On Profiling Mobility and Predicting Locations of Wireless Users

Joy Ghosh, Matthew J. Beal, Hung Q. Ngo,* Chunming Qiao†
Dept. of Computer Science and Engineering
The State University of New York at Buffalo, Buffalo, NY, U.S.A.
[joyghosh, mbeal, hungngo, qiao]@cse.buffalo.edu

ABSTRACT

In this paper, we analyze a year long wireless network users' mobility trace data collected on ETH Zurich campus. Unlike earlier work in [4, 18], we profile the movement pattern of wireless users and predict their locations. More specifically, we show that each network user regularly visits a list of places such as a building (also referred to as "hubs") with some probability. The daily list of hubs, along with their corresponding visit probabilities, are referred to as a *mobility profile*. We also show that over a period of time (e.g., a week), a user may repeatedly follow a mixture of mobility profiles with certain probabilities associated with each of the profiles. Our analysis of the mobility trace data not only validate the existence of our so-called sociological orbits [8], but also demonstrate the advantages of exploiting it in performing *hub-level location predictions*. In particular, we show that such profile based location predictions are more precise than common statistical approaches based on observed hub visitation frequencies alone.

Categories and Subject Descriptors: C.2 COMPUTER-COMMUNICATION NETWORKS: Miscellaneous

General Terms: Algorithms, Design, Human Factors, Verification

Keywords: WLAN mobility trace analysis, Sociological orbits, Mobility profiles, Location prediction, Mobile wireless networks

1. INTRODUCTION

The mobility of users forming a mobile wireless network causes changes in the network connectivity and may even lead to intermittently connected networks. On one hand, nodal mobility may increase the overall network capacity [10]. On the other hand, it is challenging to locate users and route messages within the network. Many researchers have tried to model practical mobility in various ways to achieve different goals. Earlier work on mobility modeling [3] was done mostly with Mobile Ad hoc NETworks (MANET) in mind. While the authors in [15, 19] performed physical location prediction via continuous short-term and short-range tracking of user movement, we had leveraged on our assumptions on "sociological orbits" (which however was not supported by valid evidence) to perform efficient routing within MANETs [8]. In this work, we present empirical evidence to support our prior orbital claims and illustrate its advantages in hub-level location predictions.

Delay Tolerant Networks (DTN) has received a lot of interest recently. For example, researchers [2, 21] have also suggested the concept of controlled mobility to aid in mobile ad hoc routing. However, the main focus of these projects were not on the mobility pattern of the individual users. In a recent work [9] we have also shown how to use pair-wise users' contact probability (derived from the mobility profiles) in efficient routing within Intermittently Connected Mobile Ad Hoc Networks (ICMAN).

Our study of user mobility traces is motivated by the need to extract practical mobility information, which may potentially benefit applications such as location approximation and routing within all types of wireless networks such as MANETs, DTNs, etc. More specifically, it is noted that wireless users belong to a larger social environment and as such, their movement behavior is subject to several location dependent sociological constraints (in addition to speed limits and specific walkways, as described in [1]). In particular, on any given day, each user may visit a list of places of some social importance (which we referred to as "hubs" in [8]) in some probabilistic manner, creating what we refer to in this paper as a "mobility profile".

Example applications of such profiles may be to monitor air (or, water) quality in an infrastructure-less environment and its impact on the health of the people who live or work there, or to detect and control the spread of a flu virus [20]. In these applications, some people wearing tiny sensors with limited transmission range can act as "carriers" and as a part of their "social routine" may travel near to access points for uploading (or downloading) the sensor data (or control messages). Others will only be able to send data to a person sharing the same "hub" as a part of its orbital pattern, and data is forwarded via such people to "carriers". After a remote center processes the collected data, it

*Research is supported in part by NSF CAREER Award CCF-0347565
†Research is supported in part by NSF SGER grant CNS-0553273

REALMAN'06, May 26, 2006, Florence, Italy.

may require more intensive data collection at only selected locations and/or by selected persons. Knowing the orbital patterns of the persons helps to target the right subset of people, thereby reducing unnecessary flooding of the data request, and also saves the energy/bandwidth in collecting uninterested data. Knowing the contact probability alone may not be sufficient here.

In this work, we not only validate the existence of such mobility profiles via mobility trace analysis, but also show that in practice, a user is usually associated with a probabilistic mixture of multiple profiles. The data analyzed in this paper is collected on the ETH Zurich campus and is similar in content (i.e. AP system logs) to that available from the Dartmouth campus. However, compared to the most related (and yet much different) work in [4], this paper focusses only on the *user-centric* parameters like the user mobility profiles and its applications, whereas [4] focusses more on AP-centric parameters. Our mobility profile based hub-level location prediction is shown to be more precise than common statistical methods. Note that although this work analyzes data from a campus-wide wireless access network (instead of a MANET, as data from former is more readily available), our mobility profiling and location prediction techniques are applicable to other types of networks as well, since the movement of users is ultimately influenced by their social environment.

The rest of this paper is organized as follows. In Section 2, we discuss our sociological orbit framework and its parameters. In Section 3, we study the user-centric parameters and present a clustering algorithm using a *Mixture of Bernoulli's distribution* to analyze user mobility profiles. In Section 4, we highlight the advantages of profiling users' mobility by comparing profile based hub-level location predictions to predictions based on general statistical methods. We conclude this work in Section 5.

2. SOCIOLOGICAL ORBIT FRAMEWORK

In this section, we briefly describe and enhance the sociological orbit framework we proposed in [8]. In the real world, it is observed that users routinely spend a considerable amount of time at a few specific place(s), referred to as hub(s). For example, in a WLAN scenario a hub may be just a building floor, or the entire building, depending on the network scale. Although, it is hard (and may even violate privacy) to keep track of an individual at all times, one can still take advantage of the fact that most users' movements are within, and in between, a list of hubs.

