Mining Behavioral Groups based on Usage Data in Large Wireless LANs

Wei-jen Hsu[1], Debojyoti Dutta[2], and Ahmed Helmy[1]

[1]Department of Computer and Information Science and Engineering, University of Florida

[2]Cisco Systems, Inc.

Email: [1]{wjhsu, helmy}@ufl.edu, [2]dedutta@cisco.com

Abstract

Wireless networks and personalized mobile devices are deeply integrated and embedded in our lives. Such wide adoptions of new technologies will impact user behavior and in turn will affect network performance. It is imperative to characterize the fundamental structure of wireless user behavior in order to model, manage, leverage and design efficient mobile networks.

One major challenge in characterizing user behavior stems from the significant size and complexity of user behavioral data. Without summarization and dimension reduction, the sheer amount of data does not provide much useful information. The key contribution of the paper is a novel similarity metric based on a matrix representation of mobility preferences and its decomposition. This method provides an efficient way to reduce important spatio-temporal dynamics in user mobility into a few *eigen-behavior* vectors. This also facilitates nodes to exchange their mobility summaries and determine their mutual similarity locally. Without any assumption on the properties of user population, we use unsupervised learning (clustering) techniques to classify WLAN users. Such a user grouping scheme based on learned user behavior is crucial for applications relying on the usage context of each mobile device (e.g., participatory sensing, social-relationship-aware message forwarding).

In this study, using our systematic *TRACE* approach, we analyze wireless users' behavioral patterns by extensively mining wireless network logs from two major university campuses to showcase its efficacy. While our findings partly validate intuitive repetitive behavioral trends and user grouping, it is surprising to find the qualitative commonalities and striking consistency of user behavior from the two universities. We discover multi-modal user behavior for more than 60% of the users, and there are hundreds of distinct groups with unique behavioral patterns in both campuses. The sizes of the major groups follow a power-law distribution.

I. Introduction

In recent years, we have witnessed the mass deployment of portable computing and communication devices (e.g., laptops, cellphones, PDAs) and wireless communication infrastructures. To estimate the impact of the large-scale deployment of wireless communication facilities, there is a pressing need to capture and understand the user behavioral patterns as these new technologies are adopted. This understanding will also play a crucial role in the direction of linking user behavior awareness with otherwise behavior oblivious network protocols (e.g., message delivery, participatory sensing, etc.).

Take the usage pattern of mobile devices as an example: Major work places (e.g., offices and classrooms) and the informational hubs (e.g., libraries and computer centers) would dominate users' network usage in terms of locations at which they utilize the networks. However, with ubiquitous WLAN deployments, the location from where people access information is bound to change. While the traditional "hot spots" still play an important role, we can expect that users access the network at a much more diverse set of locations to reflect their personal preferences (e.g., a small group may prefer to work at a coffee shop). As the portability of today's devices enables its users to untether themselves and follow their daily routines without being disconnected, we witness a wealth of different usage patterns from different users, depending on their affiliations and preferences. This creates a heterogeneous, complex, and dynamically evolving context in which the mobile devices can be used, and such a changing usage context is a crucial subject to understand and capture realistically to enable meaningful analysis of current mobile user behavior and design of future networks. Technique to discover usage patterns from collected data provide the focus of this paper.

While mobility is one important characteristic of wireless network devices and several previous studies provide techniques to model mobility patterns, they typically capture user mobility on a per-node basis [19], [17], [18]. Such models do not capture inter-node relationships, similarities or differences (i.e., emergent group-behavior). In large-scale deployments, such as WLANs on university campuses, it is likely to have many users displaying similar behavioral trends, due to the underlying similarities in their social affiliations (e.g., computer science graduate students) or hidden personal preferences (e.g., people who like to study in libraries). However, knowledge of such similarities among users is not easily accessible, as the trace collection process typically anonymize the data sets to ensure protection of privacy and all identity-related information of users is lost. To better understand the context of how the mobile devices are used today, it is very important to go beyond the usage patterns of seemingly unrelated individuals and mine these implicit *behavioral groups* from the anonymized traces. Such a user classification based on learned behavioral patterns is a major building block for applications that leverages the usage context of each mobile device, such as participatory sensing [29], [30], [31] or social-relationship-aware message forwarding [27], [21].

Towards the problem of understanding user behavior and classifying them as groups, the main contributions of this paper are two-folded. We first propose a generic framework for trace-based user behavior analysis, and then take on understanding the usage pattern of campus WLAN users as a case study. We choose to mainly focus on *long-run mobility preferences* as a key feature, and take a first step towards understanding and characterizing the structure of behavioral patterns of users. In the case study, we seek to answer (a) how many different behavioral modes do a typical user display? Does she repeat the same location preference everyday or could there be variances? (b) How to systematically analyze a trace with large population, and quantify users as having similar or dissimilar mobility preferences? And, finally, (c) how many groups can one identify as having distinct mobility preference from a campus environment?

1) We propose systematic procedural steps to construct proper representations of user behavior for our data sets and quantify novel similarity metrics between users, in what we refer to as the TRACE framework. The similarity metric we propose is based on the *eigen-behavior* extracted by singular value decomposition from the *association matrix* of users. This is an import step to leverage the inherent redundancy (i.e., repetition of the dominant behavior patterns) in time-variant user behavior and summarize the high-dimensional *association matrix* with few components, suppressing the noises. One major strength of the *eigen-behavior* representation is that it can be efficiently exchanged between users to calculate their similarity locally (i.e., each node could summarize its behavior and determine its similarity to another node without involving other nodes). While computational efficiency is important for efficient user clustering (in a centralized data mining application), the ability to calculate similarity locally is even more crucial for leveraging the similarity metric for decentralized, behavior-aware protocols. Lastly, while we use mobility as an example in this paper, our proposed framework is applicable to mine similarity of other features as well.
2) By applying the TRACE framework to user mobility on university campuses as a case study, surprisingly, we find qualitative commonalities in user behavior almost across the board, albeit many differences (e.g., geographical locations, sizes and structures, different student bodies, etc.) exist among the campuses. This suggests the possibility of setting up a generic model for user behavior, and adjust its details to match with various campuses. Our key findings include the following: (a) More than 60% of the WLAN users display multi-modal behavior (different mobility preferences on different days), while the mobility preferences are very stable with low dimensionality. Using Singular Value Decomposition (SVD), one can capture more than 90% of the power in the user association preference with just five components. Although it is intuitively acceptable that human beings display regular patterns in location preference in the long run, this has not been systematically measured and quantified in large scale as we do in this paper. (b) We leverage SVD to discover major trends in each user's mobility preference and design a similarity metric, which can be effectively used in unsupervised learning (i.e., clustering) [1] to classify the user population into hundreds of distinct behavioral groups. (c) The sizes of the major groups, however, are highly skewed and follow a power-law distribution.

As mobile communication devices become an integral part of our regular lives, they can be leveraged as a major venue for information distribution or collection. Incorporating user behavior in message forwarding decisions is an important direction in social behavioral-aware communication, e.g., efficient message forwarding based on roles in social networks [27], [21], [22]. We present a case study to display the efficiency improvement by limiting message delivery to relevant users with the aid of the *eigen-behavior* and similarity metric defined in this work. Similarly, user behavioral profiles developed in this work can be directly used with the emerging paradigm of participatory sensing and crowd sourcing [29]. In these classes of applications the sensing task and quality would depend heavily on the behavior of the individual participants. Having an efficient, succinct representation of the behavioral profiles using the association matrix, we can meaningfully recruit participants with specific mix of profiles to meet a sensing task. This paper provides a first step towards a framework for large-scale trace analysis to unravel user similarity efficiently, a fundamental task for many of the above applications. The techniques of summarization, classification, and comparison of user behavior proposed in the paper are essential in many group-aware tasks, such as similarity-based support groups for e-health and collaborative learning in education [40].

The rest of the paper is organized as follows. Section II discusses the related work. The *TRACE* framework and other important preliminary facts are presented in section III. We then motivate the need of a good behavior distance metric in section IV, and develop our technique for summarizing user behavior and clustering users in section V and VI, respectively. Finally, we discuss the major findings from clustering campus WLAN traces in section VII and a case study of social-aware message forwarding and the alternatives of our method in section VIII. The paper concludes in section IX.

II. Related Work

As wireless networks gain popularity, it is extremely important to understand its characteristics realistically. Along these lines, there have been great efforts to collect traces from WLAN users [9], [10], [11], [16]. Many more traces have been made available through efforts to build libraries of measurement traces [12], [13]. To protect user privacy, any identity-related information is removed from these data sets. Therefore, is becomes a challenge to understand the relationship among users based solely on usage data, without any identity or social-context related information.

WLAN usage patterns have attracted significant attention recently. Although user mobility has been one major focus in studies about WLANs, for most previous works the focus is either on aggregated statistics or on models for individual user. For the aggregate statistics, the current operating status of WLANs is studied extensively, including user association preferences and durations, mobility and hand-off, among others. There has been studies on comparison of the same campus during different time periods [9], mobility of corporate WLAN users [10], and a study specific to PDA users and their mobility [11], to name a few. These traces are compared in [16] based on aggregate statistics. However, work in this category does not attempt to differentiate between users. They do not reveal the richness of user behavioral characteristics that may vary widely among users.

For most of the mobility modeling works, the focus is to obtain particular statistics about users and to establish a model based on these quantities. In most of these modeling efforts, the users are considered as independent samples from a uniform population. The user association durations are modeled by BiPareto distributions in [20]. Tuduce et al [19] match user session lengths and hand-off probabilities between APs to generate a mobility model. In [17], [18], the authors further cluster the locations (i.e., AP) based on the number of user hand-off between them to generate a hierarchy in user hand-off model. In the TVC model, periodicity of users visiting their favorite locations has been incorporated explicitly [24]. While these mobility models provide important insights of capturing and replicating key mobility features, their goal is again not set on distinguishing users based on these features or classifying users into groups based on their behavioral similarity, which is the focus of this study.

