

# Structural Analysis of User Association Patterns in Wireless LAN

Wei-jen Hsu<sup>1</sup>, Debojyoti Dutta<sup>2</sup>, and Ahmed Helmy<sup>1</sup>

<sup>1</sup>Department of Electrical Engineering, University of Southern California

<sup>2</sup>Department of Computational Biology, University of Southern California

Email: { weijenhs, ddutta, helmy } @usc.edu

## Abstract

Due to the rapid growth in wireless local area networks (WLANs), it has become important to characterize the fine-grained structure of user association patterns. In this paper, we focus on unraveling the structure in user's daily association patterns in WLANs in the long run. The daily access pattern is defined by the fraction of time it spends with a particular location. We answer three questions: 1) Do users demonstrate *consistent behavior*? Using our novel metrics and clustering, we conclude that many users (more than 50%) are multi-modal. 2) Is it possible to represent user association patterns using a compact representation? Using eigen-decomposition, we show that the intrinsic dimensionality of the constructed user association matrices is low and only the top five eigenvalues and their corresponding eigenvectors can be used to reconstruct those association matrices with an error of 5%, in terms of the  $L_1$  and  $L_2$  matrix norms. 3) How can we decide if two users have similar association patterns? We define two new metrics, and we rigorously validate their efficacy by demonstrating that the inter and intra cluster distributions, upon clustering, have very little overlap. Our methods and observations are a first step towards systematically mining user-association patterns and could lead to new directions in network management and understanding social patterns of users.

## 1 Introduction

There has been a rapid increase in wireless LAN (WLAN) deployments, users and traffic in the recent years. Such explosive growth mandates the design of sophisticated network management schemes for a multitude of tasks ranging from better deployment of access points meeting user needs to detection of malicious activity on these WLANs. In order to investigate the above problems, one of the first fundamental challenges that needs to be addressed is to determine the user access patterns. A detailed study of how users *behave* within a WLAN environment could have far-reaching consequences, from network deployment optimization

and usage pattern detection, to new applications that push content to the end-user based on the access patterns.

Most of the current studies in WLAN access patterns have focused on usage statistics (e.g., number of users at the access points (APs), online day counts for a user, or percentage of traffic using a protocol, etc.[13, 11, 12]). Other researchers have attempted to model user association to the APs (e.g., User arrival process at APs [14], user sessions length and preferences of AP selection [9, 18, 19], etc.). However, it is important to do a fine-grained quantitative study of the structures and trends in long term user association patterns. Understanding such detailed structure is especially important for large-scale WLAN with users that re-visit the environment over a period of time (i.e., User population and its behavior does not vary drastically in the studied period), such as in university campus WLANs or corporate WLANs.

There are very few known efforts to determine and quantify the fine-grained structures and trends in user association patterns to APs in the long run, and classify users based on such patterns. In this paper, we propose novel directions to understand such trends of user associations in a campus WLAN, and ways of quantifying those trends. Our goal is to quantify repetitive and consistent structure of user association patterns.

In wireless LANs, user association to APs is a rough indicator of the user's location, and this information reveals the set of locations the user visits. The *structure* of user behavior could be defined and studied at various time-scales. In this paper, we choose to study the composition of the user's *daily* association patterns within the USC campus - for example, whether they are similar or different from day to day, and how do we represent the major trends in these association patterns. Structural study of user association behavior can be useful for the aforementioned networks in several ways. First, by understanding the typical structure in user behavior, we could establish metrics for association patterns for the users and classify users into groups based on the metrics. Also, such metrics could help us detect abnormal user behavior (e.g., significant short-term change in association pattern), and eventually help us to plan

for future network deployments based on long-term user behavior trends. Second, profiling the structures in user association patterns help the network administrator to understand and serve its users better. The administrator may group users based on structural characteristics of the user association patterns and provide location-aware services to user groups. Third, profiling the structures in user association patterns can also help users to find others with similar association patterns and provide useful information to context-aware protocols.

Specifically, we focus on the following questions in this paper:

1. Do users show single-modal or multi-modal association patterns across days? More precisely, does a user show similar association pattern each day, or does it choose the daily association pattern from one of several different association modes or classes?
2. Is it possible to summarize the user association patterns for multiple days in a concise fashion for current WLAN users, regardless of whether the users display single or multi-modal behavior?
3. How do we quantify users as having similar or dissimilar association patterns? Can we utilize such metrics to partition the whole user population into clusters with similar users?

## 1.1 Our contributions

In this paper we provide new methodologies to analyze WLAN traces. To illustrate the usage of our tools, we analyze the association patterns of 5,000 users in the campus WLAN traces at University of Southern California. The trace was collected during the spring semester of 2006 for 94 days. However, our methodologies are not limited by the choice of data set, and can be applied to study to other WLAN traces (e.g., [7], [8]).

Our primary contribution is the methodologies we use to systematically analyze the association patterns. To this effect, we define novel features that can be extracted from traces, similarity metrics using eigen decomposition and we robustly answer the questions we ask using unsupervised learning methods such as clustering, which are subsequently rigorously validated. Our specific contributions are as follows:

1. Multi-modal user behavior: We find that in our university campus WLAN, most users display multi-modal association behavior. There are only few users that remain consistent in his association pattern (i.e., single-modal user) for all the days we studied. Specifically, more than 50% of users are classified as multi-modal under intermediate inter-pattern distance and very few users are uni-modal in nature.

2. Summarize user-association patterns: We define a new association matrix by concatenating the association patterns each day into different columns. Using an eigen decomposition technique to summarize the data set, we observe that for most users, the intrinsic dimensionality of its association matrix is quite low, i.e. it has very little modes of variation. For more than 99% of users, we can use at most 7 eigenvectors and eigenvalues to capture more than 90% of power in its association matrix.
3. Metrics to compare different user-association patterns: We propose two novel metrics to quantify the similarity between the association pattern structures of different users. In the *straight-forward* approach, we derive the metric based on detailed, comprehensive comparison of the association patterns of the users. In the *feature-based* approach, we use the eigenvectors and eigenvalues obtained from the association matrix as a feature set to determine the similarity between association patterns. We show that these two metrics are closely related, with a correlation coefficient of 0.9119, but the *feature-based approach* is computationally much simpler than that of the *straight-forward approach*. Finally, we utilize these metrics to obtain a partition of the user population by clustering. We validate our metrics by demonstrating that the mean of inter-cluster distance of the resulting partition is much larger than that of the intra-cluster distance in such partition. More than 90% of intra-cluster user pairs have distance smaller than 0.5, and more than 90% of inter-cluster user pairs have distance larger than 0.9. There is a clear separation of the distributions of the inter and intra cluster distances. Thus, with the proposed metric, we are able to identify users with similar association patterns.

The rest of the paper is organized as follows. Related work is briefly outlined in section 2. In section 3, we describe our trace collection infrastructure that we use and introduce the mathematical representations of user association pattern. In section 4, we identify multi-modal user behavior using clustering techniques. In section 5 we present the matrix representation of user association patterns. The metrics for user similarity and user population partition are presented in section 6. Finally we discuss the potential of our findings in section 7, and conclude with future work direction in section 8.

## 2 Related Work

Recently, there has been numerous papers on the empirical study of wireless LANs to understand its users. Earlier papers focused on the trace collection

infrastructure[13, 11, 12, 5], presenting the user statistics and basic understanding of user behavior, and contributing those traces to the research community.

