Classifying the Mobility of Users and the Popularity of Access Points

Minkyong Kim and David Kotz

Department of Computer Science
Dartmouth College
{minkyong,dfk}@cs.dartmouth.edu

Abstract. There is increasing interest in location-aware systems and applications. It is important for any designer of such systems and applications to understand the nature of user and device mobility. Furthermore, an understanding of the effect of user mobility on access points (APs) is also important for designing, deploying, and managing wireless networks. Although various studies of wireless networks have provided insights into different network environments and user groups, it is often hard to apply these findings to other situations, or to derive useful abstract models.

In this paper, we present a general methodology for extracting mobility information from wireless network traces, and for classifying mobile users and APs. We used the Fourier transform to convert time-dependent location information to the frequency domain, then chose the two strongest periods and used them as parameters to a classification system based on Bayesian theory. To classify mobile users, we computed *diameter* (the maximum distance between any two APs visited by a user during a fixed time period) and observed how this quantity changes or repeats over time. We found that user mobility had a strong period of one day, but there was also a large group of users that had either a much smaller or much bigger primary period. Both primary and secondary periods had important roles in determining classes of mobile users. Users with one day as their primary period and a smaller secondary period were most prevalent; we expect that they were mostly students taking regular classes. To classify APs, we counted the number of users visited each AP. The primary period did not play a critical role because it was equal to one day for most of the APs; the secondary period was the determining parameter. APs with one day as their primary period and one week as their secondary period were most prevalent. By plotting the classes of APs on our campus map, we discovered that this periodic behavior of APs seemed to be independent of their geographical locations, but may depend on the relative locations of nearby APs. Ultimately, we hope that our study can help the design of location-aware services by providing a base for user mobility models that reflect the movements of real users.

1 Introduction

Wireless networks have become popular and are getting more attention as a way to provide constant connectivity over a large area in cities and as an inexpensive way to provide connectivity to rural areas. The growing popularity of wireless networks

encourages the development of new applications, including those that require quality of service (QoS) guarantees. To provide QoS, it is often useful to predict user mobility. We also need simulators of wireless network environments to test these new applications and these simulators require user mobility models.

As more mature wireless networks become available, several studies of wireless networks have been published, including studies of a campus [4, 5], a corporate environment [2], and a metropolitan area [8]. Although these studies help us to understand characteristics of different network environments and user groups, it is often difficult to apply the findings of these studies to other applications.

In this paper, we introduce a method to characterize real wireless network traces and classify different mobile users based on their mobility. We transform our traces using the Discrete Fourier Transform (DFT) to make them independent of the particular time that traces were gathered. This transform exposes periodicity in traces.

We then use AutoClass [3], an unsupervised classification tool based on Bayesian theory. Classification is important because user mobility differs widely from user to user [2]. Thus, it is difficult to describe diverse user mobility patterns with a single model. Classification breaks down this complex problem into several simpler ones, by dividing users into groups that have common characteristics and thus might be modeled similarly. Moreover, classification is important because a collection of individual cases has little predictive power for new cases.

In the second part of this paper, we focus on the behavior of access points (APs). We apply our method to extract information from real wireless network traces and classify APs. Understanding the behavior of APs is important for many applications, such as traffic engineering for APs and resource provisioning for QoS sensitive applications.

An important benefit of using the Discrete Fourier Transform is that it is easy to compute the inverse DFT to obtain the time series. After clustering instances based on the information extracted from DFT, we can construct a sequence of numbers corresponding to the power spectrum representative of each class. We can then use an inverse DFT to obtain the time series that represents that class. This method is also used by Paxson [6] to synthesize approximate self-similar networks. We leave this modeling process as future work.

2 Methodology

In this section, we describe our traces and the parameters that we have chosen to represent user mobility and behavior of APs. We then describe how we converted our traces from the time domain to the frequency domain using a Fourier Transform and how we classified users and APs using AutoClass.

2.1 Trace collection

We collected syslog traces of APs from the Dartmouth College campus-wide wireless network. The APs record client events (such as authenticating, deauthenticating, associating, disassociating, and roaming) by sending syslog messages to a central server, where the logs are timestamped with a one-second granularity. Currently, most of the

APs on our campus are Cisco 802.11b APs. Although they are in the process of being replaced by Aruba APs, we focus on Cisco APs because at the time of the study they were still the dominant set of APs and covered most of the campus.

