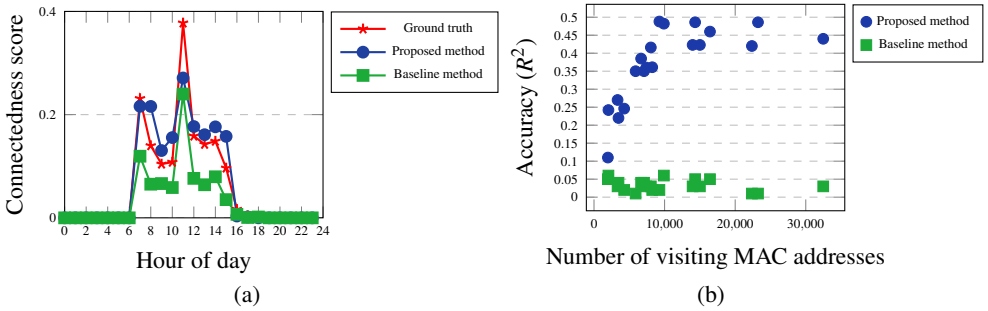

Fig. 4: (a) Aggregate connected score of one location (b) Accuracy of prediction over 20 locations

indicator close to zero. Figure 3 (b) shows the same distributions for the second feature selection round, in which each regressor uses the feature from the previous round and a newly selected feature. The main difference with the previous graph is that the feature classes related to overlap seem to improve their performance. This is consistent with the idea that once static devices without real social ties are filtered out, overlap features could indicate which social ties are stronger. Finally, Figure 4 (a) shows how aggregate social connectedness score can be extracted from one of the locations over time. As seen, compared to the baseline method the score calculated using the proposed method is closer to the actual score acquired from the indicator acquired from SSIDs.

Results on the complete dataset: In order to compare the algorithm's performance for multiple locations, we trained regressors for all 20 locations separately for the best performing feature set on the previous locations (each dot represents results on a different location). Figure 4 (b) shows the resulting scores, set out against the number of unique MAC addresses that visited the location. The figure shows a strong correlation between location popularity and the accuracy of the proposed method, but not with the accuracy of the baseline method. It also shows that the regressor performs well when a static feature set is used, instead of using the feature selection algorithm.

5.2 Sensitivity test

In order to determine how sensitive our method is to different levels of uncertainties (noise, variability of probe request frequency, etc.), in this section we perform evaluations for several modifications of the original dataset. The modified datasets are generated by applying following adjustments to the original mobility trace **MT**:

- **Adjustment 1**: Removes probe requests received from each device by decreasing the probe frequency. For example, 50% of probe requests are removed by removing every second probe, and 75% is removed by only retaining the first, fifth, ninth, etcetera probe. This adjustment reflects an environment in which devices consistently broadcast fewer probe requests. This could be caused by different implementations of the 802.11 protocol leading to different probing frequency.
- **Adjustment 2**: Removes samples before supplying them to the regressor. This change reflects a decrease of data available to the method, which could be caused

by running the initialization phase for a lower amount of time, by placing the scanner in a location where few people gather, or by a larger number of people enabling MAC address randomization.

– **Adjustment 3**: Removes probe requests from each device randomly. Each probe request is removed in a random pattern. This adjustment simulates a situation in which fewer probe requests are received. This could be caused by a noisy environment, by using scanners that are more susceptible to noise, or by making use of communication technology or protocols with less reliable transmission.

For each adjustment, a new set of samples was generated by applying the same process as the one applied for the original dataset, after which the same feature selection algorithm was used. Figures 5 (a-c) show the final performance of the proposed method, which is the accuracy of the regressor after the feature selection algorithm has completed. Each figure shows this performance for different degrees of one of the three adjustment types, for each of the locations. For example, figure 5 (a) shows the proposed and baseline performances when 0% (unadjusted), 50%, 75%, etc. of the probes have been removed by decreasing frequency. As expected, the proposed method performs better than the baseline in every case. Also, higher degrees of the adjustment decreases the performance of both methods in nearly every case.