One of the recent directions has been to understand user behavior deeply, which is also the goal of our paper. Most of them suggest models to describe aspects of user association behavior. In [9, 18, 19] the focus is to model user preferences, association durations at the access points. In [14] the authors propose a model for user arrival patterns. In this paper, we take an alternate approach and instead of modeling each user, we seek to understand user association by observing the long-term daily patterns empirically from traces and quantitatively describe the structure in user association patterns. We target at establishing a methodology to understand and explain the characteristics of user association in WLAN as a multi-variate analysis problem.

To the best of our knowledge, the questions we seek to answer in the paper are not fully addressed in the current literature. The multi-modal association behavior of WLAN users has not been investigated before, and there is no previous work on quantifying similarity metrics between user association patterns. Identifying major trends of association patterns using singular value decomposition, however, is also discussed in [22] for cell-phone users. But unlike our study, similarity metrics are not defined. More specifically, we leverage the techniques from machine learning and data mining to analyze the WLAN trace, which is a new approach that augments the existing work on the traces. We utilize mainly hierarchical clustering [1] techniques to classify users in the study.

We also utilize the eigen decomposition [4] to decompose the association matrix of users into its eigenvectors and eigenvalues, which is related to the principal component analysis (PCA) [2], which is used in [10] to find common trends in flows on network links, and in [22] to find trends of cell-phone users associations. We adopt a variant of traditional PCA, called *uncentered PCA*, which has been used in ecology to study the diversity of species at various sites[15]. Uncentered PCA is suitable if the mean is also a useful feature for comparison, which is certainly true in our study. Comparing similarity of data sets using the corresponding PCs of each set is also used in [16, 17]. However, unlike previous work, we define novel similarity metrics and use those metrics for clustering wireless users' association patterns into groups and robustly validate our metrics.

### 3 Preliminaries

In this section we first discuss about the wireless LAN trace collection process at University of Southern California (USC) and the basic facts about the trace. Then we discuss how we choose to define and represent *user*

*association pattern* in the paper.

#### 3.1 Trace collection

USC has a wide scale wireless LAN deployed on campus. We have been collecting traces from the USC WLAN from early 2004 [5, 6] and it is available for the community on our webpage [7]. Since the winter of year 2005, the user verification process for our campus network has changed significantly<sup>1</sup>, and the trace collection process is also modified. Interested readers can check out the release notes at MobiLib webpage [7]. From the traces, we derive user association history (i.e. The start and end time the user associates with particular location) at per-switch port granularity, which approximately corresponds to buildings on USC campus.

In this paper we analyze the WLAN traces collected at USC after the network policy changes. Specifically, we use the traces collected between Jan. 25, 2006 and Apr. 28, 2006, which correspond to the spring semester. There are 137 unique switch ports (corresponding to different locations) in the trace. We have seen totally 25,481 unique MAC addresses<sup>2</sup>, a significant increase as compared to 14,856 unique MAC addresses seen in fall 2004 semester or 4,258 in summer 2005. We pick the top 5,000 users among the 25,481 by ranking them in terms of total online time during the trace period. We consider a user being online whenever it is associated with one of the switch ports. Since we are interested in finding suitable methods to represent the structure in association pattern, we choose to focus on the more active users, and disregard the majority of less active users. For the 5,000 chosen users, the most active user is online for 99.9% of time during this period (almost always on), and the least active user is online for 4.2% of time. Note that even though we have disregarded about 80% of the users, the 5,000 chosen users still span across a wide range of user activeness.

#### 3.2 Representation of User Association Pattern

In this section we introduce our representation of the user association patterns using the information obtained from the traces.

The first question we address in this paper is: how to represent user association pattern. Representation of user association patterns can be described at various

<sup>1</sup>Specifically, VPN connection to a central server is no longer necessary, and the users can start using WLAN on campus upon a verification process using the USC e-mail account. This has improved the accessibility of USC WLAN to students and faculty on campus. We have seen the number of users increases significantly after the change.

<sup>2</sup>In this paper, we apply the common assumption that each unique MAC address represents a unique device, which is owned and used by a unique user.

time intervals: hours, days, or even weeks. All these representations are valid, but in this paper, we choose the *daily* association pattern. Hence, later in the paper, by *association pattern*, we implicitly mean the *daily association pattern* (to be defined later) of a user. We pick one day as the time interval because USC is a commuter campus and daily patterns capture sufficient uniqueness without representing highly averaged behavior and user group specificity. For example, students on university campuses move from class to class in 90 to 120 minutes period, but such structure is absent for staff on campus. Hence if we have chosen 90 minutes as a period for identifying consistency in association behavior, the results of study would heavily depend on to which population we perform the study, and how we apply the 90 minutes intervals (e.g. Whether it happens to align with the class schedule).

Now we can formally define the *daily association pattern* for a user: We represent the user association pattern for a day as an  $n$ -entry vector,  $(a_1, a_2, \dots, a_n)$ , where  $n$  is the number of switch ports in our trace. Each entry in the vector,  $a_i$ , represents the *fraction* of online time during the day the user spends at the switch port (i.e. the time user spends at the particular switch port divided by the user's online time of the day). We normalize the user association time with respect to his online time because we want to assess the relative importance of the locations to the user. The importance of a location for a user is better reflected by the ratio of online time the user spends at the location. In this case, the conclusions we draw is not influenced by the absolute value of online time, as this factor varies over a wide range among users. Note that the sum of the entries in the daily association pattern,  $\sum_{i=1}^n a_i$ , is always 1 if the user has been online during the day. We use a zero vector to represent the association pattern when the user is completely offline for the day.

In this paper, when we assess the similarity between two association pattern vectors, we use Manhattan distance, or the L1 norm, as the distance measure since it is more robust to statistical noise. The distance between two vectors  $a$  and  $b$ , denoted as  $d(a, b)$ , is defined as:

$$d(a, b) = \sum_{i=1}^n |a_i - b_i| \quad (1)$$

where  $a_i$  and  $b_i$  are the  $i$ -th entry in vector  $a$  and  $b$ , respectively.

## 4 Clustering of User Association Patterns

Given the daily association patterns of a user for the duration of a semester, the first question we ask is whether the user shows a single mode in its daily association

patterns, or it switches between several modes of association patterns through the course of semester. By *single-modal users*, we refer to those who display similar association patterns (i.e., the distance between association patterns for different days are small.). By *multi-modal users*, we refer to those who display several distinct groups of association patterns across different days (e.g., If a student goes to classroom buildings A and B for 2 hours each on the days he attends the classes, and stays in library for whole day if there is no class on the day, there will be two unique and well-separated association patterns - "class" mode and "library" mode - for the student). One natural way to answer the question is by applying clustering techniques [1] to the association patterns of the user. If multiple clusters can be identified from the association patterns, then the user under consideration is a multi-modal one. In this section we briefly describe the clustering technique, and apply it to the association patterns of each user.

### 4.1 Clustering Technique

The user association pattern for each day is a vector in a  $n$ -dimension space. If the user shows similar association pattern each day, the vectors should be close to each other in this space. On the other hand, if the user shows drastically different behavior each day, the vectors will be scattered in the space. Clustering is a huge area in itself, but it can be roughly classified into hierarchical or partitional schemes. In this paper we use the hierarchical clustering, in which each vector is initially considered as a cluster containing one member. Then, at each step, based on the specific distance measure between the clusters, two clusters that are closest to each other among all cluster pairs are merged into one cluster with larger membership. Therefore for each step the total number of cluster decreases exactly by one. This process continues until all vectors are merged into one cluster containing every vector.