We have been collecting syslog records since 476 Cisco APs were installed in 2001. In this paper, we focus on four weeks of traces collected from October 3 to October 30, 2004. During these four weeks, we saw 7,213 devices (i.e., MAC addresses) visiting 469 APs. In the following discussion, we refer to a MAC address as a user, although a user may own more than one device with a wireless network interface. We expect that most of the devices are laptops, based on the previous study over the traces collected at Dartmouth [4]. We saw roughly 4.5 million syslog events, of which 1.9 million events represent devices associating or reassociating with APs.

2.2 Parameter selection

To cluster users or APs we must choose an appropriate parameter.

Diameter as mobility measure. One limitation of our study is that we do not have the exact geographical location of a user. We only have the information about the location of APs on our campus and the AP which a user is associated with. Thus, we approximate a user's location using the location of the AP with which the user is associated. Because many areas are covered by more than one AP, some clients change association from an AP to another even when they do not physically move. Sometimes a client associates repeatedly with a fixed set of APs, a phenomenon we call the *ping-pong* effect.

The ping-pong effect cannot happen across two APs that are apart farther than a certain distance because APs have limited coverage, but this distance is often hard to pinpoint. The Cisco specification states that the indoor range at 11 Mbps is 39.6 meters and the outdoor range is 244 meters. Obviously, a ping-pong effect is extremely unlikely between two APs that are more than 244 meters apart, but choosing this value as the threshold is too aggressive, filtering out too many user movements. Because different APs are configured differently and located in different environments, it is hard to define a precise distance threshold to decide whether a change between two APs is due to the ping-pong effect or not. Although Henderson [4] defined the limit as 50 meters, in our traces we found that some clients ping-pong between two APs more than 50 meters apart. Thus, we do not use a threshold to filter out ping-pong effects, but choose a parameter that is less sensitive to them.

Our goal is to classify wireless network users based on their mobility patterns. Our traces list events at a particular AP with a particular mobile user. We first gathered the events associated with each user. Although the events are recorded with a one-second granularity, we aggregated them into one value for each hour. We considered several alternatives to represent this value. Because of the ping-pong effect, the total distance traveled (the sum of the distance between APs visited, in sequence) often does not reflect user mobility. A user may appear to travel a long distance if he experiences many ping-pong effects even though he did not move at all. A better measure is the *diameter*, defined as the maximum Euclidean distance (i.e., the straight line distance between two points) between any two APs visited during a fixed time period. Although

we still cannot tell whether a diameter is due to real user movements or ping-pong effects when it is short, we can at least be confident that it is caused by real movements when a diameter is longer than a certain distance.

Number of users to describe APs. For APs, we used the same set of traces, but gathered the events associated with each AP. Then, we counted the number of unique users visiting each AP during each hour. By counting the number of unique users instead of the number of user visits, we remove noise caused by ping-pong effects.

2.3 Filtering traces

We found it was necessary to filter the traces to select the most meaningful data.

Mobility. In our traces, many users do not move at all, and many others appear in the traces for a short time and disappear. Because we want to find meaningful patterns of user mobility, we need to remove these stationary and transient users. We removed any user who did not move or did not connect to wireless network for a 3-day or longer period. We chose three days based on the assumption that regular mobile users are unlikely to stay at one place for more than three days. They may stay at one place for the weekend; thus using two days as the filtering limit may be too aggressive. We also filtered out the users whose hourly diameter never exceeded 100 meters. Note that we did not filter out the *diameters* shorter than 100 meters; we filtered out the *users*. This filtering reduced the number of users from 7,213 to 246; thus our study focuses on the relatively rare "mobile users."

APs. There are many APs on our campus that are not actively used. To remove these APs, we filtered out the APs that never had more than 50 visitors during a hour. This filtering reduced the number of APs from 469 to 216.

2.4 Discovering Periodic Events

For each user, we create a 672-element vector that represents the user mobility (i.e., diameter) of each hour for four weeks. Our goal is to classify users according to their mobility patterns. Finding similar patterns by comparing these diameter vectors directly is not trivial. For example, the same mobility patterns may appear with more than one user, but they may be shifted in time or scaled. Also, we are not interested in discovering the exact value of diameter at a physical time.