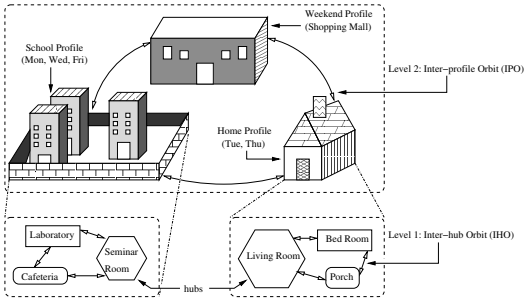


Figure 1: A hierarchical view of sociological orbits

Let us consider a graduate student with classes only on Monday, Wednesday and Friday, when he/she is found spending most of the time in either a laboratory, or a seminar room, or the cafeteria (each of which shall form a "hub" in this example) on a school campus, as shown in Figure 1. The actual list of hubs visited by the student on the same day is called a "hub list". Even if such hub lists may vary across days, that variation is only marginal (as shown later in Section 3). In most cases, a number of hub lists over a period of days may be clustered together and represented by a single "weighted hub list", where the weight associated with each hub denotes the probability of the student visiting that hub within that period. In this work, we shall refer to such a weighted hub list to be a user's "Mobility Profile", and the movement in between the hubs within a profile as an "Inter-hub Orbit" (IHO). If one wishes to locate the student on a school day, knowing this *School Profile* shall be helpful, where one can most probably find him/her in either the laboratory, or the seminar room, or the cafeteria, without having to look all over the campus.

In real life, it is observed (and later verified from the analyzed data) that a user over long periods of time is usually associated with more than one mobility profile, mixed with certain probabilities. This is shown in Figure 1 as the *Weekend Profile* and the *Home Profile* to account for the student's remainder of the week. Such a movement in between multiple profiles at a higher level is referred to as the "Inter-profile Orbit" (IPO). Over different periods of time, this mixture of profiles may change, causing what we call an "IPO Timeout". The IPO and the IHO together constitute the hierarchical *sociological orbit* at two different levels. In this paper, such orbital mobility information is shown to be helpful in improving accuracy of hub-level location predictions over statistical based methods.

Table 1: Orbital Parameters

Category	Parameters
Hub-centric	Hub Form
	Hub Visits
	Hub Stay Time
User-centric	Mobility Profiles
	Hub List Size

To formalize the sociological orbit framework, we divide the orbital parameters into two categories: *Hub-centric*, and *User-centric*, as listed in Table 1. On the *hub-centric* side, the *Hub Form* depends on the actual definition of a hub in the network being modeled; *Hub Visits* denotes the number of users visiting a hub in a given period; and the *Hub Stay Time* is the amount of time a user spends at one stretch within a hub. On the *user-centric* side, the *Mobility Profile Parameters* include a list of hubs and their corresponding weights, and the *Hub List Size* refers to the number of unique hubs visited by a user on a day. Since the AP-centric study in [4] is similar to our study of the hub-centric parameters in [7], in this paper we focus only on the analysis of the user-centric aspects related to sociological orbits.

In the following sections we analyze wireless network users' mobility trace data collected on ETH Zurich campus from 1st April, 2004 till 31st March, 2005. There were a total of 13,620 users, 43 buildings, and 391 Access Points (AP). The data was obtained as system logs from the APs which

recorded the *association, disassociation, missed polls*, and *roaming* events for users during the given period. First, to study the observed distribution of the *hub-centric* parameters of the framework, we setup an Oracle database with these traces and employed standard SQL queries[1]. Second, to analyze the *user-centric* parameters we employ a clustering algorithm using a *Mixture of Bernoulli's*. We also develop efficient methods to model and analyze mobility profiles to validate the existence of sociological orbits. Finally, we use the mobility profiles to do hub-level location predictions more precisely than common hub visitation statistics based methods.

3. USER-CENTRIC PARAMETERS

In this section, we shall analyze the *user-centric* parameters by examining individual network user's movement. First, in order to span different degrees of network activity amongst users, we divide all the users in different user groups based on the number of days they are found to be "active" within the network (i.e., associated with at least one AP in the day). In Figure 2, we plot the fraction of total population vs. the number of their active days. The x-axis shows a range of values, i.e., 25 denotes up to 25 active days, 50 denotes anywhere between 26 and 50 active days, and so on. 80% of the total population is seen to be active for only 25 days or less in an year with the number of more active users decreasing significantly. So we only consider the users active for 150 days or less creating 6 groups: G_1 (0 to 25) through G_6 (126 to 150).

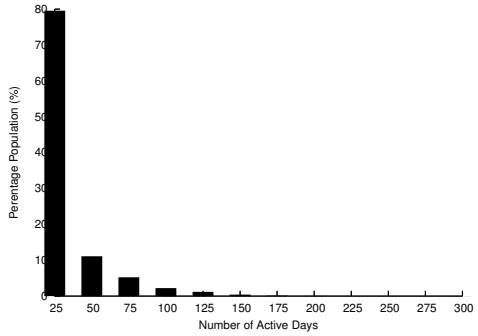


Figure 2: Number of active days for users

Second, we wish to choose one user to represent each group who is the "most active" within that group, giving us more samples. In short, we wish to maximize both the number of active days and the hub list size within each group. For a given group G_j, let $D^j_{\max}$ and $L^j_{\max}$ be the maximum number of active days and maximum average hub list size respectively, across all users in G_j. Let the pair D^j_i and L^j_i denote the number of active days and the average hub list size for a particular user i in G_j. Then, to represent group G_j we need to find a user who can *minimize*

$$\alpha.\frac{D^j_{\max} - D^j_i}{D^j_{\max}} + \beta.\frac{L^j_{\max} - L^j_i}{L^j_{\max}} \quad (1)$$

where, α and β are weights associated with each term. The results of using (1) with $\alpha = 1$ and $\beta = 1$ (we weigh both the number of active days and the hub list size equally) is summarized in Table 2. The basic intuition behind selecting the "most active" user from each group is the availability of more statistically significant mobility data for such an individual. ***Note that a user with more sample data is not necessarily more predictable, and hence this choice is not biased.*** Alternately, one may also select the sample users from each group to find users that are either "least active" or, "active on average". At the same time, studying users from different groups help represent different activation periods as seen in Figure 2.

