2017 IEEE 37th International Conference on Distributed Computing Systems

Smartphone Privacy Leakage of Social Relationships and Demographics from Surrounding Access Points

Chen Wang*, Chuyu Wang*†, Yingying Chen*, Lei Xie† and Sanglu Lu†
*Department of Electrical and Computer Engineering
Stevens Institute of Technology, Hoboken, NJ, USA
{cwang42, yingying.chen}@stevens.edu
†State Key Laboratory for Novel Software Technology
Nanjing University, Nanjing, Jiangsu, China
wangcyu217@dislab.nju.edu.cn, {lxie, sanglu}@nju.edu.cn

***Abstract*—While the mobile users enjoy the anytime anywhere Internet access by connecting their mobile devices through Wi-Fi services, the increasing deployment of access points (APs) have raised a number of privacy concerns. This paper explores the potential of smartphone privacy leakage caused by surrounding APs. In particular, we study to what extent the users' personal information such as social relationships and demographics could be revealed leveraging simple signal information from APs without examining the Wi-Fi traffic. Our approach utilizes users' activities at daily visited places derived from the surrounding APs to infer users' social interactions and individual behaviors. Furthermore, we develop two new mechanisms: the *Closeness-based Social Relationships Inference* algorithm captures how closely people interact with each other by evaluating their physical closeness and derives fine-grained social relationships, whereas the *Behavior-based Demographics Inference* method differentiates various individual behaviors via the extracted activity features (e.g., activeness and time slots) at each daily place to reveal users' demographics. Extensive experiments conducted with 21 participants' real daily life including 257 different places in three cities over a 6-month period demonstrate that the simple signal information from surrounding APs have a high potential to reveal people's social relationships and infer demographics with an over 90% accuracy when using our approach.**

I. Introduction

Wi-Fi networks are becoming increasingly pervasive, to the point where public Wi-Fi access is readily in place in numerous cities [1]. And the number of public Wi-Fi Access Points (APs) is expected to hit 340 million globally by 2018, resulting in one public Wi-Fi AP for every twenty people worldwide [2]. More commonly, retail stores, offices, universities and homes are usually Wi-Fi enabled for providing high bandwidth and cost-effective connectivity to the Internet for the mobile users. While the mobile users enjoy the anytime anywhere Internet access by connecting their mobile devices (e.g., smartphones) to the Wi-Fi networks, the surrounding APs have raised a number of privacy concerns. For example, mobile users could be located and tracked based on the ubiquitous APs, such as using Google location service [3].

In this work, we study the potential of privacy leakage caused by surrounding APs and explore to what extent the personal information, in particular users' social relationships and demographics, could be derived. Prior work in demographics inference based on Wi-Fi network mainly rely on the context information obtained from passively sniffed users' Wi-Fi traffic [4], [5]. For example, Cheng *et al.* examine users' Internet browsing activities by collecting their in-the-air traffic in public hotspots [4], whereas Huaxin *et al.* infer user demographic information by passively sniffing the Wi-Fi traffic meta-data [5]. These methods need to examine the Wi-Fi traffic and are thus not scalable to large number of users due to the high deployment overhead involved. Existing work in social relationships inference primarily depend on the encounter events detected by either bluetooth [6], Wi-Fi SSID list [7], or GPS locations [8]. These approaches can only perform coarse-grained social relationships inference by examining whether users have interactions or not instead of studying users' behaviors and how closely they interact with each other. They can neither provide fine-grained social relationships (such as advisor-student, colleagues, friends, husband-wife, neighbors) nor identify specific role of the user in the relationship.

It is known that GPS, motion sensors and contact lists on mobile devices can exhibit privacy, but how much a user's privacy could be leaked from the ubiquitous access points is unclear. In this work, we demonstrate that by examining the simple signal features of the surrounding APs it is possible to infer users' fine-grained social relationships and demographics without sniffing any Wi-Fi traffic. Specifically, the availability of surrounding Wi-Fi APs is periodically scanned by mobile devices because of their default systems purpose to optimize network service via continuously seeking better Wi-Fi signals and remembered APs [9], [10] and accessing such information only requires a common permission, which is considered with low risk [11]. Signal features such as the time-series of BSSIDs (i.e. MAC addresses) and Received Signal Strength (RSS) are then extracted from these scanned APs and analyzed to derive users' activities at daily visited places. Our system exploits the rich information of users' daily interactions and

1063-6927/17 $31.00 © 2017 IEEE
DOI 10.1109/ICDCS.2017.139

behaviors embedded in these derived activities and discloses fine-grained social relationships (including advisor-student, supervisor-employee, colleagues, friends, husband-wife and neighbors) as well as demographic information (such as occupation, gender, religion, marital status).

Our approach of using simple signal features of APs can be easily applied to a large number of users. For example, advertisers or third party companies could mine users' personal information for targeted advertising or recommending services. However, such an approach could cause significant privacy leakage if it is utilized by advertisers with aggressive business attempts, who could simply publish free apps to users while these free apps actively collect users' surrounding AP information and send back to the server to derive users' social relationships and demographics.

In particular, we describe people's daily places in three dimensions (i.e. temporal, spatial and contextual) to infer people's *activities* at each place. For users performing activities at the same place, we calculate *physical closeness* of the users (e.g., whether staying at the same room, adjacent rooms or inside the same building) and extract users' activeness (e.g., walking around or sitting) together with other features (e.g., time slots and duration) to characterize their activities at daily places. We then develop *Closeness-based Social Relationships Inference* algorithm to capture where, when and how closely people interact to derive fine-grained social relationships. We design *Behavior-based Demographics Inference* method to capture individual behavior based on users' various daily activities to reveal demographic information including occupation, gender, religion and marriage. We conduct extensive experiments with 21 participants carrying their smartphones to collect surrounding Wi-Fi AP information in their real daily life across three cities over 6 months and study to what extent we can derive these participants' social relationships and demographic information.

We summarize our main contributions as follows:

- We demonstrate that simple signal information (e.g., time-series of MAC addresses and RSS) from users' surrounding Wi-Fi APs can reveal private information including both social relationships and demographics.
- We develop statistical methods to detect and characterize users' daily visited places based on the AP signal information and further infer the context of daily places by deriving users' activity features (e.g., activeness, time slots and duration)
- We design closeness-based social relationships inference algorithm to analyze when, where and how closely users interact with each other and reveal users' detailed social relationships (e.g., advisor-student, supervisor-employee, colleagues, friends, husband-wife, customer relationship and neighbors).
- We further abstract people's various behaviors (e.g., home, working and leisure behaviors) to infer their demographic information such as occupation, gender, religion, and marital status.
- We show with experimental study of 21 participants that by using our system one can achieve over 91% accuracy of inferring social relationships and over 90% accuracy of deriving demographic information via examining the simple signal features from surrounding APs.

II. Related Work

In this work, we aim to understand the privacy leakage of smartphone users, in particular discovering users' social relationships and demographics, by analyzing only the availability of surrounding APs without sniffing any Wi-Fi traffic. Obtaining such information requires limited permission other than turning on GPS or accessing to contact lists. Our work is related to the research efforts in using various information collected from Wi-Fi network and/or smartphone for meaningful places extraction [12]–[15], social relationships inference [6], [7], [16]–[18], and demographics derivation [4], [5], [19].