To preserve the diameter but discount for shifts in absolute time, we used the Discrete Fourier Transform (DFT) to transfer our parameters from the time domain to the frequency domain. Since the Fourier Transform is well known, we only briefly describe it here, borrowing a description from *Numerical Recipes in C* [7]. Suppose that we have a function with N sampled values:

$$h_k \equiv h(t_k), \quad t_k \equiv k\Delta, \quad k = 0, 1, 2, ..., N-1. \tag{1}$$

Δ denotes the sampling period; it is one for our case. The DFT estimates values only at the discrete frequencies:

$$f_n \equiv \frac{n}{N\Delta}, \quad n = -N/2, -(N/2-1), ..., N/2-1, N/2 \tag{2}$$

where the extreme values of n correspond to the lower and upper limits of the Nyquist critical frequency range. Then, the DFT of N points h_k is defined as following:

$$H_n \equiv \sum_{k=0}^{N-1} h_k e^{2\pi i f_n t_k} = \sum_{k=0}^{N-1} h_k e^{2\pi i k n/N}. \tag{3}$$

Agrawal [1] has shown that a few Fourier coefficients are adequate for classifying Euclidean distances. He chose the first two strong, low frequency signals. Based on this study, we chose the two strongest frequency (or period) signals as our parameters for our classification of user mobility.

2.5 Clustering

To classify user mobility patterns, we use AutoClass [3], a classification system based on Bayesian theory. A key advantage of this system is that it does not need to specify the classes beforehand, allowing *unsupervised* classification. We had, and needed, few preconceptions about how our mobility data should be classified.

AutoClass takes fixed-size, ordered vectors of attribute values as input. Given a set of data X, AutoClass seeks maximum posterior parameter values $\overrightarrow{V}$ and the most probable T irrespective of $\overrightarrow{V}$, where $\overrightarrow{V}$ denotes the set of parameter values instantiating a pdf and T denotes the abstract mathematical form of the pdf. First, for any fixed T specifying the number of classes and their class models, AutoClass searches the space of allowed parameter values for the maximally probably $\overrightarrow{V}$. Second, AutoClass performs themodel-level search involving the number of classes J and alternate class models T_j. It first searches over the number of classes with a single pdf T_j common to all classes. It then tries with different T_j from class to class.

3 User Mobility

In this section, we present the result of user mobility patterns converted from the time domain to the frequency domain. We then show the classification of mobile users generated by AutoClass.

3.1 Mobility patterns

To illustrate our method, we choose one typical user from our traces. The diameters of this user in the time domain and frequency domain are shown in Figure 1 and Figure 2, respectively.

Figure 1 shows the diameter of each hour of one user and the number of unique APs visited by the user during each hour over four weeks. The x-axis shows the dates for

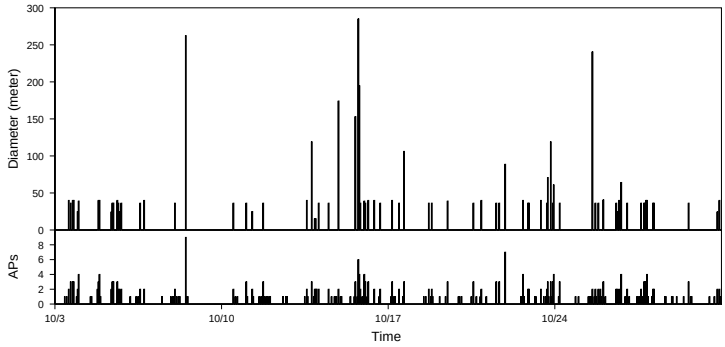


Fig. 1. Hourly diameter and APs visited by one user. This figure shows the user's hourly diameter and the number of unique access points visited by this user during each hour. Labels on the x-axis indicate the dates for Sundays.

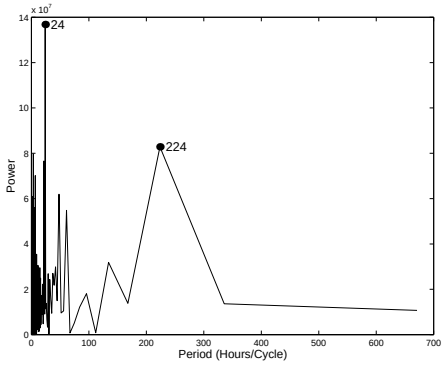


Fig. 2. Diameter in frequency domain. Two dots denote the two most strongest periods. In this example, they are approximately 24 hours and 224 hours.