A *dendrogram* can be created through the process of hierarchical clustering. It contains a tree structure showing the order of clusters merging with each other and the distance between the clusters that are chosen to merge at each step. Therefore, with the complete dendrogram, one could choose a proper distance threshold to stop the merging process if all the inter-cluster distances are larger than the threshold (i.e., all the remaining clusters are separated by distances larger than the threshold), or a cluster number threshold to stop the process when the vectors are merged into a pre-defined target number of clusters.

However, one issue regarding the above technique applied to a data set with unknown characteristics is that it is hard to pre-select adequate parameters in advance for clustering threshold or target number of clusters. An indication of a good clustering result is that the dis-

tances between vectors in the same cluster are low, the distances between vectors in different clusters are high, and there is a clear separation between inter-cluster and intra-cluster distance distributions.

The important parameters for clustering are: (1) The distance measure between the vectors, (2) the metric to calculate the distance between clusters, and (3) the clustering threshold (distance or cluster number). As discussed in last section, we use the Manhattan distance between association patterns. Regarding to calculating distance between clusters, there are also various candidate methods. In this section we adopt two different widely-used methods for cluster distance calculation: Distance between the furthest vectors in the two clusters (known as complete-link algorithm) and average distances between all vector pairs between the two clusters.

## 4.2 Multi-modal Behavior of WLAN Users

We apply clustering techniques to individual user association patterns. If the association patterns of a user can be arranged into several clusters, we say that this user displays multi-modal behavior. In other words, the user's association patterns have several distinct modes, and he chooses from one of the modes to follow on a given day. Contrarily, if there is only one cluster in the association patterns for a user, this user is referred to as a single modal user.

We apply two different ways to calculate inter-cluster distance (complete-link and average distance) and use various clustering threshold distances. We show that regardless of the clustering threshold chosen, the association patterns for most users fall into multiple clusters.

We first apply the hierarchical clustering scheme to a sample user. The result is shown in Fig. 1, in which we draw the number of clusters obtained from his association patterns with respect to the clustering threshold, under both methods of cluster distance calculation. Certainly, as the clustering threshold increases, the resulting number of clusters decreases. However, notice that the association patterns for this particular user falls into at least 2 clusters until the clustering threshold is 1.4, which is quite high given that the distance between any possible association patterns is at most 2.0 since we choose Manhattan distance and the entries in each pattern must sum up to 1.0. This leads to a strong argument that this particular user shows at least two drastically different association mode: The average distance between the patterns in the last two clusters is at least 1.4.

Now we show the distribution of number of clusters obtained from the clustering analysis for all 5,000 users with clustering threshold 0.9 in Fig. 2. The exact num-

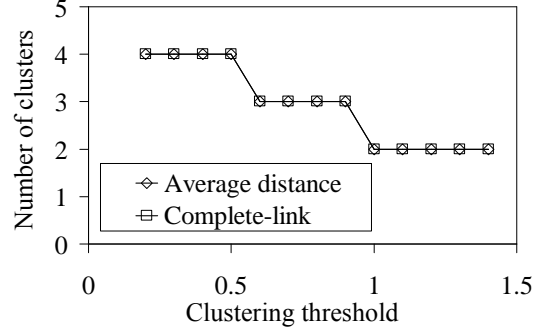


Figure 1: Number of clusters obtained by various distance measures and clustering thresholds. The user displays multi-modal behavior in most cases.

ber of clusters obtained depends on the distance calculation methods. The average distance method leads to more aggregation of clusters as compared to complete-link method. With 0.9 as the distance threshold, the patterns when the user is offline (i.e., the zero vectors) are separated from the patterns when the user is online (i.e., the vectors with entries sum up to 1), so there should be at least two clusters in the association patterns. Users with two clusters can be referred to as "on-off" users with consistent association pattern: One cluster corresponds to the patterns when the user is offline (the zero vectors), and the other one corresponds to the patterns when the user is online. These users switch between online and offline behaviors from day to day, and when it is online, the association patterns are consistent and fall in a single cluster. However, for both distance calculation methods, we also observe many multi-modal users. These users show more than two clusters, which indicates that their association patterns fall into different clusters, or association modes, when they are online.

If we consider users with more than two clusters (i.e., users with more than one association mode when it is online) as multi-modal users, the ratio of multi-modal users (out of total of 5,000) can be considered as a function of clustering threshold, and we obtain it for both distance calculation methods in Fig 3. Here we see that independent of the distance calculation method, a non-negligible portion of users are classified as multi-modal even if high clustering threshold (those above 1.0) is used. It implies that many users have at least two clusters in their association patterns when they are online. In other words, they have more than one modes of association to different set of locations when they come online.

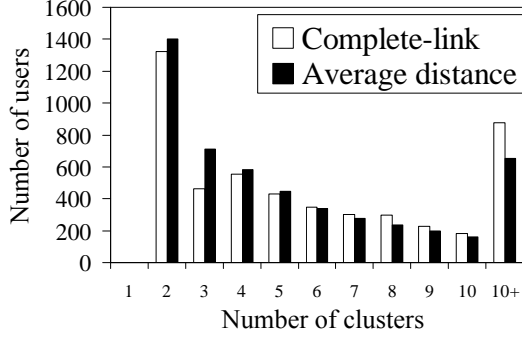


Figure 2: Distribution of number of clusters with clustering threshold 0.9 and various distance calculation methods.

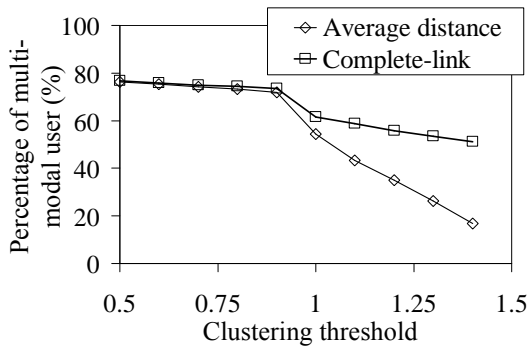


Figure 3: Percentage of multi-modal user under various clustering thresholds and distance calculation methods.

## 5 Summarizing Trends in User Association Matrix

After validating that most users display multi-modal association pattern in previous section, we move on to the second question we raised for the paper, which is designing a succinct way to express the major trend of user association patterns that dominates during the trace period. As an analogy, if one is asked to give a summary of the locations he usually stays at on campus, a sample response might be: "I usually spend eight hours in my office each day. Once a week, we have a long group meeting so I am in the meeting room for that afternoon. I go to the gym two to three times per week for an hour. I also visit the engineering library from time to time for short periods. I rarely visit other places besides the above." Note that in such summary, the desirable order of presentation is to tell the components that describe most of the visit pattern (i.e., the long duration spent in the office) first, and then explain different deviation from that in decreasing degree of importance.

We wish to have a mechanism that provides an insightful but concise summary of user association patterns similar to the example, such that it captures the major trends together with how much weight the major

trends carry as compared to the variations in the association patterns. In this section, we seek a procedure to generate a summary for the association pattern set for the user using a small number of descriptive vectors, with a quantitative measure of the importance for each vector, without the need of fine-tuning parameters for each user.