As the contextual location can be used for learning the person's interest and providing content-aware applications, there have been active studies on extracting contextual meaning of the locations people visited. For example, Kang *et al.* design a cluster-based method to extract meaningful places from traces of location coordinates collected from GPS and Wi-Fi based indoor location system [12]. Kim *et al.* propose SensLoc that utilizes a combination of acceleration, Wi-Fi, and GPS sensors to find semantic places, detect user movements, and track travel paths [13]. These existing methods however only focus on individual users' visited locations without analyzing the interactions between them. Besides, the obtained meaningful places may be not sufficient to infer the higher level personal information, such as fine-grained social relationship and demographics, due to the lack of information about the users' daily behaviors and social interactions.

Information in Wi-Fi networks and smartphones have been used in literature to infer users' social relationships. For example, Wiese *et. al* [16] use the smartphone contact list to mine personal relationships. Moreover, the similarity of smartphones' SSID lists is used to reveal users' social relationships [7]. These methods can only derive coarse-grained social relationships without analyzing the behaviors and interactions among people. Vicinity detection via Bluetooth or Wi-Fi signals opens opportunities for social interaction analysis and the strength of friendship ties can be inferred from such wireless signals [6], [18]. However, these vicinity detection methods only consider the relative interaction between people without interaction context (e.g., place context and behaviors). They are unable to differentiate the specific type of various social relationships, such as family members and friends. Our previous work focuses on extracting the social relationship from smartphone App leaked information such as GPS location, IMEI and network location [20]. It could only derive the social relationships in a coarse-grained manner. In this paper, we take a closer look and study the privacy leakage just from the surrounding APs and derive people's activities and various closeness levels of social interactions for inferring detailed relationships demographic information.

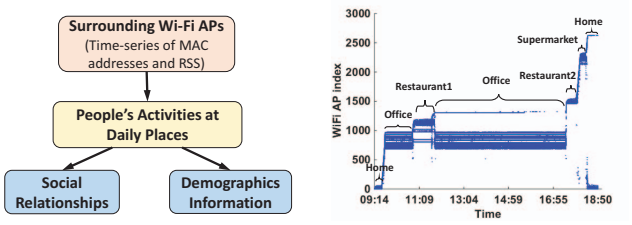
More recently, Wi-Fi traffic monitoring and smartphone Apps have been used to infer users' demographic information. For example, Cheng *et al.* examine the user's Internet browsing activities (e.g., domain name querying, web browsing) by collecting their Wi-Fi traffic in public hotspots [4]. They are able to reveal the travelers' identities, locations or social privacy. Huaxin *et al.* design an approach to infer user demographic information by sniffing the Wi-Fi traffic meta-data [5]. Seneviratne *et al.* design a system to predict various user traits by analyzing the snapshot of installed Apps [19]. Different from the above work, we study the capability of examining the simple signal information of surrounding APs to derive demographic information without sniffing any Wi-Fi traffic or examining the installed Apps.

III. System Design

A. Preliminaries

Environment-Behavior research reveals that an individual's activities such as work-related, household and leisure activities are related to the places they visit [21]. And such activities at daily visited places can be analyzed and mined to infer users' personal information such as social relationships and demographics [22]. Thus by leveraging the users' activities at daily places as a bridge, we could start from the non-contextual surrounding AP information to infer users' social relationships and demographics. This connection is depicted in Figure 1(a). The surrounding Wi-Fi APs reflect users' surrounding wireless environments, which can be utilized to determine users' daily visited places and activities. The *daily places* in our work refer to the abstract locations that users visit in their daily lives, such as home, workplace, restaurants, stores and churches. By analyzing users' activities at daily places, we could derive the social interactions between users and abstract individual's behavior. Such information is then further utilized to mine users' social relationships and demographics. Note that contrary to the existing work in social relationships and demographics inference, we only utilize the availability of surrounding APs' simple signal information without requiring to sniff any Wi-Fi traffic contents.

To study how the surrounding APs can be utilized to detect a user's daily places and activities, we conduct preliminary experiments by recording the APs on the user's smartphone at the regular rate of one scan per 15 seconds, because a Wi-Fi device usually scans every 5 - 15 seconds for providing the user non-interrupted Wi-Fi connection to cope with the user's place change [23], [24]. Figure 1(b) shows the recorded time-series of a user's surrounding APs (differentiated by BSSIDs) for one day, as well as the groundtruth of visited places. As the AP index is assigned to each unique AP in sequence, the later observed AP has larger index. The observation is that the detected AP lists have large overlaps when the user stays at the same place, while the AP lists are distinct when the user moves to a different daily place. This suggests that we may utilize the changes of the observed AP list to detect the user's daily visited places as well as the entrance/departure time and the staying duration. Moreover, the user's activities at daily places (e.g., the user's mobility at work and during



(a) Connection from surrounding APs to social relationships & demographics.

(b) Illustration of observed APs by a user's smartphone in one day.

Fig. 1. Preliminary studies.

leisure time) can be derived to reflect individual demographics. Furthermore, we observe that the same place or the places in the neighborhoods may share some APs (e.g., office and restaurant 1). Their physical closeness may be obtained by checking how many surrounding APs they share, which is useful for analyzing social interactions.

B. Challenges

Robust Daily Places and Activity Detection Using APs. Lacking the pre-knowledge of AP deployment, the accurate and robust detection of daily places and activities from ubiquitous APs is challenging. And the ubiquitous unstable and mobile APs even add to the difficulties. Additionally, the daily places need to be abstracted with sufficient spatial resolution (e.g., differentiating rooms and floors) for further deriving users' mobility and their physical closeness during interaction.

Determining the Context of Daily Places. Deriving the context of a user's daily visited places from the non-contextual AP signal information is challenging. Moreover, a place may exhibit different contexts to different users. For example, stores are leisure places to most people but the workplace to the store staff. This requires us to search for the deep implication behind the individual's activities at the place instead of relying on traditional place context based on the place function.

Fine-grained Social Relationships Inference. Fine-grained relationships inference needs the information on not only who have interactions but also on how closely they interact. Our systems needs to have the capability to define multiple closenesses between users. Furthermore, specifying the role of each user in a relationship (e.g., husband or wife) may needs the assistance from demographic information (e.g., gender).

Demography Inference without Context. Inferring a user's demographics with non-contextual simple signal information of surrounding APs is challenging. Different from the previous work relying on the content obtained from monitoring the Wi-Fi traffic, our system explores the possibility to abstract users' behaviors based on their various activities at daily places for demographic inference.