Table 2: Sample users from all Groups

Group	MAC	D_i	L_i
G_1	0004.2396.92ab	24	2.29
G_2	0001.e30d.d737	49	3.27
G_3	0004.2398.82c0	71	4.08
G_4	0020.e089.9376	98	2.46
G_5	0004.2396.8ced	119	2.13
G_6	0005.4e41.cf1d	126	2.63

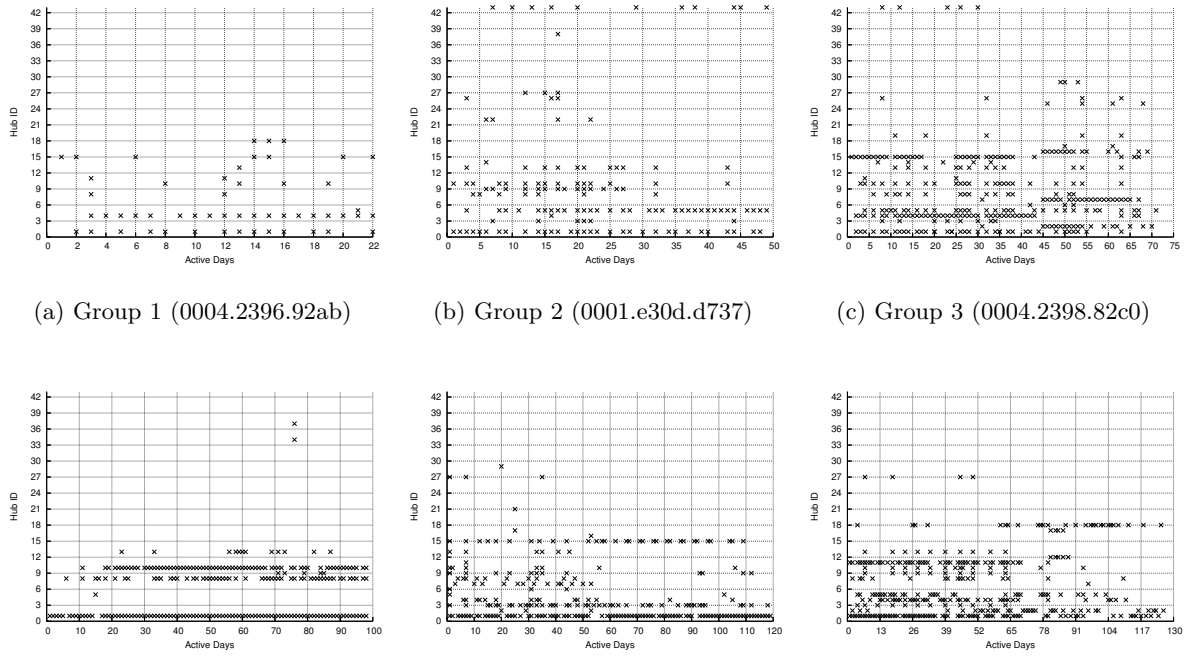
3.1 Model for Analysis of Mobility Profiles

We now present a study on the mobility of these 6 sample users. We first obtain their hub stay times in all the hubs during their active period and filter out all values of 5 minutes or less as noises (i.e., very brief hub stay durations). Then we obtain 2-D plots in Figure 3 showing only which hub(s) is(are) visited by a user (for more than 5 minutes) on a given day. We use h ($1 \leq h \leq H$) to denote the unique hub id and i ($1 \leq i \leq n$) to denote the day index, where H and n are the total number of hubs and days respectively. On each day i, we define a user's hub list to be a binary vector of hub associations $\mathbf{y}^{(i)} = [\mathbf{y}^{(i)}_1, \ldots, \mathbf{y}^{(i)}_H]$ where each element $\mathbf{y}^{(i)}_h \in \{0, 1\}$ such that $\mathbf{y}^{(i)}_h$ is equal to 1 if hub h was visited on day i, and zero otherwise. Next, we define an H dimensional space, where each dimension refers to a hub. The hub list $\mathbf{y}^{(i)}$ for a user in any given day i may then be represented by a point in this space where each element $\mathbf{y}^{(i)}_h$ in the vector represents a binary value along each dimension in the space. For a particular user, similar hub lists on different days would generate several overlapping points whereas, two hub lists that differed only in terms of one or two hubs would generate points "close" to each other in this space. We use a clustering algorithm that helps define this concept of "closeness" by considering hub lists that say only differ in a maximum of 1 or, 2 hubs to be "close" and to belong to the same cluster. The mean of the cluster, which is a weighted hub list, then represents a mobility profile, as is described in more detail below.

