Structures of User Association Patterns in WLAN

Wei-jen Hsu[1], Debojyoti Dutta[2], and Ahmed Helmy[1]
[1]Department of Electrical Engineering, [2]Department of Computational Biology, University of Southern California
Email: {weijenhs, ddutta, helmy}@usc.edu

I. Introduction

There has been a rapid increase in wireless LAN (WLAN) deployments, users and traffic in the recent years. Most of the current studies in WLAN access patterns have focused on usage statistics and user arrival patterns at access points (APs). However, little work has been done for fine-grained analysis of the structures and trends in long term user AP association patterns. Such detailed study 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. Understanding such detailed structure is especially important for large-scale WLAN with users re-visiting 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. 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 this paper, we choose to study the composition of the user's *daily* association patterns within the USC campus. Specifically, we focus on the following questions in this paper:

1) Do users show single-modal or multi-modal AP 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?

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. To illustrate the usage of our tools and methodologies, 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., [2], [3]). We represent the user association pattern for each day as an n-entry vector, $(a_1, a_2, ..., a_n)$, 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. We use a zero vector to represent the association pattern when the user is completely offline for the day.

II. Multi-modal user AP association

Given the daily association patterns of a user for the duration of a semester, we utilize clustering technique [1] to distinguish whether the user is single-modal or multi-modal in her association behavior. 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. We use Manhattan distance as the distance measure between the daily association pattern vectors. If multiple clusters can be identified from the association patterns, then the user under consideration is a multi-modal one.

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. There are only few users that remain consistent in her association pattern (i.e., single-modal user) for all the days we studied. Specifically, more than 50% of users are classified as multi-modal (i.e., having more than two clusters) under intermediate clustering threshold, as shown in Fig. 1 and very few users are uni-modal in nature.

III. Summarizing trends in user-AP association

We move on to design a succinct way to express the major trend of user association patterns that dominates during the trace period for each user. The proposed method generates 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. We concatenate user association patterns in columns of an *association matrix*, and perform singular value decomposition to the matrix. We observe that although there exists multi-modal behavior for most of the users, the intrinsic dimensionality for these association matrices is actually low, indicating that there exists *dominating* association behavior for most users. Only a few eigenvectors and corresponding eigenvalues are needed to capture a high target percentage of power in the association matrices, as

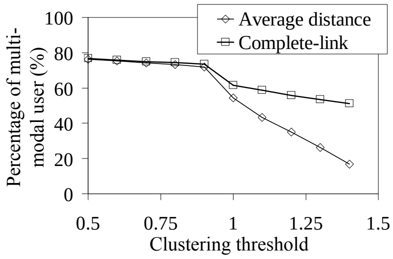

Fig. 1. Percentage of multi-modal user under various clustering thresholds and distance calculation methods.

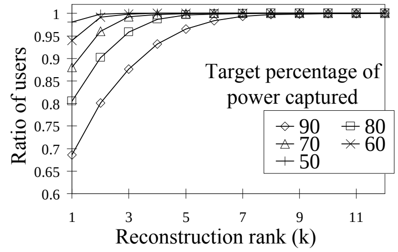

Fig. 2. Low association matrices dimensionality: A high target percentage of power is captured with low reconstruction matrix rank for many users.

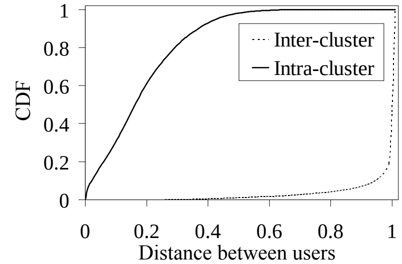

Fig. 3. CDF of distances for inter-cluster and intra-cluster user pairs

shown in Fig. 2. e.g. 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. That implies, with only a few eigenvectors and its corresponding eigenvalues, we can fully summarize the association matrices with low reconstruction errors. Only the top five eigenvalues and their corresponding eigenvectors can be used to reconstruct those association matrices with an average error of 5%, in terms of the L_1 and L_2 matrix norms.

IV. Similarity measures between users' trends

We further 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 distance metrics are intimately 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, as shown in Fig. 3 for the feature-based distance. Straight-forward distance shows similar results. There is a clear separation of the distributions of the inter and intra cluster distances.

We further verify that indeed we have grouped similar users together by composing the *joint association matrix*. We concatenate 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. Contrarily, 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. For the clusters identified with the similarity measure, we show that the cumulative power captured in the top four eigenvectors are much higher

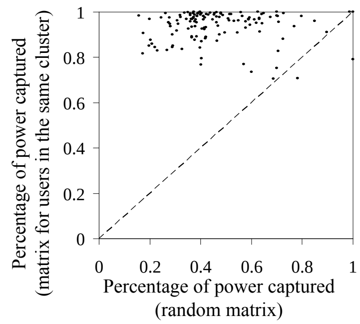

Fig. 4. Scatter graph: Cumulative power captured in top four eigenvectors of random joint association matrices (X) and matrices formed by users in the same cluster (Y). Feature-based distance is used for clustering.

than if the clusters are formed randomly in Fig. 4, indicating that the clusters contain users with similar association trends.

V. Discussion and Potential Usage

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, and check whether users have deviated from the stored major trends as a way for abnormality detection. Second, the eigenvectors can be utilized to help the administrator to provide personalized location-aware service based on users' association trends or make assessment about dominating user behaviors the network.

The summary of association patterns may also benefit the protocol designers and users. One can design a protocol for making packet forwarding decisions in an infrastructure-less network based on the similarity of user association patterns. Also, the user can profile herself based on the association pattern, and determine whether other users are similar to her. Such information may be useful in social context for finding potential friends with similar association pattern.

References

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

[2] MobiLib: Community-wide Library of Mobility and Wireless Networks Measurements. USC WLAN trace and pointers to many WLAN trace archives available at http://nile.usc.edu/MobiLib.

[3] CRAWDAD: A Community Resource for Archiving Wireless Data At Dartmouth. Many archived WLAN traces available at http://crawdad.cs.dartmouth.edu/index.php.