C. System Overview

The basic idea of our system is to analyze users' activities at daily routine-based places that are derived from users' surrounding APs for fine-grained social relationships and demographics inference. The proposed system takes as inputs the information of users' surrounding APs perceived by their smartphones at each scan, including the list of AP MAC

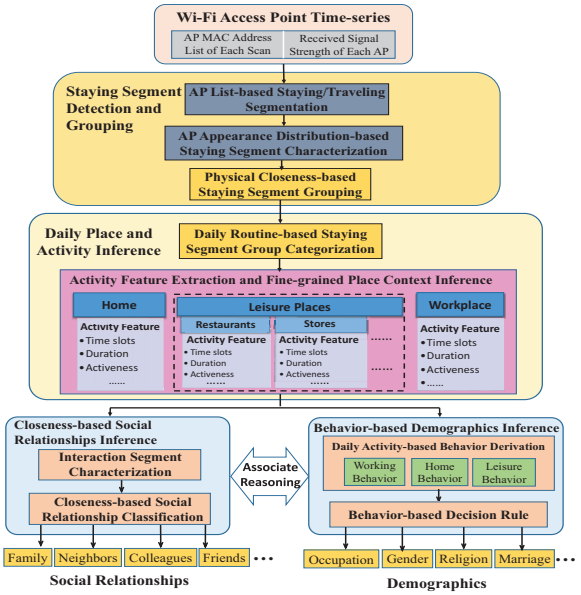


Fig. 2. Wi-Fi AP distribution-based social relationships and demographics inference framework.

addresses and RSS, to infer fine-grained social relationships and demographics. Figure 2 presents our system flow.

First, the *Staying Segment Detection and Grouping* component detects and characterizes users' daily visited places in three steps. *AP List-based Staying/Traveling Segmentation* analyzes the overlap of the AP lists over consecutive scans and divides the time-series into staying and traveling periods. *Staying Segment Characterization* estimates the significance of each surrounding AP by calculating its appearance rate within the staying segment. It then categorizes the APs by their significance to describe the spatial information of each staying segment. The spatially close-by staying segments are then grouped together as one unique place by using *Closeness-based Staying Segment Grouping*.

The next component is to derive the activities at daily places which is an important building block of social relationships and demographics inference. It is carried out by using *Daily Place and Activity Inference*, which involves *Daily Routine-based Staying Segment Group Categorization* and *Daily Activity Feature Extraction and Fine-grained Place Context Inference*. *Daily Routine-based Staying Segment Categorization* classifies the grouped staying segments (i.e. unique places) into three contextual categories (i.e. home, leisure and workplace) based on people's daily routines. At last, *Daily Activity Feature Extraction and Fine-grained Place Context Inference* derives people's activity features including the staying time slots, duration and activeness and assigns detailed contextual information to these places by leveraging the derived activity features and geo-information, such as restaurants or stores in leisure places, campus or office buildings in workplaces.

Finally, our system infers users' social relationships and demographics based on the derived activities at daily places. In particular, it first calculates the physical closenesses of the interactions between users. It then uses *Interaction Segment Characterization* and *Closeness-based Social Relationships Classification* to infer when, where and how closely people interact with each other for inferring their possible relationships such as family, neighbors, colleagues, and friends. To derive a user's demographics, *Behavior-based Demographics Inference* applies *Daily Activity-based Behavior Derivation* to abstract people's various behaviors including working behaviors, home behaviors and leisure behaviors, based on the activities at daily places. It then utilizes *Behavior-based Decision Rule* to infer users' demographic information (e.g., occupation, gender, marriage and religion) based on the behavior abstraction. At last, the *Associate Reasoning* can be applied to social relationships and demographics to improve the accuracy of inference results, such as identifying the specific role of the user in a relationship (e.g., husband-wife and advisor-student).

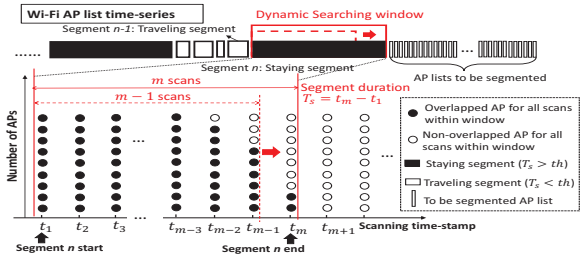


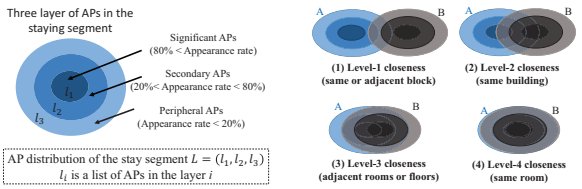
Fig. 3. Staying/traveling segmentation leveraging dynamic searching windows to analyze the overlapped AP lists over consecutive scans.

IV. Staying Segment Group Detection and Characterization

A. AP List-based Staying/Traveling Segmentation

As observed in the preliminary study of Figure 1(b), the discovered AP BSSID lists of consecutive scans have large overlaps when the user stays at the same place, while the similarity of the AP lists is rapidly diminished when the user moves to a different place. We thus take the advantage of the AP list similarity (i.e. BSSID list similarity) in consecutive scans to detect the staying and traveling segments. We define *staying segment* as the Wi-Fi AP-list time-series segment that captures the temporal and spatial information when the user stays at a location. And we analyze the overlap of the AP lists within a dynamic searching window of consecutive scans to perform staying segmentation.

In particular, Figure 3 illustrates the proposed AP List-based Staying/Traveling Segmentation in identifying the staying segment n. The dynamic searching window starts at t_1 and iteratively expands to the next scan. In each iteration, we analyze the overlapped APs of all the scans within the searching window. The number of solid dots at each scanning time $t_i (i = 1, 2, \dots)$ indicates the number of overlapped APs that are found within the window from t_1 to t_i. When the searching window iteratively expands to the next scan, the number of overlapped APs may decrease. When no overlapped AP is found in the expanded searching window (e.g., the window from t_1 to t_m), such searching window is identified as one possible staying segment. We note that because it may take several scans to travel out of an AP's range, this approach can



(a) Appearance rates and significance of the APs in a staying segment.

(b) Four kinds of closeness between staying segments A and B.

Fig. 4. AP appearance rate distribution-based staying segment characterization.

detect short staying segments even when the user is traveling. We next check whether the segment duration $T_s = t_m - t_1$ is greater than a threshold τ (e.g., $\tau = 6$ minutes) to further confirm valid staying segments and filter out the false staying segments. Meanwhile, the user's entrance/departure time and corresponding staying duration could also be obtained.

B. AP Appearance Rate Distribution-based Staying Segment Characterization

We next characterize the visited places by deriving Wi-Fi AP appearance distribution in the detected staying segments. The discovered AP BSSID list can be used to describe the wireless environment of the user in the staying segment. However, not all the APs have the same significance for characterizing the spatial information. Some APs may appear only in a few scans due to weak Wi-Fi signals, while others are more stable and appear almost in every scan. We calculate the *appearance rate* of each discovered AP to represent its significance, and then classify the APs into different categories based on their significance. In particular, the *appearance rate* of an AP is defined as $R = \frac{N_a}{N}$, where N_a is the appearance number of this AP and N is the total number of scans in the detected staying segment. The appearance rates together with BSSIDs of the discovered APs are used to characterize the spatial information of the staying segment, which has the potential to both differentiate places with good resolution but also measure people's physical closeness.

