# On Human Mobility Predictability Via WLAN Logs

Paul Y. Cao,*¶‖ Gang Li,†‖ Adam C. Champion,†‖ Dong Xuan,† Steve Romig‡ and Wei Zhao§
*Department of Computer Science and Engineering, University of California, San Diego, La Jolla, CA 92093, USA
†Department of Computer Science and Engineering, The Ohio State University, Columbus, OH 43210, USA
‡Office of the CIO, The Ohio State University, Columbus, OH 43210, USA; §The University of Macau, Macau, P.R. China
yic242@eng.ucsd.edu, {lgang, champion, xuan}@cse.ohio-state.edu, romig.1@osu.edu, weizhao@umac.mo

***Abstract*—In this research, we conduct a comprehensive measurement study on the predictability of human mobility with respect to demographic differences. We leverage an extensive WLAN dataset collected on a large university campus. Specifically, our dataset includes over 41 million WLAN entries gathered from over 5,000 students (with demographic information) during a four-month period in 2015. We observed surprising patterns on large increases of long-term mobility entropy by age, and the impact of academic majors on students long-term mobility entropy. The distribution of long-term entropy follows a bimodal distribution, which is different from previous studies. We also find that the predictability of students' short-term (daily or weekly) mobility varies on different days of the week and with student gender. Because of the large campus size, our results can mimic people's mobility patterns in metropolitan areas. We also anticipate that our results will provide insight that guides academic administrators' decisions regarding facilities planning, emergency management, etc. on campus.**

## I. Introduction

Mobile devices such as smartphones and tablets are ubiquitous in society. There are over seven billion mobile devices worldwide [1], about the size of the human population, and the ITU estimates there are nearly seven billion mobile subscriptions [2]. Since humans carry mobile devices, the devices' mobility closely approximates that of humans. Recently, mobile phone log data from cellular networks have become available at large scale [3]. Researchers have also used GPS-equipped automobiles to investigate vehicle mobility [4]. Based on these real-world datasets, human mobility models have been proposed [4]–[8].

Understanding human mobility is critical for various applications such as epidemic modeling, urban planning, and resource management for mobile communications [9], [10]. To an observer, human mobility may appear random and unpredictable. In a seminal paper, Song et al. [11] used mobility entropy to characterize the predictability of human mobility. A higher entropy for a group of people indicates that this group's mobility is less predictable on average. The entropy of a person's mobility can also be interpreted as the amount of irregularity in his/her mobility. In human mobility studies, a body of work has used this metric [12]–[14]. In this work, we use mobility entropy to investigate mobility differences across demographic categories by exploiting a new comprehensive wireless local area network (WLAN) dataset collected at a large public university.

The WLAN data are collected over 139 days from Tuesday, January 13, 2015, to Sunday, May 31, 2015. The campus size exceeds 1,500 acres and has an extensive Wi-Fi network with more than 8,000 access points (APs); the university offers several hundred majors. More than 5,000 students are included in this dataset, which represents a random sample of the overall student body. The dataset consists of the (dis)association records of all students' devices that connect to an encrypted WLAN's SSID. Students' demographic information includes their birthdays, majors, and genders, which are made available after anonymization. The WLAN log data in this paper, to the best of our knowledge, are the largest with respect to the number of APs and non-laptop mobile devices in the dataset.

This paper presents the first study of the predictability of human mobility using WLAN log data with demographic information at a large public university. We investigate the underlying mobility entropy differences among groups of students based on their ages, academic majors, and genders. By varying the sampled student size and observation length, we verify the validity of our discoveries. We conclude that the observation length and the number of students in our WLAN dataset are sufficient to infer meaningful results. We compare students' overall long-term mobility entropy across the entire observational timespan as well as short-term mobility entropy (such as daily and weekly entropy).

**Main discoveries**: Using mobility entropy as the primary measurement metric, we make the following surprising discoveries contrary to common-sense beliefs.

1) *Common-sense view: As a group with similar ages and backgrounds, it is often assumed that students' mobility follows a similar distribution with respect to age. One or two years*

¶ This work was performed while the first author was a visiting scholar at The Ohio State University collaborating with the other authors.

‖ The first three authors are co-primary authors.

This work was supported in part by the Natural Science Foundation of China (NSFC) under Grant 61373115, Grant 61402356, and Grant 61572398; by the China Postdoctoral Science Foundation under Grant 2015M572565; by the Fundamental Research Funds for the Central Universities under Grant AQ1 xkjc2015010; by the Science and Technology Fund of Macau (FDCT) Project AQ2 under Grant 061/2011/A3 and Grant 092/2014/A2; and by the University of Macau Project under Grant MYRG112-FST12-ZW and Grant MYRG2015-00165-FST. This work was also supported in part by the U.S. National Science Foundation under Grant 1117175, Grant 1350145, Grant 1116644, and Grant 0963979. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding agencies.

978-1-5090-5336-0/17/$31.00 ©2017 IEEE

*of age variation should not yield major changes in students' mobility.*

We observed that the overall long-term entropy greatly varies by age for 19–21 year-olds (the group of traditional college students). The distribution of mobility entropy noticeably increases as students' ages increase. The mean entropies for 19-year-olds and 20-year-olds are 0.89 and 1.16, respectively. This represents a 30% increase. Similarly, there is a 23% change between the entropies of 20- and 21-year-olds. The rate of entropy increase levels off after age 21. The shift in mobility entropy across age groups was not previously reported in the literature such as [11], [13].

2) *Common-sense view: The rationale why students choose a major is believed to be largely due to academics. The impact of academic majors on mobility should be minimal.*

We observe that the rate of change across age groups within each major is not uniform. The change in mobility entropy from 19- to 22-year-old age groups is markedly different for different majors. For example, the entropic change from 19- to 20-year-old engineering students is much less compared with that of business majors. Health-related majors and engineering majors have lower mobility entropies than students from other majors. When student mobility entropies are compared across majors on a daily basis, undecided majors have lower entropies than every other major.

3) *Common-sense view: The distribution of mobility for college students is expected to be similar to the distribution of the general population. Previous studies have shown that mobility distributions of people follow heavy-tail distributions such as the Pareto or Weibull distributions [8].*

We are surprised to discover that the overall long-term entropy of students yields a bimodal distribution. The two modes of observed entropies indicate that there are two groups of students with distinctive mobilities. Fig. 1 shows the observed bimodal distribution.