Sundays, and the y-axis shows the diameter and the number of APs. This user often had a diameter of 40 meters. By looking into the traces, we found that the user was visiting a fixed set of APs repeatedly: the ping-pong effect. While shorter diameters are due to ping-pong effects, longer ones represent real movements.

Note that the number of unique APs does not necessarily correlate with the diameter: although the number of APs may indicate mobility, it cannot distinguish whether an increase in number is due to real movements or due to the ping-pong effects. Even when this user associated with up to four APs, the diameter was still around 40 meters. On the other hand, in the third largest peak where the user moved around 240 meters, he only visited two unique APs. Thus, the number of APs visited by the user is not appropriate to describe mobility.

Figure 2 shows the DFT of this users' vector of diameters. The two most significant periods are 24 and 224. This implies that user mobility patterns are likely to repeat in these periods.

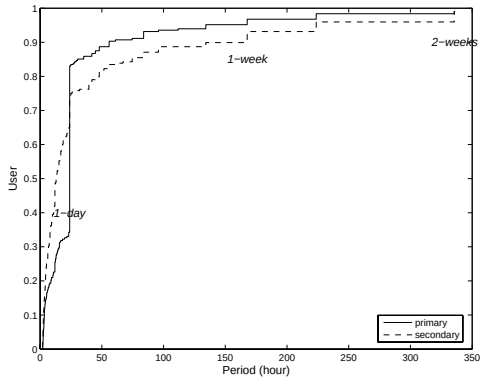


Fig. 3. Significant periods of user mobility. Cumulative distribution of the number of users versus period. From the power spectrum density graphs, we recorded the two most significant periods for each user.

We transformed all of our users' diameter vectors using the DFT and recorded the two strongest periods. Figure 3 shows the cumulative fraction of users with different periods as their first and second strongest periods. For the strongest period, the biggest jump is approximately around 24 hours. The distribution also has smaller jumps at the following hours: 84 (3 days and 12 hours), 168 (one week), 224 (9 days and 8 hours), and 336 (two weeks). Note that by using the DFT, we can observe a jump only at the period that is an integer fractions of the input length (672). We were not surprised to see users with one day, one week, or two weeks as their primary periods. But, it is interesting to observe more users with 3-days-and-12-hours than 4 days. The users with the period of 9-days-and-8-hours instead of 9 or 10 days may be an artifact from using the DFT because neither the period of 9 nor 10 days is an integer fraction of 4 weeks while that of 9-days-and-8-hours is an integer fraction; it is nonetheless interesting to observe users with this period as their primary or secondary periods.

3.2 Classification

We use the two strongest periods as our first two elements of three-element input vectors to AutoClass. In addition to these two periods that we gathered from the DFT, we also measured the maximum hourly diameter (d_{max}) observed over our traces for each user. As described in Section 2.3, we filtered out users whose d_{max} was less than 100 meters; this removed most of the stationary users.

AutoClass classified mobile users into seven classes. Table 1 shows the number of instances that fell into each class and the parameters that most influenced class assignment. The table also shows the mean and standard deviation of parameters of members within each class. Although parameters with smaller coefficient of variation (CV) often play an important role in class assignment, this is not necessarily true. It is how much the parameter value of an instance is different from those of others that determines whether the parameter plays a critical role in class assignment. Note that our third parameter d_{max} never played the major role in assigning instances to classes.

Class	Instances	Instances	Key	Period 1			Period 2			Diameter		
	(#)	(%)	Parameter	Mean	Std	CV	Mean	Std	CV	Mean	Std	CV
0	74	30.1	p2	43.1	67.8	157.3	19.4	7.8	40.2	279.1	94.1	6.0
1	75	30.5	p1	23.7	3.8	16.0	5.8	3.3	56.9	312.6	101.0	5.8
2	42	17.1	p1	23.8	4.6	19.3	41.0	34.7	84.6	184.9	90.2	8.7
3	23	9.2	p1	3.0	0.7	23.3	3.8	1.9	50.0	324.7	113.4	6.3
4	13	5.3	p2	103.9	81.7	78.6	118.2	55.9	47.3	228.7	88.5	6.9
5	15	6.1	p2	23.0	3.4	14.8	264.7	80.4	30.4	318.6	105.7	5.9
6	4	1.7	p2	5.6	0.7	12.5	209.7	28.0	13.4	255.1	118.9	8.4

Table 1. Classes of user mobility. Mean, standard deviation and coefficient of variation (%) of each parameter are listed. Period is in hours and diameter is in meters.