We empirically divide the APs of a staying segment into three layers $l_i, i = 1,2,3$ (i.e. lists of significant APs, secondary APs and peripheral APs) according to their appearance rate. As shown in Figure 4(a), the significant APs are those with appearance rate larger than 80%, the peripheral APs are the ones with the appearance rate less than 20%, and the rest of APs are secondary APs. Then the spatial information of the staying segment can be characterized by *AP set vector* $L = (l_1, l_2, l_3)$, which can tolerate the noise generated by the unstable APs, mobile APs or even missing AP scans.

C. Estimating Physical Closeness between Staying Segments

Measuring the physical closeness between different users' staying segments can capture how closely people interact with each other. It can also be used to group the same user's staying segments that are close to each other as one place. In particular, we leverage the AP set vector to measure the physical closeness between staying segments. Given two staying segments A and B and their AP set vectors L_A and L_B, we calculate the *closeness matrix* M as follows:

$$M = L_A^{-1} L_B = \begin{pmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \\ r_{31} & r_{32} & r_{33} \end{pmatrix}, \tag{1}$$

where r_{ij} is the overlapping rate between subsets l_{Ai} and l_{Bi} of AP set vectors L_A and L_B, respectively. The overlapping rate r_{ij} can be obtained by

$$r_{ij} = \frac{OverlapApNum(l_{Ai}, l_{Bj})}{min(Num(l_{Ai}), Num(l_{Bj}))}, i, j = 1,2,3. \tag{2}$$

Based on the statistical analysis with 431 staying segments collected from 167 places in 3 cities, we empirically quantify the physical closeness expressed by the closeness matrix M into five levels:

$$\begin{cases} C_0 = \{M : \sum_{i,j=1}^{3} r_{ij} = 0\}; & (Completely\ \ separated) \\ C_1 = \{M : r_{33} > 0\ and\ \sum_{i,j=1}^{3} r_{ij} - r_{33} = 0\}; & (Same\ \ street\ \ block) \\ C_2 = \{M : \sum_{i,j=1}^{3} r_{ij} - r_{33} - r_{11} > 0\ and\ r_{11} = 0\}; & (Same\ \ building) \\ C_3 = \{M : 0 < r_{11} < 0.6\}; & (Adjacent\ \ rooms) \\ C_4 = \{M : r_{11} \geq 0.6\}, & (Same\ \ room) \end{cases} \tag{3}$$

where C_1, C_2, C_3, C_4 are four mutually exclusive closeness sets with increasing closeness level as shown in Figure 4(b), representing the same street block, the same building, the adjacent rooms and the same room respectively. $C_0 = \overline{C_1 \cup C_2 \cup C_3 \cup C_4}$ means two staying segments are completely separated. We use level-i closeness to express closeness in set C_i.

D. Physical Closeness-based Staying Segments Grouping

We note that the same user's multiple staying segments may correspond to the same place as the user may pay multiple revisits. We thus combine these staying segments together by checking whether there is level-4 closeness between them and keep all the time slots. The grouped staying segments represent non-redundant places visited by the user and contains the user's activities. We can then characterize the user's activities at each unique place.

V. Daily Place and Activity Inference

In this section, we explore to what extent we can understand the contextual information of the places visited by people and their activities at the places, which facilitate the social relationships and demographics inference.

A. Daily Routine-based Place Inference

Compared to the physical information (e.g., longitude and latitude), the contextual information (e.g., name and type) of a place contains more meaningful information related to people's social relationships and demographics. To obtain such information, we exploit the simple signal information of surrounding APs (i.e., BSSIDs and RSSs) that is readily available in most mobile devices, to determine the daily place meanings of staying segments based on people's daily routines.

1) Daily Routine-based Places: Recent reports [25], [26] indicate that people's daily routines mainly consist of three categories of activities: 1) working and work-related activities (*working activities*); 2) sleeping and household activities (*home activities*); and 3) *leisure activities*. Based on the understanding of people's daily routines, we define three categories of *daily routine-based places*, namely *Workplace* (e.g., office buildings and universities), *Home*, and *Leisure Place* (e.g., stores, restaurants, and churches), to describe contextual

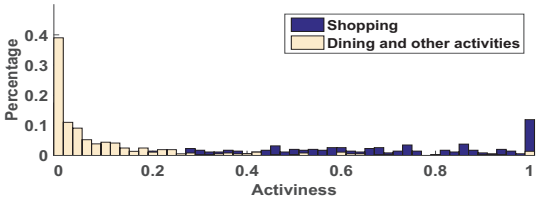
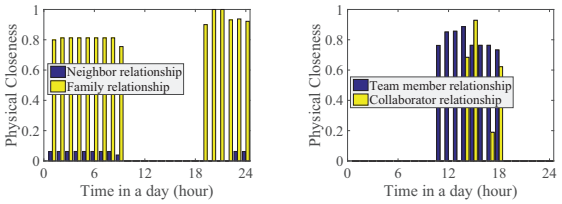


Fig. 5. Distribution of activeness score computed from each AP during staying segments when people are shopping or dinning.

information of the places. Different from categorizing daily places based on their generic nature [27], our daily routine-based categorization of daily places reflects the meaning of a place to a person instead of its function, which may vary from person to person to better describe the context of a place for every individual. For example, the same restaurant could be a workplace for waiters and waitresses, but it is a leisure place for customers. This advantage enables inferring the fine-grained social relationships and demographics.

2) Staying Segment Categorization based on Daily Routines: Next, we determine the contextual information of a place (i.e. staying segment) by categorizing it into one of the three defined daily routine-based places. The basic idea is to examine *common time spans* of the staying segments in a day with the daily routines of working and home activities, respectively. Whichever staying segment results in the longest overlapped time with the daily routine of working or home activities will be labeled as containing the Workplace or Home. The rest of staying segments are determined as containing the Leisure Places. Since people may move between different rooms for work-related activities, after determining the Workplace, we further combine the staying segments that have at least level-1 closeness with the staying segments of Workplace together to represent the whole working area. The common time spans are chosen corresponding to the majority people's daily routines from the reports [25], [26]: working activities - $8:00\text{AM} \sim 4:00\text{PM}$; home activities - $7:00\text{PM} \sim 6:00\text{AM}$; leisure activities - rest free hours of a day.

3) Fine-grained Place Context Inference: Our system is designed to derive more fine-grained place contexts (e.g. restaurants or stores in the Leisure Places and universities or office buildings in the Workplace) by leveraging Geo-information, activity features of the places and the SSID context of user associated AP. We find that the APs' BSSIDs (MAC addresses) in a staying segment generate fine-grained place contexts through certain web-based services (e.g., Google Map Geolocation API [28], Google Place API [29] and unwired labs Location API [30]). However, the place contexts obtained from the Geo-information is sometimes not unique especially in a crowded business area. Therefore, to refine the place contexts from the Geo-information, we further examine the activity features in the staying segment based on the decision rules, made from people's general time use pattern [31] and the basic knowledge of activeness at various place contexts. Moreover, if the user is associated with an AP, the semantic meaning of the AP SSID can be utilized as assistance, if available, to identify detailed contexts (e.g. company names) of the place.