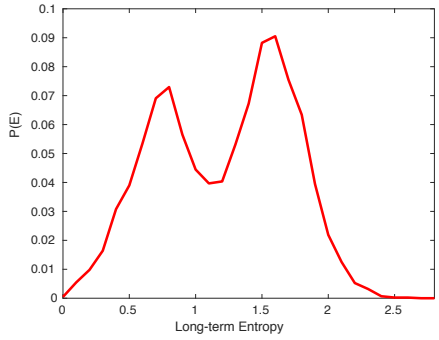


Fig. 1. Observed long-term mobility entropy distribution over a 139-day period.

In this paper, we also verify that there is no significant difference between males and females on their long-term entropies. This finding confirms the results from Song et al. [11]. However, when entropies of male and female students are compared on a daily basis, a consistent pattern appears: females' entropies are slightly larger than males' in general, especially on Tuesdays and Thursdays. This pattern persists over the 139-day observation period except during special weeks of a semester such as the spring break week and the finals week.

**Significance:** The mobility patterns discovered in our work arise on a large academic campus (over 1,500 acres in size). The campus has classrooms scattered throughout hundreds of buildings, restaurants, grocery stores, physical fitness centers, and other amenities. Thus, our findings should complement mobility patterns in a large community. The significance of this work includes the following: 1) We show the importance of demographic differences in the predictability of mobility. This observation indicates that a more accurate mobility model should incorporate demographic information. For example, the bimodal distribution of overall mobility entropy indicates that a homogeneous assumption of mobility parameterizations should be revisited. In epidemic studies, if demographic information is known, then the speed of disease spread should consider mobility predictability to reflect a more refined model. In the area of resource management, considering mobility entropy differences in demographics can yield more accurate estimates of resource distribution rates; and 2) Results presented in this paper can also help academic administrations regarding facilities planning and student retention. For example, our study shows that students with health-related and engineering majors tend to have low mobility entropy. Thus, classrooms for those majors might be intentionally allocated among different buildings considering their predictability differences. As for retention, university administrations can design better education policies and study plans that incorporate mobility differences across age groups focusing on the 19-year-old group who are most likely to be traditional college freshmen.

The rest of this paper is organized as follows. Section II reviews related work. Section III presents our dataset and processing methodology. Section IV illustrates our measurements and results. Finally, Section V concludes the paper.

## II. Related Work

The collection and analysis of datasets from mobile devices have attracted considerable attention from the research community. We refer the reader to [3], [7] for surveys of data collection methodologies, results, and applications. There are three main types of mobile device data: devices' GPS locations, devices' locations determined via cellular towers, and devices' wireless connectivity logs via Bluetooth, WLAN, or RFID [7]. The granularity of GPS and WLAN data is on the order of meters whereas that of cellular towers depends on the towers' geographic distribution, the distance among which is on the order of kilometers. GPS and cellular networks offer potential coverage over large areas whereas WLAN logs offer relatively accurate measurements of user behaviors in more localized environments. Devices' WLAN logs are mainly captured on university campuses, via public transportation, and in office buildings, cities, and conferences. Several studies

have captured devices' wireless connectivity on campuses such as Dartmouth [15], NCSU and KAIST [8], UNC [16], MIT [17], UCSD [18], and USC [19].

Public real-world WLAN logs are available from repositories such as CRAWDAD [20]. However, existing WLAN logs only contain small numbers of mobile devices, most of which are laptops [7]. These WLAN logs were collected between 1999 and 2006 using at most a few hundred APs. Besides, most work focuses on physical measurements and analysis of traces. No demographic information has been applied in conjunction with WLAN logs to infer human mobility patterns. In contrast, our study examines demographic data in much finer detail than these studies, as demographic factors are strongly correlated with human behaviors [21]. We examine the predictability of several thousand students' mobilities on a large university campus considering differences in their majors, ages, and genders.

Researchers have proposed various mathematical models to describe human mobility from mobile device datasets. González et al. [5] measure the regularity of human mobility via the radius of gyration and entropy from cellular logs for a country. Similarly, Song et al. [11] derive bounds for predictability of human mobility using cellular logs for a country and demographic data such as gender. Using entropy as a metric, Qin et al. [12] investigate the predictability of people at home and work. Lu et al. [13] use entropies measured via cellular log data to study the mobility predictability of people in West Africa. Work by Cho et al. [14] analyzes location periodicity in human mobility using entropy calculated from two location-based social network datasets and one cellular log dataset. Rhee et al. [8] show via GPS logs that human mobility patterns are statistically similar to Lévy walks, which are random walks with self-similar jumps [7]. Tuduce and Gross [22] find that WLAN traces follow a power-law model, which has been observed in many other datasets [23]. Kim et al. [24] refine this model to a lognormal model. Hsu et al. [25] propose a time-variant community mobility model that can capture characteristics of WLAN traces. In another work by Song et al. [26], the continuous-time random-walk (CTRW) was shown to conflict with empirical data, and the authors proposed an individual mobility model incorporating the number of unique locations visited and the visit frequency. Other well-known models include the random waypoint model [18] and the statistical mobility model [27]. This paper considers mobile devices as "first-class citizens" along with students' demographics. We measure the mobility entropies [11] of students' movements to infer the predictability of their movements.

## III. Dataset and Processing Methodology

### A. Dataset Description

The dataset includes the (dis)association of mobile devices with respect to APs on the academic campus of a large public university. Based on their authentication credentials, we obtain students' birthdays, majors, and genders from the university's student information system. We exclude students who use the unsecured WLAN from the dataset because these students' demographic information cannot be verified. To help protect students' privacy, we encrypt student credentials using a two-way hash function. There are 5,096 students in the dataset with 13,549 unique mobile device MAC addresses. The dataset includes association and disassociation logs from 8,420 different APs across 225 buildings.

A sample log entry has the following format:

```
timestamp,process,ap-name,student-id,
role,MAC,SSID,result
```

The fields in the log represent the event's UNIX timestamp, the process that generated the log entry, the AP name, the encrypted student ID, the role assigned to the device, the anonymized MAC address (with the OUI preserved), the SSID name, and the authentication result (success or failure), respectively. Each AP is named via the building name, floor number on which the AP resides, and a unique numeric ID for the AP on the building floor. For example, the AP name `BB-2-4` indicates that this AP is the 4th AP on the 2nd floor in building `BB`.

There are 41,006,186 log entries in the dataset, and the unit of the timestamp is seconds. Some entries do not include the student ID and the AP name; we consider these log entries invalid in this work. After we remove invalid entries from the dataset, 39,100,373 log entries remain.

