Extended Abstract: Mining Behavioral Groups in Large Wireless LANs

Wei-jen Hsu
Computer and Information Science and Engineering
University of Florida
wjhsu@ufl.edu

Debojyoti Dutta*
Cisco Systems, Inc.
dedutta@cisco.com

Ahmed Helmy
Computer and Information Science and Engineering
University of Florida
helmy@cise.ufl.edu

ABSTRACT

Recent years have witnessed significant growth in the adoption of portable wireless communication and computing devices (e.g., laptops, PDAs, smart phones) and large-scale deployment of wireless networks (e.g., cellular, WLANs). We envision that future usage of mobile devices and services will be highly personalized. Users will incorporate these new technologies into their daily lives, and the way they use new devices and services will reflect their personality and lifestyle. Therefore it is imperative to study and characterize the fundamental structure of wireless user behavior in order to model, manage, leverage and design efficient mobile networks and services.

In this study, using our systematic *TRACE* approach, we analyze wireless users' behavioral patterns by extensively mining wireless network logs from two major university campuses. We represent the data using *location-preference vectors*, and utilize unsupervised learning (clustering) to classify trends in user behavior using novel similarity metrics. Matrix decomposition techniques are used to identify (and differentiate between) major patterns. We discover multi-modal user behavior and hundreds of distinct groups with unique behavioral patterns in both campuses, and their sizes follow a power-law distribution. Our methods and findings might provide new directions in network management and behavior-aware network protocols and applications, to name a few.

Categories and Subject Descriptors: C.2.1 [Network Architecture and Design]: Wireless Communication, I.5.3 [Clustering]: Similarity measures

General Terms: Measurement, Human Factors.

Keywords: Mobility, Wireless LAN, User Classification, Similarity Measure.

1. INTRODUCTION

Due to the rapid adoption of portable computing and communication infrastructure, the locations from where people access information are changing rapidly. There is a pressing need to take a holistic view of this adoption, beyond mere technology, and seek for a better understanding of user behavior patterns when WLANs are deeply incorporated in our lives. The techniques to discover such patterns from wireless measurements is the main focus of our paper.

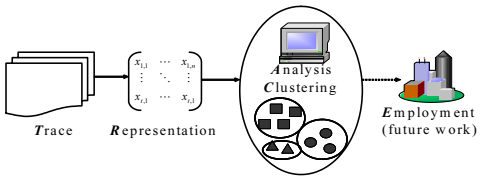

Figure 1: Illustration of the *TRACE* approach.

We use Fig. 1 to illustrate the conceptual flow of our approach in the paper, which we refer to as the *TRACE* framework. The work starts with the wireless LAN (WLAN) *Traces* that capture realistic user behavior. We focus on a specific *Representation* distilled from the traces, *normalized association vectors*, in this paper. We conduct extensive *Analysis* on the user association patterns, and propose summarization methods to capture its major trend. We then utilize a feature-based approach to achieve meaningful user *Clustering* and further discuss its interpretation. Finally, we briefly discuss future *Employment* of the methods and findings.

We analyze the association patterns in wireless LAN (WLAN) measurements obtained from two large-scale trace archives [10, 11]. User association and mobility is one of the defining characteristics of WLANs, and a deep understanding of the same would facilitate a multitude of applications, including user modeling, network management, and design of behavior-aware protocols. Existing work addresses such needs by extensive trace analysis with focuses mostly on aggregated statistics (e.g., [8, 9]) or on association models for individual users (e.g. [12, 13]). In either approach, the users are mostly considered as independent samples from a uniform population, and little effort has been directed to understanding the relationships between users. In [6], Kim et al. classify users based on the range and periodicity of their mobility, but not according to the detailed preference in their mobility processes. We provide a further step towards this understanding by classifying users into groups of similar association preferences. This provides a different and important perspective to understand user association patterns.

We leverage unsupervised learning (e.g., hierarchical clustering) techniques (for an introduction, see [2]) to determine groups of users displaying similar behavior. While clustering has been widely-applied to several other problems, the main contribution of the paper is to construct proper representations for our data sets and design novel distance metrics between users. These two aspects are

*The work was done when the author was at University of Southern California.

MobiCom'07, September 9–14, 2007, Montréal, Québec, Canada.
ACM 978-1-59593-681-3/07/0009.