Figure 4 shows how classes are clustered in three dimensions in different perspectives for a better view. We first notice that there are many users tightly clustered around one day as their primary period. At the same time, there are many others for which one day was not their strong period. The first group of people with a strong one-day period make up classes 1, 2, and 5, while the second group of people make up the rest of classes.

First, we consider the group of users that have a strong one-day period. This group of people are divided into three classes based on the secondary period; classes 1, 2, and 5 correspond to small, mid-range, and big secondary periods as shown in Figure 4(c). Class 1 represents users who have one day as their strongest period and a small secondary period. Students who have regular classes may exhibit this kind of mobility behaviors. The average second period for class 2 is close to two days. The average for class 5 is close to 11 days, but this value is misleading; secondary periods of this class are bimodal around one week and two weeks. Thus, class 5 can be described as a cluster of users with one day and either one or two weeks as their strong periods. Note that mobile users with one day as their strongest period and a small secondary period are most prevalent—Class 1 is the biggest class.

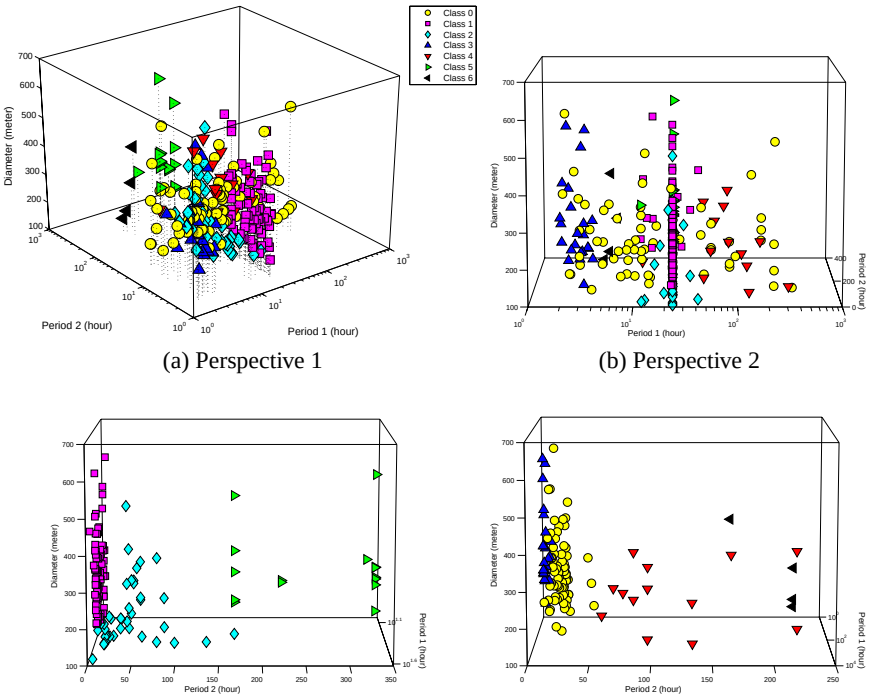
Second, we look into the group of users whose primary period is not one day. These users are divided into four classes. As shown in Figure 4(d), classes 3, 0, 4, and 6 have smallest to biggest secondary periods, respectively. Class 6 consists of users with the very small primary periods and 9-days-and-8-hours as the secondary period. It is interesting to note that most of the users whose primary period is not one day have their secondary period close to one day—Class 0 is the biggest class among these four classes.

4 Access Points

We now use the same method to classify APs based on how busy they are.

4.1 Periodicity

Figure 5 shows the cumulative distribution of the number of APs with primary and secondary periods: 85% of APs had their primary period at one day (24 hours); 25% of APs had their secondary period at 1 week (168 hours). Compared to the mobility traces, more APs have their primary period at one day and the secondary period at one week.



(a) Perspective 1

(b) Perspective 2

(c) Perspective 3 with classes 1,2,5

(d) Perspective 3 with classes 0,3,4,6

Fig. 4. Clustered users

Class	Instances (#)	Instances (%)	Key Parameter	Period 1 Mean	Period 1 Std	Period 1 CV	Period 2 Mean	Period 2 Std	Period 2 CV
0	99	45.8	p2	23.8	1.7	7.1	158.6	67.9	42.8
1	68	31.5	p2	24.0	0.0	0	11.6	2.3	19.8
2	28	13.0	p2	25.4	10.4	40.9	28.3	6.9	24.4
3	21	9.7	p1	165.1	97.4	59.0	90.0	97.7	108.6

Table 2. Classes of access points

4.2 Classification

As input to AutoClass, we used three parameters: the period at which power is maximum, the period at which the power is second to maximum, and the maximum number of users that an AP serviced during any hour, u_{max}.