(a) Spatial closeness difference. (b) Temporal closeness difference.

Fig. 6. Illustration of social relationships classification derived from temporal and spatial closeness based on one day's data.

B. Activity Feature Extraction

We determine three activity features (i.e., including *activeness*, *visiting time slots* and *staying duration*) that can capture the users' mobilities and the differences between activities at the daily routine-based places. *Activeness* (i.e. active or static) describes the person's status at a place, e.g., shopping in a store is active while dinning in a restaurant is static. *Visiting time slots*, including the person's one or multiple entrance/departure time at a daily routine-based place, captures the person's specific pattern of visiting the place, e.g., faculties may leave office several times in one day for teaching, conference, lunch *et al*. *Staying duration* captures the time nature of the activities such as buying coffee for 10 minutes or doing hair cut for one hour. We note that all the other activity features, except the activeness, can be easily obtained by examining the temporal information of the staying segments. Therefore, we discuss how to derive the activeness for each staying segment in detail.

Activeness Estimation. We devise a unique activeness estimation approach to determine the activeness of the user at a place by only utilizing the RSS of APs observed in the staying segment (This is the only place we apply RSS in this paper). The intuition behind this approach is that the user's position changes within a place result in changing distances to every surrounding AP and thus unstable RSS from each AP. From the time series of RSS in a staying segment, we derive a time series of *RSS stability* of the i^{th} AP, denoted as $\Lambda_i = \{\lambda_1, \ldots, \lambda_j, \ldots, \lambda_t\}$, where λ_j is the standard deviation of RSS calculated based on a sliding time window W. Then we further derive the *activeness score* of a staying segment by using the equation:

$$\psi_i = \frac{\sum_{j=1}^{t-w+1} v_j}{t-w+1},\ v_j = \begin{cases} 1, \lambda_j > \lambda_{th} \\ 0, otherwise, \end{cases} \tag{4}$$

where the λ_{th} is a threshold of standard deviation of RSS. To ensure the robustness, we only consider significant APs ($80\% \leq$ appearance rate) in each staying segment for deriving the activeness score, because the significant APs can capture the person's activeness in the entire staying segment. Thus, the activeness score is the ratio of active period over entire duration at the place. As an illustration, Figure 5 shows the distribution of the activeness score of all significant APs in the staying segments, when a user is dinning at a restaurant (i.e. sitting statically) or shopping in a store (i.e., walking actively), respectively. We observe more APs of dinning have lower activeness scores (less than 0.2) compared with shopping, indicating that the activeness score can well differentiate

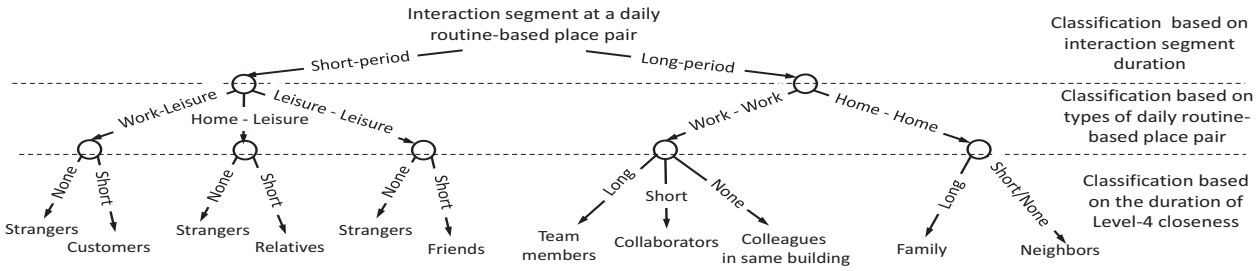


Fig. 7. Decision tree of closeness-based social relationships classification.

people's static and active status. We empirically set a threshold to the activeness score of each significant AP and further determine the activeness (i.e., active or static) of a staying segment based on the majority vote over all significant APs.

VI. Social Relationships and Demographics Inference

In this section, we present how our system utilizes the activity features provided by staying segments to derive the user's fine-grained social relationships and demographics.

A. Closeness-based Social Relationships Derivation

The social relationship is about how two people interact with each other in their daily lives, including both face-to-face interaction and event the hidden interaction without encountering. Therefore, to infer social relationships, we need to understand not only a person's activities at a place, but also how the person interacts with other people at different places. Towards this end, we define the *interaction segment* based on the staying segments between two people to capture the temporal and spatial characteristics of their interactions. The basic idea is that, we first extract and characterize the interaction segments between a target user and other people based on their staying segments and corresponding activity features. Then we utilize the temporal and spatial patterns of the closenesses of the interaction segments as well as the individual daily place contexts to derive fine-grained relationships.

1) Interaction Segment Characterization: We generate interaction segments based on the staying segments of two people in the same day. Specifically, we first find the temporally overlapped segments between the daily staying segments from the two people. Then we estimate the physical closeness between every two overlapped segments by using the Equation 1. Only long overlapped segments (i.e., time duration is longer than 10min) with at least level-1 closeness are considered as valid interaction segments. Each overlapped segment is described by three characteristics: 1) *interaction time slot*, 2) *daily routine-based place pair* based on the two users' same or different personal daily place contexts at the interaction place (e.g., Home-Home or Work-Leisure), and 3) *physical closeness*, which correspond to *when, where* and *how closely* the two people interact, respectively. Finally, the characterized interaction segments represent users' interaction at the place.

2) Closeness-based Social Relationships Classification: After determining the interaction segments, we classify the user's social relationships leveraging the temporal and spatial patterns of the physical closeness in the interaction segments. Our approach is based on the intuition that different types of social relationships show different temporal patterns for various levels of physical closeness in the overlapped daily routine-based place, which reveal different degrees of interactions between two people. Figure 6 illustrates this intuition by comparing the interaction segment characteristics for two pairs of social relationships (i.e., neighbor and family, and team member and collaborator), which can be differentiated from spacial closeness degree or temporal pattern difference.

We design a triple-layer decision tree for relationships classification based on examining the characteristics of the interaction segments between two people (i.e., the temporal and spatial patterns of their physical closeness). Figure 7 illustrates the flow of the decision tree. In the first layer, the decision tree takes the detected interaction segment of two people in one day as input, and classifies it into two classes (i.e., Short-period and long-period interaction segment) by examining the duration of the *interaction time slot* in the interaction segment. The intuition behind this layer is that people usually spend most time at several places (e.g., homes, offices, or schools) and shorter time at other places (e.g., diners, grocery stores, and post office) and so as their interactions at these places. In the second layer, we make finer decisions from the result of the first layer. In particular, we examine the *daily routine-based place pair* of the interaction segment to further classify the interaction based on the people's individual daily place contexts. Because the short-period interaction should happen at least at one person's leisure place in logic, the short-period interaction segment leads to three possible branches: workplace-leisure, home-leisure and leisure-leisure. And the long-period interaction segment leads to the pairs of workplace-workplace and home-home. In the last layer, we further detail the classification of the interaction by analyzing the *physical closeness* of the interaction segment to infer fine-grained relationships. Specifically, we examine whether the level-4 closeness of the interaction segment is non-zero or not, which suggest the two people have or not have the face-to-face interaction in the place. The duration of the face-to-face interaction allows the decision tree to further distinguish social interaction into 8 categories of fine-grained relationships: Customers, Relatives, Friends, Team members, Collaborators, Same-building Colleagues, Family and Neighbors, as well as excluding strangers.