3.2 Using a Mixture of Bernoulli's distribution

A suitable choice to model the binary hub visitation vectors is a *Mixture of Bernoulli's distribution*. In this mixture model there shall be more than one mixture component where, each component is considered an unique mobility profile represented by the component mean. Thus, a profile is nothing but a distribution over the hub visitation probabilities (i.e., a weighted hub list). We refrained from using

[1]We thank Nirmal Thangaraj for developing the SQL queries



(a) Group 1 (0004.2396.92ab) (b) Group 2 (0001.e30d.d737) (c) Group 3 (0004.2398.82c0)

(d) Group 4 (0020.e089.9376) (e) Group 5 (0004.2396.8ced) (f) Group 6 (0005.4e41.cf1d)

Figure 3: Daily hub visitation patterns of all sample users

the commonly used *Mixture of Gaussian* model because the domain of the Gaussian variable, being $(-\infty, \infty)$, is clearly not suitable for binary valued vectors. Assuming that the current mobility profile of a user is known, we model each hub visitation by a user as an independent event. On the other hand, if the current profile is not known, the general probability of a user visiting a hub is dependent on the probability associated with each mobility profile. The latter fact is crucial, since it allows for the knowledge of a user's hub visits to help infer the current mobility profile and therefore the probabilities of visits to other hubs on the same and future days, as shown later in Section 4.

More formally, we denote the complete trace of hub visits across all n days with the symbol Y, which is the collection $Y = \{\mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(n)}\}$. The total probability of Y is given by the product of a mixture of independent Bernoulli distributions as follows: $p(Y) = \prod_{i=1}^{n} p(\mathbf{y}^{(i)})$, where, $p(\mathbf{y}^{(i)}) = \sum_{j=1}^{k} p(j) \prod_{h=1}^{H} p(\mathbf{y}_h^{(i)}|\rho_{j,h})$. Here, k is the number of mixture components (or, mobility profiles); $p(j)$ is the probability of following profile j; $\rho_{j,h}$ is the probability of visiting hub h on a day when following profile j. This framework is a generative Bayesian model in the sense that it defines a probability to every possible outcome, or pattern, that can be produced for Y.

This mixture model is trained using the Expectation- Maximization (EM) algorithm of Dempster, Laird and Rubin [6]. By employing consecutive Expectation (E)- and Maximization (M)- steps, the probability of the entire data set Y is guaranteed to monotonically increase (or, remain the same). The E-step consists of computing the posterior probability of membership of a datum (or, hub list) across the k mixture components (or, mobility profiles). Intuitively, at this E-step we look at each hub list and try to guess the mobility profile being followed on that particular day. Formally, this corresponds to computing the *responsibilities* of each component in the mixture, denoted by $r_j^{(i)}$, such that $\sum_{j=1}^{k} r_j^{(i)} = 1$, and are found using Bayes' theorem:

$$\textbf{E-step} \qquad r_j^{(i)} \equiv p(j|\mathbf{y}^{(i)}) = \frac{p(j)p(\mathbf{y}^{(i)}|j)}{p(\mathbf{y}^{(i)})} \qquad (2)$$
$$\forall i = 1, \ldots, n \quad \text{and} \quad \forall j = 1, \ldots, k \ .$$

The M-step of the EM algorithm updates the parameters of each of the k components of the mixture model, in light of the responsibilities $r_j^{(i)}$ computed in the E-step. In other words, at this M-step we look at the probabilistic associations of the hub lists with each profile computed in the E-step, and update both the probabilities associated with each profile (i.e., mixing proportions), and the probabilities associated with each hub visitation within a profile. Thus, formally the parameters of the mixture model are: the mixing proportions, denoted by vector $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_k)$ where $\pi_k = p(k)$ such that $\sum_{j=1}^{k} \pi_j = 1$; and for each mixture component j, there is a vector of dimension H of probabilities of each hub being used, denoted by $\boldsymbol{\rho}^{(j)} = (\rho_{j,1}, \ldots, \rho_{j,H})$. Thus each component in the mixture represents a mode of a user's interaction with a subset of the H hubs available (i.e., each profile is nothing but a weighted hub list). The updates to the parameters in the M-step are as follows:

$$\textbf{M-step}, \boldsymbol{\pi} \qquad \pi_j = \frac{1}{n} \sum_{i=1}^{n} r_j^{(i)} \qquad (3)$$
$$\forall j = 1, \ldots, k \ .$$

and

$$\textbf{M-step}, \boldsymbol{\rho} \qquad \rho_{j,h} = \frac{\sum_{i=1}^{n} r_j^{(i)} \mathbf{y}_h^{(i)}}{\sum_{i=1}^{n} r_j^{(i)}} \qquad (4)$$
$$\forall j = 1, \ldots, k \quad \text{and} \quad \forall h = 1, \ldots, H \ .$$

In this study, for each user we choose the number of components k (i.e, profiles) for each mixture model by visual inspection of the data distribution and initialized the *mixing proportions* and *component means* at random such that each profile has moderate associativity with hub lists. An alternate approach may include approximate Bayesian model selection techniques, e.g. via the Bayesian Information Criterion (BIC; [14]) or, other criteria. Figure 4 shows the pattern of mobility profiles over all the days. Table 3 lists both the probability that a user is in a given profile, and the probability that a hub is visited when following a particular profile. As an example, from Figure 4(a) we find that the sample user from group G_1 is following his/her mobility profile 1 on day 14. From Table 3 we see that given profile 1 for that user, the hub visitation probabilities indicate *definite* visits to hubs $1, 4, 15$ and 18 on day 14, which may then be verified from his/her actual hub list distribution shown in Figure 3(a).

3.3 Hub List Size Distribution

The results in Table 3 may seem to indicate that several users tend to visit many hubs in any given day as their mobility profiles include multiple hubs. Hence, to study the distribution of the hub list sizes of our sample users we generate daily hub lists for each of them over their individual activation period based on their mobility profiles. More specifically, for each day we first choose one of their possible profiles at random following the mixing proportions, and then generate visits to each hub individually following the hub visitation probabilities in that chosen profile. We then obtain the aggregated (i.e., across all sample users) Hub List Size distribution and compare it with the actual distribution observed in the trace data. As seen in Figure 5, both the observed and the generated hub list sizes are distributed almost identically, and shorter (≤ 3) hub list sizes occur most often.