Table 2 shows the number of cases that resulted in each class. AutoClass classified the input cases into four classes. The last parameter (u_{max}) did not make any difference in classifying the input cases. Thus, we do not include it in the table. The determining parameter for the first three classes was the secondary period (p2). This is because the primary period (p1) was equal to 24 hours for most of the cases, and therefore did not play a critical role in determining to which class a case belongs.

Figure 6 shows each instance in three dimensions in two different perspectives. Because u_{max} did not play a major role for classification, we do not include it in this

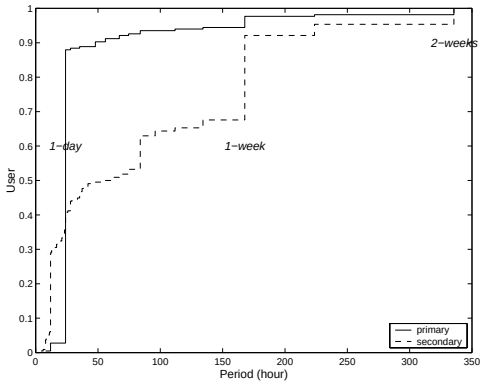
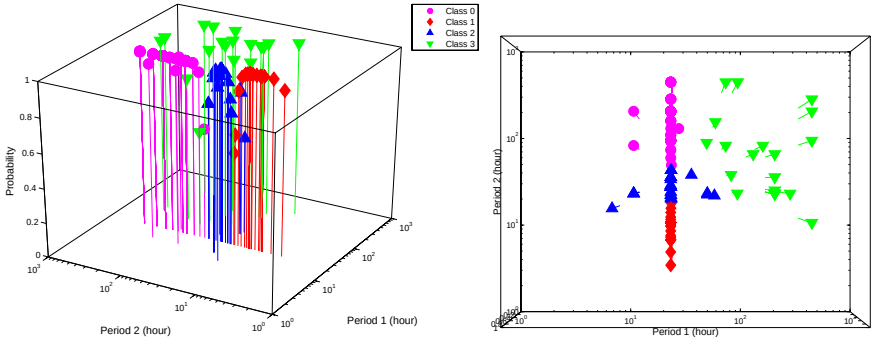


Fig. 5. Significant periods of APs. Cumulative distribution of APs versus period. From the power spectrum density graphs, we recorded the two most significant periods for each AP.



(a) Perspective 1 (b) Perspective 2

Fig. 6. Clustered access points

graph. Instead, we include the probability of an instance being in a particular class as the third axis. AutoClass computes this probability, for each instance, which indicates the likelihood that an instance is a member of a class. If this probability is one, that instance is a strong member of the class. Not surprisingly, the probability drops for the instances in the regions where different classes meet.

Figure 6 shows that most APs had their primary period at one day. It is also clear that classes 0, 1, and 2 had distinct secondary periods. Note that among these three classes, class 0 had the most instances; this means that APs with one day as their primary period and around one week as their secondary period were the dominant category. Class 3's primary period is much bigger than one day; its secondary period is also big.

Figure 7 shows the geographical location of the APs on our campus. Many of the Cisco APs on our campus have recently been replaced by Aruba APs. Because we focus only on Cisco APs, many APs on the map did not appear in our traces and therefore were not classified. Also, APs who never had more than 50 users per hour are not classified.

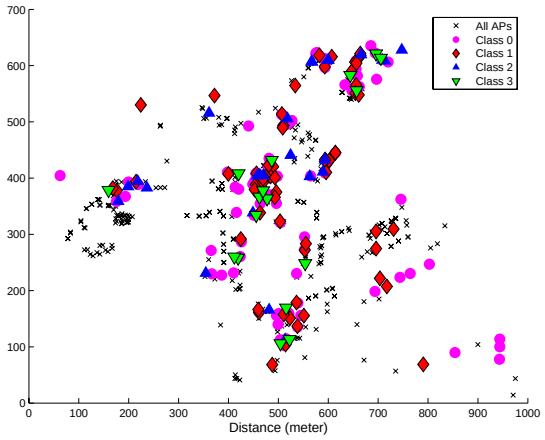


Fig. 7. Map of access points on campus

There are two things to note in Figure 7. First, APs within a small geographical location, even within the same building, often had different patterns of behavior. Thus, characterizing APs based on their geographical locations or type of building may be erroneous. Second, class 0 and 1 are located all over the campus, but class 2 and 3 are located only where many APs are deployed. We do not have a clear explanation of why this is happening, but it is still interesting to note that deploying too many APs within a limited space sometimes prevents APs from having a strong period of one day.