The decision tree infers the possible relationships between two people based on their one-day social interactions. But

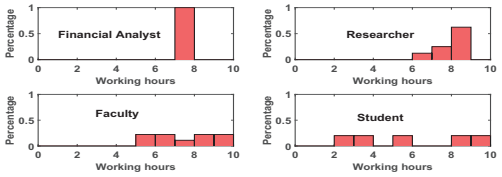


Fig. 8. Histogram of people's working duration in a week.

making relationships inference based on one-day observation may sometimes be opportunistic. For instance, students in the same school may be regarded as strangers or classmates depending on whether a face-to-face interaction is detected in one day. In order to reduce the opportunistic inferences, we propose to infer the relationships in a relative long time period (e.g., multiple days, one week or several weeks) and utilize a majority-vote approach to make the final decision.

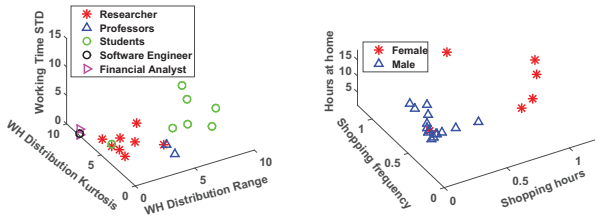
B. Behavior-based Demographics Inference

Next, we discuss how to utilize the activity features to further capture people's behavior characteristics at various daily places and infer people's demographics (e.g., occupation, gender, religion and marriage).

1) Behavior Derivation at Daily routine-based Places: In this work, we define the *behavior* as the mannerisms made by an individual in the daily routine-based place during a period of time (e.g, several days). A behavior usually consists of a series of activities, and thus can be described by the temporal and spatial statistics of the activity features extracted from the staying segments across different days. In particular, we define three kinds of behaviors: 1) *home behavior*, 2) *working behavior*, and 3) *leisure behavior* based on three daily routine-based place categories. We utilize the activity features of the same daily routine-based place across multiple days to derive the features that can characterize the three behaviors. We note that the leisure behavior can be further specified according to the fine-grained daily routine-based places in Section V-A3.

2) Occupation Inference: Occupation is the job or profession of the user, which is related to the working behavior. The inference approach is based on the fact that people of different occupations have different working time slots and duration at Workplace (may include single or multiple nearby places), which reveals different working behaviors in temporal and spacial. Figure 8 illustrates the intuition by showing the working duration histogram of 4 users with different occupations in a week. We find that office staff has the most concentrate working duration, followed by Researchers, Faculties and Students, because company office uses more regular timetable compared with school. Meanwhile, Faculties need to leave office for teaching and faculty meeting, which leads to wider working duration distribution compared with Researchers. On the other hand, Students have the most scattered working durations because they have different number of classes for each day and flexible hours at library for study.

We derive three specific working behavior features to differentiate working behaviors for multiple days at working place. *Working hour(WH) Distribution range* describes the range of the working duration histogram, which shows the flexibility



(a) Working behavior-based occupation inference results.

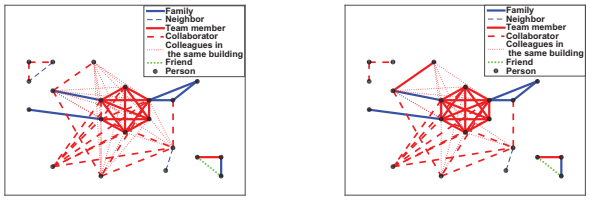
(b) Shopping and home behavior-based gender inference.

Fig. 9. Illustration of behavior-based occupation and gender inference results.

of working hours. *Working time STD* is the average standard deviation of the start and ending time of working across multiple days and *WH Distribution Kurtosis* is a descriptor of the distribution shape, which represents how concentrate the working duration is distributed. Figure 9(a) illustrates that the three working behaviors can well separate different types of occupations, which suggests that we can utilize a threshold-based approach to determine people's occupations by using these features. We note that different occupations may have similar working behaviors, such as financial analyst and software engineer, we can further narrow the choices for the occupation inference by leveraging the supplementary place contexts from Geo-information and user associated AP SSIDs as in Section V-A3.

3) Gender Inference: The information of user gender is more implicit compared with occupation, because there is no information from surrounding APs, which directly links to this biological characteristic. However, we find that males and females usually behave differently in some specific scenarios. For example, females tend to spend more time on housework and in-store shopping, while males tend to work for longer hours [32]. Such behavior difference shows the trend of the majority people and exists in many countries according to the survey. Thus our basic idea is to examine a person's behavior characteristics at home or in shops. From activity features, we derive three behavior features for gender inference: *shopping duration*, *shopping frequency* and *home duration*, which mainly capture the behavior patterns at home and leisure behavior at shops. Figure 9(b) illustrates that the three devised behavior features can well capture the differences between males and females in their behaviors at home and in shops. Additionally, we also check the user's associated AP SSIDs at leisure places, if any, to look for the particular leisure places that can differentiate gender, such as nail spa and beauty salon.

4) Religion Inference: We further demonstrate that it is possible to infer people's religion status (i.e. Christian or Non-Christian) from surrounding APs. The intuition is that Christian usually goes to church every Sunday and shows a regular pattern of leisure behavior around the church. Therefore, we extract three religion behavior features: *church attendance days*, *church attendance duration* and *church attendance frequency*, and apply a threshold-based method to decide Christian. We note that, by including more religion activities, we can also cover other religions or religious sects.



(a) Social relationships inference. (b) Social relationships groundtruth.

Fig. 10. Social relationships comparison between inference results and the groundtruth.

5) Relationships and Demographics Refinement: We find that the inferred relationships and demographics results can be mutually complementary. We then adopt several rules for the relationship and demographics refinement. For example, the family relationship between a male and a female is refined as the couple relationship or married; the collaborator between a faculty and a student (or a company supervisor and a software engineer) is refined as the advisor-student (or supervisor-employee) relationship.

VII. Performance Evaluation

A. Experiment Methodology

1) Data Collection: Due to the limitation of the man power, we choose the representative occupations, working hours and age groups for experiments to evaluate the feasibility of our approach. We recruit 21 volunteers (i.e., 6 females and 15 males) across three cities to collect surrounding APs information in their daily lives for over 6 months. The volunteers age from 20 to 40 and are mainly from six occupations, including financial analyst, Ph.D. candidate, Master student, undergraduate, assistant professor, and software engineer. We ask the volunteers to install a tool developed for data collection on their own phones and run it in the background throughout every day during the experiments. The users are asked to fill a questionnaire to input the groundtruth. The IRB is approved.

