Pers Ubiquit Comput
DOI 10.1007/s00779-006-0093-4

ORIGINAL ARTICLE

Periodic properties of user mobility and access-point popularity

Minkyong Kim · David Kotz

Received: 11 August 2005 / Accepted: 16 November 2005

Abstract Understanding user mobility and its effect on access points (APs) is important in designing location-aware systems and wireless networks. Although various studies of wireless networks have provided useful insights, it is hard to apply them to other situations. Here 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 reveal important periods and chose the two strongest periods to serve as parameters to a classification system based on Bayes' theory. Analysis of 1-month traces shows that while a daily pattern is common among both users and APs, a weekly pattern is common only for APs. Analysis of 1-year traces revealed that both user mobility and AP popularity depend on the academic calendar. By plotting the classes of APs on our campus map, we discovered that their periodic behavior depends on their proximity to other APs.

Keywords Wireless network · User mobility · Popularity of access points · Periodicity

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. Thus, we aim to understand mobility of mobile devices in Wi–Fi networks.

As more mature wireless networks become available, several studies of wireless networks have been published, including studies of a campus [7, 8, 11], a corporate environment, and a metropolitan area. Henderson et al. [7] analyzed the characteristics of wireless network usage on the Dartmouth campus using traces collected during the Fall 2003 and Winter 2004 terms. Balazinska and Castro [2] traced 1,366 corporate users on 117 APs over 4 weeks. Tang and Baker [12] studied a 7-week trace of the Metricom metropolitan-area packet radio wireless network, containing 24,773 mobile radios. 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. So we set out to develop methods to extract mobility characteristics from network traces, allowing anyone to obtain model parameters from traces of their network (or a network similar to the desired network).

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

M. Kim (✉) · D. Kotz
Department of Computer Science, Dartmouth College, Hanover, NH, USA
e-mail: minkyong@cs.dartmouth.edu

D. Kotz
e-mail: dfk@cs.dartmouth.edu

independent of the particular time that traces were gathered. This transformation exposes periodicity in traces.

We then use AutoClass [5], an unsupervised classification tool based on Bayes' 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.

We then focus on the behavior of access points (APs). We apply our method to extract periodicity from wireless network traces and to 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.

We first use a 1-month trace to understand short-term periodicities in user mobility and access-point popularity, and then analyze a 1-year trace to discover long-term seasonal effects. Both short-term and long-term effects are essential components of modeling. For example, a short-term effect would be a drop in mobility during the night, while a long-term effect would be an increase in mobility during certain academic terms on university or college campuses.

An important benefit of using the DFT 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 [9] 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 the Fourier transform and how we classified users and APs using AutoClass.

2.1 Collecting traces

At the Dartmouth College campus-wide wireless network, we have been collecting syslog records since 2001, when 476 Cisco APs were installed. 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 1 s granularity. As of December 2004, most of the APs on our campus were Cisco 802.11b APs. Although they were in the process of being replaced by Aruba APs, we focused on Cisco APs because they were still the dominant set of APs and covered most of the campus during the time when our traces were collected.

2.2 Selecting parameters

To cluster users or APs we must choose an appropriate parameter. In particular, we seek a simple measure of users' mobility within a time interval.

2.2.1 Diameter as mobility measure

One limitation of our study is that we do not have the exact geographical location of a user. We do know the location of each AP on our campus and the APs where a user is associated over time. Thus, we work around this limitation and 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 m and the outdoor range is 244 m. Obviously, a ping–pong effect is extremely unlikely between two APs that are more than 488 m 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 [7] defined the limit as 50 m, in our traces we found that some clients ping–pong between two APs more than 50 m 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 1 s 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 [7]. 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.

2.2.2 *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 removed noise caused by ping–pong effects. This measure gives a broad sense of the population's mobility about campus from hour to hour.

2.3 Discovering periodic events

For each user, we created a vector that represents the user mobility (i.e., diameter) of each hour during the length of traces. 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 briefly describe it here, borrowing a description from *Numerical Recipes in C* [10]. Suppose that we have a function with N sampled values:

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

Here Δ denotes the sampling period; it is 1 h for our case. The DFT estimates values only at the discrete frequencies:

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

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:

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

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.

2.4 Clustering

To classify user mobility patterns, we used AutoClass [5], a classification system based on Bayes' 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.

A Bayesian classification model consists of T, which denotes the abstract mathematical form of the model, and $\vec{V}$, which denotes the set of parameter values for the variables appearing in T. AutoClass takes fixed-size, ordered vectors of attribute values as input. Given a set of data X, AutoClass seeks maximum posterior parameter values $\vec{V}$ and the most probable T irrespective of $\vec{V}$. AutoClass performs two levels of search: parameter-level search and model-level search. 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 probable $\vec{V}$. Second, given the parameter values, AutoClass performs the model-level search involving the number of classes J and alternate class models T_j. It searches over the number of classes with a single probability density function T_j common to all classes. It then repeats this process with different T_j from class to class.

3 Short-term effects and classification

In this section, we analyze periodicities in 4 week[1] trace and present the result of classification generated

[1] We discuss the reason that we used a trace of 4 weeks instead of 1 month in Sect. 3.4.

by AutoClass. To study the short-term effects, we focus on 4 weeks of traces collected from October 3 to October 30, 2004. During these 4 weeks, 7,213 devices (i.e., MAC addresses) visited 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 [7]. The 4-week trace contains roughly 4.5 million syslog events, of which 1.9 million events represent devices associating or reassociating with APs.