5 Lessons Learned

In the Fourier Transform, it is important to truncate data so that the input data is a multiple of the period of the signal. This is the reason that we used 4-week traces instead of one-month; we truncated data to be multiple of one week (i.e., 168). For access points, we tried both 4-week traces and one-month traces. With 4-week traces, an AP had one day as the strongest period and one week as the second. When we used one-month traces, we got the same value of one day for the first maximum, but got *one week and 12 hours* for the second maximum instead of exactly one week.

After clustering data, it was important to visualize the result. Visualization helped understanding how classes are divided and how each parameter contributes in distinguishing instances. But, it was not trivial to find the 'right' way to present clustered data. We expect it will even be harder with longer traces and more input parameters for classification.

6 Conclusion and Future Work

In this paper, we present a method to extract information from real wireless network traces and transform the time series to the frequency domain using the Fourier Transform. We then extracted the two most significant periods and clustered instances using a Bayesian classification tool. Our study is unique in using Fourier Transform and Bayesian theory to provide insights into user mobility and behavior of access points.

This paper presents ongoing work, and we plan to pursue several extensions. First, we would like to try our method with longer traces. We expect the trend will be similar to our study presented here although there may be varieties depending on the long-term academic schedules, such as when a term starts and ends. Second, we want to expand our study of APs to include the newly deployed Aruba APs, but we must first update our map data. Third, we plan to build generalized models for user mobility and activities of APs. We believe that our method will help us build models by identifying the most significant characteristics, by clustering users into groups that need different models or different parameters, and by abstracting traces. Finally, after successfully modeling user mobility based on our real traces, we would like to build a simulator for wireless network environments using our mobility model.

Acknowledgments

This project was supported by Cisco Systems, NSF Award EIA-9802068, and Dartmouth's Center for Mobile Computing. We are grateful for the assistance of the staff in Dartmouth's Peter Kiewit Computing Services in collecting the data used for this study. We would like to thank Songkuk Kim for the insightful suggestions throughout the process of developing our method. We also thank Tristan Henderson for commenting on draft versions of this paper.

References

1. Rakesh Agrawal, Christos Faloutsos, and Arun N. Swami. Efficient similarity search in sequence databases. In D. Lomet, editor, *Proceedings of the 4th International Conference of Foundations of Data Organization and Algorithms (FODO)*, pages 69–84, Chicago, Illinois, 1993. Springer Verlag.
2. Magdalena Balazinska and Paul Castro. Characterizing mobility and network usage in a corporate wireless local-area network. In *Proceedings of MobiSys 2003*, pages 303–316, San Francisco, CA, May 2003.
3. Peter Cheeseman and John Stutz. Bayesian classification (AutoClass): Theory and results. In Usama M. Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth, and Ramasamy Uthurusamy, editors, *Advances in Knowledge Discovery and Data Mining*, Philadelphia, PA, USA, 1996. AAAI Press/MIT Press.
4. Tristan Henderson, David Kotz, and Ilya Abyzov. The changing usage of a mature campus-wide wireless network. In *MobiCom '04: Proceedings of the 10th Annual International Conference on Mobile Computing and Networking*, pages 187–201, Philadelphia, PA, USA, 2004. ACM Press.
5. Ravi Jain, Anuparma Shivaprasad, Dan Lelescu, and Xiaoning He. Towards a model of user mobility and registration patterns. *MC^2R*, 8(4):59–62, October 2004. MobiHoc 2004 poster abstract.
6. Vern Paxson. Fast approximation of self similar network traffic. *Technical Report LBL-36750*, 1995.
7. William H. Press, Saul A. Teukolsky, William T. Vetterling, and Brian P. Flannery. *Numerical Recipes in C: The art of scientific computing*. Cambridge University Press, Cambridge, 1992.
8. Diane Tang and Mary Baker. Analysis of a metropolitan-area wireless network. *Wireless Networks*, 8(2-3):107–120, 2002.