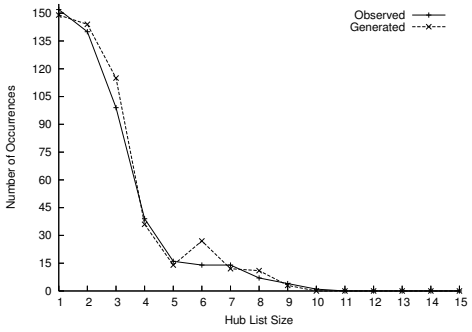


Figure 5: Observed vs. Generated Hub List Size

This work presenting a formal method of building mobility profiles constitutes one of our major contributions.

4. LOCATION PREDICTIONS

In this section, we highlight another important contribution of our work by showing how the mobility profiles may be useful in making hub-level location predictions with more accuracy than general statistical methods based on hub visitation frequency information alone. More specifically, we first show an efficient way to apply the clustering algorithm described in Section 3.2 and identify the right mixture of mobility profiles for each user. We then focus on two types of profile based predictions: *Unconditional Prediction*, where given the hub visit information over a window of n days, we wish to predict the hub visit patterns for the next window of n days; *Conditional Prediction*, where given that we can identify the current mobility profile of a user (based on available information about a hub a user either visited, or plans to visit), we wish to find the probability of that user visiting another hub in that same day.

4.1 A Mixture of Mobility Profiles

From Figure 4, it becomes lucid that the seemingly random movements of a user as seen from Figure 3 can now be systematically described via a mixture of mobility profiles over a period of time. However, since this mixture will eventually change, we still need an efficient method to identify the right mixture of profiles describing the user's movement pattern over a given period. One may use the mobility traces of hub visits collected over 7 days (i.e., a week) to determine the possible mobility profiles and their corresponding mixing proportions using the *Mixture of Bernoulli's* described in Section 3.2. It is then possible to identify the appropriate mixture to include all the profiles with a corresponding mixing proportion greater than some specified threshold. One may then choose to only consider this specific mixture for the next 7 days (or even more), when the next mixture update is performed (only if there is a substantial change in movement pattern). Later in this section, we show that even with such infrequent updates our mobility profiles are able to predict daily hub-level locations with more accuracy than common statistical methods.

4.2 Unconditional Prediction

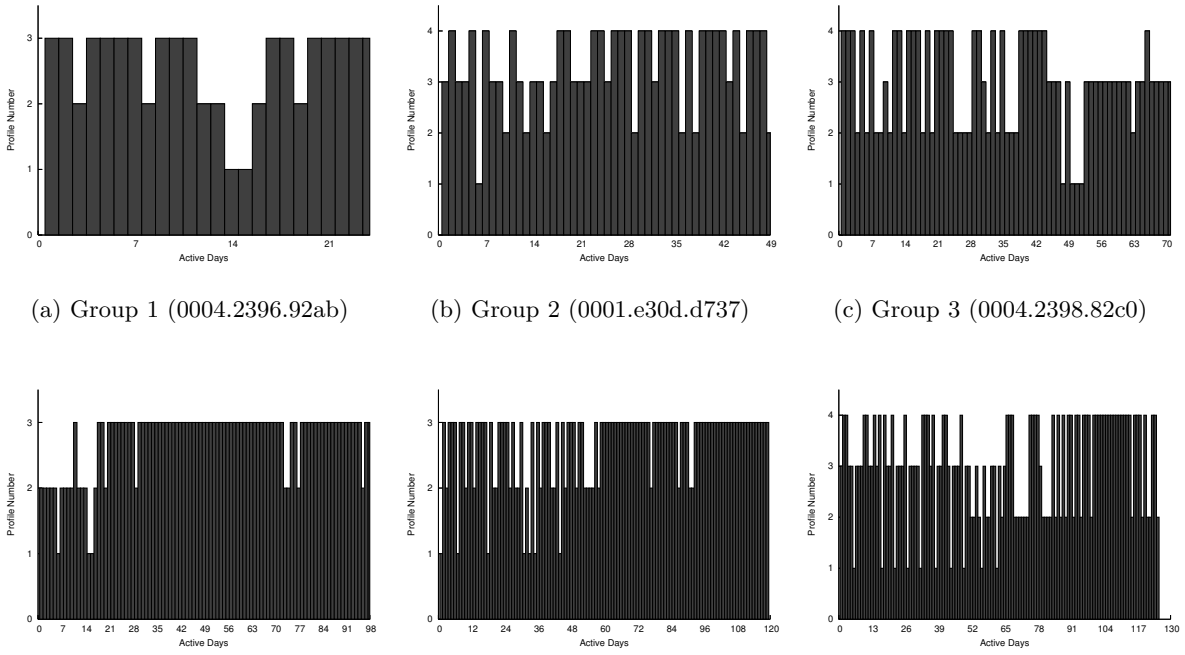
In this part, we study the accuracy of the unconditional profile based hub-level location prediction, and compare it with that made from statistical observation alone. We again consider only the sample users.