To represent the major trend of user association concisely, one intuitive candidate would be the average of the user's daily association patterns. Indeed, taking average would be suitable if a user shows only single mode in his daily association patterns (i.e. Having only one cluster in the analysis in the last section). However, as this is not the case for most of the users, using only the average could be sometimes misleading, as it shows an average association pattern that falls in between the common patterns for the user, and it is not at all a representative vector for the user.

In this section we present a novel way to summarize the user association patterns. Instead of taking the average, we arrange user association patterns in an *association matrix*, and we perform singular value decomposition to the matrix. Interestingly, although there exists multi-modal behavior for most of the users, the intrinsic dimensionality for these association matrices is actually low. That implies, with only a few eigenvectors and its corresponding eigenvalues, we can fully summarize the association matrices with low reconstruction errors. We will introduce our definition of *association matrix* for the users and the necessary background of singular value decomposition in subsection 5.1 and our finding by applying this technique to users in the traces in subsection 5.2.

### 5.1 Association Matrix and Singular Value Decomposition

For a detailed description of users for the studied period, all of his daily association patterns for each day during this period are necessary. In this paper, we construct the *association matrix* for a user by concatenating his daily association patterns in a single matrix. In the *association matrix*, each column corresponds to a daily association pattern within the studied period, and each row corresponds to the ratios of online time the user associates with a given location for all the days. If there are  $n$  distinct locations and the trace period is  $d$  days, the *association matrix* for a single user is an  $n$ -by- $d$  matrix. A zero column vector corresponds to a day in which the user is never online.

From linear algebra [4], we know that for any  $n$ -by- $d$  matrix  $X$ , it is possible to perform singular value decomposition, such that:

$$X = U \cdot \Sigma \cdot V^T \quad (2)$$

where  $U$  is an  $n$  by  $n$  matrix,  $\Sigma$  is an  $n$  by  $d$  matrix with  $r$  non-zero entries on its main diagonal, and  $V^T$  is an  $d$  by  $d$  matrix where the superscript  $T$  in  $V^T$  indicates the transpose operation to matrix  $V$ .  $r$  is the rank of the original association matrix  $X$ .

From the properties of singular value decomposition (SVD) [4], we know that the columns of matrix  $V$  are the eigenvectors of the covariance matrix  $X^T X$ , and  $\Sigma$  is a diagonal matrix with the corresponding singular values to the eigenvectors on its diagonal, denoted as  $\sigma_1, \sigma_2, \dots, \sigma_r$ . These singular values on the main diagonal of  $\Sigma$  are ordered by their values (i.e.  $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_r$ ), and they are also the square-roots of the eigenvalues of matrix  $X^T X$ . We denote the eigenvalues as  $\lambda_1, \lambda_2, \dots, \lambda_r$ . The eigenvalues are the measure of power captured by its corresponding eigenvectors, in columns of matrix  $V$ .

We can re-write Eq. (2) by taking column vectors,  $u_i$  and  $v_i$ , from matrices  $U$  and  $V$ , and use them as building blocks to reconstruct the original matrix,  $X$ , with the following relationship (since  $\Sigma$  is a diagonal matrix):

$$\tilde{X}_k = \sum_{i=1}^k u_i \sigma_i v_i^T \quad (3)$$

This yields a rank- $k$  approximation to the original matrix  $X$ . If the intrinsic dimensionality of the original matrix  $X$  is low, then by applying SVD to the matrix and storing the most important eigenvectors and singular values (e.g.  $u_i$ 's and  $v_i$ 's correspond to large  $\sigma_i$ 's), a significant less amount of storage is required as compared to storing the original matrix  $X$ . The percentage of power in the original matrix  $X$  captured in the rank- $k$  reconstruction in Eq. (3) can be calculated by

$$\frac{\sum_{i=1}^k \sigma_i^2}{\sum_{i=1}^r \sigma_i^2} \quad (4)$$

where  $r$  is the rank of the original matrix  $X$ . If the percentage of power captured in the rank- $k$  reconstruction is large (i.e., close to 1) with small value of  $k$ , we say that the original matrix  $X$  has a small intrinsic dimensionality. In some cases, many non-zero singular values and the corresponding eigenvectors are not important for reconstruction of the original matrix because the relative weight for the component is low. Following Eq. (4), the relative weight (or the importance) of an eigenvector  $v_j$  is represented by  $\sigma_j^2 / \sum_{i=1}^r \sigma_i^2$ .

SVD can be viewed as calculating the eigenvalues and eigenvectors of the covariance matrix,  $X^T X$ . This is the procedure typically used to perform Principal Component Analysis (PCA) for matrix  $X$ . In our case, we do not remove the mean of each dimension (i.e. each row) before performing the SVD. Such a technique is known as *uncentered PCA*, which is applicable if the origin of the data set is an important point of reference [2]. In

our case, we want to capture how the average trend in the association patterns (captured in the first principal components (PC) if uncentered PCA is applied) relates to the variations (captured in PCs other than the first component), so we choose not to perform SVD on a zero mean centered matrix. Furthermore, by adopting uncentered PCA, the reference point (origin) remains the same for all the association matrices, and it enables us to compare PCs obtained from different association matrices. This point will be utilized further in the next section. Using our notation, the principal components (PCs) are the column vectors,  $v_1, v_2, \dots, v_r$ , in the matrix  $V$ , and the corresponding eigenvalues are the squares of singular values, i.e.  $\lambda_i = \sigma_i^2$ . The PCs show the trends in the user's association patterns in decreasing importance, with the first PC corresponding to the average association pattern and the other PCs corresponding to variation around the average.

## 5.2 Low Dimensionality of Association Matrices

Following the procedure outlined in the previous subsection, we create the *association matrices* for the 5,000 chosen users and apply singular value decomposition to them. In this section we explain the observations we make from the results of SVD.

From the SVD results, the first property we look into is whether the *association matrices* can be decomposed into a small number of representative eigenvectors. In other words, do these *association matrices* have low intrinsic dimensionality? Low dimensionality will correspond to few intrinsic modes of association and will also help us store the association pattern in a compact space. From Eq. (3), we see that if the matrices have low dimensionality, the original matrices can be summarized with only the vectors corresponding to high singular values. Discarding the vectors corresponding to low singular values would lead to only marginal reconstruction error for the original matrices.

Interestingly, although in the last section we show that most users display multi-modal behavior, from the results of SVD we find the major trend in association patterns can be captured in a few eigenvectors for most users, and the dimensionality for the *association matrices* are actually low. To visualize this, we calculate the percentage of power in the association matrices captured by the rank- $k$  reconstruction using Eq. (4). In Fig. 4, we show the ratio of users that a certain percentage of power in its association matrix can be captured in the rank- $k$  reconstruction. From the graph we observe most users have high percentage of power in its association matrices captured by low-rank reconstructions. For example, if we use rank-1 reconstruction, it captures 50% or more of power in the association matri-

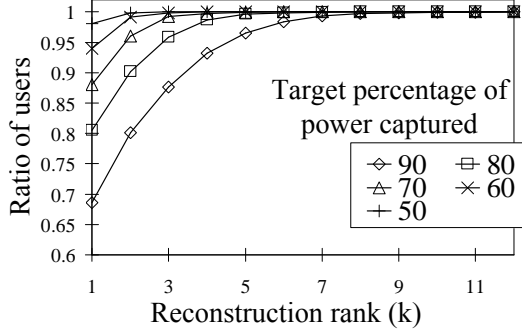


Figure 4: Low association matrices dimensionality: A high target percentage of power is captured with low reconstruction matrix rank for many users.

ces for more than 98% of users, and rank-3 reconstruction is sufficient to capture more than 50% of power in association matrices for all users. Even if we consider an extreme requirement, capturing 90% of power, it is achievable for 68% of users using the rank-1 reconstruction, and for more than 99% of users using at most rank-7 reconstruction. Comparing the findings we have by the matrix decomposition approach to the clustering approach in the last section, we could conclude that although some of the users show multi-modal association patterns, for most users the top eigenvectors are relatively much more important than the remaining ones. That implies, the eigenvectors with high corresponding eigenvalues obtained from SVD capture the important trend in the association matrices. In other words, although there are several clusters in user association patterns, only a few are important.