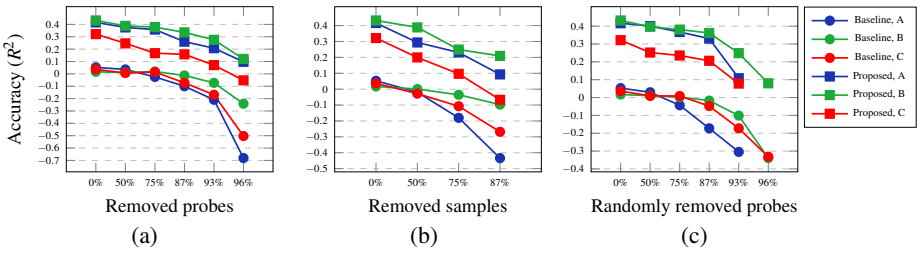

Fig. 5: Performance in different degrees of (a) adjustment 1 (b) adjustment 2 (c) adjustment 3

6 Conclusion and Future Work

In this study, we proposed a new method to characterize the social connectedness of a location based on the mobility traces acquired from it. The proposed method works by extracting and choosing from a large subset of mobility features. This method was evaluated on a real-world dataset and a number of modified versions of that dataset, in order to determine the method's sensitivity to various parameters that degrade the quality of such datasets. Our results show that it is possible to characterize the social connectedness of a location from the presence pattern of its visitors. While our evaluations were performed using a Wi-Fi dataset, the proposed method is also applicable to other types of mobility data (e.g. check-in records in location-based social networks and GPS tracks). Our future research entails studying the relationship between social connectedness and location attributes such as category of the location.

Appendix A

In this section, we provide algorithmic details on the procedures of the initialization and utilization phase presented before in Sections 4.4 and 4.5.

Algorithm 1: Initialization Phase, Sample Generation

```
   Data: <mobility trace MT, social tie strengths ST>
   Result: collection of samples SC
1  SC = [];
2  D = determineDevices(MT);
3  DP = computePairs(D);
4  forall pair in DP do
5  |  MF = computeAllMobilityFeatures(pair, MT);
   |  /* set of values of all mobility features for this pair  */
6  |  st = ST[pair]; /* the pairwise social tie strength        */
7  |  append(SC,⟨st, MF⟩);
8  return SC;
```

Initialization phase: Algorithm 1 shows the pseudo code for the first part of the initialization phase, in which the samples are generated. The inputs of this algorithm are the mobility trace (a collection of detections with form $\langle d,t \rangle$, with d being a device and t being a timestamp) and a collection of pairwise social tie strengths between some of the devices in the mobility trace. Its output is a collection of samples, each of which is a tuple $\langle st, \mathbf{MF} \rangle$, with st being a pairwise social tie strength for some pair of devices, and **MF** being the set of values of all the mobility features for the same device pair, calculated in lines 4-7. Algorithm 2 shows the pseudo code for feature selection and learning a regressor. The input of this algorithm is the collection of samples generated in Algorithm 1 and its output is a regressor trained using the combination of features as selected during feature selection. After computing the mobility features for each pair of devices, we need to determine which features should be supplied to the regressor. The feature selection algorithm performs as follows [15]. Initially the set of selected features is empty. The algorithm moves through the search space in a greedy manner by evaluating features (lines 10-21) and it halts when no new features improve the regression performance (line 18). The performance of the regressor is evaluated using 10-folded cross validation and determining the average of their mean squared errors (line 15). Once the features are selected we proceed to learning the regressor (line 22).

Utilization phase: Algorithm 3 shows the pseudo code for the utilization phase. The algorithm takes the regressor generated in Algorithm 2, the mobility trace, and a timestamp and outputs the aggregate social connectedness score for this mobility trace at that specific timestamp. The algorithm first determines which devices were present at the location at the specific moment in time (line 2). It again uses device timestamps to determine visit starts and ends. After doing so, the method determines the value of mobility features that were given by the feature selection algorithm for each pair of devices present (line 5). Then, these feature values are supplied to the regressor, which predicts the tie strength for each pair (line 6). Finally, the tie strengths are averaged in order to obtain a score of aggregate social connectedness (line 8).

