Principal Component Analysis of User Association Patterns in Wireless LAN Traces

Wei-jen Hsu and Ahmed Helmy

Department of Electrical Engineering, University of Southern California
Email: {weijenhs, helmy}@usc.edu

I. Introduction

Wireless networks have gained its popularity quickly in recent years. As the usage increases, there is also an increasing need to understand the characteristics of wireless users. Among all the properties to describe user behaviors in 802.11-based wireless LANs, their association patterns to access points (APs) play a very important and fundamental part. In this poster, we apply principal component analysis (PCA) to unearth the common pattern of user association in wireless networks.

The questions we seek to answer by applying the PCA technique are: (1) Are users similar to one another in their association pattern in long run? (2) Does individual user show consistent daily association pattern across multiple days? (3) If the answer to question 2 is yes, then how do we find some summarized presentation of the daily association pattern of a user? (4) Can users be grouped using the summarized presentation obtained in question 3, leading to groups that show similar association pattern?

Throughout the analyses we find that for university campuses, the whole user population is diverse enough that the major common trends of association, even if it may exist, is fairly insignificant. This observation is consistent from the traces we studied about generic users. However, if we focus on a user group in which the individual users have some common attributes, the common trends in association patterns become much stronger for the group. We further investigate the individual user association pattern across days and find most users show a clear consistent trend in its daily association patterns. The principal components (PCs) of the daily individual association data set can be used to characterize individual users and summarize their association behaviors.

Based on the principal components of individual user association matrices, we further propose a way to group users. We define the similarity index between two users by performing a weighted sum of inner products of the PCs from each of the user pairs. Using this definition, we provide a method to distinguish users that display similar association patterns from the others, and identify such users as a sub-group from the whole population. We show that by grouping users based on the similarity index, the members of sub-groups have much more significant common trends in the association patterns than randomly generated group, or the whole user population.

In this paper we use three WLAN traces collected from university campuses, including University of Southern California (USC) [2], Dartmouth College [4], and University of California at San Diego (UCSD) [3]. The USC and Dartmouth trace are collected from all types of wireless devices on campus. UCSD trace is from a specific project targeting at PDA users. We select to analyze the traces for one whole semester/quarter from the studied universities, including fall quarter 2002 for UCSD, spring quarter 2004 for Dartmouth, and summer semester 2005 for USC.

II. Matrix representation of user association patterns

To facilitate PCA, we define *group association matrix* and *individual association matrix*, for which we apply PCA to identify its underlying common patterns. The objective of the *group association matrix* is to find the similar structure in association patterns among multiple users over some period of time. Therefore, we choose to represent the association pattern of each user for this studied period in a single column vector. Each column consists of m entries, where m is the total number of APs in the corresponding trace. The value in each entry of the column vector is the total amount of time the MN associated with each AP during the time period.

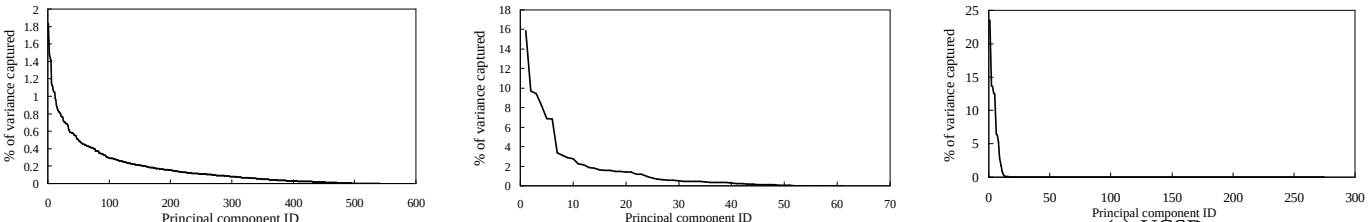
The objective of the *individual association matrix* is to find the characterizing daily association pattern of a single user. Therefore we choose to represent the association pattern of the user for *each single day* as a vector. In *individual association matrix*, each column is a $(m+1)$-entry vector. The first m entries are the amount of time the MN associated with each AP. The last, $(m+1)$-th entry, represents the total time the MN is not associated with any AP (i.e. in the offline state). We need to add the offline entry because in some days the MN can be completely offline, and a column with all 0's imposes an obstacle to perform singular value decomposition.