4.2.1 Statistical based prediction

In this method, we assume no knowledge of mobility profiles and hence no clustering by the mixture of Bernoulli's distribution is performed. One simply collects the mobility traces of a sample user for a few days, and then based on the hub visit frequencies determines the user's hub visit probabilities in future, which can then be used for hub visit predictions and may be compared with the observed hub lists to compute the Statistical based Prediction Error (SPE) rate as

$$SPE = \frac{Incorrect\ number\ of\ hub\ predictions}{Total\ number\ of\ hubs} \qquad (5)$$

We consider 2 variations of this statistical approach. In the first one (***SPE-ALL***), prediction for day $n+1$ is done based on the history of past n days, after which the hub visit probabilities are recomputed based on the past $n+1$ days



(a) Group 1 (0004.2396.92ab) (b) Group 2 (0001.e30d.d737) (c) Group 3 (0004.2398.82c0)

(d) Group 4 (0020.e089.9376) (e) Group 5 (0004.2396.8ced) (f) Group 6 (0005.4e41.cf1d)

Figure 4: Daily distribution of mobility profiles

Table 3: Mobility Profile Parameters

G_1 Profiles j	Mix. Prop. $\boldsymbol{\pi}_j$	Hub ID h (Hub Visitation Probability $\rho_{j,h}$)
1	0.08	1(1.0), 4(1.0), 15(1.0), 18(1.0)
2	0.31	1(1.0), 4(0.83), 8(0.27), 10(0.54), 11(0.27), 13(0.13), 18(0.13)
3	0.61	1(0.38), 4(0.81), 5(0.07), 15(0.34)
G_2 Profiles j	Mix. Prop. $\boldsymbol{\pi}_j$	Hub ID h (Hub Visitation Probability $\rho_{j,h}$)
1	0.02	1(1.0), 9(1.0), 14(1.0)
2	0.14	1(0.14), 4(0.14), 5(0.49), 43(1.0)
3	0.31	1(1.0), 3(0.27), 5(0.8), 8(0.4), 9(0.53), 10(1.0), 13(0.6), 26(0.13), 27(0.2), 38(0.07), 43(0.07)
4	0.53	1(0.54), 3(0.08), 5(0.68), 8(0.04), 9(0.19), 13(0.08), 43(0.11)
G_3 Profiles j	Mix. Prop. $\boldsymbol{\pi}_j$	Hub ID h (Hub Visitation Probability $\rho_{j,h}$)
1	0.06	1(1.0), 2(1.0), 5(0.75), 6(0.5), 7(0.75), 8(0.75), 10(0.5), 13(0.25), 16(1.0), 17(0.25), 29(0.25)
2	0.25	1(1.0), 3(0.22), 4(1.0), 5(1.0), 7(0.06), 8(0.66), 10(0.94), 11(0.06), 13(0.28), 15(0.89), 19(0.22), 26(0.11), 43(0.11)
3	0.32	1(0.09), 2(0.62), 4(0.05), 5(0.28), 7(0.75), 10(0.04), 15(0.05), 16(0.44), 17(0.04), 19(0.04), 25(0.18), 26(0.04), 29(0.09)
4	0.37	1(0.53), 3(0.12), 4(0.83), 5(0.18), 6(0.04), 8(0.04), 10(0.08), 14(0.19), 15(0.6), 43(0.11)
G_4 Profiles j	Mix. Prop. $\boldsymbol{\pi}_j$	Hub ID h (Hub Visitation Probability $\rho_{j,h}$)
1	0.03	5(0.33), 8(1.0)
2	0.17	1(1.0)
3	0.80	1(1.0), 8(0.65), 10(0.9)
G_5 Profiles j	Mix. Prop. $\boldsymbol{\pi}_j$	Hub ID h (Hub Visitation Probability $\rho_{j,h}$)
1	0.06	1(1.0), 3(1.0), 4(0.51), 5(0.14), 6(0.43), 8(0.29), 9(0.85), 10(0.58), 11(0.14), 13(0.58), 25(0.43)
2	0.20	1(0.14), 2(0.12), 3(0.22), 4(0.04), 7(0.54), 10(0.04), 15(0.26), 16(0.04), 30(0.04)
3	0.74	1(1.0), 3(0.36), 4(0.12), 5(0.01), 8(0.09), 9(0.06), 10(0.02), 15(0.23), 17(0.01), 21(0.01)
G_6 Profiles j	Mix. Prop. $\boldsymbol{\pi}_j$	Hub ID h (Hub Visitation Probability $\rho_{j,h}$)
1	0.08	1(1.0), 3(1.0), 4(0.3), 5(0.4), 8(0.2), 9(1.0), 10(1.0), 11(0.9), 13(1.0), 18(0.1), 27(0.4)
2	0.21	1(0.03), 2(0.92), 12(0.19), 17(0.15)
3	0.26	1(0.87), 2(0.22), 4(0.7), 5(0.13), 8(0.31), 10(0.03), 11(1.0), 18(0.11)
4	0.45	1(0.56), 2(0.03), 3(0.045), 4(0.17), 5(0.35), 8(0.02), 10(0.05), 11(0.11), 18(0.36)

for predicting day $n+2$ and so on. In the second version (***SPE-W7***), the past history of a window of size $n = 7$ days (e.g., days 1 till n) is considered to predict the hub list for not only day $n+1$ but for the entire next window of n days (e.g., days $n+1$ till $2*n$). After this the learning window shifts over the days $n+1$ till $2*n$ to predict hub lists for days $2*n+1$ till $3*n$, and so on till the end of the activation period for the sample user is reached.

4.2.2 Profile based prediction

This approach assumes knowledge of mobility profiles. One initially collects a user's mobility traces for a window of size $n = 7$ days (e.g., days 1 till n), and then applies the clustering algorithm described in Section 3.2 to find out a mixture of mobility profiles and their associated probabilities. Based on this profile information, one predicts the hub list for the entire next window of n days (i.e., day $n+1$ till $2*n$, similar to ***SPE-W7***. This process is repeated by shifting the learning window n days ahead. To be more precise, for each day, one first randomly chooses one mobility profile out of the mixture of profiles based on their mixing proportions and then predicts the day's hub list based on that chosen profile. For each day within the window, one compares the hub visit predictions with the observed hub visit values to compute the daily Profile based Prediction Error (PPE) rate similar to that shown for SPE in (5). Since an empirical value of 7 is chosen for the window size, this error rate is referred to as ***PPE-W7*** from now on.