Our dataset has wireless data for 5,096 students with demographic information, which is 12% of the overall student population at the University. Each student is represented via an anonymized ID associated with the student's gender, birthday, and major. We also obtained a list of all students at the university including their demographics. A careful examination of the distribution of students in the dataset reveals that students in our dataset are representative of the overall student population at the university. Nearly all demographic categories are within $(12 \pm 2)\%$, which limits potential student under- or over-representation. To verify that our results are significant with respect to student size, we also sample 50–80% of students from the dataset and perform the same calculations. We observe that our original findings remain when over 60% of students are sampled. To ensure the best mobility entropy measurement, we eliminate laptop computers from consideration. This requirement eliminated 720 students whose only device that appeared in the dataset is a laptop. We verify that these students are uniformly distributed across demographic categories, so this adjustment does not distort demographics. We believe that this group is likely to have set up their real mobile devices to use the university's unsecured WLAN.

Table I shows student data. We group student majors into seven categories based on the university registrar's major descriptions. The age range of students is 19–58 with over 80% of students in the 19–22-year-old groups. The number of students decreases in the "24 and older" group (24+) as students' ages increase. In this study, we treat the 24+ group as a single age category. This group of students includes senior citizens who take tuition-free college courses.

| Majors | # of Students |
|---|---|
| Business | 893 |
| Education | 218 |
| Engineering | 840 |
| Health | 645 |
| Science | 840 |
| Social Science | 803 |
| Undecided | 137 |

| Gender | # of Students |
|---|---|
| Female | 2,039 |
| Male | 2,335 |
| Unknown | 2 |

| Age | # of Students |
|---|---|
| 19 | 706 |
| 20 | 1,116 |
| 21 | 1,014 |
| 22 | 676 |
| 23 | 309 |
| 24 and older | 125 (age 24) |
| | 105 (age 25) |
| | 325 (older than 25) |

| Total | 4,376 |
|---|---|

TABLE I
STUDENT DEMOGRAPHIC DATA.
TOTAL REFLECTS THE LAPTOP REDUCTION MENTIONED ABOVE.

### B. Data Processing Methodology

The primary metric used in this paper, *mobility entropy*, is calculated using a person's trajectory $T$ that is defined as

$$T = (L_1, t_1, ST_1) \rightarrow (L_2, t_2, ST_2) \rightarrow \cdots \rightarrow (L_N, t_N, ST_N),$$
$$t_1 < t_2 < \cdots < t_N,$$

where $L_i$ is the $i$th location in trajectory $T$, $t_i$ is the arrival time of the person at location $L_i$, and $ST_i$ is the stay time of the person at location $L_i$.

The length of a trajectory, $N$, is the total number of locations a person visits between times $t_1$ and $t_N$. The maximum observation length of our WLAN log is 139 days but by varying $t_1$ and $t_N$, appropriate time intervals of interest can be selected. In this paper, we define *long-term* as the entire time span of 139 days and *short-term* as daily or weekly.

To build the trajectory of each user for entropy calculation, we need to extract the sequence of locations for all students as well their stay times from the WLAN log entries. Each log entry provides information about a user's identity (user/MAC), the name of the AP with which the student was interacting, and a timestamp $t$. Since a student is likely to have several mobile devices (the number usually ranges from 2 to 5), each MAC address is treated independently when the trajectory is constructed. Thus each log entry can be viewed as a triple ($t$, user/MAC, $AP$). We sort log entries based on the timestamp $t$ to ensure sequential order.

– *Location Granularity:* Since an AP can only provide accuracy within its range, and a device's connection to an AP does not imply that this AP is the closest one to the device, we use a building as the base unit in the trajectory. We believe this approximation is appropriate since this work focuses on human mobility and buildings are naturally the base units for outdoor mobility. Thus, all sequential connections to APs inside a building are treated as a single data point in a student's trajectory with accumulated stay time. In the example shown in Fig. 2, a device interacts with three APs in building $B_1$ before connecting to $AP_4$ in $B_2$ followed by two APs in $B_3$. Thus, the sequence of locations this device has visited is $B_1, B_2, B_3$. The stay time of a device with an AP is the length of the time interval between its association time and disassociation time. Most devices are not logged when they disassociate with APs. Hence, the stay time of each user inside the same building is

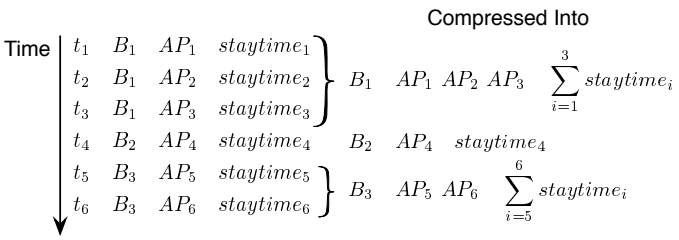


Fig. 2. Location granularity as buildings

calculated as the difference of association times between the current AP and the next AP in the log. For example, in Fig. 2, $AP_1$'s stay time is calculated as $t_2 - t_1$. Thus, the overall stay time of a user in building $B_i$ is $\sum_{j=i}^{k} staytime_{AP_j}$, where $AP_j$ denotes APs in the same building.

– *Stay Time Adjustment:* We observe that users normally connect to several APs in the same building before leaving it and students' speeds are inconsistent. The latter observation is similar to that of Kim et al. [24]: speeds calculated between consecutive buildings in the log can be large. We believe that this is due to mobile devices' ability to connect to a remote AP even when people carrying the devices remain far from it. Fig. 3 shows our process of stay time estimation.

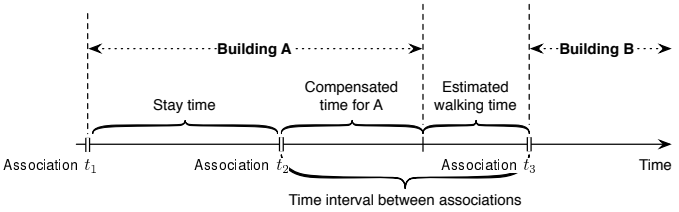


Fig. 3. Stay time estimation for inter-building AP connections in a person's trajectory

For a device, suppose $t_1$ is the time of the first association between itself and an AP in building $A$, $t_2$ is the time of the last association between this device and an AP in building $A$, and $t_3$ is the time of the first association between this device and an AP in building $B$. We apply the following rules to compensate a building's stay time based on the analysis of inter-building travel time.

1) If the time interval from building $A$ to building $B$, $t_3 - t_2$, is smaller than the estimated travel time, we judge that a fast remote connection has been established between the mobile device and an AP in a building. There is no need to compensate building $A$'s stay time.

2) If the time interval $t_3 - t_2$ exceeds the estimated travel time, we judge that the user stays in building $A$ for a longer period before moving to building $B$. Thus, the total stay time of building $A$ is ($t_3 - t_1$−estimated travel time).