3.1 Filtering traces

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

3.1.1 Mobility

In our traces, many users do not move at all, and many others appear in the traces only for a short duration. Because we want to find meaningful patterns of user mobility, we need to remove these stationary and transient users. Figure 1a shows the hourly diameter of all 7,213 users. For a given user and a given hour, ''white'' denotes a diameter of zero, while ''black'' represents a diameter greater than zero. Note that users are sorted based on the time when they first connected to the network. This figure shows that there were many users who joined the network several days after the beginning of our trace. There are also many users who rarely moved. We eliminated any user who did not move or did not connect to the wireless network for a period of 3 days or longer. We chose 3 days based on the assumption that regular mobile users are unlikely to stay in one place for more than 3 days. They may stay in one place for the weekend; thus using 2 days as the filtering limit may be too aggressive. After the filtering, we ended up with 360 users. Figure 1b shows the hourly diameter of these 360 users.

In analyzing the periodicity in mobility, we do not want to consider stationary users. Thus, we divided the 360 users into two groups: mobile and stationary. The users whose hourly diameter never exceeded 100 m belong to the stationary set, while the rest of the users belong to the mobile set. If a user was mobile, it is very likely that she had at least 1-h diameter value bigger than 100 m since it only takes a little over 1 min to walk that distance (with the average human-walking speed of 3 miles/h). Among the 360 users, 246 users (68%) belong to the mobile set. We focus on these mobile users in analyzing the periodicities.

3.1.2 APs

There are many APs on our campus that are not actively used. We divided the 469 APs into two groups: active and inactive. The APs that were never visited by more than 50 users per hour belong to the inactive set, while the rest of APs belong to the active set. Among 469, 216 APs (46%) belong to the active set. We focus on these actively used APs in our analysis.

3.2 User mobility

We first present the result of user mobility patterns converted from the time domain to the frequency domain and then show the classification of mobile users.

3.2.1 Mobility patterns

To illustrate our method, we chose one typical user from our trace. The diameters of this user in the time domain and frequency domain are shown in Figs. 2 and 3, respectively.

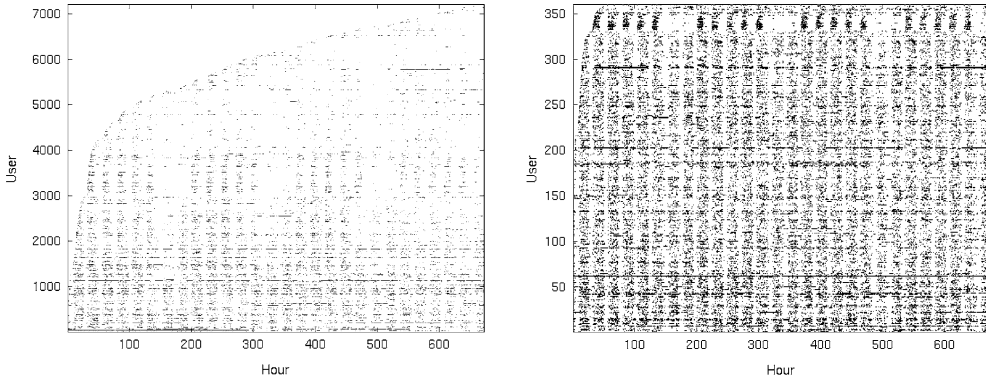

Fig. 1 This figure shows the hourly diameter of individual users over the 4 week trace. ''White'' denotes a diameter of zero, while ''black'' represents a diameter greater than zero. There are total of 7,213 users, among which 360 users do not have any 3-days gaps

(a) All users **(b)** Users without a three-day gap

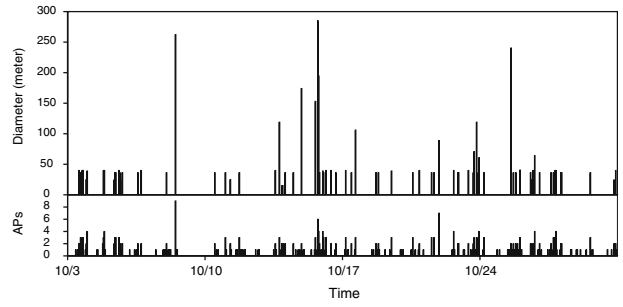

Fig. 2 Hourly diameter and APs visited by one user. This figure shows the user's hourly diameter and the number of unique access points visited by this user during each hour. Labels on the *X*-axis indicate the dates for Sundays

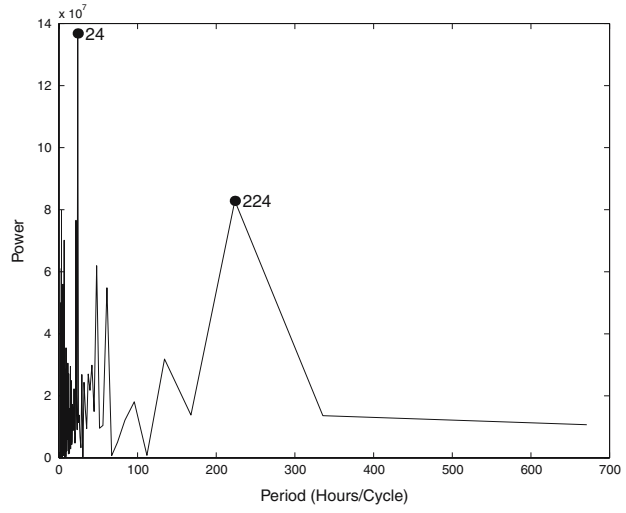

Fig. 3 Diameter in frequency domain. *Two dots* denote the two most strongest periods. In this example, they are approximately 24 and 224 h

Figure 2 shows the diameter of each hour of one user and the number of unique APs visited by the user during each hour over 4 weeks. The *X*-axis labels indicate the dates for Sundays, and the *Y*-axis shows the diameter and the number of APs. This user often had a diameter of 40 m. By looking into the trace, we found that the user was visiting a fixed set of APs repeatedly due to 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, we cannot distinguish whether an increase in this number is due to real movements or due to the ping–pong effect. Even when this user associated with up to four APs, the diameter was still around 40 m. On the other hand, in the third largest peak where the user moved around 240 m, he only visited two unique APs. Thus, the number of APs visited by the user does not accurately describe mobility.