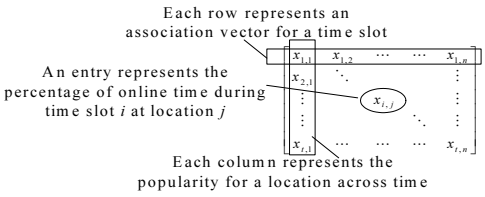

Figure 2: Illustration of association matrix representation.

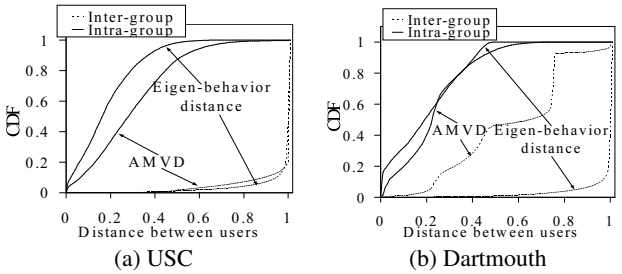

Figure 3: Cumulative distribution function of distances for inter-cluster and intra-cluster user pairs when the whole user population is divided into 200 clusters. Comparing between the curve pairs labeled as $AMVD$ and eigen-behavior distance, the latter leads to better separation between the distance distribution curves.

fundamental in the application of clustering techniques and determine the quality of the results we obtain. The key challenge in designing a good distance metric is to accurately and succinctly summarize the trends in the data, so the distances are not influenced by noise and can be evaluated efficiently. We show that a singular-value decomposition (SVD) based scheme not only provides the best summary of the data, but also leads to a distance metric that is robust to noise and computationally efficient. Furthermore, we validate our methods and explain its significance.

We introduce the data sets and the representation in section 2 and our techniques for user clustering in section 3. We further discuss the validation and interpretation of the results and its application in section 4 and 5, respectively. Section 6 concludes.

2. DATA SET AND REPRESENTATIONS

University campus WLANs have attracted high adoption from its community. Due to its high penetration and diversity in users, university campus networks are good platforms to study the behavioral pattern of WLAN users. We choose two WLAN traces collected from USC [10] and Dartmouth [11] for the study. The duration of the analyzed traces are 94 and 61 days, respectively, and the users included in the analysis are $5,000$ and $6,582$, respectively. For more details of the selected traces see [1].

To understand user behavior from wireless network traces, the first fundamental task is to choose a representation of the raw data and construct useful features. We choose to analyze the patterns of users visiting various locations in a large-scale WLAN. The visiting pattern is important to WLANs as mobility is one of its defining characteristics. From a social context, the places a person visits regularly and repeatedly usually have a strong connection with the affiliation and lifestyle. It is perhaps one of the important distinguishing factors for people with different social attributes.

We represent a user's visiting pattern by what we refer to as *normalized association vectors*, or just *association vector* in short. The association vector is a summary of a user's association preference during a given time slot. We choose to use a day as the time slot since it represents the most natural behavior cycle in our lives. The *association vector* for each time slot is an n-entry vector, $(x_1, x_2, ..., x_n)$, where n is the number of unique locations (e.g., buildings) in the given trace. Each entry in the vector, x_i, represents the *fraction* of online time the user spends at the location during the time slot, i.e. we normalize the user association time with respect to her online time. With this representation, the conclusions we draw are not influenced by the absolute value of online time, which varies across a wide range among different users and different time slots of a given user. To represent a user's association preference for the long run, we construct the *association matrix* X for the user, as illustrated in Fig. 2. Note that entries in each row in the association matrix sum to *one* if the user has been online during the time slot. A row of zeros represents a time slot the user is completely offline.

Note that there are potentially many ways to represent user behavior from a rich data set. Different representations certainly provide different insights. While our chosen representation emphasizes location preferences, Kim et al. emphasize the degree of mobileness and the periodicity in user mobility[6]. Kang et al. apply clustering technique to the trace of location coordinates of a user to discover significant places for a given user[7]. Due to space limitations, we leave discussions about other alternatives to an extended technical report[1].

3. CLUSTERING USERS

3.1 Need for a Distance Metric

The most important step in clustering is to define the *distance metric* between users. An intuitive distance metric between association patterns of two individuals is to consider all the association vector pairs, and take the average of distances from each association vector of user A to the corresponding closest association vector of user B (this *average minimum vector distance ($AMVD$)* is discussed in more details in [1]). We apply the hierarchical clustering algorithm to classify users with this distance metric, and as shown in Fig. 3, the separation between inter and intra cluster distance distributions indicated with label $AMVD$ is not clear for the Dartmouth users (i.e., distances between users in the same cluster is not much smaller than distances between users in different clusters), indicating this is not a very reliable distance metric.