2) Hardware and Software: We include a variety of Android mobile devices in the real experiments including Samsung, Huawei, LG and Xiaomi. We develop a tool on Android platform to collect information of surrounding APs at a given frequency, i.e., 4 scans/min, which is the AP scanning frequency of many android systems [23]. For each scan, our tool collects the simple information of surrounding APs, including BSSIDs, SSID, scanning time stamp and RSS.

3) Evaluation Metrics: We use the following two metrics to evaluate the performance of our inference: *Detection Rate.* The ratio of correctly identified results over the total numbers in groundtruth. *Inference Accuracy.* The ratio of correct inference results over the total number of inference results.

B. Evaluation of Social Relationships Inference

We first examine the performance of social relationships inference from surrounding Wi-Fi APs. Figure 10 shows the comparison between the inferred social relationships (i.e., Figure 10(a)) among the 21 volunteers and the groundtruth from the questionnaire (i.e., Figure 10(b)) in graphs of relationships.

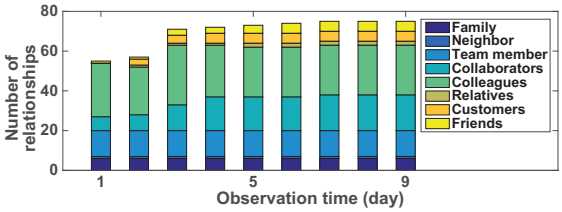


Fig. 11. Social relationships inference results based on different length of observation time.

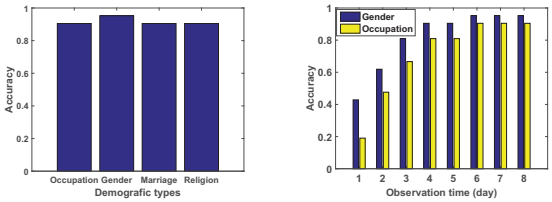
Each point in the graph represents a volunteer and different types of lines between points represent the different relationships between two volunteers. Compared to the groundtruth, the overall detection rate of social relationships inference is 91%, suggesting that our system can efficiently detect various relationships from surrounding AP information. In addition, our system also detects *hidden relationships*, which represent the potential relationship that is recognizable by our system but unknown to the two volunteers due to the lack of face-to-face interactions. We find that certain relationships (e.g., colleagues and neighbors) may contain such hidden relationship.

Table I shows the detailed statistics of our social relationships inference results. We observe that we achieve 100% detection rate for Relatives, Family and Neighbor, whereas achieve 83.3%, 94.1%, 89.5% and 87.5% detection rate for Friends, Team members, Collaborators and Colleagues, respectively, indicating that our method can accurately detect different relationships based on interaction features characterized from surrounding APs. For the misclassified relationships, one team-member relation is classified as collaborators due to irregular working time; two collaborators are classified as colleagues in the same building due to low interaction frequency. The overall inference accuracy is 95.8% when we compare the detected relationships with the groundtruth. We further detect 10 hidden relationships (i.e., 9 colleagues and 1 neighbor), while these relationships are not realized by the volunteers but can be derived from their questionnaires, indicating our system can accurately detect most relationships in daily life.

Figure 11 shows the relationships inference results under different length of observation time. We observe that most regular relationships (i.e., family, neighbor, team member) can be detected in the first day. As for other relationships, since their interactions do not occur every day, we need to observe for more days to make a decision. The relationship inference results become stable after $5 \sim 7$ days, indicating that our system can detect most relationships in people's daily life based on their social interactions in one week.

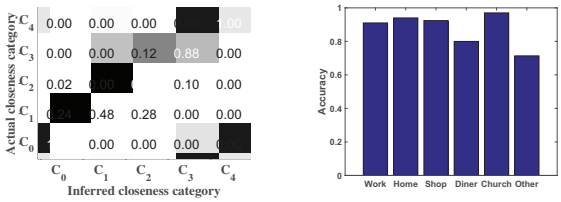
TABLE I
Social relationships inference.

Relationships	Groundtruth	Inference	Correct	Hidden
Relatives	2	2	2	0
Friends	6	5	5	0
Team members	17	16	16	0
Collaborators	19	18	17	0
Colleagues	24	23	21	9
Family	6	6	6	0
Neighbor	1	1	1	1



(a) Demographics Inference Results. (b) Demographics Inference with different observation time.

Fig. 12. Accuracy of behavior-based demographics inference.



(a) Classification confusion matrix of 4 kinds of physical closeness. (b) Classification of detailed daily routine-based places.

Fig. 13. Classification accuracy of physical closeness and daily routine-based places.

C. Evaluation of Demographics Inference

1) Accuracy of Demographics Inference: Figure 12(a) shows the overall accuracy of inferring demographics. For all the demographics in our study, our system achieves over 90.5% accuracy for Occupation, Religion and Marriage, whereas the accuracy of gender inference is 95.2% for the 21 volunteers, suggesting that it is possible to accurately infer people's demographics from surrounding AP information. We further study the performance of gender and occupation inference with different length of observation time as shown in Figure 12(b). The inference results converge after 5 days, suggesting that people's behavior features derived in a short period (i.e., one week) can accurately infer the demographics.

2) Fine-grained Social Relationships Derived from Demographics: By leveraging the derived demographics information, we further obtained refined relationships. Based on the gender information, we successfully detect all the two couples from the 21 volunteers. Besides, from the occupation inference, we specify the relationship of collaborators, e.g. who is superior and who is subordinate. In specifically, we correctly differentiate 4 superior-subordinate from 5 collaborator pairs. These results show it is possible to accurately infer fine-grained social relationships and demographics from surrounding AP information.

D. Performance of Daily Place Extraction

We randomly select 100 staying segments to examine whether our different levels of physical closeness can reflect the true relations between their physical locations. Figure 13(a) presents the confusion matrix of the inferred four kinds of closenesses and the results show that our system can achieve over 88% accuracy for measuring most levels of closeness except for C_1, whose inference relies on the remote APs or unstable signals. We note that the lowest level C_1 does not affect the social relationships and demographics inference as both of them mainly rely on C_4 and C_3.

Finally, we evaluate the accuracy of the contextual meaning inference with 594 detected places. Figure 13(b) shows we can achieve over 90% accuracy for Workplace and Home and over 80% accuracy for detailed Leisure places (e.g., Shop, Diner, Church and Other). The results demonstrate the possibility to measure the physical closeness between places and infer complex contextual meaning of daily places only from user's surrounding APs.

VIII. Discussion

Due to the limited manpower and shortage of public available data sources (i.e., containing the scanned AP signal information in large-scale areas), we evaluate our system by recruiting 21 volunteers with representative occupations and social relationship types. Furthermore, the study is based on the users' daily life activities across three cities without being restricted in a confined area. Since the participants' activities at daily places are employed as the inference basis in this work, we believe our system has the capability to successfully infer fine-grained social relationships and demographics in larger areas when given the opportunity. We demonstrate that the privacy leakage from the simple signal information of surrounding APs is significant and should arouse public attention. For the future work, we will continue our efforts to enlarge the Wi-Fi AP dataset and investigate more potential privacy leakages from such simple radio signals surrounding our daily lives.