We calculate the estimated walking time between two buildings using the Google Maps API [28]. We obtain the latitude and longitude of every building appearing in the log and calculate the estimated travel time between each pair of buildings with the "walking" travel mode.

There is no association record following a device's last AP association entry at the end of the observation period. Thus, the aforementioned formulas cannot be applied to find the stay time of the last AP. Since we focus on non-laptop mobile devices' trajectories, it is very unlikely that a mobile device simply "disappears" from the extensive campus WLAN. The most plausible explanation is that the device has left the WLAN after a very short connection. Hence, we assume that the last AP's connection is small but nonzero. When entropy is calculated, we assume that the last AP's stay time is less than a predefined constant (i.e., $\Delta t$ as defined in Section IV-A).

After data processing, each user/MAC's trajectory becomes a time series of buildings and their corresponding stay times.

## IV. Measurements and Results

In this section, we define mobility entropies that are used to capture student mobility patterns. Next, we present our findings based on different student demographics.

### A. Entropy Calculation

Entropies addressed in this research are mobility entropies. This paper refers to mobility entropies as *entropies* for short. Given the trajectory $T$ of a person, several types of entropies can be calculated as follows [11]:

1) *Random entropy:* This entropy ignores the spatial and temporal relationship of locations and is only concerned with the number of unique locations that a person visits. Thus, this entropy is $S_1 = \log_2(N)$, where $N$ is the number of unique locations to which a person has traveled.

2) *Non-sequential entropy:* This entropy considers the frequency of visited locations to determine the randomness of a person's mobility. Thus the formula for this entropy is $S_2 = -\sum_i p(i)\log_2(p(i))$, where $p(i)$ is the frequency of location $i$ in a person's trajectory.

3) *Real entropy:* This entropy considers the spatial and temporal information of a person's trajectory. Suppose the location sequence of a person's trajectory is $L = B_1 \to B_2 \to \cdots \to B_N$. The real entropy can be calculated as

$$S_3 = -\sum_{L'_i} p(L'_i)\log_2(p(L'_i)), \tag{1}$$

where $L'_i$ is a subsequence of $L$ and $p(L'_i)$ is the probability of $L'_i$ appearing in $L$. Kontoyiannis et al. [29] proposed a fast approximation as

$$S_3 \approx \left(\frac{1}{N}\sum_{i=1}^{N}\frac{\Lambda_i}{\log_2(n)}\right)^{-1}, \tag{2}$$

where $N$ is the length of the trajectory and $\Lambda_i$ is the length of the shortest substring starting at location $i$ that does not appear as a continuous substring between positions 1 and $i-1$. Kontoyiannis et al. [29] prove that when $N$ is large, $S_3$ rapidly approaches the real entropy.

In this paper, we choose real entropy as our main measurement instead of random or non-sequential entropies. Our rationale is that real entropy incorporates a person's spatial and temporal features. We also calculate students' random and non-sequential entropies. In Section IV-B, we show that our findings are consistent with [11].

The location sequence needed for Eq. (1) can be constructed in two ways. The first approach does not consider students' stay times inside buildings; hence, the trajectory is a list of buildings that a student visited. The real entropy calculated from such a trajectory is defined as *time-independent entropy* ($S_{ti}$). It shows the randomness of a person's mobility given only sequential geographic constraints. The other approach considers both the sequence of locations that a student visits and the time the student stays at each location. Starting from the time of a user's first association entry on a day, we can "slice" the time of this student on that day into intervals of $\Delta t$ each. Slicing incorporates the student's temporal randomness in the sequence of locations. We call this type of entropy *time-dependent entropy* ($S_{td}$). A student's trajectory $(B_1, ST_1) \to (B_2, ST_2) \to \cdots \to (B_N, ST_N)$ is changed to $L = \underbrace{(B_1) \to (B_1) \to (B_1)}_{\lceil ST_1/\Delta t \rceil} \to \cdots \to \underbrace{(B_N) \to (B_N) \to (B_N)}_{\lceil ST_N/\Delta t \rceil}$.

For example, suppose a student's trajectory and stay time (in minutes) in buildings are $(B_1, 60) \to (B_2, 30) \to (B_3, 34) \to (B_4, 19)$ and $\Delta t$ is 30 min. The location sequences for Eq. (1) are $B_1, B_2, B_3, B_4$ and $B_1, B_1, B_2, B_3, B_3, B_4$ for $S_{ti}$ and $S_{td}$, respectively.

Students may have several mobile devices such as smartphones and tablets. We construct each mobile device's trajectory and calculate its entropy according to Eq. (2). We select the device with the largest entropy to represent the owner's entropy. We calculate students' overall long-term time-dependent entropy ($S_{td}$) from cumulative trajectories of all visited locations with a time slice $\Delta t = 30$ min throughout the 139-day data collection period. As students' log entries terminate before 11:59 p.m. each day, we add a "dummy" marker to each student's last log entry on that day to indicate the end of the day. We treat the marker as a "virtual location" in the student's trajectory. We investigate short-term entropy on daily and weekly bases to examine students' predictability and regularity with respect to age, major, and gender. We calculate short-term entropies $S_{td}$ and $S_{ti}$ by limiting the time interval to a shorter scale such as a day or a week instead of the entire 139-day period. In addition, we calculate entropies based on different time intervals from 1 to 9 weeks. We observe that overall entropies ($S_{td}$) for a 6-week time interval converge statistically to the overall long-term entropy measured over the 139-day period. Thus, the saturation point of our calculated entropy is $\sim$6 weeks, which is shorter than the 12-week saturation time interval using cellular data [11].

### B. Results

This section presents students' mobility patterns using the entropy defined in Eq. (1). We choose $\Delta t$ as 30 min for $S_{td}$ as the measured average stay time of locations in our dataset is 52.9 min; therefore, 30-min slices provide temporal repetition for locations with longer stay times. In addition, we

test other $\Delta t$ values in our calculations and report the results in Section IV-B4.

First, we present the demographic differences of entropy categorized by age, academic major, and gender and we summarize the overall patterns. For each demographic category, we discuss interesting findings based on long-term and short-term entropies.

*1) Differences Among Age Groups:*

– *Long-Term Entropy:* Long-term $S_{td}$ across age groups shows marked differences. Fig. 4 shows the distribution of student entropies stratified by age groups.