Since a few important eigenvectors capture most of the power in the association matrices, a good reconstructed version of the original association matrices should be built with a few eigenvectors and eigenvalues. We verify the goodness of low-rank reconstruction by calculating the matrix reconstruction error using best rank- $k$  reconstruction,  $\tilde{X}_k$ , for association matrix  $X$ . We define the L- $p$  norm of the relative error as:

$$\frac{\|X - \tilde{X}_k\|_p}{\|X\|_p} \quad (5)$$

where  $\|X\|_p$  is the L- $p$  norm for matrix  $X$ , defined as

$$\|X\|_p = \sqrt[p]{\sum_{\forall(i,j)} |X_{(i,j)}|^p} \quad (6)$$

where  $X_{(i,j)}$  is the entry at  $i$ -th row and  $j$ -th column of matrix  $X$ . L- $p$  norm of relative error is a standard way to quantify relative errors in matrices [3]. In this paper, we calculate the cases for  $p = 1, 2$  with Eq. (5) (i.e., The L1 and L2 norm of relative error) to capture the reconstruction error with low-rank matrices. We show the average L1 and L2 norm of relative error for all the

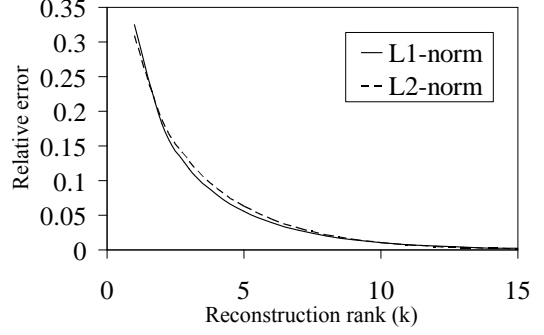


Figure 5: L1-norm and L2-norm of relative error with rank- $k$  reconstruction.

users in Fig. 5. We can see from the figure that the reconstruction error is also small under these standard metrics. Using 5 eigenvectors, it is possible to achieve relative error around 0.05.

The relationship between multi-modal association pattern sets and low dimensionality of association matrices can be investigated further. We show the percentage of power captured by the first eigenvector of association matrix with respect to the user's ranking by its *activeness* (i.e., the total online time for the trace period). In Fig. 6 (a), we observe that there is an interesting relationship between the user activeness ranking and the percentage of power in its association matrix captured by the first eigenvector. Comparing with the online time fraction for users (the total online time divided by the trace period) shown in Fig. 6 (b), we see that for the most active users who are almost online all the time, the power captured by the first eigenvector alone is relatively high. This is due to the fact that most of them are stationary users using WLAN as a substitute for the wired network. Hence, these users do not have much variation in its association pattern, and the first eigenvector is a good descriptor for their association matrices. As the activity of user decreases, those users who use the WLAN sporadically are more dynamic in their association pattern, and hence the power captured in the first eigenvector is lower.

In Fig. 6 (c) we show the number of clusters found from the association patterns of users using average distances between cluster with clustering threshold 0.9. We observe that the less active users also tend to have more clusters in their association patterns than the highly active users. These conclusions can be linked together: For the current WLAN users on USC campus, the highly active users mostly remain stationary at a single location so the dimensionality for their association matrices are very low, and the association patterns tend to form fewer clusters. Meanwhile, the users with higher variation in the association patterns are those who are less active, and their association patterns form more clusters. However, the bottom line is, for almost all users,



the intrinsic dimensionality of their association matrices is much lower than the columns and rows of the matrices.

## 6 User Clustering with Similar Association Pattern

After understanding the composition of individual user association patterns and its characteristics, we further ask the following question: How do we quantify users as having similar or dissimilar association patterns? Can we define a good metric to describe the similarity between user association patterns, and utilize such a metric to partition the whole user population into clusters with similar users per cluster?

In this section we first propose two different approaches to quantify similarity between users: In the first, *straight-forward approach*, we directly perform complete pair-wise comparison of association patterns between two users to evaluate their similarity. This approach, albeit intuitive, requires significant computation to evaluate similarity between users. Targeting to reduce the computation requirement, we come up with the second, *feature-based approach*, in which we only compare some carefully selected features that capture the essence of user association patterns, and hopefully the computation will be reduced if the selected feature has a smaller cardinality than the original association pattern set. From the results in previous section, using the eigenvectors and the eigenvalues would be a good candidate.

Based on the two different approaches above, we could define distances between users based on their similarity. Using these distances, we can group similar users into clusters. We further verify that the proposed distances and clustering techniques are appropriate, as the inter-cluster and intra-cluster distance distributions show a clear separation. We also show that the users in the same cluster indeed have similar association patterns by verifying the dimensionality of the *joint association matrices* of these clusters.

### 6.1 Similarity Measure of Association Pattern Sets

#### 6.1.1 The Straight-forward Approach

We first introduce the straight-forward approach to compare similarity between two sets of association patterns (each set of association patterns belongs to a single user). Intuitively, the sets are similar if they satisfy the following property: For each pattern from a set, we can find a pattern in the other set while the distance between these two patterns from different sets is small. To quantify this intuitive definition, we propose the *average*