Figure 3 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.

We transformed all of our users' diameter vectors using the DFT and recorded the two strongest periods. Figure 4 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 h. The distribution also has smaller jumps at the following hours: 84 (3 days and 12 h), 168 (1 week), 224 (9 days and 8 h), and 336 (2 weeks). Note that by using the DFT, we can observe a jump only at a period that is an integer fraction of the input length (672). We were not surprised to see users with 1 day, 1 week, or 2 weeks as their primary periods. But, it is interesting to observe more users with 3-days-and-12-h than 4 days. The users with the period of 9-days-and-8-h 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-h is an integer fraction; it is nonetheless interesting to observe users with this period as their primary or secondary period.

3.2.2 *Classification*

We used the two strongest periods as our first two elements of three-element input vectors to AutoClass.

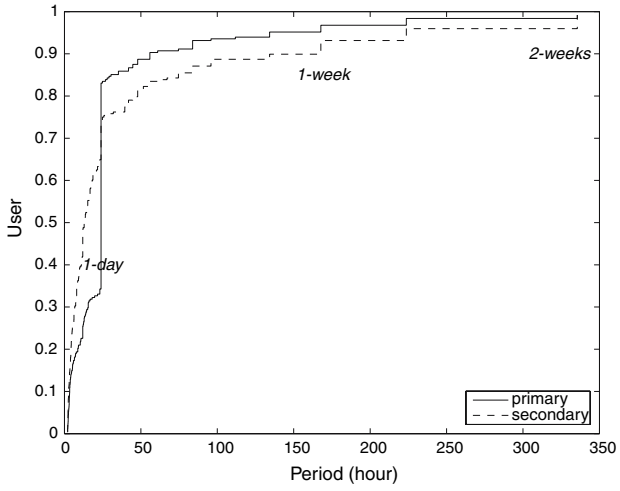

Fig. 4 Significant periods of user mobility. Cumulative distribution of the number of users versus period. From the power spectrum density graphs, we recorded the two most significant periods for each user

In addition to these two periods that we gathered from the DFT, we also measured the maximum hourly diameter ($d_{\max}$) observed over our trace for each user. As described in Sect. 3.1, we focus on the mobile set of users whose $d_{\max}$ was greater than or equal to 100 m.

AutoClass used these three parameters to classify the mobile set of 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.

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

First, we considered the group of users that have a strong 1-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 Fig. 5c. Class 1 represents users who have 1 day as their strongest period and a small secondary period. Students who have regular classes may exhibit this kind of mobility behavior. The average second period for class 2 is close to 2 days. The average for class 5 is close to 11 days, but this value is misleading; secondary periods of this class are bimodal around 1 and 2 weeks. Thus, class 5 can be described as a cluster of users with 1 day and either 1 or 2 weeks as their strong periods. Note that mobile users with 1 day as their strongest period and a small secondary period are most prevalent—Class 1 is the biggest class.

Second, we looked into the group of users whose primary period is not 1 day. These users are divided into four classes. As shown in Fig. 5d, classes 3, 0, 4, and 6 have smallest to biggest secondary periods, respectively. Class 6 consists of users with 9-days-and-8-h as the secondary period and the very small primary periods. It is interesting to note that most of the users whose primary period is not 1 day have their secondary period close to 1 day—Class 0 is the biggest class among these four classes.

3.3 Access points

We used the same method to classify APs based on how many visitors they had each hour, and particularly the periodicity of that metric.

3.3.1 Periodicity

Figure 6 shows the cumulative distribution of the number of APs with primary and secondary periods: 85% of APs had their primary period at 1 day (24 h); 25% of APs had their secondary period at 1 week (168 h). Compared to the user mobility (see Fig. 4), more APs have their primary period at 1 day and the secondary period at 1 week.

3.3.2 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

Table 1 Classes of user mobility

Class	Instances (No)	Instances (%)	Key parameter	Period 1 (h)			Period 2 (h)			Diameter (m)		
				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