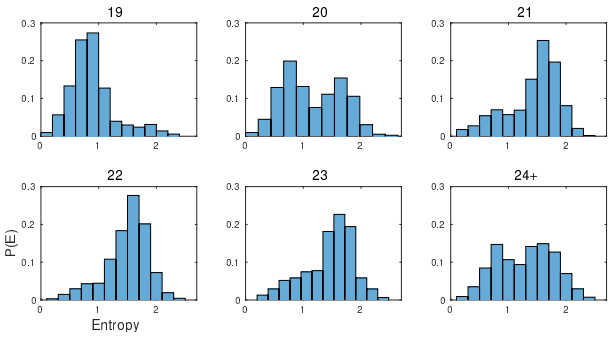


Fig. 4. Histograms of long-term entropy $S_{td}$ for different age groups.

Traditional college students should be in age groups between 19 and 22. Students' class ranks can be inferred as follows: 19-year-olds are freshmen, 20-year-olds are sophomores, 21-year-olds are juniors, and 22-year-olds are seniors. As can be seen from Fig. 4, 19- and 22-year-olds' entropies follow a unimodal distribution with 22-year-olds' average entropy markedly higher than that of 19-year-olds. 20- and 21-year-olds' entropies are bimodal with one peak centered near that of 19-year-olds and the other peak centered around that of 22-year-olds. Notably, the mean mobility entropy increases with age until age 22. The 23-year-old group has a similar distribution as that of the 22-year-old group. The 24-and-older (24+) group also follows a bimodal distribution.

*Discussion:* Both spatial and temporal factors may cause the entropic increase as students age. Thus, we also investigate long-term the $S_{ti}$ of students across different age groups. Results show that the distributions of $S_{ti}$ are all unimodal across age groups and the entropic increase as students age is much subtler. For students in the 19–22-year-old age groups, there is a 6% increase for $S_{ti}$ compared with the 67% increase for $S_{td}$. Thus, temporal differences across age groups should have contributed more to the observed phenomenon.

– *Short-Term Entropy:* We focus on students' entropies on each day of the week when short-term entropies are studied. There are $\sim$19 weeks throughout the observation period. We used two approaches to determine the average entropy for every day of the week: 1) For any day of a week, the average entropy for that day is calculated only using students who appeared at least once on that day of the week throughout the observation period. This requirement ensures that a student must be in the log at least once for each day of the week. 1,934 students fit this requirement which we call "strict"; 2) Each student in the log contributes to the average entropy of a day of a week in which the student appears. This approach includes all students since every student in the dataset appears at least once for a day. We call this method of entropy calculation "inclusive." We calculate entropies using both approaches and obtain similar results.

Daily entropies of students with different ages show consistent differences. Fig. 5 shows the differences of $S_{td}$ on ages using the two aforementioned approaches. As can be seen from both figures, the daily average $S_{td}$ strictly increases as students' ages increase from 19 to 22 using both the "inclusive" and the "strict" approaches. This finding confirms the pattern found in the overall entropy results.

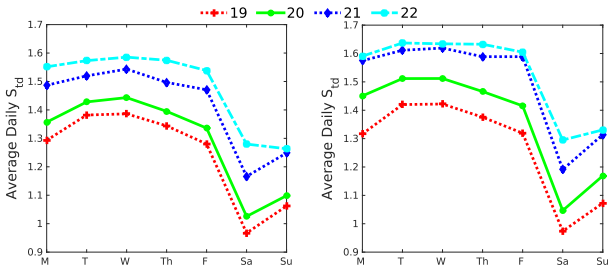


Fig. 5. Average entropies across age groups for each day of the week. The figure on the left is calculated using the "inclusive" approach, and the one on the right uses the "strict" approach.

*2) Differences Among Academic Majors:*

– *Long-Term Entropy:* Long-term entropy $S_{td}$ is diverse across academic majors whereas long-term $S_{ti}$ follows similar patterns. Table II shows the mean entropies for each major.

| **Majors** | Engineering | Science | Health | Business | Social | Education | Undecided |
|---|---|---|---|---|---|---|---|
| **Mean** $S_{td}$ | 1.22 | 1.29 | 1.18 | 1.30 | 1.33 | 1.33 | 1.09 |
| **Mean** $S_{ti}$ | 2.19 | 2.24 | 2.13 | 2.25 | 2.24 | 2.25 | 2.17 |

TABLE II
LONG-TERM METRICS FOR EACH ACADEMIC MAJOR

$S_{ti}$ values vary within 4% of each other while variations of $S_{td}$ exceed 22%. Undecided majors have the smallest $S_{td}$ and $S_{ti}$ compared to all other majors.

Another interesting observation is that the rate of entropic increase from 19-year-olds (freshmen) to 20-year-olds (sophomores) differs significantly among majors. For example, this rate of increase for engineering majors is small compared with that of business majors. Fig. 6 shows histograms among six different majors for their 20-year-old group (sophomores). In contrast to the rate of change from age 19 to age 20, the rate of entropic change from age 20 to age 21 is similar among majors. Undecided majors are not shown due to limited data after dividing into individual age categories.

*Discussion:* The most dramatic increase occurs with business majors whose entropy mode shifts from left to right when students' ages increase from 19 to 20.

The entropy mode shift for engineering and science majors is much smaller. We calculate the average stay time of students with respect to their ages (19 and 20) and majors and find that for business majors, the average stay time decreased from

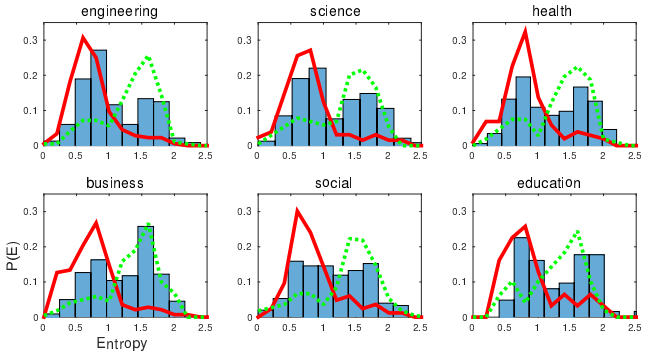


Fig. 6. Histograms of 20-year-olds (sophomores)'s $S_{td}$ for six different groups of students based on their majors. The two contours represent the long-term entropy of 19-year-olds (freshmen, solid red lines) and 21-year-olds (juniors, dashed green lines)

72.9 min (age 19 group) to 52.5 min (age 20 group), yielding a 28% drop. For engineering majors, their average stay time decreased from 68.8 min to 59.6 min, yielding a 14% drop.

We calculate the time-independent entropy $S_{ti}$ for all students and observe that there is no dual mode. The observed shifts between 19- and 20-year-olds across majors are minimal. For example, Fig. 7 shows 20-year-olds' (sophomores') $S_{ti}$ for different majors.