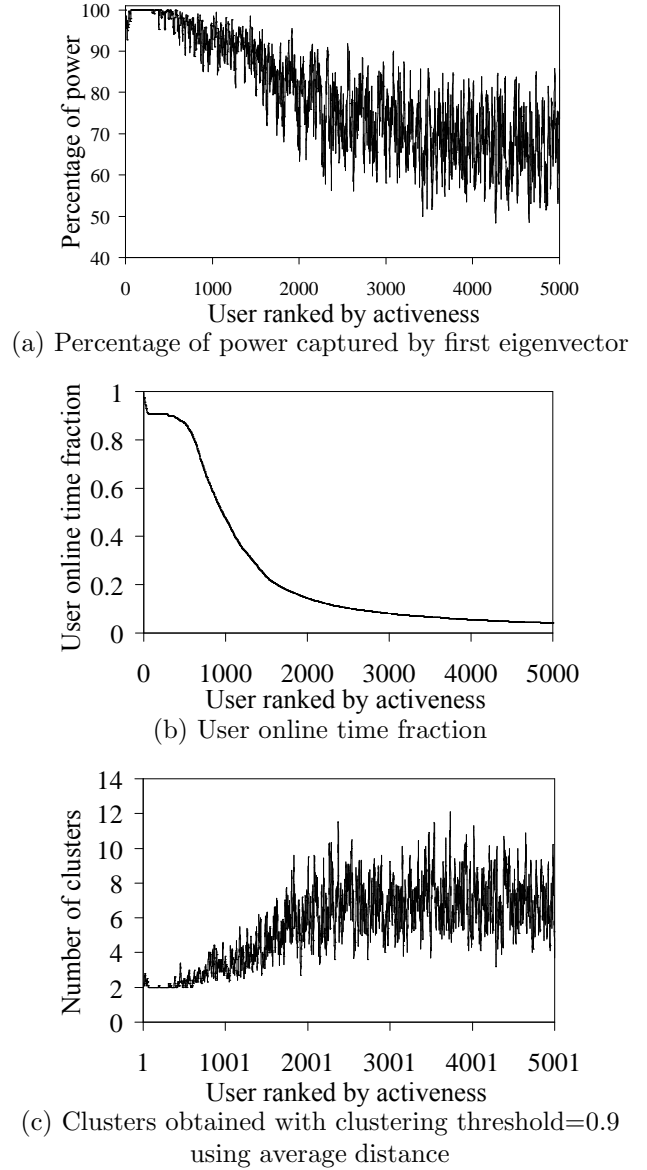


Figure 6: Metrics of users ranked by online time fraction. The curves show averaged value for 10 close-by points for better visualization

*minimum pattern distance (AMPD)* between sets  $A$  and  $B$  of association patterns,  $AMPD(A, B)$ , calculated by the following:

$$AMPD(A, B) = \frac{1}{|A|} \sum_{\forall a \in A} \arg \min_{\forall b \in B} d(a, b) \quad (7)$$

where  $a$  and  $b$  denote a single association pattern in set  $A$  and  $B$ , respectively.  $d(a, b)$  denotes the Manhattan distance between the patterns as defined in Eq. (1).  $|A|$  denotes the cardinality of set  $A$ . The *average minimum pattern distance* between sets  $A$  and  $B$  is the average of distances from each of the patterns in set  $A$  to the closest pattern in set  $B$ . If this distance is small, then these two association pattern sets are similar to each other. Note that, with this definition,  $AMPD(A, B)$  is not necessarily equal to  $AMPD(B, A)$ . To obtain a symmetric distance measure between association pattern sets  $A$  and  $B$ , we further normalize the *average minimum pattern distances* from set  $A$  to all other sets between  $(0, 1)^3$ , and define the straight-forward distance between set  $A$  and  $B$  as  $D(A, B) = (AMPD(A, B) + AMPD(B, A))/2$ . Now we have  $D(A, B) = D(B, A)$  for any given sets  $A$  and  $B$ . We will utilize this distance measure to perform cluster analysis of users in section 6.2.

### 6.1.2 The Feature-based Approach

Identifying similar users by comparing the complete set of association patterns is computationally very expensive. Hence, it is desirable to have a computationally simpler measure for association pattern set similarity. From last section we know that using the eigenvectors and eigenvalues is a good way to summarize the association matrices, so we propose to leverage them as the features based on which we compare the similarity between two association pattern sets. Since the eigenvectors are unit length vectors and orthogonal to each other, the problem of comparing the eigenvectors between users is equivalent to comparing the similarity between two sets of orthonormal vectors, and each of these vectors is associated with some weight to indicate its relative importance in its set. To carry out such comparison, we propose the following procedure.

Suppose  $u_i$ 's and  $v_j$ 's are eigenvectors of two users,  $i = 1, \dots, r_u$  and  $j = 1, \dots, r_v$  where  $r_u$  and  $r_v$  are the ranks of the corresponding association matrices. The similarity of any pair of unit length vectors among these two sets can be obtained by the absolute value of their inner product,  $|u_i \cdot v_j|$ . This is equivalent to calculating  $|\cos \theta|$ , where  $\theta$  is the angle between vectors  $u_i$  and  $v_j$ . The similarity of the two sets can be calculated by the sum of pair-wise inner products of individual vectors

$u_i$ 's and  $v_j$ 's, adjusted by the weights of  $u_i$  and  $v_j$ . We use  $w_{u_i}$  to represent the weight of eigenvector  $u_i$  in its set, calculated by  $w_{u_i} = \sigma_i^2 / \sum_{k=1}^{r_u} \sigma_k^2$ . The weights  $w_{u_i}$ 's sum up to 1, and  $w_{v_j}$ 's are defined similarly. After considering the weights, we propose the similarity index between two sets of eigenvectors,  $U = \{u_1, \dots, u_{r_u}\}$  and  $V = \{v_1, \dots, v_{r_v}\}$ , as:

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

Higher similarity index  $Sim(U, V)$  indicates that the eigenvector sets  $U$  and  $V$  are more similar, and hence the corresponding users have similar association patterns. Since in most cases, only a few eigenvectors capture most of the power in the association matrices, the above equation can be reduced to comparing only these important eigenvectors, in order to simplify the calculation. In the following computations, we consider only the eigenvectors that capture at least 0.1% of total power.

Now we define the distance between association pattern sets of users  $U$  and  $V$  based on the similarity index between their eigenvector sets. We normalize the similarity indexes from set  $U$  to all other sets between  $(0, 1)^4$ , and define the feature-based distance between set  $U$  and  $V$  as  $D'(U, V) = 1 - (Sim(U, V) + Sim(V, U))/2$ . Note that the similarity indexes are higher for similar users, so we need to subtract the similarity indexes from 1 to get a distance measure, in which larger value implies the two users are further separated.

We need to verify that the two distance measures, straight-forward distance and feature-based distance, are correlated to each other. For this purpose we calculate the correlation coefficient between the two distance measures for all the possible user pairs among the 5,000 chosen users. The correlation coefficient is very high with the value 0.9119. The strong correlation indicates that the distance measures generated by the two methods are consistent with each other. In other words, the feature-based distance is a valid substitute for straight-forward distance between user association patterns, with much less computation required.

## 6.2 Clustering Users based on the Similarity Measures

Now we apply the clustering techniques introduced in section 4 to the users using the distance metric developed in the previous subsection. In this subsection we discuss the results of clustering and show that we can obtain meaningful clusters using both straight-forward distance and feature-based distance.

<sup>3</sup>Among all sets, we find the set  $X$  such that  $AMPD(A, X) = \max_{\forall N} AMPD(A, N)$ . We then normalize  $AMPD(A, B) = AMPD(A, B)/AMPD(A, X)$  for all sets  $B$ .

<sup>4</sup>Among all sets, we find the set  $X$  such that  $Sim(U, X) = \max_{\forall N} Sim(U, N)$ . We then normalize  $Sim(U, V) = Sim(U, V)/Sim(U, X)$  for all sets  $V$ .

A good clustering solution must be robust to the choice of clustering parameters. One way to justify the choices of clustering threshold is to compare the distributions of inter-cluster distance and intra-cluster distance. If the clustering of users is a meaningful one, then the users fall into the same cluster should have much smaller distances between each other, as compared to users in different clusters. We generate the distance distributions for both categories. If there is clear separation between the two distributions, then the clustering is meaningful.