Between aggregated mobility statistics and mobility modeling for individual users, we identify a compelling need to discover groups of users from the whole population, *based on their behavioral patterns* (in our case we use mobility preference as one example). There are hardly any studies on quantifying the level of similarity between users based on the collected WLAN traces to establish their mutual relationships in the literature. We provide a first step towards this understanding by classifying users into groups of similar behavior. This provides a different and important perspective to understand user's usage patterns. A preliminary short version of this work appeared as an extended abstract in MobiCom 2007 [26]. We include much more depth in this detailed version, in terms of the understanding of multi-modal association pattern of individual users, the comparison between summarization techniques, and the discussions of user behavior in the discovered clusters.

There are several papers in the literature that also use clustering techniques. One with a similar goal to ours is done by Kim et al. [5], which classifies users based on a different *representation*. The authors look into the range of user movement, and classify users based on the periodicity of the movement range. With this representation, users are classified based on the dominant periods in their movement (e.g. those who display strong daily or weekly movement patterns) and their longest movement ranges, but not based the location preferences. Hence the results have different interpretations to ours (e.g., users with similar movement ranges and dominant periods, regardless of the actual locations visited, would be classified the same in [5]. We, on the other hand, focus on classifying users based on the main locations and the frequency of their visits). Ghosh et al. classify WLAN users based on the locations each user visit in the concept known as sociological orbits [36]. Our association matrix representation enhances this concept as we further consider the amount of time a user spends at each location and extract the dominant locations in terms of visit duration. Barabasi et al. study user mobility based on cell phone user data [25]. Their focus is mainly on understanding the mobileness (e.g., movement distance and radius) of users, and classify users into groups with different levels of mobility. In [7] the authors apply clustering technique to the trace of location coordinates of a user to discover significant places for the user, but they have not focused on classifying users.

The technique we utilize to obtain association features from users, singular value decomposition [3], is widely-applied to discover linear trends in large data sets. It is closely related to principal component analysis (PCA) [2]. In [8], the authors utilized PCA to decompose the traffic flow matrices for ISP networks and understand the major trends in the traffic. Our application of SVD to individual user association matrices is similar in spirit to their work. Note that it is typical for people to follow dominant routines in lives, hence we expect the SVD approach to be applicable to various human behavioral data sets. In [4], the authors also use PCA to discover trends in a cellphone user group. In this paper, in addition to analyzing much larger data sets, we further compare user similarity and define distance metrics to classify wireless network users into groups with robust validation. Note that in order to make the eigen-behavior vectors obtained from all users comparable, we need to keep the origin fixed among all association matrices. Hence we adopt a variant, called *uncentered PCA* [2] where the mean of each dimension is not subtracted. It has been used to study the diversity of species at various sites [15]. This is important as the behavioral features can be calculated by each node in a distributed fashion, and then exchanged or compared efficiently – a crucial property for the behavior summary to be used also in distributed network protocols, not only in centralized data analysis.

While the goal of this paper is mainly classifying mobile devices based on mobility trends displayed in usage patterns instead of social relationships, a recent work by Crandall et al. indicates that repeated geographical coincidences suggest a social tie between people [37]. As such, similarity in mobility identified by our methods could potentially be leveraged as an indicator of social relationship.

III. Preliminaries

In this section, we first introduce the *TRACE* framework, a systematic approach to analyze large scale traces we follow in this paper. We then introduce the traces we analyze in the paper and the *normalized association vector* representation we choose. We also briefly introduce the necessary background knowledge about clustering.

A. *TRACE Framework*

In this study, we develop systematic steps to handle the task of analyzing large-scale traces. The framework is generic and widely applicable, not necessarily tied to the analysis of user mobility preference in the paper. We use Fig. 1 to illustrate the conceptual flow of our approach, which we refer to as the *TRACE* framework. The five major components in the framework are: *T*race, *R*epresentation, *A*nalysis, *C*lustering, and *E*mployment (or application). The descriptions of the raw *traces* and the *representation* we choose to base our work on are introduced later in this

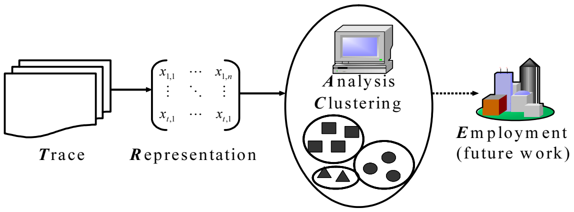


Fig. 1. Illustration of the *TRACE* approach.

section. The choice of *representation* is a crucial step in the framework, as different representations reveal different aspects of user behavior, and one has to pick a representation suitable for the task at hand. We will briefly show the drawback of an improper choice of distance metric in the following section.

We then conduct *analysis* and *clustering* upon the users based on the chosen representation, *normalized association vectors*. To devise a proper distance metric of user behavior, we conduct further *analysis* to understand the nature of user mobility preferences, and evaluate and contrast various summaries to capture its major trend in section V. We then utilize a SVD-based mobility feature summary to achieve meaningful user *clustering* in section VI and discuss its interpretation in section VII. On top of achieving good clustering results, the *eigen-behavior* based distance metric we use can also be leveraged for decentralized user-matching or *behavior-aware* message dissemination protocols, as we will discuss in section VIII-A.

B. Wireless Network Measurement Data and the Association Matrix Representation

The widespread deployments of large-scale wireless LANs on university campuses have attracted high adoption from its community. These deployments have outgrown experimental networks and become commodities. Due to its high penetration and diversity in users (as compared to corporate WLANs), campus networks are good platforms to study the behavioral pattern of WLAN users. To our benefit, great efforts have already been made to collect the user traces from several large WLAN deployments [12], [13]. We elect two extensive WLAN traces collected from large populations for long durations for the study. The details for the selected traces are listed in Table I. Details of pre-processing of these two traces can be found in the first papers using the traces, in [16], [9].

While university WLAN traces are suitable for the study of user behavior, there are also shortcomings in these traces. The most important ones are (1) Users are not always online and many of them access the network sporadically. (2) Most WLAN users access the network with laptops, which are not always easily portable and limit the mobility of users while accessing the network. However, these WLAN traces are by far the most extensive publicly available traces and we can indeed discover interesting patterns, as we shall show, if a proper *representation* is chosen. Due to the nature of sporadic, non-continuous usage, we choose not to focus on the movement path (i.e., the sequences of roaming from one location to another) in user mobility as the devices are mostly shut down when moving anyway. Hence, the mobility we discover from the trace is really more about the *usage pattern* of

TABLE I

FACTS ABOUT STUDIED TRACES

Trace source	USC [12]	Dartmouth [14]
Time/duration of trace	2006 spring semester (94 days)	2004 spring quarter (61 days)
Start/End time (Time frame)	01/25/06 - 04/28/06 Spring 2006 semester	04/05/04 - 06/04/04 Spring 2004 quarter
Location granularity	Building	Access point
Unique locations	137 buildings	545 APs/ 162 buildings
Unique MACs analyzed	5,000	6,582

the devices (i.e., where are the devices being used?) rather than the true *movement path* of their owners (i.e., how do people move?). This *usage pattern* provides a good representation of the location(s) at which the user mostly use the devices, which would help in finding users with desired mobility preference in network protocol designs. However, note that our methods are not limited to the specific data sets we choose, and it would be of great interest to study traces from other mobile devices (e.g., cellphones, iPods), if available for a large population.

To understand user behavior from wireless network traces, the first fundamental task is to choose a representation of the raw data. We choose the patterns of users visiting WLAN access points (APs) for the analysis. Visiting pattern is important to WLANs as mobility is one of its defining characteristics. When a WLAN user moves across campus, the set of APs with which the user associates is considered an indicator of the user's physical location. From a social context, the places a person visits regularly and repeatedly usually have a stronger connection to one's interest, identity and affiliation. It is perhaps one of the important distinguishing factors for people with different social attributes [7].

We represent a user's location visiting pattern by what we refer to as *normalized association vectors*[1]. The association vector is a summary of a user's AP association during a given time slot T. The *association vector* for each time slot is an n-entry vector, $(x_1, x_2, ..., x_n)$, where n is the number of unique locations (i.e., buildings[2]) in the given trace. Each entry in the vector, x_i, represents the *fraction* of online time the user spends at the location during the time slot, i.e., we normalize the user association time with respect to his online time (the total time the user appears in the trace during the time slot). With this representation, the conclusions we draw are not influenced by the absolute value of online time, which varies across a wide range among different users and different time

[1] For brevity, we sometimes use the shortened term *association vector* for *normalized association vector* unless stated otherwise.

[2] We aggregate APs in the same building as a single location for better interpretation of user behavior. Note that we have also applied the same technique at AP level and similar results hold. This location aggregation is only done for better understanding and interpretation of the results. We choose to aggregate APs in the same building into one location to mitigate the harm caused by so-called "ping-pong effect" (nodes changing association back and forth among close-by APs due to signal perturbation).

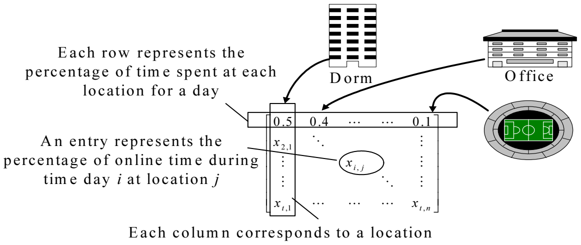

Each row represents the percentage of time spent at each location for a day
Dorm
Office
0.5
0.4
...
...
0.1
An entry represents the percentage of online time during time day i at location j
$x_{2,1}$
$x_{i,j}$
$x_{t,1}$
$x_{t,n}$
Each column corresponds to a location

Fig. 2. Illustration of the association matrix representation. This can be maintained by each user/device locally, without the need for a centralized server or third party.

slots of a given user. Note that the sum of the entries in the association vector, $\sum_{i=1}^{n} x_i$, is always 1 if the user has been online during the time slot. We use a zero vector to represent the association vector when the user is completely offline for the time slot. To represent a user's association preference for the long run, we construct the *association matrix* X for the user, as illustrated in Fig. 2, i.e., we concatenate the association vectors for each time slot. If there are n distinct locations and the trace period consists t time slots, the *association matrix* for a user is a t-by-n matrix.