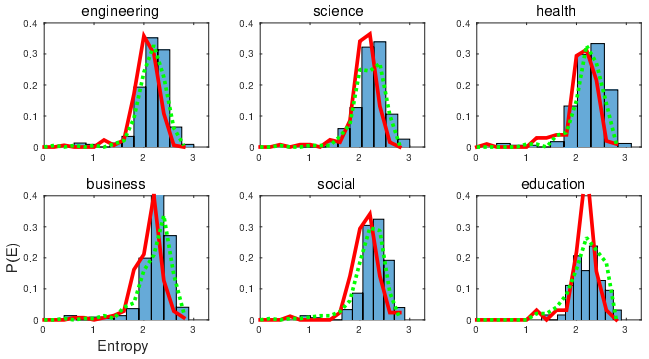


Fig. 7. Histograms of 20-year-olds (sophomores)'s $S_{ti}$ for six different groups of students based on their majors. The two contours represent the long-term entropy of 19-year-olds (freshmen, red solid lines) and 21-year-olds (juniors, green dashed lines)

– *Short-Term Entropy:* For every day of the week, undecided majors have low entropies compared with those of every other major. Table III shows $S_{td}$ for each day of the week for all seven majors using the "inclusive" approach. The "strict" approach yields similar results.

| Majors | Mon | Tue | Wed | Thu | Fri | Sat | Sun |
|---|---|---|---|---|---|---|---|
| **Business** | 1.47 | 1.54 | 1.52 | 1.52 | 1.47 | 1.14 | 1.25 |
| **Education** | 1.56 | 1.58 | 1.57 | 1.51 | 1.50 | 1.12 | 1.22 |
| **Engineering** | 1.53 | 1.53 | 1.56 | 1.50 | 1.52 | 1.10 | 1.21 |
| **Health** | 1.40 | 1.45 | 1.47 | 1.43 | 1.38 | 1.12 | 1.21 |
| **Science** | 1.50 | 1.56 | 1.58 | 1.54 | 1.51 | 1.09 | 1.18 |
| **Social** | 1.47 | 1.56 | 1.55 | 1.54 | 1.50 | 1.11 | 1.25 |
| **Undecided** | **1.36** | **1.45** | **1.42** | **1.39** | **1.35** | **0.95** | **1.07** |

TABLE III
SHORT-TERM DAILY $S_{td}$ FOR EACH ACADEMIC MAJOR

*Discussion:* We observe that undecided majors have the lowest entropies for every day of the week compared with other majors. This finding is interesting as we would expect undecided majors to have high mobilities as they explore classes around the campus, which would increase entropy. Examining the average stay time reveals that undecided majors have the highest stay time (on average) than any other group. We believe that undecided majors are generally from lower class ranks (in our dataset, over 73% of undecided majors are between ages 19 and 20, i.e., this group has more freshmen and sophomores). Hence, this group has lower entropy than those of other groups. Also, health-related majors have relatively low entropies compared with other majors. The pattern for $S_{ti}$ is similar, but undecided majors do not have the lowest $S_{ti}$. There are no distinct patterns across academic majors for daily $S_{ti}$.

*3) Differences Between Genders:* We compare the long-term overall entropies of male and female students and observe no statistical difference between them. This finding confirms the discovery in previous work [11].

– *Short-term entropy:* When students' entropies are compared on different days of the week, gender differences appear. Females have relatively larger daily $S_{td}$ while this pattern is not apparent for $S_{ti}$. For each day of the week, males' and females' average entropies follow very similar shapes, but there are differences between genders primarily between Tuesday and Thursday.

Fig. 8 shows that females' $S_{td}$ entropies are slightly higher than males' (at least 2% on Tuesdays and Thursdays for both measurements).

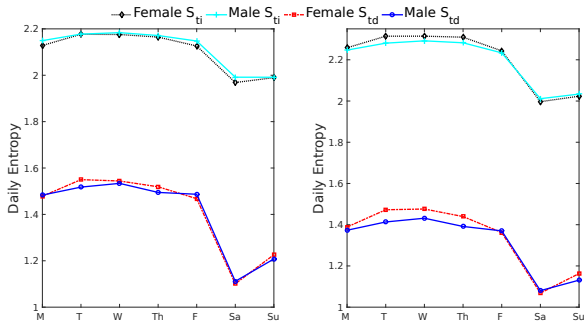


Fig. 8. The average entropy for each day of the week for males and females. The figure on the left shows the result using the "inclusive" approach and the one on the right is the "strict" approach.

There are some nuanced differences between the two approaches as the "strict" approach yields slightly more observable differences between male and females during the middle of a week.

We use the non-parametric Kolmogorov-Smirnov test to determine if males' and females' entropies on different days of the week follow the same distribution. The test yielded statistical differences ($p < .001$) on Tuesday, Thursday, and Sunday with both the "strict" and the "inclusive" approaches. *Discussion:* Though the difference on certain days of a week between males and females is small, we argue that this difference is significant as this pattern holds throughout the entire observation period. Fig. 9 shows daily average entropies

for males and females throughout the 139-day observation period. Females' daily entropies are slightly larger most of the time, especially during "normal" weeks (week 9 is spring break, around day 60; week 15 and thereafter are summer break, around day 100).

However, the overall entropies of females and males are statistically similar as stated earlier in this subsection. This discrepancy is quite intriguing since females' larger entropies on a daily basis would imply that females have higher cumulative entropies throughout the 139-day period.

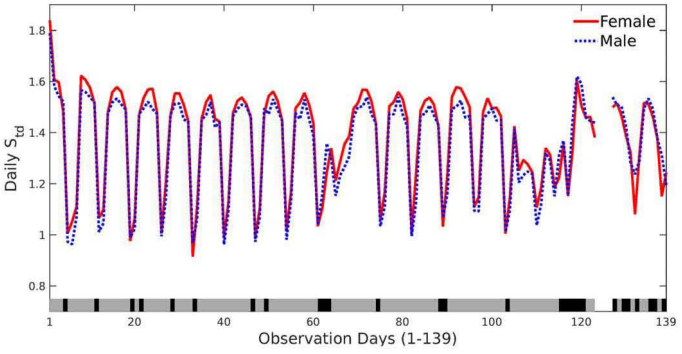


Fig. 9. Males' and females' average daily entropies during the 139-day period. The heat map near the bottom of the figure indicates if females' average entropy is higher than males' on that day (gray) or not (black). Days 124-126 mostly contain laptop devices due to infrastructure problems. The average entropies for these 3 days are omitted.