IX. Conclusion

In this paper, we show that by analyzing the information from surrounding Wi-Fi Access Points (APs), the users' fine-grained social relationships and demographics could be disclosed. We present a scalable inference system that has the potential to derive people's activities at daily visited places leveraging surrounding APs and utilize such information to infer fine-grained social relationships and demographics. This implemented system only uses the simple signal features of surrounding APs such as MAC addresses and Received Signal Strength without requiring to obtain the context information by sniffing the Wi-Fi traffic. In particular, we describe people's daily places in three dimensions (i.e. time, space and context) to infer people's activities and extract their activity features as well as their physical closeness at same places. Our *Closeness-based Social Relationships Inference* algorithm further analyzes people's physical closeness to capture when, where and how closely people interact to reveal fine-grained social relationships, while the *Behavior-based Demographics Inference* method extracts people's various individual behavior from their activity features to infer demographics. By using the data collected by 21 participants in their daily lives over 6 months, our system confirms the possibility of using surrounding APs to infer people's social relationships and demographics with over 90% accuracy.

Acknowledgment

This work is supported in part by the NSF grants *CNS*1514436, *CNS*1409767, the NSF of China grant 61472185 and the JiangSu Natural Science Foundation grant *BK*20151390.

References

[1] "Municipal wireless network," https://en.wikipedia.org/wiki/ Municipal_wireless_network, 2016.

[2] "The global public wi-fi network grows to 50 million worldwide wi-fi hotspots," https://www.ipass.com/press-releases/the-global-public-wi-fi-network-grows-to-50-million-worldwide-wi-fi-hotspots/, 2015.

[3] "Google maps apis," https://developers.google.com/maps/, 2016.

[4] N. Cheng, X. Wang, W. Cheng, P. Mohapatra, and A. Seneviratne, "Characterizing privacy leakage of public wifi networks for users on travel," in *IEEE INFOCOM*, 2013, pp. 2769–2777.

[5] H. Li, Z. Xu, H. Zhu, D. Ma, S. Li, and K. Xing, "Demographics inference through wi-fi network traffic analysis," in *IEEE INFOCOM*, 2016.

[6] V. Sekara and S. Lehmann, "The strength of friendship ties in proximity sensor data," *PLoS ONE*, vol. 9, pp. 1–8, 2014.

[7] N. Cheng, P. Mohapatra, M. Cunche, M. A. Kaafar, R. Boreli, and S. Krishnamurthy, "Inferring user relationship from hidden information in wlans," in *IEEE MILCOM*, 2012, pp. 1–6.

[8] R. B. Braga, A. Tahir, M. Bertolotto, and H. Martin, "Clustering user trajectories to find patterns for social interaction applications," in *W2GIS*, 2012, pp. 82–97.

[9] "Wifimanager," https://developer.android.com/reference/android/net/wifi/WifiManager.html, 2016.

[10] "Understanding wireless scanning," http://www.juniper.net/documentation/en_US/network-director1.5/topics/concept/wireless-scanning.html, 2013.

[11] P. Sapiezynski, A. Stopczynski, R. Gatej, and S. Lehmann, "Tracking human mobility using wifi signals," *PLOS ONE*, vol. 10, pp. 1–11, 2015.

[12] J. H. Kang, W. Welbourne, B. Stewart, and G. Borriello, "Extracting places from traces of locations," in *ACM WMASH*, 2004, pp. 110–118.

[13] D. H. Kim, Y. Kim, D. Estrin, and M. B. Srivastava, "Sensloc: sensing everyday places and paths using less energy," in *ACM Sensys*, 2010, pp. 43–56.

[14] Z. Chen, S. Wang, Y. Chen, Z. Zhao, and M. Lin, "Inferloc: calibration free based location inference for temporal and spatial fine-granularity magnitude," in *IEEE CSE*, 2012, pp. 453–460.

[15] A. K. Das, P. H. Pathak, C.-N. Chuah, and P. Mohapatra, "Contextual localization through network traffic analysis," in *IEEE INFOCOM*, 2014, pp. 925–933.

[16] J. Wiese, J. I. Hong, and J. Zimmerman, "Challenges and opportunities in data mining contact lists for inferring relationships," in *ACM UbiComp*, 2014, pp. 643–647.

[17] M. Cunche, M.-A. Kaafar, and R. Boreli, "Linking wireless devices using information contained in wi-fi probe requests," *Pervasive and Mobile Computing*, vol. 11, pp. 56–69, 2014.

[18] B. Han, J. Li, and A. Srinivasan, "Your friends have more friends than you do: Identifying influential mobile users through random-walk sampling," *IEEE/ACM Transactions on Networking*, vol. 22, pp. 1389–1400, 2014.

[19] S. Seneviratne, A. Seneviratne, P. Mohapatra, and A. Mahanti, "Predicting user traits from a snapshot of apps installed on a smartphone," *SIGMOBILE Mob. Comput. Commun. Rev.*, vol. 18, pp. 1–8, 2014.

[20] Y. Wang, Y. Chen, F. Ye, J. Yang, and H. Liu, "Towards understanding the advertiser's perspective of smartphone user privacy," in *Proceedings of the 35th IEEE International Conference on Distributed Computing Systems (ICDCS)*, 2015, pp. 288–297.

[21] D. Seamon, "A way of seeing people and place: Phenomenology in environment-behavior research," *Theoretical Perspectives in Environment-Behavior Research*, pp. 157–78, 2000.

[22] X. Lu and E. I. Pas, "Socio-demographics, activity participation and travel behavior," *Transportation Research Part A: Policy and Practice*, vol. 33, no. 1, pp. 1 – 18, 1999.

[23] J. Shi, L. Meng, A. Striegel, C. Qiao, D. Koutsonikolas, and G. Challen, "A walk on the client side: Monitoring enterprise wifi networks using smartphone channel scans," in *Proceedings of the IEEE International Conference on Computer Communications (IEEE INFOCOM)*, 2016.

[24] "Wifi scanning every 5 seconds," http://androidforums.com/threads/wifi-scanning-every-5-seconds.631388/, 2012.

[25] "American Time Use Survey: Work and employment," http://www.bls.gov/TUS/CHARTS/WORK.HTM, 2014.

[26] "American Time Use Survey: Students," http://www.bls.gov/TUS/CHARTS/STUDENTS.HTM, 2014.

[27] B. Krämer, "Classification of generic places: Explorations with implications for evaluation," *Journal of Environmental Psychology*, vol. 15, no. 1, pp. 3 – 22, 1995.

[28] "Google places api," https://developers.google.com/places, 2016.

[29] "The google maps geolocation api," https://developers.google.com/maps/documentation/geolocation/intro, 2016.

[30] "Unwired labs location api," https://unwiredlabs.com/, 2016.

[31] "United states department of agriculture economic research service," http://www.ers.usda.gov/data-products/.aspx, 2016.

[32] "Charts: How american men and women spend their time," http://www.usnews.com/news/articles/2013/06/24/charts-how-american-men-and-women-spend-their-time, 2016.