Note that there are potentially many ways to represent user behavior from a rich data set. We choose to use a day as the time slot (i.e., T = one day)[3]. Different representations certainly provide different insights. Due to space limitations, we focus on the *normalized* representation for *daily association vectors* to illustrate our analysis, and briefly discuss about other alternatives in section VIII-B. In the remainder of the paper, the term *user behavior* is used to explicitly refer to a user's *normalized association vectors*.[4] While this representation captures only a partial picture of generic user behavior, we will show that a properly defined summarization and distance metric leads to robust yet meaningful classification of users based on this representation.

C. Preliminaries of Clustering Techniques

Clustering (one of the key methods in unsupervised learning) is a widely-applied technique to discover patterns from data sets with unknown characteristics. It can be roughly classified into hierarchical or partitional schemes [1]. In this paper we use the hierarchical clustering, in which each element is initially considered as a cluster containing one member. Then, at each step, based on the distances between the clusters[5], two clusters that are the closest to each other among all cluster pairs are merged into one cluster with larger membership. This process continues until a *clustering threshold* has been reached, when all the inter-cluster distances for the remaining clusters are larger than a given distance threshold, or the remaining cluster number reaches a given target.

[3] We choose one day as the time slot of analysis as daily behavior cycle is a naturally meaningful time frame for almost all human beings. If one wants to understand user behavior based on the time-of-day factor (i.e., users can behave differently during the day versus during the night), our framework is extensible by constructing multiple association matrices for one user, where each matrix is used for a particular part of a day.

[4] Furthermore, the term *behavioral group* is used to refer to users with similar trends in their association vectors.

[5] Among several alternatives, we use the average distance of all element pairs between the clusters. Use of other methods does not change the results significantly.

One major issue in applying clustering to a data set with unknown characteristics is that it is hard to pre-select a proper clustering threshold in advance. The indication of a good clustering result is that the distances between elements in the same cluster are low, and the distances between elements in different clusters are high. (i.e., there is a clear separation between inter-cluster and intra-cluster distance distributions.) Usually the clustering threshold comes from the domain knowledge or trial-and-error. Often the decisive factor for the quality of the clustering results is the selection of the *distance metric*, which is one main contribution in this paper.

IV. Challenges

As mentioned previously, the most important step in clustering is to define the *similarity* or *distance metric* between users[6]. We highlight the challenges in selecting a proper distance metric with an example in this section.

An intuitive distance metric between two individuals is to consider all the association vector pairs. Formally, we define the *average minimum vector distance (AMVD)* between users A and B, $AMVD(A, B)$, as

$$AMVD(A, B) = \frac{1}{|A|} \sum_{\forall a_i \in A} \arg \min_{\forall b_j \in B} d(a_i, b_j), \tag{1}$$

where a_i and b_j denote an association vector of user A and B, respectively. $|A|$ denotes the cardinality of set A. $d(a_i, b_j)$ denotes the Manhattan distance[7]. $AMVD(A, B)$ is the average of distances from each of the vectors in set A to the closest vector (or the nearest neighbor) in set B[8].

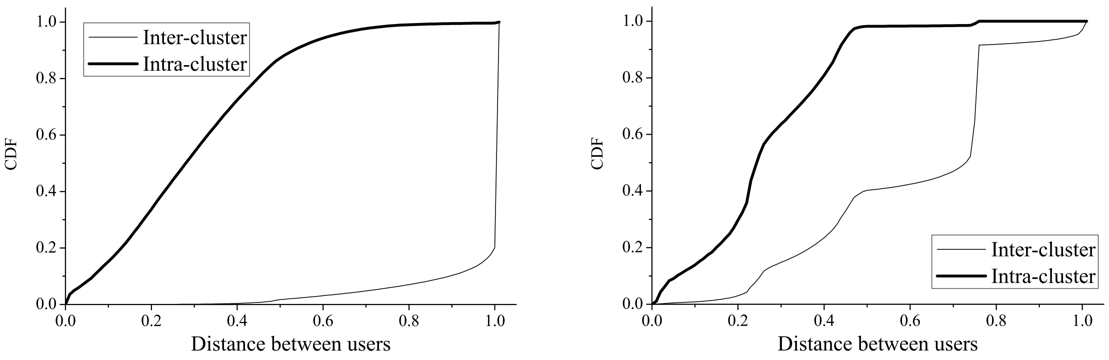
We apply the hierarchical clustering algorithm discussed in section III-C to users with the distance metric derived from AMVD. As mentioned earlier, a clustering algorithm requires properly chosen thresholds, and the particular choice is data-dependent. We experiment with various thresholds, and discover that for the USC trace, we can group the users into 200 clusters with a clear separation between inter and intra cluster distance distributions (Fig. 3 (a)), which is a qualitative indicator for a valid clustering result. However, the distance metric works poorly for the Dartmouth trace, as shown in Fig. 3 (b). The separation between inter and intra cluster distance distributions is not clear, *regardless* of cluster thresholds.

One problem with the $AMVD$ metric is that it considers all association vectors with equal weights. A meaningful distance metric should capture the major trends of user behavior and be robust to noise and outliers. Another problem is its computation complexity. We have to calculate the distances between all t^2 pairs of association vectors for each user pair. For N users the computation requirement is of order $O(N^2t^2)$. Thus we would like to design a distance metric that is both (1) robust to noise and (2) computation and storage efficient. In order to achieve both

[6] $d(x, y)$ is a distance function if $d(x, x) = 0$ and $d(x, y)$ is small if x and y are similar and large otherwise. Similarity can be considered to be the opposite of distance, i.e., $sim(x, y) = 0$ means x,y are dissimilar.

[7] We use Manhattan distance, or the $L1$ norm, since it is robust to statistical noise. Note that by our representation, $0 \leq d(a, b) \leq 2$ for normalized association vectors a and b.

[8] Note that, with this definition, $AMVD(A, B)$ is not necessarily equal to $AMVD(B, A)$. We define a symmetric distance metric between users A and B as $D(A, B) = (AMVD(A, B) + AMVD(B, A))/2$.



(a) USC.

(b) Dartmouth

Fig. 3. Cumulative distribution function of distances for inter-cluster and intra-cluster user pairs (AMVD distance).

goals, we start by studying the characteristics of the association patterns of a single user to validate the repetitive patterns or modes of behavior. We show that this study leads us to the appropriate distance metric.

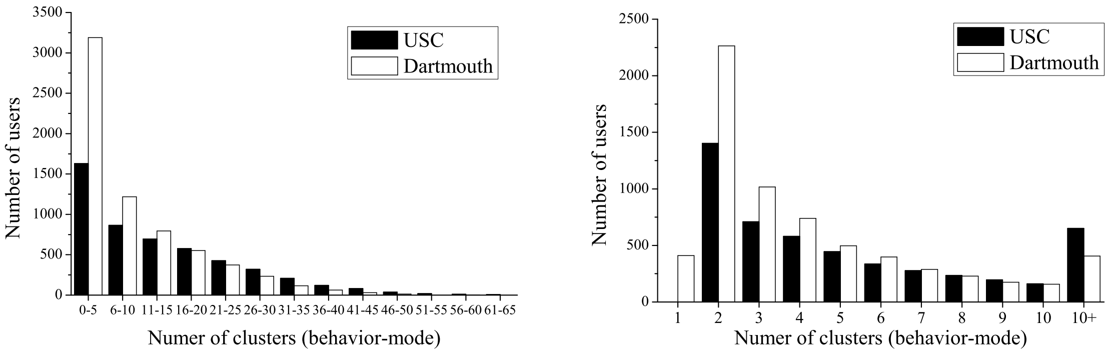
V. Summarizing the Association Patterns

In this section, we first analyze the characteristics in mobility preferences displayed by individual users. The goal of such understanding is to evaluate proper summarization techniques, and construct a compact representation of the association matrix, which is suitable for distance computations used in clustering.

A. *Characteristics of Association Patterns*

We first understand the repetitive trend in a single user's associations pattern, and how dominant the trend is (i.e., among all *association vectors* obtained each day, are there dominant *behavioral modes*?). We obtain this upon clustering the association vectors of a single individual. In this step, we consider the association vectors, X_i for $i = 1, ..., t$ (i.e., row vectors of an association matrix X) of a single user and apply clustering to these vectors. The identified clusters represent distinct *behavioral modes* of the user. Similar association vectors will be merged into a cluster and its size indicates its dominance – large clusters imply that the user follows consistent association patterns on many different days as its major behavioral modes.