We also calculate the average weekly entropies for males and females for each of the 19 complete weeks in the dataset. The weekly averages are almost identical between males and females (Kolmogorov-Smirnov test with $p = 0.9563$). Thus, females have larger daily entropies, yet the gender difference disappears as the measurement time interval increases to weeks or longer. One possible explanation for this discrepancy is as follows: females' mobility is more random than males' on a daily basis, but females tend to repeat their daily routines more strictly than males. Thus, when entropies are measured over a longer period, the randomness of females' daily mobility is countered by their mobility's long-term regularity.

*4) Overall Patterns:*

– *Long-term entropy:* Long-term entropy $S_{td}$ follows a bimodal distribution. As shown in Fig. 10(A), the left mode is centered at $\sim$0.8 and the right mode is centered at $\sim$1.65, twice the left mode. This means that students whose entropies are near the left peak have (on average) less than $2^{0.8} \approx 1.7$ locations as possible choices if students randomly pick their locations, whereas students whose entropies are near the right peak have $2^{1.65} \approx 3$ locations as possible choices. The overall mean entropy is 1.2655 for all students.

*Discussion:* The overall mean entropy calculated from this WLAN dataset is slightly larger than the result from Song et al. [11]. We hypothesize this numerical difference is due to the finer granularity of WLAN data compared with the call detail records (CDR) data.

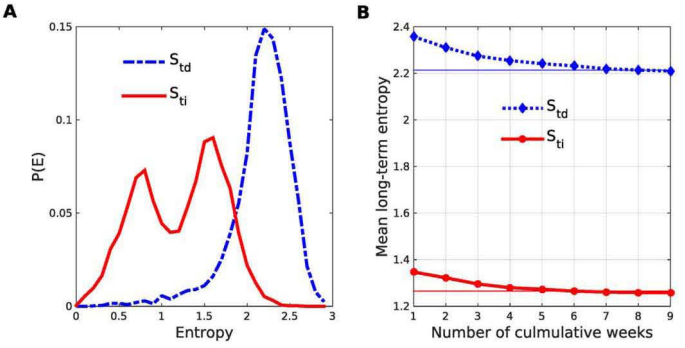


Fig. 10. (A). Overall long-term entropies $S_{td}$ and $S_{ti}$, of students based on cumulative trajectories across the 139-day span. (B). The convergence result of long-term overall entropies ($S_{ti}$ and $S_{td}$).

To the best of our knowledge, bimodal entropy distribution has not been reported in the literature, although Pappalardo et al. [30] find two distinct profiles in human mobility through the study of CDR and GPS data. We investigate possible reasons for the bimodal distribution and test the correlation between entropy and other metrics such as the radius of gyration, the number of unique locations students visit, and the total number of locations they visit. None are strongly correlated with the entropy distribution's bimodal pattern. We also find that students' average stay time correlates with their entropy. The average stay time for students whose entropies are near the left and right modes are 80.5 min and 29.1 min, respectively. This finding reaffirms the concept of entropy. Students with longer stay time (on average) have less random mobility, yielding smaller entropies. Similarly, students with shorter stay time (on average) change locations along their trajectories more actively, yielding larger entropies.

We increase time slice length $\Delta t$ from 30 min to 110 min with 10 min increments. We discover that once the time slice $\Delta t$ reaches 110 min, the bimodal pattern no longer holds in the time-dependent entropy ($S_{td}$) distribution. This observation indicates that the unit of temporal measurement is an important parameter in the measurement of human mobility.

We also investigate the convergence time for mobility entropies. Fig. 10(B) shows the convergence of mean $S_{td}$ and $S_{ti}$ when the measurement length spans 1–9 weeks. $S_{td}$ converges when the measurement length is 6 weeks while $S_{ti}$ converges at 8 weeks.

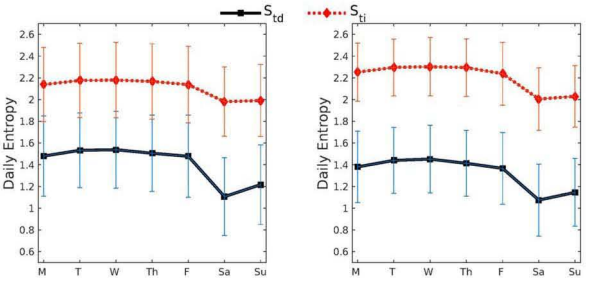


Fig. 11. Average entropy for each day of the week. The figure on the left uses the "inclusive" approach and the one on the right uses the "strict" approach.

– *Short-Term Entropy:* Patterns of average daily entropy (both

$S_{td}$ and $S_{ti}$) follow hook shapes similar to those in Figs. 5 and 8. Fig. 11 shows the average daily entropy for all students.

*Discussion:* This hook-shaped pattern for days of a week is confirmed by examining weekly patterns during the 139-day period of data collection. The only exceptions are the special weeks of a semester such as spring break and final exams during which student mobility patterns change dramatically.

Changing mobility patterns on weekends can be partially attributed to students lingering longer at locations such as libraries and dormitories, resulting in lower mobility. However, the slightly larger mobility entropy from Tuesday to Thursday is an interesting phenomenon.

We investigate location-related metrics including the average number of (unique and total) locations visited by students as shown in Table IV. As can be seen, both the number of unique and total locations visited by students on each day of the week follow a very similar trend to mobility entropies.

| **Days** | Mon | Tue | Wed | Thu | Fri | Sat | Sun |
|---|---|---|---|---|---|---|---|
| **Mean** $S_{td}$ | 1.48 | 1.53 | 1.54 | 1.51 | 1.48 | 1.10 | 1.22 |
| **Mean** $S_{ti}$ | 2.14 | 2.18 | 2.18 | 2.17 | 2.14 | 1.98 | 1.99 |
| **Unique Locations** | 11.79 | 12.55 | 12.70 | 12.42 | 12.10 | 9.18 | 8.83 |
| **Total Locations** | 17.56 | 18.64 | 19.00 | 18.39 | 17.86 | 13.92 | 13.22 |

TABLE IV
SHORT-TERM METRICS FOR EACH DAY OF A WEEK

## V. CONCLUSION

This paper presented the results of a human mobility predictability study across demographics (age, gender, and academic major). The results were inferred from a comprehensive WLAN dataset containing over 41 million log entries for over 5,000 students across several hundred buildings. The size of the campus enabled our results to complement mobility patterns at a large community level. We analyzed the predictability of students' mobility via long-term and short-term entropies. Our study showed that demographic information warrants consideration in mobility research as mobility patterns vary across demographic categories. We observed large increases in mobility entropy with respect to age. Students' overall long-term mobility entropy showed a bimodal distribution with differences across academic majors. Average daily entropies showed a hook-shaped pattern throughout the week for everyone. In general, females' daily entropies were slightly higher than those of males, but females' entropies were countered by their potential regularity over a longer period. Students with undecided majors had lower entropies than other students throughout the week. We offered potential explanations for the observed phenomena.