As a case study, we present the clustering of users using the feature-based distance,  $D'(U, V)$ . We consider average distance between clusters when we merge clusters in hierarchical clustering, and we proceed until 5,000 users are merged into 200 clusters. In Fig. 7 (a), we show the pdf's for inter-cluster and intra-cluster distances. We see from the figure that the peaks of these two pdf's are well separated (i.e., The pdf of intra-cluster distance is almost zero for distance larger than 0.5, and the pdf of inter-cluster distance is almost zero for distance before 0.9). We show the cdf's for the same distributions in Fig. 7 (b). From the figure we observe that more than 90% of intra-cluster user pairs have distance smaller than 0.5, and more than 90% of inter-cluster user pairs have distance larger than 0.9. We could use a cut-off threshold around 0.6 to 0.8 to separate most intra-cluster user pairs from inter-cluster user pairs. The separation between the two distributions is clear, hence we have a meaningful result of clustering. Once we find out the cut-off threshold, it can also be applied to evaluate the similarity between two users - If their distance is lower than the threshold, they should be considered similar and belong to the same cluster. The users hence are able to directly judge if they belong to the same cluster without the global knowledge of all user distances. We have to emphasize once again the actual number of clusters obtained from the data set is data-dependent, and currently there is no standard method to tell how many clusters exist in the data set beforehand. However, with our proposed distance measure, well-separated clusters can be obtained from the data set. We also try using the other straight-forward distance,  $D(A, B)$ , and get similar results as those shown in Fig. 7.

We need to further verify that indeed we have grouped similar users together in those clusters. The verification process we propose is the following: We compose the *joint association matrix* by concatenating the daily association patterns of a cluster of  $m$  similar users in a larger  $n$ -by- $md$  matrix, where  $n$  is the number of locations and  $d$  is the number of days. The percentage of power captured by the top eigenvectors of this *joint association matrix* should be high, as the association patterns in the matrix follow similar trends. On the

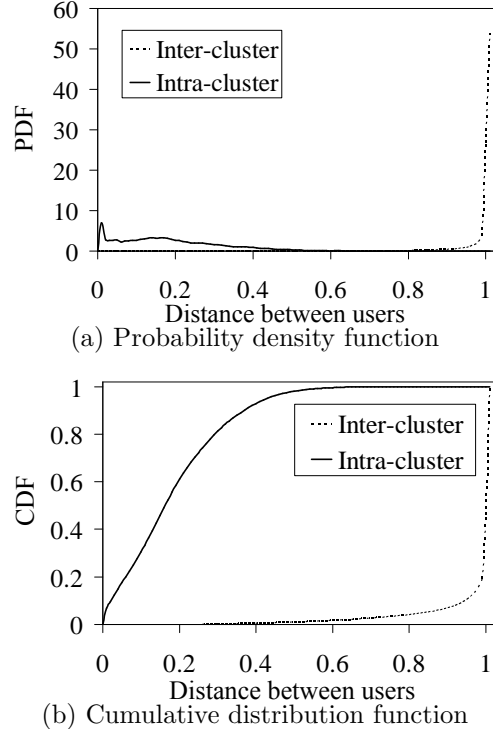
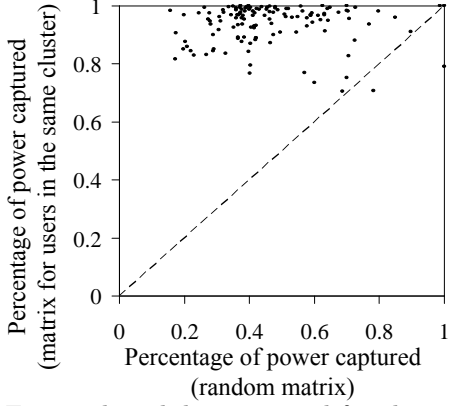


Figure 7: Distributions of distances for inter-cluster and intra-cluster user pairs

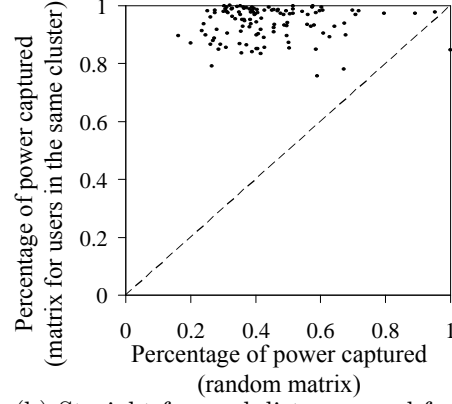
other hand, if association patterns of users with different trends are put in one *joint association matrix*, the percentage power captured by its top eigenvectors should be much lower. We carry out this verification process first for some sample clusters. We form three different *joint association matrices*: The first joint association matrix contains the users in one of the clusters formed with the feature-based distance. There are totally 20 users in this cluster. The second matrix contains the users in one of the clusters formed with the straight-forward distance. There are totally 37 users in this group. For the third matrix we randomly pick 20 users from the whole population and put their association patterns in the matrix.

The ratio of cumulative power captured by the top eigenvectors for these three matrices is shown in Fig. 8. From the figure we see clearly that both clusters formed with the proposed distance measures in last subsection lead to *joint association matrices* with low dimensionality. The power captured by the first eigenvector is above 75% for both matrices. On the other hand, although the matrix with random users has the same size as the first matrix, its dimensionality is much higher, as more than 9 eigenvectors are required to capture more than 75% of its power. Clearly, the users in clusters identified by the proposed distance measures are indeed similar in their association trends.

We further carry out a large-scale verification of the



(a) Feature-based distance used for clustering



(b) Straight-forward distance used for clustering

Figure 9: Scatter graph: Cumulative power captured in top four eigenvectors of random joint association matrices (X) and matrices formed by users in the same cluster (Y).

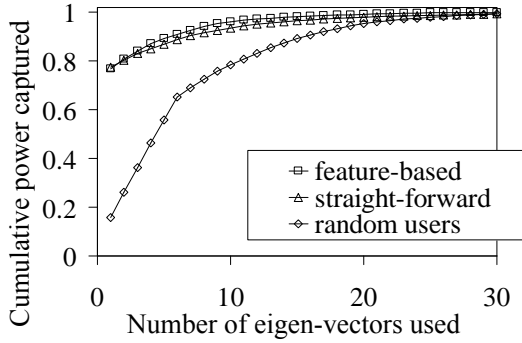


Figure 8: Cumulative ratio of power captured by the top eigenvectors of joint association matrices for sample clusters under various user clustering methods.

dimensionality of all clusters. Among the 200 clusters formed, we pick those with more than five users in the cluster, and check whether these clusters indeed contain users with similar association patterns. There are totally 130 such clusters when the feature-based distance is used, and 117 such clusters when straight-forward distance is used. We form the joint association matrices for users in these clusters. At the same time, we also form random matrices consist of the same number of randomly picked users. We compare dimensionality of matrices formed with users in the clusters to the matrices formed with random users by plotting the cumulative power captured in the top four eigenvectors of the matrices in scatter graphs, Fig. 9. In these graphs, each dot represent the cumulative ratio of power captured by the top four eigenvectors of the joint association matrix for users in a cluster as its Y-coordinate, and the ratio of cumulative power captured by the top four eigenvectors of the corresponding random matrix is its X-coordinate. Clearly, most the dots are above the 45-degree line regardless of either feature-based distance (Fig. 9 (a)) or straight-forward distance (Fig. 9 (b)) is used, indicat-

ing that similar users can be found with our definition of distances between users.