We apply clustering to the association vectors of each user in the USC and the Dartmouth traces using various clustering thresholds. The distribution of number of clusters (each of them corresponds to a different *behavioral mode* where the user spends her online time at a distinct set of locations different from other days) obtained are shown in Fig. 4. In Fig. 4(a), we use a small clustering threshold (0.2), with which only very similar association vectors are merged. We see that for the USC and the Dartmouth traces, respectively, about 50% and 67% of users have less than 10 different clusters or behavioral modes (much fewer than total number of time slots, 94 and 61) with this low clustering threshold. This indicates that *the users have distinct repetitive trends in its association*



(a) Clustering threshold = 0.2

(b) Clustering threshold = 0.9

Fig. 4. Distribution of number of clusters (behavioral modes) for users.

vectors. On the other hand, if we consider a moderate clustering threshold (0.9), we see in Fig. 4 (b) that users still show multiple behavioral modes. On average, with 0.9 as the clustering threshold, the number of behavioral modes for USC and Dartmouth users are 5.57 and 4.32, respectively. This indicates, *different behavioral modes of a given user are not due to small variances of time spent at various locations on different days. There is indeed significant difference of visited locations among those behavioral modes.*

Most of those users with only two behavioral modes have a consistent association pattern: One mode corresponds to the association vectors when the user is offline, and the other one corresponds to the association vectors when the user is online. These users switch between online and offline behavior from day to day, and when they are online, the association vectors are consistent and fall in a single behavioral mode. We refer to these users as *single-modal* users. On the other hand, we also observe many *multi-modal users*. These users show a more complex behavior: their association vectors form more than two clusters, which indicate that they display distinct behavioral modes when they are online. 71.9% of users in USC and 59.4% of users in Dartmouth are classified as multi-modal when the clustering threshold is 0.9. Hence, we conclude that even we consider only significantly different association vectors (i.e., with Manhattan distance larger than 0.9 while the largest possible distance is 2.0), many users do display significantly different behavioral modes (e.g., connecting to very different set of locations) over the studied time periods in both traces.

To examine the degree of dominance of the most important behavior modes of users, we compare the most important behavioral mode and the second most important one in terms of their sizes. In Fig. 5 we plot the size (i.e. number of vectors) distributions of the first and the second behavioral modes under clustering threshold 0.2 (solid lines) and 0.9 (dotted lines) for USC users. We see that there is a clear separation between the sizes of these two behavioral modes. (i.e., the most dominant behavioral mode is much more important than the second most important one for most users.) Different clustering thresholds do not change the results much, since we observe for most users, their important behavioral modes are typically distinct and well-separated (i.e., have large

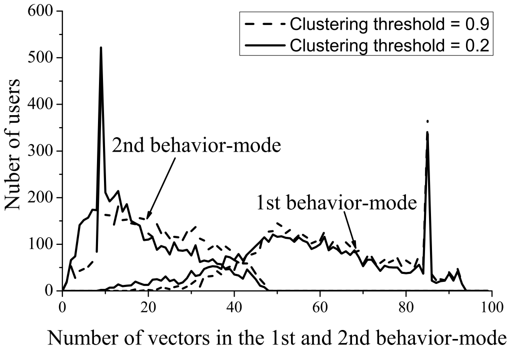


Fig. 5. Distribution of association vectors in the first and the second behavioral modes for the USC trace. Right: the first cluster, Left: the second cluster.

distances among each other). In other words, observations of the most dominant behavioral mode could reveal user characteristic to a good extent for many users. Similar observations also hold for Dartmouth users.

We note that this dominance of the first behavioral mode is related to the observed stability of user mobility profile. We use the term *stability of user mobility profile* to describe how the association vectors of a given user remain similar when the same user is considered after certain time gap. We find that the association vectors are relatively stable even with a time gap up to several weeks [34]. This stability implies that the behavioral modes of a given user do not change noticeably over time. This property is especially important if the association vectors are to be used as an intrinsic property defining the user behavior in behavior-aware protocols. In the remainder of this paper, we focus on methods to derive a succinct summary of user association vectors.

Finally, we point out that looking at the most dominant cluster exclusively could be sometimes misleading and leaving out information about the user's detailed behavior. We show the distribution of the size ratio between the largest and the second largest cluster in Fig. 6. For USC and Dartmouth, respectively, 36% and 31% of users follow their second behavior modes for more than one half of the days they follow their first behavior modes (i.e., with size ratio smaller than 2.0). It is therefore desirable to have a summary that takes not only the dominant behavioral mode, but also the subsequent ones into account.

B. Summarization Methods

Now we investigate various ways to summarize the association vectors. In order to quantitatively compare the quality of the summarization techniques, we propose to measure the *significance score* of a summary vector with respect to a user by summing the projections of all association vectors on the summary vector, normalized by the online days of the user.

$$SIG(Y) = \frac{\sum_{i=1}^{t} |X_i \cdot Y|}{\sum_{i=1}^{t} \parallel X_i \parallel_1}, \tag{2}$$

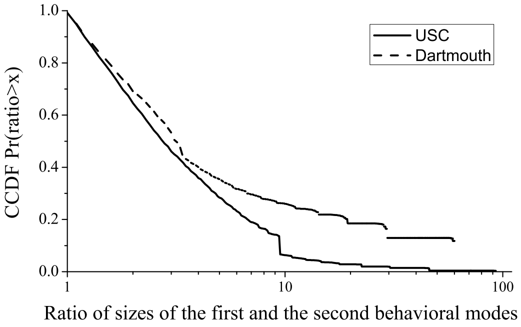


Fig. 6. Complementary CDF for the ratio of the first behavioral mode size to the second behavioral mode size. Note that the X-axis is in log scale to make the graph more visible.

TABLE II

THE AVERAGE SIGNIFICANCE SCORE FOR VARIOUS SUMMARIES OF USER ASSOCIATION VECTORS

	X_{onavg}	$X_{centroid1}$ threshold 0.5	$X_{centroid1}$ threshold 0.9	SVD
USC	0.646	0.716	0.702	0.764
Dartmouth	0.690	0.757	0.747	0.789

where $\| X_i \|_1$ is the L1 norm of vector X_i (recall that for online days, the elements in association vectors sum to 1) and Y is any summary vector. The physical interpretation of the *significance score* is the percentage of power in the association vectors X_i's explained by the summary vector Y.

Based on this definition, we pose the choice of the most representative summary of association vectors as an optimization question: Given the association vectors X_i's, what is the best possible summary vector Y to maximize its significance? Mathematically, we want the vector Y to be

$$Y = \arg \max_{\|v\|=1} \sum_{i=1}^{t} |X_i \cdot v|. \tag{3}$$

This is exactly the procedure to obtain the first singular vector if we perform singular value decomposition (SVD) [3] of the association matrix X. In other words, if we want the summary vector Y to capture the maximum possible power in the association vector X_i's, the optimal solution is to apply singular value decomposition to extract the first singular vector. We apply this technique and calculate the *significance score* in the last column in Table II, in comparison with other simple summary techniques, such as the average of all association vectors when the user is online (X_{onavg}) or the average of the centroid of the most dominant behavioral mode ($X_{centroid1}$).

Indeed, SVD provides the best summary among the compared methods. Hence we use mainly the SVD-based summary, and defer the discussion of other summary techniques to section VIII-B.

C. Interpreting Singular Value Decomposition

In this subsection we explain other important properties of SVD as applied to the association matrices.

From linear algebra [3], we know that for any t-by-n matrix X, it is possible to perform singular value decomposition, such that

$$X = U \cdot \Sigma \cdot V^T, \tag{4}$$

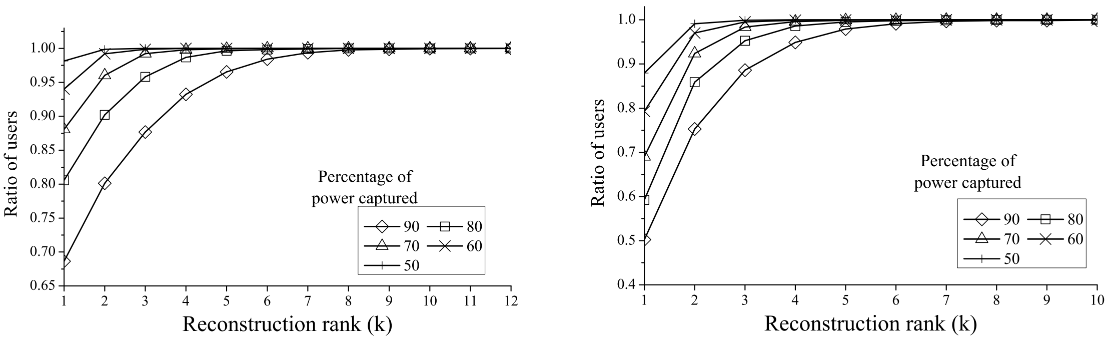
where U is a t-by-t matrix. Σ is a t-by-n matrix with r non-zero entries on its main diagonal where r is the rank of the original association matrix X. V^T is an n-by-n matrix where the superscript ^{T} in V^T indicates the transpose operation to matrix V. The column vectors of the matrix V are the eigenvectors of the covariance matrix $X^T X$, and Σ is a diagonal matrix with the corresponding singular values to these eigenvectors on its diagonal, denoted as σ_1, σ_2, ..., σ_r. These singular values are ordered by their values (i.e. $\sigma_1 \geq \sigma_2 \geq ... \geq \sigma_r$). We can re-write Eq. (4) in a different form:

$$\tilde{X}_k = \sum_{i=1}^{k} u_i \sigma_i v_i^T. \tag{5}$$

Here u_i's and v_i's are the column vectors of matrix U and V. They are used as the building blocks to reconstruct the original matrix X. With this format, SVD can be viewed as a way to decompose a matrix: It breaks the matrix X into column vectors u_i, v_i and real numbers σ_i. If we retain all these components (i.e., $k = rank(X) = r$ in Eq. (5)), SVD is a lossless operation and the matrix X can be reconstructed accurately. However, in practical applications, SVD can be treated as a lossy compression and only the important components are retained to give a rank-k approximation of matrix X. The percentage of power in the original matrix X captured in the rank-k reconstruction in Eq. (5) can be calculated by

$$\frac{\sum_{i=1}^{k} \sigma_i^2}{\sum_{i=1}^{Rank(X)} \sigma_i^2}. \tag{6}$$

For our data sets, users have much fewer behavioral modes than the number of association vectors, and for most users the dominant behavioral modes are much stronger than the others. Hence we expect SVD to achieve very good data reduction on the association matrices. This is indeed the case, as we show in Fig. 7: Most of the users have a high percentage of power in association matrix X explained by a relatively low-rank reconstruction – For example, in the USC trace (Fig. 7(a)), if we use a rank-1 reconstruction matrix, it captures 50% or more of power in the association matrices for more than 98% of users. Even if we consider an extreme requirement, capturing 90% of power, it is achievable for 68% of users using a rank-1 reconstruction matrix, and for more than 99% of users using at most a rank-7 reconstruction matrix. Similar observations can be made for Dartmouth users (in Fig. 7(b)). For both campuses, five components are sufficient to capture 90% or more power for most (i.e., more than 90%) of the users. This indicates although users show multi-modal association pattern, for most users the top behavioral



(a)USC.