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

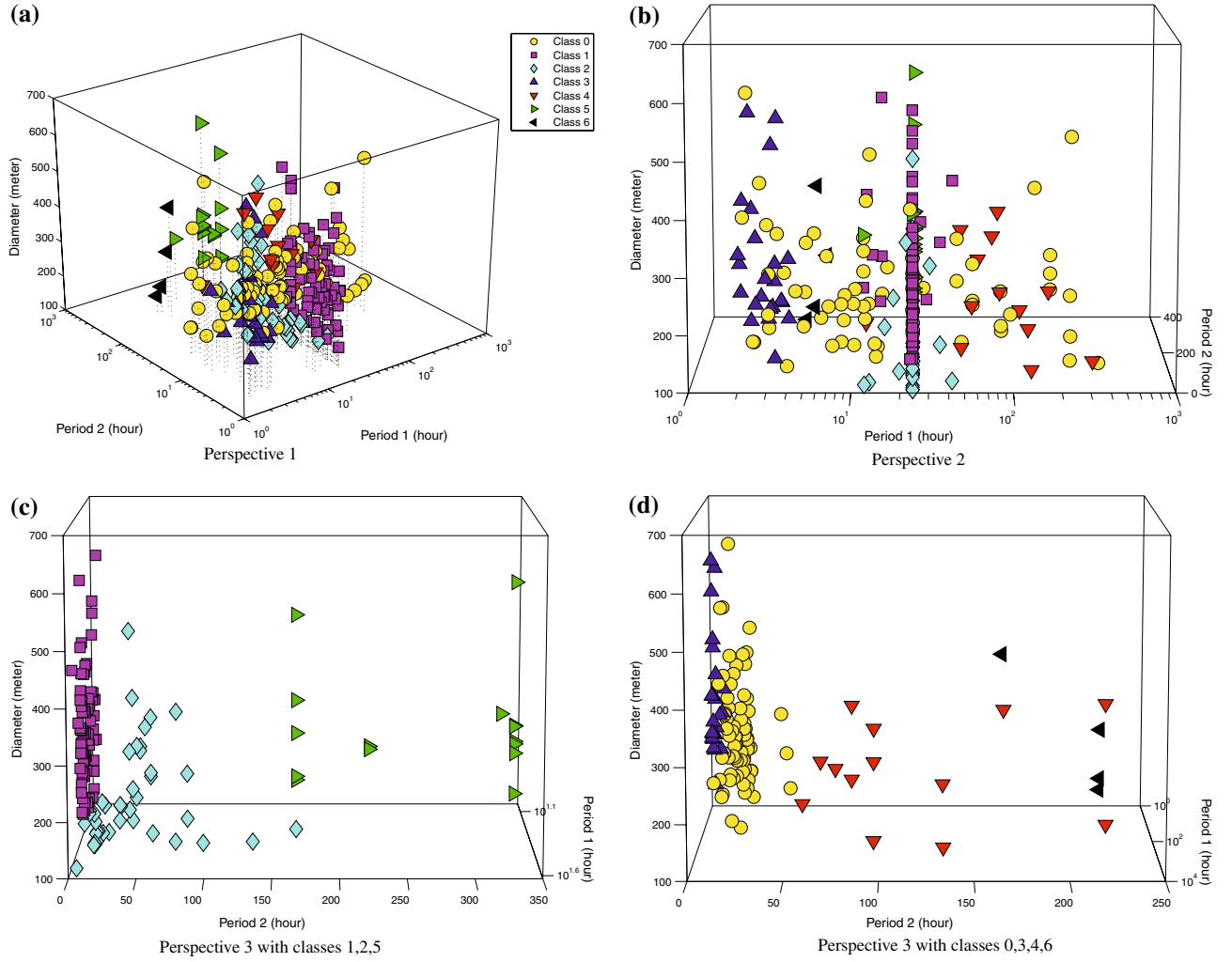

Fig. 5 Clustered users

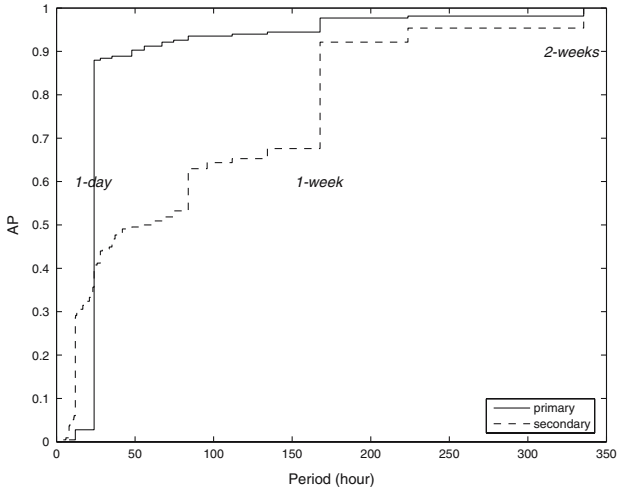

Fig. 6 Significant periods of APs. Cumulative distribution of APs versus period. From the power spectrum density graphs, we recorded the two most significant periods for each AP

classes. The last parameter ($u_{\max}$) did not make any difference in classifying the input cases. Thus, we did 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 h for most of the cases, and therefore did not play a critical role in determining to which class a case belongs.

Figure 7 shows each instance in three dimensions in two different perspectives. Because $u_{\max}$ did not play a major role for classification, we did not include it in this graph. Instead, we included 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.

Table 2 Classes of access points

Class	Instances (No)	Instances (%)	Key parameter	Period 1 (h)			Period 2 (h)		
				Mean	Std	CV	Mean	Std	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

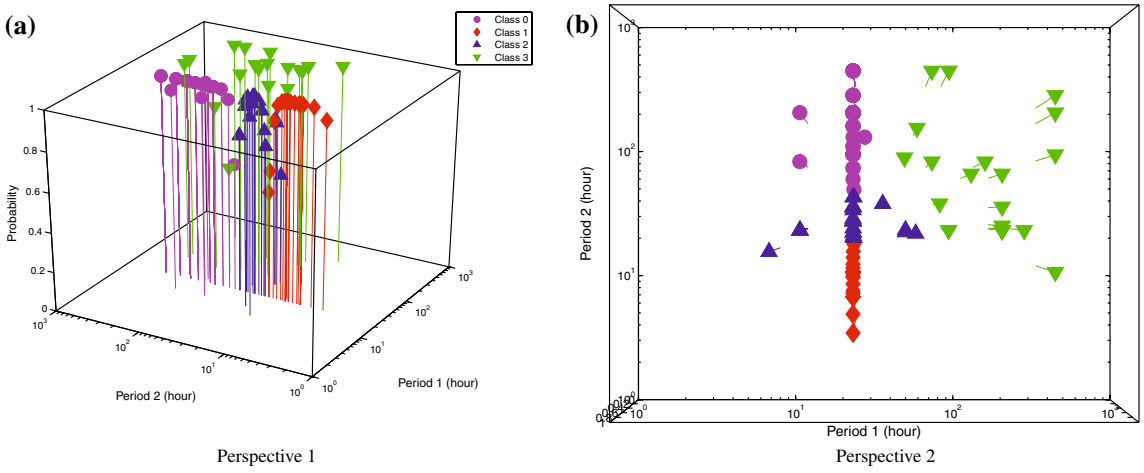

Fig. 7 Clustered access points

Figure 7 shows that most APs had their primary period at 1 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 1 day as their primary period and around 1 week as their secondary period were the dominant category. Class 3's primary period is much bigger than 1 day; its secondary period is also big.