## REFERENCES


[1] Z. D. Boren, "There are officially more mobile devices than people in the world," 7 Oct. 2014, http://www.independent.co.uk/life-style/gadgets-and-tech/news/there-are-officially-more-mobile-devices-than-people-in-the-world-9780518.html.

[2] ITU, "Mobile subscriptions near the 7-billion mark; Does almost everyone have a phone?" Jul.–Aug. 2013, https://itunews.itu.int/En/3741-Mobile-subscriptions-near-the-78209billion-mark-Does-almost-everyone-have-a-phone.note.aspx.

[3] D. Naboulsi, M. Fiore, S. Ribot, and R. Stanica, "Mobile Traffic Analysis: a Survey," Université de Lyon, Tech. Rep. hal-01132385, 2015.

[4] F. Giannotti, M. Nanni, D. Pedreschi, F. Pinelli, C. Renso, S. Rinzivillo, and R. Trasarti, "Unveiling the complexity of human mobility by querying and mining massive trajectory data," *VLDB J.*, vol. 20, no. 5, pp. 695–719, 2011.

[5] M. C. González, C. A. Hidalgo, and A.-L. Barabási, "Understanding individual human mobility patterns." *Nature*, vol. 453, no. 7196, pp. 779–782, 2008.

[6] D. Brockmann, L. Hufnagel, and T. Geisel, "The scaling laws of human travel." *Nature*, vol. 439, no. 7075, pp. 462–465, 2006.

[7] N. Aschenbruck, A. Munjal, and T. Camp, "Trace-based mobility modeling for multi-hop wireless networks," *Comput. Commun.*, vol. 34, no. 6, pp. 704–714, 2011.

[8] I. Rhee, M. Shin, S. Hong, K. Lee, S. J. Kim, and S. Chong, "On the Levy-Walk Nature of Human Mobility," *IEEE/ACM Trans. Netw.*, vol. 19, no. 3, pp. 630–643, 2011.

[9] D. Balcan, V. Colizza, B. Gonçalves, H. Hu, J. J. Ramasco, and A. Vespignani, "Multiscale mobility networks and the spatial spreading of infectious diseases," *PNAS*, vol. 106, no. 51, pp. 21 484–21 489, 2009.

[10] J. Yuan, Y. Zheng, and X. Xie, "Discovering regions of different functions in a city using human mobility and pois," in *KDD*, 2012, pp. 186–194.

[11] C. Song, Z. Qu, N. Blumm, and A.-L. Barabási, "Limits of predictability in human mobility." *Science*, vol. 327, no. 5968, pp. 1018–1021, 2010.

[12] S.-M. Qin, H. Verkasalo, M. Mohtaschemi, T. Hartonen, and M. Alava, "Patterns, Entropy, and Predictability of Human Mobility and Life," *PLoS One*, vol. 7, no. 12, Dec. 2012.

[13] X. Lu, E. Wetter, N. Bharti, A. J. Tatem, and L. Bengtsson, "Approaching the limit of predictability in human mobility," *Scientific Reports*, vol. 3, 2013.

[14] E. Cho, S. A. Myers, and J. Leskovec, "Friendship and mobility: User movement in location-based social networks," in *KDD*, 2011, pp. 1082–1090.

[15] T. Henderson, D. Kotz, and I. Abyzov, "The Changing Usage of a Mature Campus-Wide Wireless Network," in *MobiCom*, 2004.

[16] F. Hernandez-Campos and M. Papadopouli, "A Comparative Measurement Study of the Workload of Wireless Access Points in Campus Networks," in *PIMRC*, 2005.

[17] M. Balazinska and P. Castro, "Characterizing Mobility and Network Usage in a Corporate Wireless Local-Area Network," in *MobiSys*, 2003.

[18] M. McNett and G. M. Voelker, "Access and Mobility of Wireless PDA Users," *Mobile Comput. Commun. Rev.*, vol. 9, no. 2, pp. 40–55, 2005.

[19] W.-J. Hsu, D. Dutta, and A. Helmy, "Structural Analysis of User Association Patterns in University Campus Wireless LANs," *IEEE Trans. Mobile Comput.*, vol. 11, no. 11, pp. 1734–1748, 2012.

[20] CRAWDAD, http://crawdad.org.

[21] L. Pappalardo, M. Vanhoof, L. Gabrielli, Z. Smoreda, D. Pedreschi, and F. Giannotti, "An analytical framework to nowcast well-being using mobile phone data," *Int. J. Data Sci. Analytics*, vol. 2, no. 1, pp. 75–92, 2016.

[22] C. Tuduce and T. Gross, "A Mobility Model based on WLAN Traces and Its Validation," in *INFOCOM*, 2005.

[23] A. Clauset, C. R. Shalizi, and M. E. J. Newman, "Power-Law Distributions in Empirical Data," *SIAM Rev.*, vol. 51, no. 4, pp. 661–703, Nov. 2009.

[24] M. Kim, D. Kotz, and S. Kim, "Extracting a mobility model from real user traces." in *INFOCOM*, vol. 6, 2006, pp. 1–13.

[25] W.-J. Hsu, T. Spyropoulos, K. Psounis, and A. Helmy, "Modeling Spatial and Temporal Dependencies of User Mobility in Wireless Mobile Networks," *IEEE/ACM Trans. Netw.*, vol. 17, no. 5, pp. 1564–1577, 2009.

[26] C. Song, T. Koren, P. Wang, and A.-L. Barabasi, "Modelling the scaling properties of human mobility," *Nat. Phys.*, vol. 6, no. 10, pp. 818–823, Oct 2010, http://dx.doi.org/10.1038/nphys1760.

[27] J. Yoon, B. D. Noble, M. Liu, and M. Kim, "Building Realistic Mobility Models from Coarse-Grained Traces," in *MobiSys*, 2006.

[28] Google Maps, http://maps.google.com.

[29] I. Kontoyiannis, P. H. Algoet, Y. M. Suhov, and A. J. Wyner, "Nonparametric entropy estimation for stationary processes and random fields, with applications to English text," *IEEE Trans. Inf. Theory*, vol. 44, no. 3, pp. 1319–1327, 1998.

[30] L. Pappalardo, F. Simini, S. Rinzivillo, D. Pedreschi, F. Giannotti, and A.-L. Barabási, "Returners and explorers dichotomy in human mobility," *Nature Commun.*, vol. 6, p. 8166, Sep 2015, http://dx.doi.org/10.1038/ncomms9166.