(b)Dartmouth.

Fig. 7. Low association matrices dimensionality: A high target percentage of power is captured with low rank reconstruction matrix for many users.

modes are relatively much more important than the remaining ones.

The existence of multi-modal users can also be observed from the eigen-behavior vectors. For both campuses, while the top eigen-behavior explains high percentage of power in the association matrices, there are also many users (30% for USC and 50% for Dartmouth) for which the top eigen behavior captures less than 90% of power. This indicates the non-dominance of the top-1 eigen behavior, or the multi-modal nature of the mobility patterns.

If a low-rank reconstruction of the association matrix is achievable, it is natural to ask for the representative vectors for the behavioral modes of a user. For this purpose, SVD can be viewed as a systematic procedure to obtain representative vectors that capture the most remaining power in the matrix. Mathematically[9],

$$v_k = \arg \max_{\|v\|=1} \|(X - \sum_{i=1}^{k-1} X v_i v_i')v\|. \tag{7}$$

We can interpret the singular vectors, v_j's, as the vectors that describe the user's behavioral modes in decreasing order of importance in the association matrix X, with its relative weight (or the importance) quantified by $\sigma_j^2 / \sum_{i=1}^{r} \sigma_i^2$, following Eq. (6). In the paper we refer to these vectors as *eigen-behavior* vectors for the user.

The *eigen-behavior* vectors, v_j's, are unit-length vectors. The absolute values of entries in an *eigen-behavior* vector quantify the relative importance of the locations in the user's j-th behavioral mode. For example, suppose a given user visit location l almost exclusively, then in his first eigen-behavior vector, the entry corresponds to location l would carry a high value (i.e. close to 1), and the weight of the first eigen-behavior vector, $\sigma_1^2 / \sum_{i=1}^{r} \sigma_i^2$, shall be high. With a set of *eigen-behavior* vectors and their corresponding weights, we can capture and quantify the relative importance of a user's behavioral modes.

In sum, there are several benefits of applying SVD to obtain the summary as compared to other schemes: (1) SVD provides the optimal summary that captures the most remaining power in the original matrix with each additional

[9] SVD on matrix X can be viewed as calculating the eigenvalues and eigenvectors of the covariance matrix, $X^T X$. This is also the procedure typically used to perform Principal Component Analysis (PCA) for matrix X.

component. (2) The components can be used to reconstruct the original matrix, while the calculation of average or centroid vectors are non-reversible. Thus SVD provides a way to compress user association vectors and helps us save storage space. (3) Not only the most important behavioral mode, but also the subsequent ones can be systematically obtained with SVD, with a quantitative notion of their relative importance.

VI. Clustering Users by Eigen-behavior vectors

In this section, we first define our novel distance metric based on the *eigen-behavior* vectors and then use it for clustering.

A. Eigen-behavior Distance

Given two users and their respective association matrices, we would like to determine their similarity based on the SVD-based summary. This similarity indicates the relative position of users in the space of their mobility preferences. Users with similar mobility preferences are identified with high similarity scores, calculated from their eigen-behavior vectors. Suppose u_i's and v_j's are the eigen-behavior vectors of two users, $i = 1, ..., r_u$ and $j = 1, ..., r_v$ where r_u and r_v are the ranks of the corresponding association matrices. The similarity between the two users can be calculated by the sum of pair-wise inner products of their eigen-behavior vectors u_i's and v_j's, weighted by w_{u_i} and w_{v_j}[10]. Our measure of similarity between two sets of eigen-behavior vectors, $U = \{u_1, ..., u_{r_u}\}$ and $V = \{v_1, ..., v_{r_v}\}$, is defined as:

$$Sim(U,V) = \sum_{i=1}^{r_u} \sum_{j=1}^{r_v} w_{u_i} w_{v_j} |u_i \cdot v_j|. \tag{8}$$

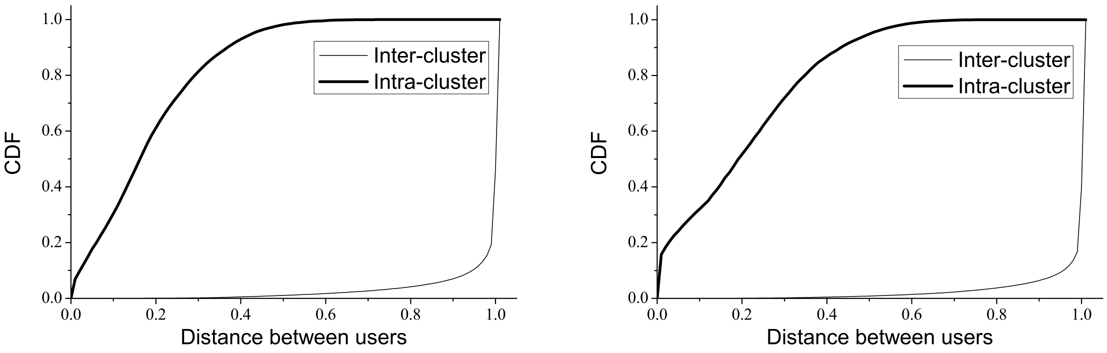
Higher similarity index $Sim(U,V)$ indicates that the eigen-behavior vectors U and V are more similar, and hence the corresponding users have similar association patterns. We define the *eigen-behavior distance* between users U and V as $D'(U,V) = 1 - (Sim(U,V) + Sim(V,U))/2$.

Using the *eigen-behavior distance* also reduces the computation overhead. If we use only the top-5 components (which captures more than 90% power in the association matrices as shown in Fig. 7), instead of going through t-by-t pairs of original association vectors as in section IV, we reduce the distance calculation to 5-by-5 pairs. Since we have at least 61 days in the traces, this is at least a $(61/5)^2 \approx 148$ fold saving for all N^2 pair of users. By paying the pre-processing (i.e., SVD for all N users) overhead of $O(Nt^2)$ [11] , we can reduce the distance calculation complexity from $O(N^2t^2)$ to $O(c \cdot N^2)$[12]. In the following study where we classify users into clusters

[10] w_{u_i} represents the weight of the eigen-behavior vector u_i, calculated by $w_{u_i} = \sigma_i^2 / \sum_{k=1}^{r_u} \sigma_k^2$. w_{v_j}'s are defined similarly.

[11] Note that in our notation, each association matrix is a t-by-n matrix. SVD of this matrix is of complexity $O(4t^2n + 8tn^2 + 9n^3)$, however, here we are more interested in how the computational complexity scale with t (length of the trace) and N (total users in the trace), since n (number of AP) is a constant in the environment.

[12] Since users follow repetitive trends (as indicated by the result of few behavioral modes in section V-A) in the association patterns, its total *eigen-behavior* vectors would not grow with the number of time slots, t. If we consider longer traces, the reduction can be even more significant.



(a) USC.

(b) Dartmouth.

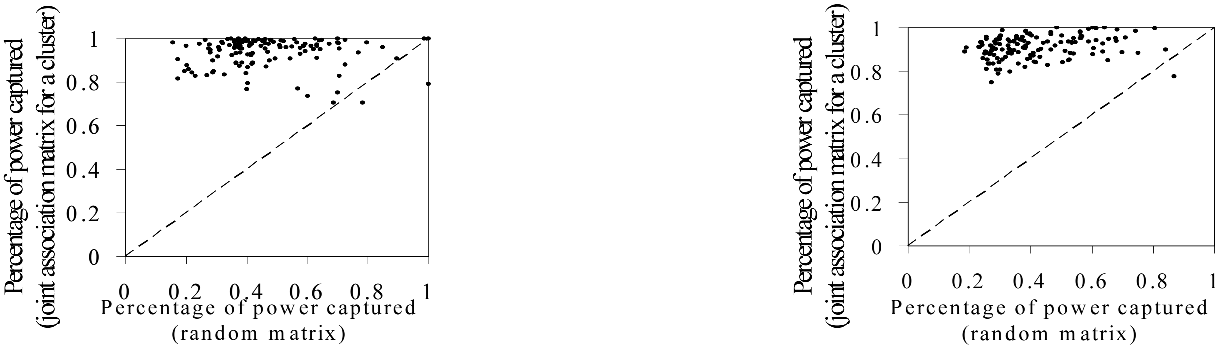
Fig. 8. Cumulative distribution function of distances for inter-cluster and intra-cluster user pairs (eigen-vector distance).

with similar behavior, we consider only the eigen-behavior vectors that capture at least 0.1% of total power.

B. Significance of the Clusters

We now reconsider the user clustering (grouping users with similar mobility trend) we have considered in section IV based on the eigen-behavior distance discussed in the previous section. Again, we validate the results by plotting the intra-cluster and inter-cluster distance distributions, when we group all users into 200 clusters. With the eigen-behavior distance, for both USC and Dartmouth traces, there is a better separation between the CDF curves (Fig. 8) as compared to the results with the AMVD distance (Fig 3), indicating a meaningful clustering. This shows the eigen-behavior distance is a better metric than the AMVD distance as it helps us to group users into well-separated behavioral groups based on their WLAN association preferences, for both campuses.