We apply singular value decomposition (SVD) to both *group association matrix* and *individual association matrix*. After performing SVD, we obtain the PCs and corresponding eigenvalues for the matrices. The relative importance of each PC in its set can be determined by the corresponding eigenvalue. The most important PCs (i.e., those with high weights) of *group association matrix* are unit-length vectors in m-dimension space that capture highest power (or strongest trend) of association patterns in the group. The most important PCs of *individual association matrix* are unit-length vectors in $(m+1)$-dimension space that capture strongest trend of daily association patterns for the user.

III. Principal Component Analysis of Association Matrices

The first property we look into is that whether the association trend of whole user population can be characterized by a few PCs. This check is done by observing the distribution of eigenvalues of the *group association matrix*, since they represent the total variation of original data set captured by the corresponding principal components (PCs). We show the percentages of variation in *group association matrix* captured by each PC in Fig. 1.

From Fig. 1 we observe for the traces that record the network activities of a diverse population (i.e. Dartmouth and USC), the variation in *group association matrix* is distributed across a large set of PCs. In both cases, about one third of the PCs are carrying some non-negligible variation (i.e., more than 10% of the variation captured in the most important PC) of the original data set. This is an indication that the columns of *group association matrix* do not vary



(a) Dartmouth (b) USC (c) UCSD

Fig. 1. Percentage of variation in group association matrices captured by each Principal Component (PC). The PCs are ordered in decreasing importance. Note the X-axes and Y-axes are in different scale in the graphs.

with just a few common pattern. By contrast, the observation from UCSD trace is different. From this data set, the variation captured in first few PCs are relatively high. The top-10 PCs capture 98.36% of total variation together. The association patterns of users in UCSD trace do show common trends. This phenomenon may be related to the experiment setup: The users are not randomly chosen, but are all freshmen in an anonymized college in UCSD [3]. Therefore we could expect some commonality in their association patterns, as validated by the result of PCA.

In light of the different results of PCA observed from the *group association matrices*, we further ask the following questions: How do we identify "sub-groups" among the diverse user population, such that users in each sub-group display common trend in their association pattern? Again, PCA provides a useful tool to serve the purpose.

We propose the *individual association matrix* for each MN as a description of its daily association pattern, and use PCA to obtain the major trends of its variation. The questions to answer from this operation are: (1) Can we find a few important PCs to capture a single user's association pattern? (2) How can we utilize these PCs to group them? In this section we answer questions (1) and defer the discussion of question (2) to the next section.

We apply the same PCA technique to test the dimensionality of individual association matrices. For each individual association matrix, we determine the number of PCs required to capture a certain percentage of its variation. If the required number of PCs to capture a high-percentile of variation is small for most individual association matrices, we can claim that the dimensionality for individual association matrices are low, and in other words, individual users show similar association patterns day by day.

We perform PCA on individual association matrices of users in Dartmouth trace, and show the CDF of number of PCs needed to capture various percentage of variation in Fig. 2 (a). From the graph we observe that the dimensionality for individual association matrices are smaller than that of group association matrix. By using only 10% of PCs (i.e. the top-6 PCs out of 61), we could capture more than 70% of variation for more than 99% of MNs. Even if we consider a more extreme requirement, capturing 90% of variation, it can be done with top-6 PCs for more than 92% of users. For USC-05su and UCSD-02f traces, the dimensions of individual association matrices are even smaller; see Fig. 2 (b)(c).

Hence, the key distinction between PCA of group association matrix and individual association matrix is the following: Although the whole user population displays a diverse pattern of association to APs, the source of this diversity comes from the fact that users have different major trends in their association patterns. The association behavior for a single user, however, is quite consistent across days in most cases, as the variation in individual association matrices can be captured using few PCs.