Figure 8 shows the location of the APs on our campus. Many of the Cisco APs on our campus have recently been replaced by Aruba APs. Because we focused only on Cisco APs, Aruba APs were not included in the map. Out of 469 Cisco APs, we did not know the locations of ten APs. Thus, only 459 APs are marked. Because we did not classify the APs who never had more than 50 users per hour, only 216 APs were classified. Note that 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.

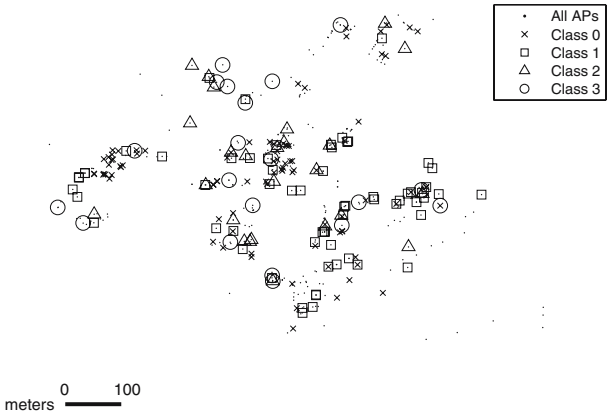

Fig. 8 Map of access points on campus

3.4 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. Because we expected some weekly periodicity, we used a 4-week trace instead of 1 month; we truncated the data to be multiple of 1 week (i.e., 168 h). For access points, we tried both a 4-week trace and 1-month trace. With the 4-week trace, an AP had 1 day as the strongest period and 1 week as the second. When we used the 1 month trace, we got the same value of 1 day for the first maximum, but got 1 week and 12 h for the second maximum instead of exactly 1 week.

Visualizing clustered data is important to understand results. 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 if more input parameters are used for classification.

4 Seasonal effects

To understand long-term seasonal effects, we used a longer trace, lasting a year, from 2 November 2003 to 30 October 2004. We chose 364 days instead of 365 days to make the length of our trace a multiple of a week. Our trace consists of syslog records collected at both Cisco and Aruba APs. We observed 17,522 MAC addresses visiting 780 access points (APs). Note that the total number of APs does not represent the number of active APs at a certain time. Because Cisco APs on the campus were replaced by Aruba APs during this time period, the number of APs observed during the year (780) is probably higher than the number of active APs at any given time.

On the Dartmouth campus, there are two types of always-on mobile devices: Cisco VoIP mobile phones and Vocera communicators. While laptops tend to be turned off while carried from place to place, these always-on devices are connected to the network even when users are moving. Thus, we can get a relatively accurate picture of how much these users move. These always-on VoIP devices constitute 1% of all devices; we observed activities of 39 VoIP phones and 128 Vocera communicators in our 1-year trace.

We applied the same method that we used for the 1-month trace to understand the cycles in the 1-year trace. To analyze user mobility, we computed the hourly diameter for each user. To analyze AP popularity, we counted the number of unique users visiting each AP during each hour. For each user or AP, we had a 8,736-entry vector, each entry representing the value for that hour. We then used the DFT to discover seasonal periodicities. From the DFT result, we chose the five strongest periods.

Analyzing periodicities of random patterns makes little sense. Thus, we identified users and APs with random behaviors using the randomness test in Sect. 4.1. We then explored seasonal effects in user mobility in Sect. 4.2 and analyzed the popularity of APs in Section 4.3.

4.1 Randomness

To test whether a series can be considered random, we used the autocorrelation test [3]. This test relies on notion that a random series should not be similar to its shifted versions. We computed autocorrelation of series (vectors) and analyzed the plot of the autocorrelation function as a function of lag, called the *correlgram*. For a random series, lagged values are uncorrelated and thus $r_k \cong 0$, where r_k is the autocorrelation coefficient at lag k. We decided that a series was random if all r_k were within the 95% confidence limits, except for $k = 0$.

Table 3 shows the result of this randomness test. Of the total 17,522 users, 31.3% never visited more than one AP during an hour. Among the 12,040 users who ever moved during the year, 75.9% of users show non-random behavior. 97.4% of all APs show non-random patterns. In analyzing periodicities, we focused on these non-random users and APs.

4.2 User mobility

Figure 9 shows an example of the hourly diameter computed for a typical user. The X-axis shows the calendar month and the Y-axis shows the diameter in a log scale. The black horizontal bars at the bottom graph show the time when the school was in session. We can see that this user's movement closely matches the academic calendar, except for the Summer term. This user was probably away from the campus during the summer. From the device's MAC address, we know that the device is not an always-on device; it is likely to be a laptop since most devices on the Dartmouth campus are laptops [7]. We expect that movements of other laptop users follow similar patterns.

4.2.1 Maximum of hourly diameter

To get a rough idea of users' movement patterns, we considered the maximum hourly diameter over the year for each user. Figure 10 shows the cumulative fraction of the maximum diameter over all users. Of all devices, 31.3% had a maximum diameter of zero; these devices never associated with more than one AP within each hour. Another interesting observation is that there is a knee towards 244 m, which is the outdoor signal range for Cisco APs. The number of devices increases slowly towards this value because devices are not affected by the ping–pong effect after this limit.