In our experiment, we empirically choose a window size of $n = 7$ and compute the percentage values for SPE-ALL, SPE-W7, and PPE-W7. To quantify the improvement in location prediction achieved by our profile based method over that by the statistical methods, we define

$$Prediction\ Improvement\ Ratio(PIR) = \frac{SPE - PPE}{SPE}$$

where PIR-ALL indicates improvement of PPE-W7 over SPE-ALL, and PIR-W7 indicates the improvement of PPE-W7 over SPE-W7 and present its distribution parameters in Table 4. As seen, the mean values (considering the standard errors) are all positive, indicating a much better overall performance of our profile based hub-level location predictions as compared to the statistical approaches with similar (when compared to SPE-W7) or better (when compared to SPE-ALL) cost of location updates. This is one of the most critical contributions of our concept of profiling mobility based on sociological orbits.

Table 4: The Distribution of PIR (%)

Group	Mean ± Standard Error	
	PIR-ALL	PIR-W7
G_1	20.6 ± 2.3	24.3 ± 3.2
G_2	18.9 ± 2.0	21.4 ± 2.1
G_3	12.9 ± 2.0	14.5 ± 1.9
G_4	27.2 ± 3.0	27.9 ± 3.8
G_5	21.5 ± 1.4	24.6 ± 1.6
G_6	21.2 ± 1.5	22.6 ± 1.6

4.3 Conditional Prediction

In this section, we show how the current mobility profile information may improve the performance of certain hub-level predictions. The authors in [4, 5] have shown that a common statistical approach (similar to the one described in Section 4.2) is capable of keeping track of a user's visits to different locations (via the system logs on APs). Consequently, it is possible to provide a probabilistic view of finding the user in any location at any time based on the past history of that user's hub visits. *Let us assume that this mobility behavior for all our sample users repeats itself the next year, such that their future visits in the next activation period may be validated by the data present.* Taking the user from group G_2 as an example, we find that he/she visited hub 43 on 11 days in a 49 day activation period within the year the data was collected, as seen in Figure 3(b). If we were to consider that this mobility pattern over 49 days is going to repeat itself the next year following our assumption above, then the general probability of finding that user in hub 43 on any day during his next activity period would be $\frac{11}{49} = 0.22$. From within our profile based framework, we not only are capable of providing similar general information but, given the current mobility profile, also can be much more specific. For instance, given the same example and assumption as above, the general statistical probability $P(h)$ of finding the user in hub h on any given day may be calculated equivalently through our approach as

$$P(h) = \sum_{j=1}^{k} \boldsymbol{\pi}_j * \rho_{j,h} \tag{6}$$

Using (6) and the data in Table 3, the general probability of finding the user from group G_2 in say "target hub" $H_t = 43$ on any given day of his next period of activity would be given as: $(0.02) * (0.00) + (0.14) * (1.0) + (0.31) * (0.07) + (0.53) * (0.11) = 0.22$ (which is the same as that noted before). However, on a specific day $D = 16$ for example this general probability may be improved with additional profile based information as follows. On this day 16, as soon as the user ventures into say "identifier hub" $H_i = 4$ (see Figure 3(b)), our method shall identify the current profile (P_{now}) to be 2, as it is the only one with hub 4 in it. With this additional knowledge, our approach would then be able to re-compute the probability of finding the user in hub 43 on day 16 to be $\rho_{2,43} = 1$. From Figure 3(b), we see that under our assumption of repeated period of activation, the user would indeed visit hub 43 on day 16 of his next activation period (i.e. $\mathbf{y}_{43}^{(16)} = 1$), which makes our profile based prediction more precise. Several similar cases for each user type are listed in Table 5, where we find that the conditional probability ρ_{j,H_t} (obtained based on mobility profiles) is closer to the actual event $\mathbf{y}_{H_t}^{(D)}$ than the general probability $P(H_t)$ (obtained from the common statistical approach). In particular, as seen in the cases for the users from groups G_2, G_3, G_4 and G_6 our predictions would be completely accurate, whereas those from the statistical method are far from correct.

Essentially, the mobility profiles help us group the hubs in separate (but, potentially overlapping) sets of hubs on the basis of visits occurring to them within the same period of time (i.e., following some mobility pattern), unlike in the statistical method where all the hubs are treated independently and identically. Note that in practice it may not always be possible to uniquely identify the current mobility profile based on the hubs visited so far (i.e., *identifier hubs*), as one hub could belong to 2 (out of say 4) profiles. However, as shown earlier in Section 4.2, as long as the *iden-*

Table 5: Conditional Prediction Comparison

Group	H_t	D	$P(H_t)$	H_i	P_{now}	ρ_{j,H_t}	$\mathbf{y}_{H_t}^{(D)}$
G_1	11	13	**0.08**	8	2	**0.27**	**1**
G_2	43	16	**0.22**	4	2	**1**	**1**
G_3	7	7	**0.3**	14	4	**0**	**0**
G_4	1	15	**0.97**	5	1	**0**	**0**
G_5	7	53	**0.11**	2	2	**0.54**	**1**
G_6	3	63	**0.1**	9	1	**1**	**1**

tifier hub is able to suggest a proper subset (or, a mixture) of the user's mobility profiles for a given period, we are able to predict hub visits more precisely than common statistical methods based on hub visitation frequencies alone.