One problem with the $AMVD$ metric is that it considers all association vectors, i.e. it includes not only the important trends, but also the noise vectors when the users deviate from their typical behavior, leading to bad clustering results. Another problem of the $AMVD$ metric is its computation complexity. We have to calculate the distances between all t^2 pairs (t denotes the number of time slots) of association vectors for each user pair. If there are N users the computation requirement is of order $O(N^2t^2)$. Furthermore, it requires significant space to store t association vectors for all N users. Thus we would like to design a metric that is both (1) robust to noise and (2) computation and storage efficient. In order to achieve both goals, we start by studying the proper summary of user association patterns. We show that this study leads us to the appropriate distance metric.

3.2 Summarizing the Association Patterns

In this section, we identify association trends of an individual and construct a compact representation of her association patterns. We expose the dominant association trends by clustering of the as-

sociation vectors, X_i for $i = 1, ..., t$ (i.e., row vectors of an association matrix X) of a single user, using Manhattan distance as the distance metric between association vectors. The identified clusters represent distinct *behavioral modes* of the user. Similar association vectors will be merged into a cluster in the process and the cluster size indicates its dominance - large clusters imply that the user follows consistent association patterns on many different days as her major behavioral modes.

From both USC and Dartmouth traces, we discover that most users show multi-modal association behavior. Even if we consider a moderate clustering threshold (i.e., association vectors with Manhattan distance less than 0.9 are merged into one cluster), on average, the number of behavioral modes for USC and Dartmouth users are 5.57 and 4.32, respectively, much less than the total number of vectors (days). This implies although users in WLANs are not extremely mobile, they do move and display various association patterns over a period of time. Moreover, we observe that for USC and Dartmouth, respectively, 36% and 31% of users have the two most important behavior modes with comparable sizes (i.e., the largest cluster of association vectors is less than twice the size of the second largest cluster). Hence looking at the most dominant cluster exclusively could still be sometimes misleading and we might be ignoring information about the user's detailed behavior.

In order to quantitatively compare summary techniques, we propose the *significance score* of a summary vector as the sum of the projections of all association vectors on the summary vector, normalized by the online days of the considered user:

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

The physical interpretation of the *significance score* is the percentage of power in the association vectors X_i's explained by the summary vector Y. If one wants the summary vector Y to capture the most of power in vectors X_i's, mathematically,

$$Y = \arg \max_{\|v\|=1} \sum_{i=1}^{t} |X_i \cdot v|. \quad (2)$$

This is exactly the procedure to obtain the first singular vector if we perform singular value decomposition (SVD) [4] of the association matrix X. We have validated that SVD indeed generates a summary vector (i.e., the first singular vector) with higher significance score as compared to average of all association vectors or the centroid of the largest behavioral mode[1].

There are two ways to understand the application of SVD to the association matrices. Mathematically, SVD can be viewed as a procedure to generate a low-rank reconstruction of the original matrices[4]. We have shown that such low-rank reconstructions are possible for most association matrices due to the high repetition in user's daily association vectors (i.e., the existence of dominant behavioral modes). Similar observations have been made for cellphone user association with cellphone towers [5]. For more than 90% of users in both traces, we are able to reconstruct their association matrices with 90% of the power captured in at most top-*five* singular vectors and corresponding singular values. On the other hand, for behavioral analysis purpose, SVD can be viewed as a procedure to obtain representative vectors that capture the most remaining power in the association matrix, defined by

$$\begin{aligned} u_1 &= \arg \max_{\|u\|=1} \|X \cdot u\| \\ u_k &= \arg \max_{\|u\|=1} \|(X - \sum_{i=1}^{k-1} X u_i u_i')u\| \quad \forall k \geq 2. \end{aligned} \quad (3)$$

We can interpret the vector u_j's as the vectors that describe the user's association behavior in decreasing order of importance , with its relative weight quantified by the ratio of the corresponding singular values, $\sigma_j^2 / \sum_{i=1}^{Rank(X)} \sigma_i^2$. We refer to these unit-length vectors as *eigen-behavior* vectors for the user. The absolute values of entries in an *eigen-behavior* vector quantify the relative importance of the locations in the user's j-th behavioral mode.

To sum up, SVD provides the optimal summary that captures the most remaining power in the original matrix with each additional component, with a quantitative notion of their relative importance. In addition, the components can be used to reconstruct the original matrix, while the calculation of average or centroid vectors are non-reversible.