We further validate whether the resulting clusters indeed capture users with similar behavioral trends. We compose the *joint association matrix* by concatenating the daily association vectors of a cluster of m similar users in a larger mt-by-n matrix, where n is the number of locations and t is the number of time slots. When we perform SVD to the *joint association matrix*, the top eigen-behavior vectors represent the dominant behavioral patterns within the whole group. If the users in the group follow a coherent behavioral trend, the percentage of power captured by the top eigen-behavior vectors should be high. On the other hand, if association vectors of users with different association trends are put in one *joint association matrix*, the percentage of power captured by its top eigen-behavior vectors should be much lower. Among all clusters, we pick those with more than five users, and compare the cumulative power captured by the top four eigen-behavior vectors of these clusters with random clusters of the same size (i.e., randomly pick the same number of users from the whole population) in scatter graphs, in Fig. 9. Clearly, most the dots are well above the 45-degree line for both campuses. This indicates the users in the same cluster follow a much stronger coherent behavioral trend than randomly picked users, pointing to the significance of our clustering results.



(a) USC(129 clusters)

(b) Dartmouth(136 clusters)

Fig. 9. Scatter graph: Cumulative power captured in top four eigen-behavior vectors of random matrices (X) and joint association matrices formed by users in the same cluster (Y). Only clusters with 5 or more members are included.

TABLE III

THE AVERAGE SIGNIFICANCE SCORE OF THE FIRST EIGEN-BEHAVIOR OF EACH CLUSTER (GROUP WITH USERS OF SIMILAR BEHAVIOR)

	for its own group	for other groups
USC	0.779	0.005
Dartmouth	0.727	0.004

We would also like to see if each cluster from the population shows a distinct behavioral pattern. To quantify this, we obtain the first eigen-behavior vector from each group and calculate its *significance score*, defined in Eq. (2), for all the groups. The results confirm with our goal of identifying groups following different behavioral trend (see Table III): For both campuses, the first eigen behavior of each group has high significance score for its own group, but very low score for any other group.

We conclude that we have designed a distance metric that effectively partitions users into groups based on behavioral patterns. In addition, these clusters are unique with respect to behavioral trends. The eigen behavior distance metric shows several benefits: (1) The dominant behavior mode(s) get higher weights in distance calculation, thus noises in user mobility is suppressed and this leads to better clustering results. (2) Computation overhead is reduced as explained in section VI-A. (3) The eigen-behavior vectors form a compact summarization of user mobility. It is more efficient to store and exchange, an important benefit in decentralized applications.

VII. INTERPRETATION OF THE CLUSTERING RESULTS

In this section we analyze and interpret the results of clustering for both university campuses from social perspective. We first understand the distribution of cluster sizes from the trace, and then discuss the detailed user behavior of some groups of interest.

Distribution of Cluster Sizes

First we analyze the group size distribution, as shown in Fig. 10. We observe the distributions of group sizes are highly-skewed for both campuses. There are dominant behavioral groups that many users follow: the largest

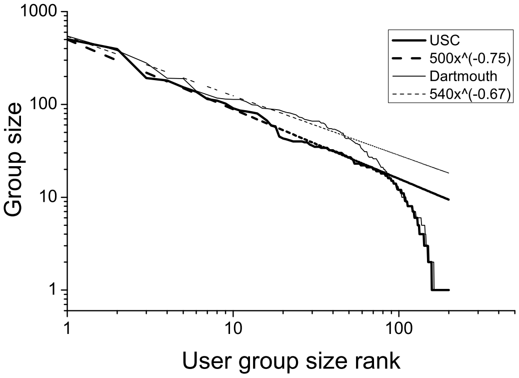


Fig. 10. Rank plot (group size ranking v.s. group size) in log-log scale. User group size follows a power-law distribution.

groups in the two campuses include 504 and 546 members, out of the population of 5000 for USC and 6582 for Dartmouth, respectively. The ten largest groups combined account for 39% and 33% of the total population, respectively. On the other hand, there are also many small groups, or even singletons, for both populations: out of the 200 clusters, there are 68 and 57 of them with less than five members, respectively, and in both campuses about half of the groups have less than 10 members. More interestingly, we observe that besides these small clusters, the distribution of the cluster size follows a power-law distribution, for the top half of clusters. In Fig. 10, we plot the straight lines that illustrate the best power-law fits, using

$$Groupsize(x) = Cx^{-s}, \tag{9}$$

where x is the ranking of the group in terms of its size, C is a constant, and s is the slope of the power-law fit. The slopes for these lines are -0.67 for Dartmouth and -0.75 for USC, respectively [13]. The power law distribution of group sizes may be related to the skewed popularity of locations on campuses – it has been shown that the number of patrons to various locations differ significantly[9]. However, the link between the distributions of number of patrons and the distribution of group sizes is not direct. While the most-visited locations on both campuses easily attract thousands of patrons, these people are broken into different behavioral groups depending on their association preferences. This finding suggests, in a model that captures user groups with different behavior, one should definitely consider group size variances for major groups and minor ones, as they differ in several orders of magnitude in realistic scenarios.

Important Behavioral Trends in Major Clusters

We now study the detailed behavior within a cluster. The largest clusters on both campuses include the library visitors, as expected, since libraries are still the most visited area on university campuses. For the USC campus, the

[13] We have performed Kolmogorov-Smirnov [38] test to examine the quality of the power-law distribution fit. The resulting D-statistics for both traces are below 0.03, indicating a good fit.

largest user cluster visits the library (the first eigen-behavior vector has a single high-value entry corresponding to the library, and this eigen-behavior vector captures 83% of the power in the joint association matrix for the group), followed by a couple locations around the Law school (4.45%) and the school of Communication (4.5%), both are popular locations on campus. For the Dartmouth campus, the largest user cluster visits LibBldg2 (72.85%), followed by LibBldg1 (5.13%), SocBldg1 (3.56%), and LibBldg3 (1.93%). It seems this group consists of library patrons who mainly move about the public area on the campus and access the WLAN from these locations.

While libraries are popular WLAN hot spots, we also discover many user clusters that rarely visit these locations. The second largest cluster for USC consists of users visiting mostly the Law school (89.73% of power), school of accounting (6.37%), and a couple of locations close to the Law school (0.59%). For Dartmouth, the second largest cluster visits AcadBldg18 (56.38%), AcadBldg6 (13.4%), ResBldg83 (10.15%), AcadBldg31 (3.5%), AcadBldg7 (3.12%), which seems to be a group of students going to classes at multiple academic buildings. We have also observed various clusters featured different dorms and classrooms as their most visited location from both campuses.

In general, for most of the larger groups, the association behavior can be described by a sequence of eigen-behavior vectors of decreasing importance with a clear ordering. This observation matches with the current status of WLAN usage: people tend to access WLAN at only a limited number of locations, and the preference of visiting locations is skewed [16]. For such users, just its most visited locations might be sufficient to classify them.

Clusters with Multiple Frequent Locations

On the other hand, we have also discovered groups with multiple high-value entries in its top eigen-behavior vectors from both campuses. One prominent example from USC trace consists of 32 users, who visit buildings VKC and THH, two major classrooms on the USC campus. The top two eigen-behavior vectors of the cluster both consist of two high-value entries corresponding to these two buildings[14], and they capture 63.14% of power in the joint association matrix. This cluster consists of users who visit these two locations with similar tendency, according to the eigen-behavior vectors. This cluster is a good example to show why it is not sufficient to merely use the most dominant behavioral mode (or the most-visited location) of a user to classify it. If only the centroid of the dominant behavioral mode is used to classify users, the behavioral trend of visiting multiple locations with similar tendency will not be revealed (these users will then be either classified as visiting VKC or THH the most often, while the fact that the other location also plays a significant role in their location preferences being ignored). As portable wireless devices gain popularity, we expect to see more users displaying diverse behavioral trends in terms of network usage. To fully capture such behavior, averaging-based summary is not sufficient, and this is where SVD shows its strength the most.

Small Groups with Distinct Behavior

[14] One of the eigen-behavior vectors has positive values for both entries, and the other has one positive and one negative value, in order to adjust the ratio between these two locations in the association vectors.

Interestingly, we also discover many small clusters with unique behavioral patterns that deviate from the "main stream" users in both traces. For example, in the USC trace, there is a small cluster of eight users who visit exclusively a fraternity house. Probably these are the people who live there. In the Dartmouth trace, there is a cluster of eight users who visit mostly athlete buildings (AthBldg5 (90.9%), AthBldg10 (4.62%), AthBldg2 (3.14%), AthBldg3 (0.8%), and ResBldg26 (0.54%)). These are probably either athletes or management staffs of the athlete facility. Such findings substantiate our motivation of the study: as the wireless technology prevails, we can expect users to display diverse behavioral patterns that reflect to their personal preferences, and it is important to capture such behavioral trend and quantify its significance.

Please note that while we observe the above mentioned mobility preferences through WLAN trace analysis as a mean, some of the findings may reflect more intrinsic human behavior. For example, multi-modal user mobility might be a consequence of people following different daily schedule on different days of week, and it holds regardless of wireless technology. We do believe that with ubiquitous network coverage, the observation from network traces will reflect the intrinsic human behavior better than before, as users are not limited by the technology to connect only at certain locations. However, it is not our intention to link the observations to human behavior in this paper since much more in-depth study is required.

To sum up, we have demonstrated a systematic way to identify distinct behavioral groups within on-campus populations. Using unsupervised clustering, our method does not rely on a pre-defined assumption about user behavior, and is able to partition the users into distinct groups, each capturing users with similar mobility preferences. The method and findings are useful for various applications, as we discuss next.