To sum up, in this section we propose two metrics for distance between users in terms of their similarity of association pattern. Such metrics can be utilized to form well-separated clusters in the user population. This further implies characteristic of association pattern is a distinguishing feature of users - they can be classified into groups with different association pattern feature. Such information can be useful for the network operator in several ways, as we discuss more in the following section.

## 7 Discussion

The eigen-decomposition based representation of user association patterns can help the network administrator in several ways. First, due to the low dimensionality of user association matrices, the network operator can efficiently store the major trends in user association patterns. By checking whether the new association patterns generated by the same user can be represented as a linear combination of the eigenvectors obtained from its previous association matrix, the network administrator could determine whether the user has changed its association behavior. If a user with consistent previous association behavior suddenly changes its association behavior, it could be viewed as an abnormal activity: either it is a significant change in its usage pattern, or a potential ongoing impersonation attack. Depending on the policy, the administrator may want to look into the actual reason for the change. Second, if the administrator wants to deploy some personalized services in the network (e.g., storing user's email and files on separate machines close to the user's frequently visited locations), the eigenvectors can be utilized to determine where to store the user's data. The clustering of users

based on similarity of association patterns can further help the network operator to enumerate users with different association patterns, and find out which type of association trend dominates its network, and plan accordingly. Thirdly, although not discussed in detail in this paper, it is also possible for the network administrator to incorporate the single-day association patterns from all its users into a matrix, and apply the same singular value decomposition based grouping technique to observe the trends of all users on campus for the given day, and perhaps create some daily norm (i.e., typical behavior on Monday, Tuesday, etc.) for the network. Creating such reference data would help the administrator to understand its user and network better.

The summary of association patterns may also benefit the users. For example, the user can profile herself based on its association pattern, and determine whether other users are similar to her in this aspect. As shown in section 6, the users have a simpler way to express, exchange, and compare their association patterns using the eigenvectors. In particular, the insights developed in our study are essential in efficient protocol design. Identifying users with similar association or movement patterns is useful, for example, in making better forwarding decisions to deliver packets in an infrastructure-less network. In [21], the authors proposed a mechanism in which packets are delivered towards the destination node by forwarding to nodes with increasing similarity in movement patterns to the final destination. Our proposal of summarizing user association pattern with the eigenvectors and the similarity index proposed in Eq. (8) provides a good way to serve this purpose. In [20], the authors proposed to spread multiple copies of a packet to several intermediate nodes to expedite the delivery. This multi-copy strategy is better suitable when the sender does not know the exact association pattern of the destination. In this scheme, it should be better to spread the copies to nodes with different association pattern trends to improve the probability that one of these copy-carriers meets with the destination node. Again, our proposal is a good candidate method to determine similarity between nodes in this case.

## 8 Conclusion and Future Work

In this paper, we systematically unearth the underlying structure in user's daily association patterns. Through analysis of WLAN traces of 5,000 users collected at USC campus, we are able to identify the overall association trends in WLAN users in a university campus network, and answer the three questions proposed in the introduction with the following:

1. Most WLAN users show multi-modal association behaviors. The vectors representing their daily as-

sociation behavior fall into separate clusters. For each user there are at least two clusters, one corresponds to the days it is offline, and the other corresponds to the days it is online. Furthermore, for many users the association patterns for online days fall into multiple clusters, suggesting multi-modal association patterns exist.

2. Regardless of the multi-modal association patterns, the association matrix for the user usually has low dimensionality. This property suggests that usually one or only few association patterns dominate the behavior of the user. Due to the low dimensionality of association matrix, it is possible to summarize the matrix with its eigenvectors and eigenvalues. Such representation summarizes the association patterns of the user with eigenvectors in decreasing importance.
3. We propose the metrics for distances between the association pattern sets of the users and the similarity index between the eigenvector sets. These metrics are utilized to identify users with similar association trends. The feature-based distance calculated from eigenvectors leads to significant amount of saving in calculation due to the low dimensionality of most user's association matrices.

We believe that our methodical approach to mining the WLAN user-association patterns along with these findings pave a new way to understand the structure in user association patterns in WLANs. Although we have chosen the USC trace to illustrate the usefulness of the methodology and metrics, these methods are not limited by the choice of data set, and can be applied to study other traces available to the research community. We intend to apply the techniques to other traces in the research community (e.g., [7, 8]) as the next step. We also plan to leverage the understanding of the WLAN user behavior in designing better mobility models and ad hoc routing protocols.

Finally, this paper is a small first step to the systematic mining of information from WLAN traces, especially user association patterns. We asked basic questions in mining and to answer those, we robustly motivated, defined and validated the choice of our similarity metrics, which is the key to most mining tasks. Investigating better metrics is an ongoing work. Also we are working on leveraging more sophisticated clustering algorithms and statistical tools for validating our metrics.

## 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] G. Strang, *Linear Algebra and Its Applications*, third ed., Brooks Cole, published 1988.
- [5] 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.
- [6] W. Hsu and A. Helmy, "On Nodal Encounter Patterns in Wireless LAN Traces," the Second International Workshop On Wireless Network Measurement (WiNMee 2006), April 2006.
- [7] MobiLib: Community-wide Library of Mobility and Wireless Networks Measurements (Investigating User Behavior in Wireless Environments). USC WLAN trace and pointers to many WLAN trace archives available at <http://nile.usc.edu/MobiLib>.
- [8] CRAWDAD: A Community Resource for Archiving Wireless Data At Dartmouth. Many archived WLAN traces available at <http://crawdad.cs.dartmouth.edu/index.php>.
- [9] 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.
- [10] 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.
- [11] M. Balazinska and P. Castro, "Characterizing Mobility and Network Usage in a Corporate Wireless Local-Area Network," In *Proceedings of MobiSys 2003*, May 2003.
- [12] 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.
- [13] 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.
- [14] M. Papadopoulou, H. Shen, and M. Spanakis, "Modeling Client Arrivals at Access Points in Wireless Campus-wide Networks," 14th IEEE Workshop on Local and Metropolitan Area Networks, Chania, Crete, Greece, September 2005.
- [15] C. J. F. ter Braak, "Principal Components Biplots and Alpha and Beta Diversity," *Ecology*, vol. 64, pp. 454-462, 1983.
- [16] W. J. Krzanowski, "Between-groups Comparison of Principal Components," *J. Amer. Statist. Assoc.*, vol. 74, pp. 703-707, 1979.
- [17] K. Yang and C. Shahabi, "A PCA-based Similarity Measure for Multivariate Time Series," *ACM International Workshop On Multimedia Databases*, November 2004.
- [18] C. Tudeuce and T. Gross, "A Mobility Model Based on WLAN Traces and its Validation," in *Proceedings of IEEE INFOCOM*, March 2005.
- [19] M. Papadopoulou, 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.
- [20] T. Spyropoulos, K. Psounis, and C. S. Raghavendra, "Spray and Wait: An Efficient Routing Scheme for Intermittently Connected Mobile Networks," workshop on delay tolerant networking and related networks (WDTN-05), August, 2005.
- [21] J. Leguay, T. Friedman, and V. Conan, "Evaluating Mobility Pattern Space Routing for DTNs," in *Proceedings of IEEE INFOCOM*, April, 2006.
- [22] N. Eagle, "Machine Perception and Learning of Complex Social Systems", Ph.D. Thesis, Program in Media Arts and Sciences, Massachusetts Institute of Technology, June, 2005.