Algorithm 2: Initialization Phase, Feature Selection

```
   Data: collection of samples SC
   Result: regressor r
 1 FRC = range(0, length(SC[0])); /* range of all feature indices    */
 2 FIC = []; /* current best feature indices overall               */
 3 BTFIC = []; /* best feature indices for the current round       */
 4 s = 0; /* overall best score overall                            */
 5 bts = 0; /* best score for the current round                    */
 6 sib = true; /* boolean indicating whether score has improved    */
 7 ftb = true; /* boolean indicating generation of best score      */
 8 frb = true; /* boolean indicating completion of first round     */
 9 while sib do
10     forall index in FRC do
11         if index in FIC then
12             continue;
13         TFIC = union(FIC, [index]); /* feature indices to test  */
14         TS = selectByIndices(SC, TFIC); /* samples to test      */
15         ts = 10FoldCrossValidateRegressor(TS); /* score from test */
16         if ts >bts or ftb then
17             bts = ts,BTFIC = TFIC,ftb = false;
18     if bts >s or frb then
19         s = bts, FIC = BTFIC,frb = false;
20     else
21         sib = false;
22 r = trainRegressor(SC, FIC); return r;
```

Algorithm 3: Utilization Phase

```
   Data: <regressor r, mobility trace MT, timestamp t>
   Result: aggregate social connectedness score as
 1 D = computePresentDevices(MT, t);
 2 DP = computePairs(D);
 3 PSC = [];
 4 forall pair in DP do
 5     MF = computeMobilityFeatures(pair, MT);
 6     ps = predictTieStrength(regressor, MF);
 7     append(PSC,ps);
 8 as = computeAggregateTieStrength(PSC);
 9 return as;
```

Bibliography

[1] Seeman, T.E.: Social ties and health: The benefits of social integration. Annals of epidemiology 6(5), 442–451 (1996)

[2] Kawachi, I., Berkman, L.F.: Social ties and mental health. Journal of Urban health 78(3), 458–467 (2001)

[3] Jylhä, M., Aro, S.: Social ties and survival among the elderly in tampere, finland. International Journal of Epidemiology 18(1), 158–164 (1989)

[4] Baratchi, M., Heijenk, G., van Steen, M.: Spaceprint: a mobility-based fingerprinting scheme for public spaces. arXiv preprint arXiv:1703.09962 (2017)

[5] Petre, A.C., Chilipirea, C., Baratchi, M., Dobre, C., van Steen, M., WiFi tracking of pedestrian behavior, in Smart Sensors Networks: Communication Technologies and Intelligent Applications, Elsevier (2017)

[6] Cunche, M., Kaafar, M.A., Boreli, R.: Linking wireless devices using information contained in wi-fi probe requests. Pervasive and Mobile Computing 11, 56–69 (2014)

[7] Barbera, M.V., Epasto, A., Mei, A., Perta, V.C.: Signals from the crowd: uncovering social relationships through smartphone probes. In: Proceedings of the 2013 conference on Internet measurement conference. pp. 265–276. ACM (2013)

[8] McPherson, M., Smith-Lovin, L., Cook, J.M.: Birds of a feather: Homophily in social networks. Annual review of sociology pp. 415–444 (2001)

[9] Mashhadi, A., Vanderhulst, G., Acer, U.G., Kawsar, F.: An autonomous reputation framework for physical locations based on wifi signals. In: Proceedings of the 2nd workshop on Workshop on Physical Analytics. pp. 43–46. ACM (2015)

[10] Cheng, N., Mohapatra, P., Cunche, M., Kaafar, M.A., Boreli, R.: Inferring user relationship from hidden information in wlans. In: MILCOM 2012-2012 IEEE Military Communications Conference. pp. 1–6. IEEE (2012)

[11] Scellato, S., Noulas, A., Mascolo, C.: Exploiting place features in link prediction on location-based social networks. In: Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 1046–1054. ACM (2011)

[12] Eagle, N., Pentland, A.S., Lazer, D.: Inferring friendship network structure by using mobile phone data. Proceedings of the national academy of sciences 106(36), 15274–15278 (2009)

[13] Baratchi, M., Meratnia, N., Havinga, P.J.M.: On the use of mobility data for discovery and description of social ties. In: Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. pp. 1229–1236. ASONAM '13, ACM, New York, NY, USA (2013), `http://doi.acm.org/10.1145/2492517.2500263`

[14] Di Luzio, A., Mei, A., Stefa, J.: Mind your probes: De-anonymization of large crowds through smartphone wifi probe requests. In: Computer Communications, IEEE INFOCOM 2016-The 35th Annual IEEE International Conference on. pp. 1–9. IEEE (2016)

[15] Blum, A.L., Langley, P.: Selection of relevant features and examples in machine learning. Artificial intelligence 97(1), 245–271 (1997)
[16] Misra, B.: ios8 mac randomization analyzed! http://blog.mojonetworks.com/ios8-mac-randomization-analyzed/ (2014), [Online; accessed 21-November-2016]