3.3 Eigen-behavior Distance

We propose to use the eigen-behavior vectors of two users to measure the similarity of their corresponding association matrices. Suppose u_i's and v_j's are the eigen-behavior vectors of two users, $i = 1, ..., r_u$ and $j = 1, ..., r_v$ where r_u and r_v are the ranks of the corresponding association matrices. The similarity between the two users is defined as the sum of pair-wise weighted inner products of their eigen-behavior vectors u_i's and v_j's:

$$Sim(U, V) = \sum_{i=1}^{r_u} \sum_{j=1}^{r_v} w_{u_i} w_{v_j} |u_i \cdot v_j|, \quad (4)$$

where $w_{u_i} = \sigma_{u_i}^2 / \sum_{k=1}^{r_u} \sigma_{u_k}^2$ and w_{v_j} is defined similarly. Higher similarity index $Sim(U, V)$ indicates the corresponding users have similar association patterns. We define the *eigen-behavior distance* between users U and V as $D'(U, V) = 1 - (Sim(U, V) + Sim(V, U))/2$. As shown in Fig. 3, the eigen-behavior distance leads to a better separation between the CDF curves as compared to the $AMVD$, indicating a meaningful clustering of users into well-separated behavioral groups.

Using the *eigen-behavior distance* also reduces the computation overhead. If we use only the top-5 components (which captures more than 90% power), instead of going through t-by-t pairs of original association vectors with $AMVD$, we reduce the distance calculation to 5-by-5 pairs. By paying the pre-processing (i.e., SVD for all N users) overhead of $O(Nt^2)$, we can reduce the distance calculation complexity from $O(N^2t^2)$ to $O(c \cdot N^2)$[1].

4. INTERPRETING BEHAVIORAL GROUPS

4.1 Significance of the Clusters

We validate the resulting clusters obtained with eigen-behavior distance indeed capture users with similar behavioral trends by comparing the clustering results with randomly formed clusters. When association vectors from all m users in one cluster are concatenated in a larger mt-by-n *joint association matrix*, its top eigen-behavior vectors represent the dominant behavioral patterns within the group. We verify that for the clusters obtained through eigen-behavior distance, its top eigen-behavior vectors capture much more power in the joint-association matrices than if the clusters are formed randomly, pointing to the significance of our clustering results. In addition, to quantify each cluster corresponds to a group of users with distinct behavioral pattern, we obtain the first eigen-behavior vector from each group and calculate its *significance score*, defined in Eq. (1), for all the groups. The results confirm with our goal

[1]Since users follow repetitive trends in the association patterns, its total *eigen-behavior* vectors would not grow with the number of time slots, t.

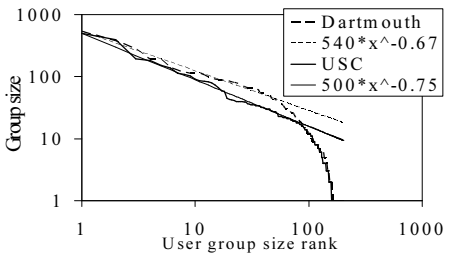

Figure 4: Rank plot (group size ranking v.s. group size) in log-log scale. User group size follows a power-law distribution.

of identifying groups following different behavioral trend: For the USC trace, the first eigen-behavior vectors obtained from the *joint association matrices* have an average *significance score* of 0.779 for their own groups and an average score of 0.005 for other groups. The corresponding numbers for the Dartmouth trace are 0.727 and 0.004, respectively.

4.2 Interpretation

In this section we analyze and interpret the results of clustering for both university campuses from a social perspective. We observe that the distributions of behavioral group sizes are highly-skewed for both campuses: The ten largest groups combined account for 39% and 33% of the total population, respectively. On the other hand, in both campuses about half of the groups have less than 10 members. More interestingly, we observe that the distribution of the group size follows a power-law distribution. In Fig. 4, we plot the straight lines that illustrate the best power-law fits. The slopes for these lines are -0.67 for Dartmouth and -0.75 for USC, respectively.

We are also able to pin-point the detailed behavioral pattern for each group by a close inspection of its eigen-behavior vectors. We discover that for most of the large groups, their top eigen-behavior vectors dominate, and there is a clear ordering in the importance of eigen-behavior vectors. Hence the association behavior of the group can be described by a sequence of locations of decreasing importance with a clear ordering. The largest behavioral groups on both campuses include the visitors of the library, as expected, since libraries are still the most visited area on university campuses. We also discover various groups featured different dorms and classrooms as their most visited location from both campuses.