VIII. Discussions

In this section, we first present a case study of how the user similarity metric proposed above is useful in social-behavior-aware message dissemination, to illustrate the practical applicability of the analysis and proposed similarity metric. We then briefly discuss alternative representations we have considered.

A. Evaluation of behavior-aware protocols

One particular use case for social networking is to discover and share information with people similar to a pre-specified behavioral pattern or interest. This is in general useful for discussions and advertisements among people with common interest. Here, using the *eigen-behavior* defined above, we demonstrate the efficiency improvement such behavior awareness can result in for message dissemination protocols, in a scenario when mobile users exchange messages among peers, without reliance on communication infrastructure. To showcase the application utility of such idea, we use a concrete example of message dissemination in mobile networks. Here we evaluate a dissemination scheme using behavioral pattern matches, where the messages can be transferred between the users in peer-to-peer (as in ad hoc and delay tolerant networks) without using the infrastructure, as in [6], [21], [22]. We simulate users trying to group-cast messages to other people displaying similar *eigen-behavior* to themselves, using the WLAN trace as the input for user mobility pattern.

In our simple similarity-based group-cast protocol, each node has only a limited view about other nodes' behavior. We assume each node has a limited, localized knowledge – whenever a node has a message to send and encounters with another node, they exchange the summarized eigen-behavior and calculate their similarity based on Eq. (8). If their similarity is higher than a pre-defined threshold, the message is copied to the other node and it also becomes a sender. We compare this simple *similarity-based* protocol with the following alternatives: (1) *Flooding*: The nodes in the network are all oblivious to user behaviors. Hence, the sender and other nodes that have received the message just blindly send out copies of the message to nodes who have not received it yet. This scheme is also known as *epidemic routing* [39]. (2) *Centralized*: In this ideal scenario, all nodes acquire global knowledge about the user groups based on similar mobility trend (i.e., each node has centralized view of user grouping results as we described in section [**?**]). Note that such knowledge requires the processing of eigen-behavior from all nodes and typically is not available in decentralized networks. We use this as an ideal case where nodes only propagate the message to others if they are in the same group. (3)*Random-transmission (RTx)*: In order to show the benefit of utilizing the knowledge gained from the user behavior analysis, we also include another group-oblivious protocol with the objective of reducing message transmissions. In the RTx protocol, the current message holder sends the message to another node with probability p when they encounter. This one copy of message is passed around among users. Loops are avoided by not sending to the nodes who have seen the same message before. This process continues until a pre-set hop limit is reached.

We evaluate the performance of the aforementioned schemes with the following metrics: (1) *Delivery ratio*: The number of nodes receiving the message over the number of intended receivers. (2) *Delay*: The average time taken for a scheme to deliver the messages to recipient nodes. (3) *Overhead*: The total number of transmissions involved in sending messages to intended receivers. We use the USC trace as an example and split the WLAN trace into two halves – The first half of the trace is used to determine the grouping of users based on their mobility preferences, and we evaluate the group-casting protocol performances using the second half of the same trace. For each group with more than 5 members, we randomly pick 20% of the members as the source nodes sending out a *one-shot message* to all other members in the same group.

We choose *epidemic routing* (or *flooding*) as the baseline for our evaluation. We show the relative performance of all other group-casting protocols relative to that of epidemic routing in Fig. 11. In the figure we see that *flooding* has the lowest delay and the highest delivery ratio as it utilizes all the available opportunities to propagate the message. However, it also incurs significant overhead, as all nodes in the network receive a copy of the message, regardless

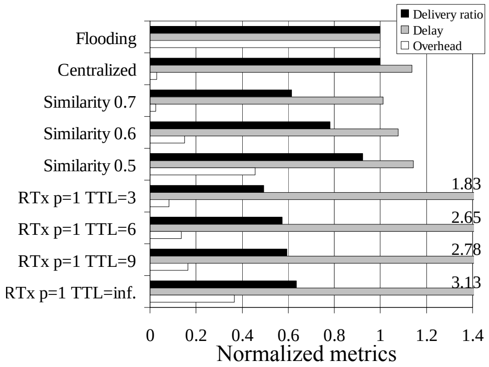


Fig. 11. Relative performance metrics of the group-casting schemes normalized to the performance of *flooding*.

of whether they need it, as long as the nodal encounter patterns allow for message propagation. The average delay, which is the lowest possible under the given encounter patterns, is in the order of days (3.56 days in this particular case). Group-casting based on *centralized* clustering information, the ideal scenario with centralized knowledge, significantly reduces the overhead (since messages are delivered only to the target nodes) while maintains almost perfect delivery ratio, with a little extra delay. However, it is not realistic to assume such centralized knowledge.

For the *similarity-based* forwarding, the aggressiveness of the protocol can be fine-tuned with the forwarding threshold of the similarity index. Experiment results show a significant reduction of overhead (only 2.5% of *flooding*) at the cost of delivery ratio if we set high threshold such as 0.7. Refer to Fig. 8, setting similarity threshold of 0.7 (or, equivalently, feature-based distance 0.3) leads to sending messages almost exclusively within the group of similar mobility preference. Setting a low threshold (e.g., 0.5) leads to little reduction in delivery ratio (92% of *flooding*) and increase in delay but still cuts the overhead to 45% of *flooding*. For the *RTx* protocol, although the overhead can be controlled with the hop-limit (which we set as TTL times of the group size), we see that the delivery ratio is lower than that of the *similarity-based* protocol with comparable overhead (comparing similarity 0.6 with RTx $p = 1.0$ $TTL = 9$, the former has 30% higher delivery ratio than the latter) because in many cases the message is transmitted to some node out of the desired group and there is no knowledge to direct its propagation. In addition, we try the *RTx* protocol with various p and TTL values and find it is not as flexible as the *similarity-based* protocol in which the parameters can be tuned to trade overhead for better delivery ratio.

Based on the case study, we see that the similarity index proposed in this paper has direct impact on improving message delivery efficiency by incorporating social behavior awareness. While we use mobility as one aspect of social behavior, we envision the framework proposed in this paper similarly applicable to other aspects of mobile user behavior as well. Finally, the proposed similarity metric is also useful in the framework of participatory sensing [31] in order to recruit users representing a diverse behavior set for more holistic view of the information collected.

The proposed similarity-based scheme here is a simple and direct utilization of understandings in user behavior, to show case its utility. A more in-depth discussion of this direction can be found in [34], where we formally establish the stability of user mobility preference, and leverage the implicit social behavior space and different roles of nodes in this space to design more generic protocols. Finally, leveraging social behavior patterns can also help other tasks in mobile networks, such as identifying important locations for individuals [7] for location-aware services, and mitigating privacy attacks [32] or establishing trust [33]. We will further explore this direction in the future.

B. Alternative Representations and Metrics

We have evaluated our *TRACE* approach extensively with other distance metrics and representations of the data. Due space constraints, we only briefly discuss them here.

We design distance metrics with other types of summaries in section V-B and they lead to user partitions similar to that from the SVD-based summary, since in current WLANs, most users have a dominant behavioral mode so simple summaries suffice to capture the trends. However, SVD is able to discover repetitive trends when users have a complicated pattern of visiting access points, while other methods cannot, as we argue in section VII.

We also consider several other representations. Without normalization (i.e., the entries in the association vectors represent the absolute duration of association), the most active users are classified similarly as in the case of normalized representation, but the less active users fall in different clusters due to their sporadic usage of WLAN. Our idea is to view a user's behavior based on the fraction of time she spends at a location. We also explore the clustering by using both finer location granularity (i.e., use each AP as a unique location) as well as finer time granularity (i.e., one association vector for each three-hour period). The resulted clusters are similar to what we obtain, indicating the time-scale of daily vectors with per-building location granularity provide sufficient information. Due to space constraint, more details are left to the appendix in technical report[28]. We would like to note, if one is interested in classifying users based on their behavior during certain time frames (e.g., consider only working hours during weekdays), our technique is still applicable and we only need to change the construction of association matrices to include only interested time periods. Overall, different representations manifest themselves in clustering results with different emphasis and interpretations. One needs to pick a proper representation understanding the particular part of user behavior to focus on.

On a different note, it may be of interest to use similar techniques in other domains in different type of networks. For example, in encounter-based networks [21], a representation of encounter probability or duration would be appropriate. We plan to investigate this in our future work.

IX. Conclusion

In this paper, we classify groups of WLAN users based on the trends in their association patterns in two major university campuses by leveraging clustering techniques and our systematic *TRACE* approach. We design a novel distance metric between users based on the similarity of their *eigen-behavior* vectors, obtained through singular value decomposition (SVD) of the association matrices. We have shown although many (at least 60%) users display multi-modal behavioral modes, SVD is able to capture at least 90% of power in association matrices for most users with at most five components. This also leads to space and time efficient computations.

The eigen-behavior distance leads to a meaningful partition of users. We establish that WLAN users on university campuses form a diverse community, which includes hundreds of distinct behavioral groups in terms of association patterns. The size of the groups follows a power-law distribution on both campuses. While the top ten groups account for at least 33% of population in our data sets, there exist many small groups with unique association patterns. In spite of the very different location and demography of the two university campuses studied, it is surprising to find out the qualitative commonalities of the user behavior trends.

While distance metrics based on simple summaries (e.g., average or centroid of the dominant behavior mode) suffice for most current WLAN users, our study indicates that SVD is capable to capture user association trends in complex situations, e.g., when users visit several distinct locations on different time slots. As personalized wireless devices become more popular, WLANs become ubiquitous and their powerful combination impacts our daily lives, a powerful tool to understand the user behavior is essential. Such understanding could enable more efficient behavior-aware protocols and applications.

REFERENCES

[1] A. Jain, M. Murty, and P. Flynn, ”Data Clustering: A Review,” ACM Computing Surveys, vol. 31, no. 3, September, 1999.

[2] I.T. Jolliffe, Principal Component Analysis, second ed., Springer series in statistics, published 2002.