Table 3 Randomness test result

	Total	Mobile	Non-random	Random
Users	17,522	12,040	9,140 (75.9%)	2,900
APs	780	–	760 (97.4%)	20

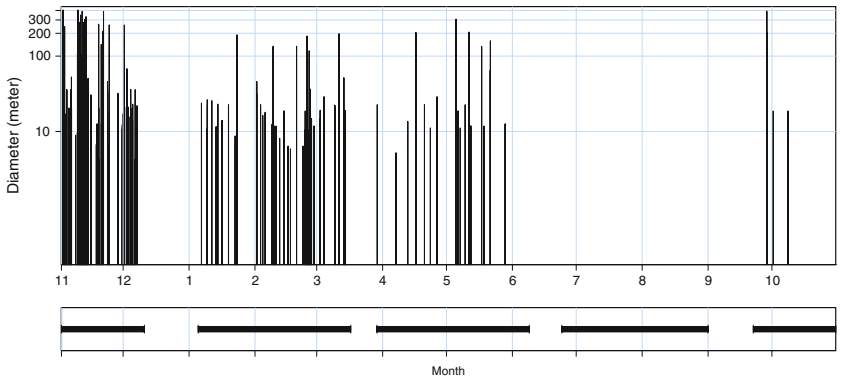

Fig. 9 Example of a user's hourly diameter. The *X*-axis shows the calendar months from November 2003 to October 2004. The *Y*-axis shows the diameter in a log scale. Each vertical line corresponds to a diameter for each hour. The horizontal bars denote the time when the school was in session

The median of the maximum diameters over all devices is 46.5 m.

Figure 10 shows the the cumulative fraction of maximum diameter over Cisco VoIP phones and Vocera communicators. Always-on devices have bigger maximum diameters than the diameters of the whole population of wireless devices. The medians of the maximum diameters for VoIP phones and Vocera communicators are 539.5 and 538.9 m, respectively. One interesting observation is that the line for Vocera communicators shows a plateau approximately from 50 to 260 m. This indicates that Vocera devices are divided into two groups: stationary and mobile. Most stationary Vocera devices have a diameter close to zero, while some stationary devices have a diameter greater than zero due to the ping–pong effect, which causes devices to associate with more than one AP even when they are not moving.

To understand whether user mobility changes across different academic terms, we divided our trace into four academic terms based on the Dartmouth's academic calendar. Figure 11a shows the cumulative fraction of the maximum diameter over all users for each academic term. Users are most mobile in the Fall term; Spring and Winter terms follow in order. Many users are not mobile during the Summer term because they are not on the campus.

We removed the users who were not on the campus for certain terms, as well as those users who never visited more than one AP per hour. We computed the sum of hourly diameter for each term and considered only the users whose sum is bigger than zero for all four terms; this reduced the number of user from 17,522 to 1,423. We considered a different metric, the average diameter over all hours with a non-zero diameter; this average gives a sense of how much people move when they do move. Figure 11b shows the cumulative fraction of this average diameter over 1,423 users. Users' mobility increased from Winter to Fall, Fall to Spring, and Spring to Summer. This difference may be due to different weather conditions, though there are many possible explanations.

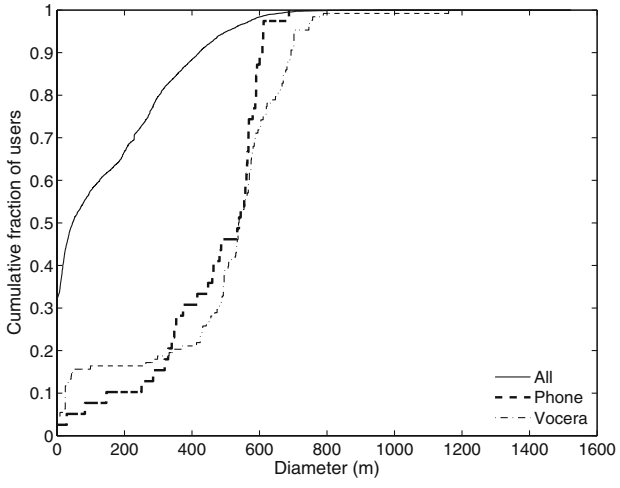

Fig. 10 Cumulative fraction of maximum diameter over users for 1 year. "All", "Phone" and "Vocera" denote every device, Cisco VoIP phones and Vocera communicators, respectively

4.2.2 *Periodicities*

We used the DFT to discover seasonal periodicities. We first considered the DFT result of our sample user from Fig. 9. This user has peaks (not shown)—in order from the strongest to the weakest—at 1 day, 3, 6, 12 and 2 months. Note that this user's secondary period is 3 months, which corresponds to Dartmouth's quarter calendar.

Figure 12a shows the cumulative fraction of each peak period over 9,140 users. Note that we excluded users who never moved during the year and whose behavior is random. Peaks 1 through 5 are the five strongest periods in descending order of strength. All five peaks have many users with the period of 24 h, while the period of 168 h (1 week) is negligible. It is interesting to note that the lines are roughly in order,

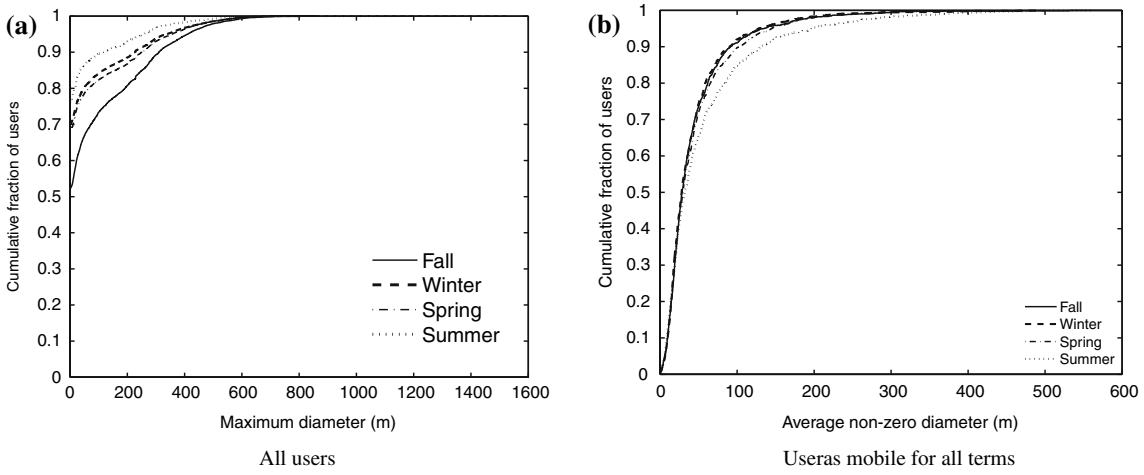

Fig. 11 Mobility for each academic term. **a** Shows the cumulative fraction of maximum diameter over all 17,522 users. **b** Shows the average non-zero diameter for 1,423 users who have diameters bigger than zero for all four terms