On the other hand, we have also observed groups with multiple high-value entries in its top eigen-behavior vectors from both campuses. This indicates the existence of behavioral groups which visit multiple locations with equal importance to them. One prominent example from USC trace consists of 32 users, who visit buildings VKC and THH, two major classrooms on the USC campus. The top two eigen-behavior vectors of the group both consist of two high-value entries corresponding to these two buildings, and they capture 63.14% of power in the joint association matrix. It is a good example to show why it is not sufficient to merely use the most dominant behavioral mode (or the most-visited location) of a user to classify it. We have tried to use the centroid vector of the dominant behavioral mode to classify users, and it fails to identify a group with such behavior.

To sum up, we have demonstrated a systematic way to identify distinct behavioral groups within on-campus populations, by using clustering based on association features obtained from large-scale wireless network traces.

5. POTENTIAL APPLICATIONS

The insight obtained from clustering of users obtained from our analysis can be applied in many different ways. For example, (1) Our decomposition approach provides two pieces of important information: the distribution of group sizes follows a power-law distribution and the detailed eigen-behavior vectors of the groups. These results could help us to propose more realistic models for WLAN users, which is a challenge and a necessity for evaluating network protocols and services. (2) SVD-based behavioral analysis could be applied to discovery norms of user mobility preferences, and serves as a baseline for abnormal behavior detection schemes. (3) The similarity measure (i.e., (4)) between users could enable a new service we name as *profile-casting*: delivering a message to a group of users with similar long-run association characteristics [1]. In many cases, similar mobility characteristics infers similar affiliation, and hence such application would be important for social networking. Note that, this service is different from geo-casting, which delivery messages based on the *current* locations of users, as it compare users based on the *long-run history* of association patterns.

Our *TRACE* approach could be applied to various representations with different data sets. For example, in encounter-based networks [14], a representation of encounter probability or duration would be appropriate. We plan to investigate this in our future work.

6. CONCLUSION

In this paper, we classify groups of WLAN users based on the trends in their WLAN association patterns with the *TRACE* approach. We design a novel distance metric between users based on the similarity of their *eigen-behavior* vectors, obtained through singular value decomposition (SVD) of the association matrices. The eigen-behavior distance leads to a meaningful and distinct partition of users. It also leads to space and time efficient computations.

7. REFERENCES

[1] Longer version of technical report available at http://arxiv.org/abs/cs/0606002
[2] A. Jain, M. Murty, and P. Flynn, "Data Clustering: A Review," ACM Computing Surveys, vol. 31, no. 3, September, 1999.
[3] I.T. Jolliffe, Principal Component Analysis, second ed., Springer series in statistics, published 2002.
[4] R. Horn and C. Johnson, Matrix Analysis, Cambridge University Press, published 1990.
[5] N. Eagle and A. Pentland, "Reality mining: sensing complex social systems," in Journal of Personal and Ubiquitous Computing, vol.10, no. 4, May 2006.
[6] M. Kim and D. Kotz, "Periodic properties of user mobility and access-point popularity," Journal of Personal and Ubiquitous Computing, 11(6), August, 2007.
[7] J. Kang, W. Welbourne, B. Stewart, and G. Borriello, "Extracting places from traces of locations," in SIGMOBILE Mobile Computing and Communication Review, vol. 9, no. 3, July 2005.
[8] 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.
[9] M. Balazinska and P. Castro, "Characterizing Mobility and Network Usage in a Corporate Wireless Local-Area Network," In Proceedings of MobiSys 2003, May 2003.
[10] W. Hsu and A. Helmy, MobiLib USC WLAN trace data set. Download from http://nile.cise.ufl.edu/MobiLib/USC_trace/
[11] D. Kotz, T. Henderson and I. Abyzov, CRAWDAD data set dartmouth/campus/movement/01_04 (v. 2005-03-08). Downloaded from http://crawdad.cs.dartmouth.edu/dartmouth/campus/movement/01_04
[12] 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.
[13] C. Tuduce and T. Gross, "A Mobility Model Based on WLAN Traces and its Validation," in Proceedings of IEEE INFOCOM, March 2005.
[14] A. Chaintreau, P. Hui, J. Crowcroft, C. Diot, R. Gass and J. Scott, "Impact of Human Mobility on the Design of Opportunistic Forwarding Algorithms," in Proceedings of INFOCOM 2006, Barcelona, Spain, April 2006.