[3] R. Horn and C. Johnson, Matrix Analysis, Cambridge University Press, published 1990.

[4] N. Eagle and A. Pentland, ”Reality mining: sensing complex social systems,” in Journal of Personal and Ubiquitous Computing, vol.10, no. 4, May 2006.

[5] M. Kim and D. Kotz, ”Periodic properties of user mobility and access-point popularity,” Journal of Personal and Ubiquitous Computing, 11(6), August, 2007.

[6] W. Hsu and A. Helmy, ”On Nodal Encounter Patterns in Wireless LAN Traces,” in IEEE Trans. on Mobile Computing, vol. 9, issue 11, Nov. 2010.

[7] J. Kang, W. Welbourne, B. Stewart, and G. Borriello, ”Extracting places from traces of locations,” in SIGMOBILE Mobile Computing and Communication Review, vol. 9, no. 3, July 2005.

[8] A. Lakhina, K. Papagiannaki, M. Crovella, C. Diot, E. D. Kolaczyk, and N. Taft, ”Structural Analysis of Network Traffic Flows,” ACM SIGMETRICS, New York, June 2004.

[9] T. Henderson, D. Kotz and I. Abyzov, ”The Changing Usage of a Mature Campus-wide Wireless Network,” in Proceedings of ACM MobiCom 2004, September 2004.

[10] M. Balazinska and P. Castro, ”Characterizing Mobility and Network Usage in a Corporate Wireless Local-Area Network,” In Proceedings of MobiSys 2003, May 2003.

[11] M. McNett and G. Voelker, ”Access and mobility of wireless PDA users,” ACM SIGMOBILE Mobile Computing and Communications Review, v.7 n.4, October 2003.

[12] MobiLib: Community-wide Library of Mobility and Wireless Networks Measurements. http://nile.cise.ufl.edu/MobiLib.

[13] CRAWDAD: A Community Resource for Archiving Wireless Data At Dartmouth. http://crawdad.cs.dartmouth.edu

[14] D. Kotz, T. Henderson and I. Abyzov, CRAWDAD data set dartmouth/campus/movement/01_04 (v. 2005-03-08). Downloaded from http://crawdad.cs.dartmouth.edu/dartmouth/campus /movement/01_04, March 2005.

[15] C. J. F. ter Braak, ”Principal Components Biplots and Alpha and Beta Diversity,” Ecology, vol. 64, pp. 454-462, 1983.

[16] W. Hsu and A. Helmy, ”On Important Aspects of Modeling User Associations in Wireless LAN Traces,” the Second International Workshop On Wireless Network Measurement (WiNMee 2006), April 2006.

[17] R. Jain, D. Lelescu, and M. Balakrishnan, ”Model T: An Empirical Model for User Registration Patterns in a Campus Wireless LAN,” in Proceedings of ACM MobiCom 2005, August 2005.

[18] D. Lelescu, U. Kozat, R. Jain, and M. Balakrishnan, ”Model T++: an empirical joint space-time registration model,” in Proceedings of ACM MobiHoc 2006, May 2006.

[19] C. Tuduce and T. Gross, ”A Mobility Model Based on WLAN Traces and its Validation,” in Proceedings of IEEE INFOCOM, March 2005.

[20] M. Papadopouli, H. Shen, and M. Spanakis, ”Characterizing the Duration and Association Patterns of Wireless Access in a Campus,” 11th European Wireless Conference 2005, Nicosia, Cyprus, April, 2005.

[21] A. Chaintreau, P. Hui, J. Crowcroft, C. Diot, R. Gass and J. Scott, ”Impact of Human Mobility on the Design of Opportunistic Forwarding Algorithms,” in Proceedings of INFOCOM 2006, Barcelona, Spain, April 2006.

[22] E. Daly and M. Haahr, ”Social Network Analysis for Routing in Disconnected Delay-Tolerant MANETs,” In Proceedings of ACM MOBIHOC, Sep. 2007.

[23] X. Hong, M. Gerla, G. Pei, C. and Chiang, ”A group mobility model for ad hoc wireless networks,” in Proceedings of the 2nd ACM international workshop on modeling, analysis and simulation of wireless and mobile systems, August, 1999.

[24] W. Hsu, T. Spyropoulos, K. Psounis, and A. Helmy, ”Modeling Time-variant User Mobility in Wireless Mobile Networks,” in IEEE/ACM Transactions on Networking, vol. 17, no. 5, Oct. 2009.

[25] M. Gonzalez, Cesar Hidalgo, and A. Barabasi, ”Understanding individual human mobility patterns,” Nature 453, pp. 779 - 782, June 2008.

[26] W. Hsu, D. Dutta, and A. Helmy, ”Extended Abstract: Mining Behavioral Groups in Large Wireless LANs” In Proceedings of ACM MOBICOM, Sep. 2007.

[27] W. Hsu, D. Dutta, and A. Helmy, ”Profile-Cast: Behavior-Aware Mobile Networking,” in Proceedings of IEEE WCNC, Las Vegas, NV, Mar. 2008.

[28] W. Hsu, D. Dutta, and A. Helmy, ”Mining Behavioral Groups in Large Wireless LANs,” Technical report available at http://arxiv.org/abs/cs/0606002

[29] S. Reddy, D. Estrin, and M. Srivastava, ”Recruitment Framework for Participatory Sensing Data Collections,” In proceedings of the Eighth International Conference on Pervasive Computing, Pervasive, Helsinki, Finland, May 17-20, 2010.

[30] K. Shilton, J. Burke, D. Estrin, R. Govindan, and J. Kang, ”Designing the Personal Data Stream: Enabling Participatory Privacy in Mobile Personal Sensing,” In proceedings of the 37th Research Conference on Communication, Information and Internet Policy (TPRC), Arlington, VA, 25-27 September 2009.

[31] S. Reddy, K. Shilton, J. Burke, D. Estrin, M. Hansen, M. Srivastava, ”Using Context Annotated Mobility Profiles to Recruit Data Collectors in Participatory Sensing,” in proceedings of the 4th International Symposium on Location and Context Awareness, LOCA, Tokyo Japan, May 2009.

[32] U. Kumar and A. Helmy, ”Human Behavior and Challenges of Anonymizing WLAN traces,” Ad Hoc Sensor and Mesh Networking Symposium, IEEE GLOBECOM 2009.

[33] U. Kumar, G. Thakur, and A. Helmy, ”PROTECT: Proximity-based Trust-advisor using Encounters for Mobile Societies”, ACM IWCMC, June 2010.

[34] W. Hsu, D. Dutta, and A. Helmy, ”CSI: A Paradigm for Behavior-oriented Profile-cast Services in Mobile Networks,” accepted and to appear in Elsevier Ad Hoc Networks.

[35] G. S. Thakur, A. Helmy, and W. Hsu, ”Similarity Analysis and Modeling in Mobile Societies: The Missing Link, ” in ACM Workshop on Challenged Networks (CHANTS ’10), Sep. 2010.

[36] J. Ghosh, M. J. Beal, H. Q. Ngo, and C. Qiao, ”On Profiling Mobility and Predicting Locations of Wireless Users,” in Proceedings of ACM REALMAN, May 2006.

[37] D. Crandall, L. Backstrom, D. Cosley, S. Suri, D. Huttenlocher, and J. Kleinberg, ”Inferring Social Ties from Geographic Coincidences,” Proc. National Academy of Sciences, Dec. 2010.

[38] R. Hogg and E. Tanis, ”Probability and Statistical Inference,” sixth edition, Prentice Hall, 2001.

[39] A. Vahdat and D. Becker, ”Epidemic Routing for Partially Connected Ad Hoc Networks,” Technical Report CS-200006, Duke University, April 2000.

[40] A. Helmy et al., ”Behavior-based Mobile Communities”, UF CISE Tech Report, Aug 2011.

Wei-jen Hsu was born in Taipei, Taiwan, in March 1977. He received the B.S. degree in Electrical Engineering and the M.S. degree in Communication Engineering from National Taiwan University in June 1999 and June 2001, respectively. He received the Engineer degree in Electrical Engineering from University of Southern California in August 2006, and the Ph.D. degree in Computer Science from University of Florida in August 2008. He joined Cisco Systems, Inc. in 2008. His main research interest involves the utilization of realistic measurement data in various tasks in computer networks, including user modeling and behavior-aware protocol design.

Debojyoti Dutta received a Btech in Computer Science and Engineering from Indian Institute of Technology (IIT), Kharagpur, India and a PhD in computer science from the University of Southern California (USC), Los Angeles, USA. Before joining Cisco Systems, San Jose, USA, he was a postdoc in Computational Biology at USC. His current interests include inferring models of human behavior from diverse networked measurements, applied data mining and network security.

Ahmed Helmy received Ph.D. in Computer Science 99 from the University of Southern California (USC), M.S. in Electrical Engineering (EE) 95 from USC, M.S. Eng. Math 94 & B.S. in EE 92 from Cairo University. He was a key researcher in the NS- 2 and PIM projects at USC/ISI from 95-99. Starting 99 has been on the EE Dept faculty at USC and the director of the wireless networking Lab. Since Fall 06, he became an associate professor and the director of the mobile networking laboratory in the CISE Dept at the University of Florida. He received the NSF CAREER Award in 02, the 00 Zumberge Research Award, and the best paper award at IEEE/IFIP MMNS 02. In 04 & 05, he got the best merit ranking in the EE-USC faculty. In 08 & 07, he was a finalist and winner in the ACM MobiCom SRC. He is an area editor of the Ad Hoc Networks Journal Elsevier and IEEE Computer, and the workshop coordination chair for ACM SIGMOBILE. His research interests include design, analysis & measurement of wireless adhoc, sensor & mobile social networks, mobility modeling, multicast protocols, IP mobility & network simulation.