with the strongest period at the bottom. This is partly because stronger peaks have more users around 24 h while weaker ones have diverse periods as their values. This diversity makes slopes around 24 h steeper for weaker peaks. Also, note that Peak 1, 2, 3, 4 and 5 have big jumps at 12, 6, 4, 3 and 2.5 months, respectively. We currently do not have an explanation why five peak periods are in a descending order with these specific values.

To see clearly significant periods with a large number of users, we aggregated data of the five peaks. Figure 12b shows the cumulative fraction of strong periods over users. The *X*-axis shows the period in hours in a log scale. ''All'' corresponds to aggregated values from Peaks 1 to 5. We also extracted users with diameters greater than 100 and 244 m. The threshold of 100 m corresponds to the value that we used earlier in Sect. 3.1, and 244 m denote the outdoor signal range for Cisco APs. This filtering reduced the number of users from 9,140 to 6,320 and to 4,289 for the threshold of 100 and 244 m, respectively. All three lines have big jumps at 1 day, 3, 4, 6, and 12 months; jumps at 1 week are relatively small. Compared to the ''All'' case, results filtered with thresholds contain a fewer number of users with the peak period less than 24 h, and big jumps around 3 months. Jumps around 3 months reflect the Dartmouth college's quarter calendar.

4.3 Popularity of access points

To understand seasonal effects on the popularity of access points (APs), we applied the same technique used for user mobility. We first observed the maximum of hourly visitors and then considered periodicities.

Figure 13 shows an example of hourly visitors to a typical AP. The *X*-axis is the calendar month and the *Y*-axis is the number of visitors in a log scale. The horizontal bars show the time when the school was in session. APs experienced more visitors while the school was in session. During the breaks (except for Christmas), this AP still had some visitors although the number of visitors was reduced. This is because some graduate students and faculty are on the campus during the breaks although most undergraduate students leave the campus. Another periodic pattern is the weekly repetition; the number of visitors reduced to zero during most weekends.

Figure 13 shows a gap from March 18 to April 19 of 2004; this gap is due to Cisco AP failures while upgrading OS from VxWorks to IOS. The duration of failures for many APs was elongated because APs did not work properly on the radio side while they were alive on the wired side. Thus, it took network administrators a while to discover malfunctioning APs. During this time period, only 185 APs out of 780 APs worked normally. We expect that understanding periodic behavior of APs can help in detecting anomalies in the future.

4.3.1 Maximum of hourly visitors

Figure 14 shows the cumulative fraction of the maximum of the hourly number of visitors over the year for each AP. Unlike the user diameter distribution (see Fig. 10), this graph does not have any knee. The graph is heavy tailed; 2% of APs have a maximum value of over 1,000 visitors per hour. The median of the maximum visitors over all APs is 118 visitors, and the maximum is 3,344 visitors.

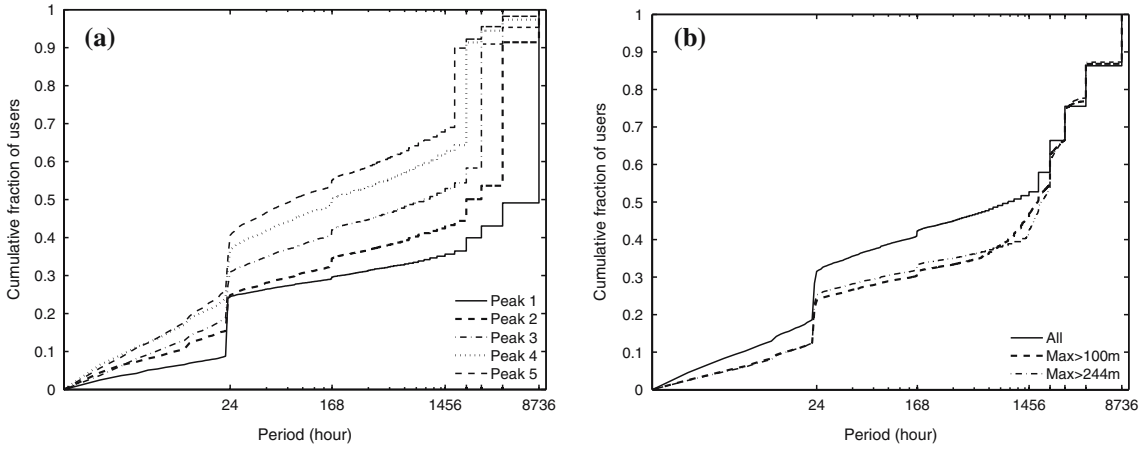

Fig. 12 Periodicity in user mobility. The *X*-axis shows the period in hours in a log scale. The *Y*-axis shows the cumulative fraction of mobile users. **a** Shows the cumulative fraction of five strongest periods of hourly diameters over users. Peaks 1 though 5 are the five strongest periods in descending order of strength. The graphs are roughly in order, with the strongest period (peak 1) at the bottom. **b** Shows the cumulative fraction of the aggregated data of the five peak periods over users. ''All'' corresponds to all 9,140 users. ''Max > 100 m'' and ''Max > 244 m'' denote users whose maximum is greater than 100 and 244 m, respectively