IV. Similarity of principal component sets and grouping

In this section, we propose a way to quantify the similarity of PCs among users. Since PCs summarize the important dimensions of association for individual users, they provide a more efficient way to compare similarity between individual user association patterns. Principal components are of unit length and orthogonal to each other. So, the problem of comparing the PCs of users is equivalent to comparing the similarity between two sets of orthogonal vectors with unit lengths, while each of these vectors is associated with some weight (i.e. the eigenvalue) to indicate its relative importance in its set. To carry out such comparison, we propose to use the sum of pair-wise inner product normalized by their corresponding weights in their sets. The similarity index between two sets of PCs, $U = \{u_1, ..., u_{r_u}\}$ and $V = \{v_1, ..., v_{r_v}\}$, is defined as:

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

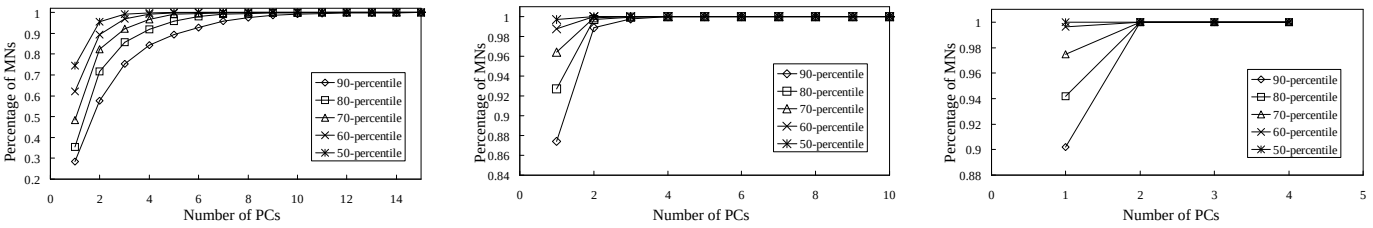
where w_{u_i}'s are defined as the percentage of variation captured by the PC u_i. The weights w_{u_i}'s sum up to 1. w_{v_j}'s are defined similarly. We say two MNs have similar sets of PCs if the similarity index is beyond a threshold.

To show the proposed similarity index provides a reasonable heuristic to group MNs with similar association behaviors, we obtain the *group association matrices* for the groups suggested by the similarity indexes, and perform PCA to these matrices. We show the grouping suggested by the similarity indexes increases the variation captured in its top PCs, and hence indeed we have put MNs with similar association patterns in the same group.

As an example, we obtain the following groups using 0.8 as the grouping threshold from Dartmouth trace: (A) A group including MNs similar to a MN with middle-ranked activeness. This group contains $1,328$ MNs based on the similarity indexes. (B) A group including MNs similar to the least active MN. This group contains $1,619$ MNs. (C) A randomly generated group containing $1,619$ MNs. (D) The whole user group containing $6,599$ MNs. For each of the groups, we perform PCA and show curves for the cumulative variation captured in its top PCs in Fig. 3.

From Fig. 3 we see for the groups suggested by similarity index (i.e. Group A and B), the top PCs capture more variation in their group association matrices than the other two groups. For example, the top-10 PCs capture 76% and 45% of total variation in group A and B, respectively, as compared to 16% and 13% in group C and D.

In addition to the above, we also check the overall performance of the grouping suggested by the similarity indexes by the following experiment. From Dartmouth trace, for all MNs with 100 or more similar nodes (using grouping threshold 0.8), we perform PCA to the group association matrix of the suggested group, and to a



(a) Dartmouth (b) USC (c) UCSD

Fig. 2. Number of PCs needed to capture the given percentage of variation in individual association matrices. Note the X-axes and Y-axes are in different scale in the graphs.