5. CONCLUSION

Knowing users' mobility patterns is crucial to the efficient design and operation of many wireless networks and applications that need to be scalable and QoS-capable. In this paper, we have analyzed the year-long mobility trace data of 13,620 WLAN users collected on the campus of ETH Zurich with 391 Access Points (APs). We not only validate the so-called sociological orbits exhibited by mobile wireless users, but also profile the user movements to help in location prediction. Unlike previous work on analyzing similar mobility trace data which focus on AP-centric parameters, our focus has been on user centric-parameters such as the number of hubs visited by a user in a day and mobility profiles of a user.

This work is the first to propose an efficient method to determine the main mobility profiles of a user using a mixture of Bernoulli's distribution as the clustering algorithm, and then make either unconditional or conditional hub-level location predictions. More specifically, our results are shown to predict around 10% to 30% more accurately than general statistical approaches that simply rely on hub visitation frequencies. This illustrates the strength of our sociological orbit aware approach, and in particular, the usefulness of the mobility profiles of a user.

Note that although this work is based only on the mobility trace data from ETH Zurich, it is expected that the data analysis, mobility profiling and location prediction techniques we have developed, as well as the conclusions we have drawn in this paper that validate the existence and usefulness of the sociological orbits are in general applicable to other university and corporate campuses, as well as other public/private environments (there certainly isn't a sufficient amount of mobility trace data available except from a couple of places). In addition, we expect that this work will inspire additional innovative work on social influence aware and user-centric designs and operations of not only wireless access networks, but also mobile ad hoc and peer-to-peer networks, as well as intermittently connected or, delay tolerant networks.

6. REFERENCES

[1] Bai, F., Sadagopan, N., and Helmy, A. Important: a framework to systematically analyze the impact of mobility on performance of routing protocols for adhoc networks. *Proceedings of IEEE INFOCOM '03 2* (March 2003), 825–835.

[2] Burns, B., Brock, O., and Levine, B. N. Mv routing and capacity building in disruption tolerant networks. *In Proceedings of IEEE INFOCOM '05 1* (March 2005), 398–408.

[3] Camp, T., Boleng, J., and Davies, V. A Survey of Mobility Models for Ad Hoc Network Research. *Wireless Communications and Mobile Computing (WCMC): Special issue on Mobile Ad Hoc Networking: Research, Trends and Applications 2*, 5 (2002), 483–502.

[4] Chen, G., Huang, H., and Kim, M. Mining frequent and periodic association patterns. *Dartmouth College, Coomputer Science and Engineering, Tech Report: TR 2005-550* (July 2005).

[5] Chinchilla, F., Lindsey, M., and Papadopouli, M. Analysis of wireless information locality and association patterns in a campus. *In Proceedings of IEEE INFOCOM '04 2* (March 2004), 906–917.

[6] Dempster, A., Laird, N., and Rubin, D. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society, Series B 39*, 1 (1977), 1–38.

[7] Ghosh, J., Beal, M. J., Ngo, H. Q., and Qiao, C. On profiling mobility and predicting locations of campus-wide wireless network users. *CSE Dept. TR-2005-27, State University of New York at Buffalo* (December 2005).

[8] Ghosh, J., Philip, S. J., and Qiao, C. Sociological orbit aware location approximation and routing in manet. *In Proceedings of IEEE Broadnets '05, Boston, MA* (October 2005), 688–697. *Also presented as a Poster in ACM MobiHoc '05, Champaign, IL* (May 2005).

[9] Ghosh, J., Ngo, H. Q., and Qiao, C. Mobility Profile based Routing within Intermittently Connected Mobile Ad hoc Networks (ICMAN) *Accepted for publication in IWCMC 2006 Delay Tolerant Mobile Networks workshop, Vancouver, Canada* (July 2006).

[10] Grossglauser, M., and Tse, D. N. C. Mobility increases the capacity of ad hoc wireless networks. *IEEE/ACM Transactions on Networking 10*, 4 (August 2002), 477–486.

[11] Kim, M., Kotz, D., and Kim, S. Extracting a mobility model from real user traces. *To appear in IEEE INFOCOM'06, Barcelona, Spain* (April 2006).

[12] Lai, K., Roussopoulos, M., Tang, D., Zhao, X., and Baker, M. Experiences with a mobile testbed. *In Proceedings of The Second International Conference on Worldwide Computing and its Applications (WWCA98)* (March 1998).

[13] Samal, S. Mobility pattern aware routing in mobile ad hoc networks. *MS Thesis, Virginia Polytechnic Institute and State University* (May 2003).

[14] Schwarz, G. Estimating the dimension of a model. *The Annals of Statistics 6* (1978), 461–464.

[15] Su, W., Lee, S.-J., and Gerla, M. Mobility prediction and routing in ad hoc wireless networks. *International Journal of Network Management 11*, 1 (February 2001), 3–30.

[16] Tang, D., and Baker, M. Analysis of a metropolitan-area wireless network. *In Proceedings of ACM MOBICOM '99* (August 1999), 13–23.

[17] Tang, D., and Baker, M. Analysis of a local-area wireless network. *In Proceedings of ACM MOBICOM '00* (August 2000), 1–10.

[18] Tuduce, C., and Gross, T. A mobility model based on wlan traces and its validation. *In Proceedings of IEEE INFOCOM '05 1* (March 2005), 664–674.

[19] Wang, W., and Akyildiz, I. F. On the estimation of user mobility pattern for location tracking in wireless networks. *Proceedings of IEEE Globecom '02* (November 2002), 619–623.

[20] Wang, Y., and Wu, H. DFT-MSN: The Delay/Fault-Tolerant Mobile Sensor Network for Pervasive Information Gathering. *To appear in Proceedings of IEEE INFOCOM '06, Barcelona, Spain* (April 2006).

[21] Zhao, W., Ammar, M., and Zegura, E. Controlling the mobility of multiple data transport ferries in a delay-tolerant network. *In Proceedings of IEEE INFOCOM '05* (March 2005).