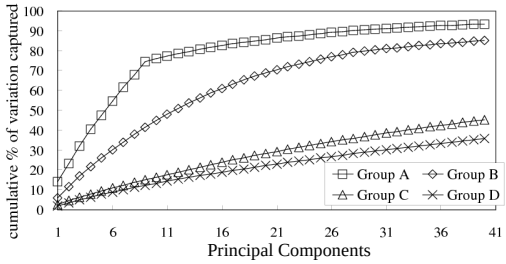


Fig. 3. Cumulative variation captured in top PCs. Groups suggested by similarity index have more variations captured by top PCs.

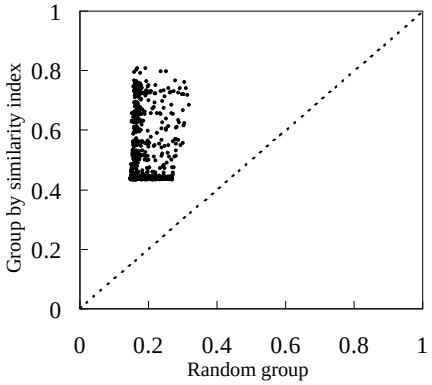


Fig. 4. Variation captured by top-10 PCs in groups suggested by similarity index (with grouping threshold 0.8) versus random groups

random group with the same number of MNs. We calculate the total percentage of variation captured by top-10 PCs in both groups, and show it as a dot on the scatter plot shown in Fig. 4. In the figure, a dot above the 45-degree line indicates for that MN, the grouping suggested by similarity index has more significant common trends in the group association matrix than the random group. As we see from the figure, the proposed similarity index indeed generates group with similar association pattern in most of the cases. We have also checked the variation in *group association matrices* captured by top-k PCs for k values other than 10, and the trend is similar to those shown in Fig. 4. Similar trends are also observed for the USC traces.

V. Conclusion and Future Work

By applying principal component analysis techniques, we unearth the underlying trends of group association matrices and individual association matrices from WLAN traces. For the whole user population, the dimensionality of the group association matrices are typically high, unless the users are from a population with some inherent similarities in behavior. For individuals, most power of its association patterns can be captured with a small set of PCs for most users. We further propose to use the similarity index obtained by comparing the important PCs of users to put similar users in sub-groups. We show that grouping by our proposed similarity index provides sub-groups with lower dimensionality than those of random sub-groups or the whole user population, indicating the grouped users are indeed following common trends of association with each other.

The findings from the analysis of WLAN traces point to shortcomings in earlier mobility modeling work. Mobility modeling has been an important subject in evaluation of wireless network performances. However, in most existing mobility models, the MNs are assumed to be homogeneous in the sense that all the node should be *identical* to each other in its behavior pattern in long run. This contradicts our finding that the group association matrices have high-dimensionality for the traces coming from a generic user group. In other words, some typical mobility scenarios (e.g. random waypoint model) are only suitable when MNs in the user population are inherently similar to one another.

The potential directions of future work are: (1) Since each MN is able to obtain its own association patterns, and summarize its association patterns using only a small set of PCs, it provides an efficient way for MNs to convey, exchange, and compare their association patterns. Such technique can be utilized to compare whether two MNs are similar, and helps to design context aware information diffusion protocols. (2) By inspecting the weights of PCs, one could tell whether a MN displays significant bi-modal behavior (e.g. The top PC stands for association patterns mainly for weekdays, while the second PC stands for weekends.), or whether a MN changes its association pattern significantly at a point of time (e.g. New association patterns deviate from linear combinations of PCs obtained from previous association patterns). Such identification could be used by a network operator to better understand its users, and may be useful for abnormal user-behavior detection.

References

[1] I.T. Jolliffe, Principal Component Analysis, 2nd ed., Springer series in statistics, published 2002.
of Network Traffic Flows," ACM SIGMETRICS, New York, June 2004.
[2] W. Hsu and A. Helmy, "IMPACT: Investigation of Mobile-user Patterns Across University Campuses using WLAN Trace Analysis," Technical report, USC-05-858, Available at http://nile.usc.edu/MobiLib/Trace_analysis_TR.pdf
[3] 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.
[4] 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.
[5] Longer version technical report of this poster is available at online http://nile.usc.edu/∼weijenhs/PCA-TR.pdf