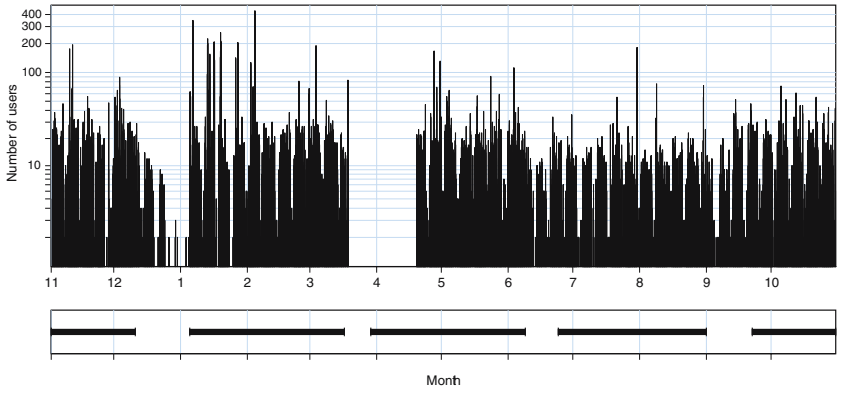

Fig. 13 Example of an AP's number of visitors. The *X*-axis shows the calendar months from November 2003 to October 2004. The *Y*-axis shows the number of visitors in a log scale. Each vertical line corresponds to a number of visitors during each hour. The horizontal bars denote the time when the school is in session

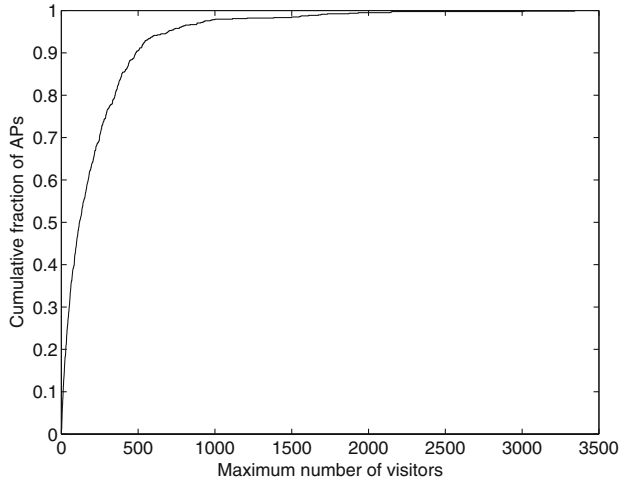

Fig. 14 Cumulative fraction of the maximum number of visitors over all APs

Figure 15 shows the cumulative fraction of the maximum number of visitors over APs for each academic term. APs are busiest during the Fall term and least busy during the Summer term. (There are fewer faculty and students on campus in Summer term.) Winter and Spring terms show similar distributions.

To understand whether the geographical distribution of popular APs changes across different terms, we plotted the APs on geographical coordinates, with six different symbols for ranges of maximum-hourly-visitors values: [0], (0,50], (50, 100], (100, 150], (150, 200], and (200,∞). Figure 16 shows coded APs for Fall, Winter, Spring, and Summer terms. Arrows in Figure 16a denote areas that show distinct differences across terms. The top arrow points a computer science building. This building is popular during the Spring and Fall terms, due to classes held in the building. The other two arrows point areas that are mostly under-

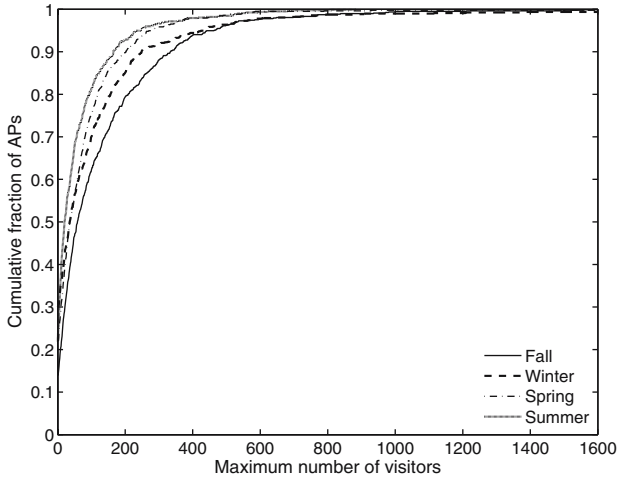

Fig. 15 Cumulative fraction of the maximum number of visitors over all APs for each academic term

graduate student housing. The area denoted by the left arrow (Tree houses) is busiest during the Fall and least active during the Winter. The area denoted by the right arrow (Ripley–Woodward–Smith Cluster) is not active during the summer. In short, while the popularity of most APs does not change across terms, that of some APs changes dramatically; types of buildings where APs are located may affect these seasonal variations.

4.3.2 *Periodicities*

We used DFT to observe periodicities in the number of hourly AP visitors. We start with our example AP (see Fig. 13). This AP has peaks (not shown) at 1 day, 1 week, 3 months, 21 and 84 h. While both the user diameters and the AP visitors have the strongest period of 1 day, they have different secondary periods: 3 months for the user diameter and 1 week for the AP visitors.

Figure 17a shows the cumulative fraction of five strongest periods of hourly visitors over all APs. The *X*-axis shows the period in hours in a log scale. Peaks 1 though 5 are the five strongest periods in descending order of strength. Most peaks have big jumps at 24 h; Peak 1's jump is especially big. All of them also have smaller jumps at 168 h (1 week). As with the user diameter (see Fig. 12a), Peak 1, 2, 3, 4 and 5 have big jumps at 12, 6, 4, 3 and 2.5 months, respectively.

We aggregated values of the five peaks to observe significant periods. Figure 17b shows cumulative frac-

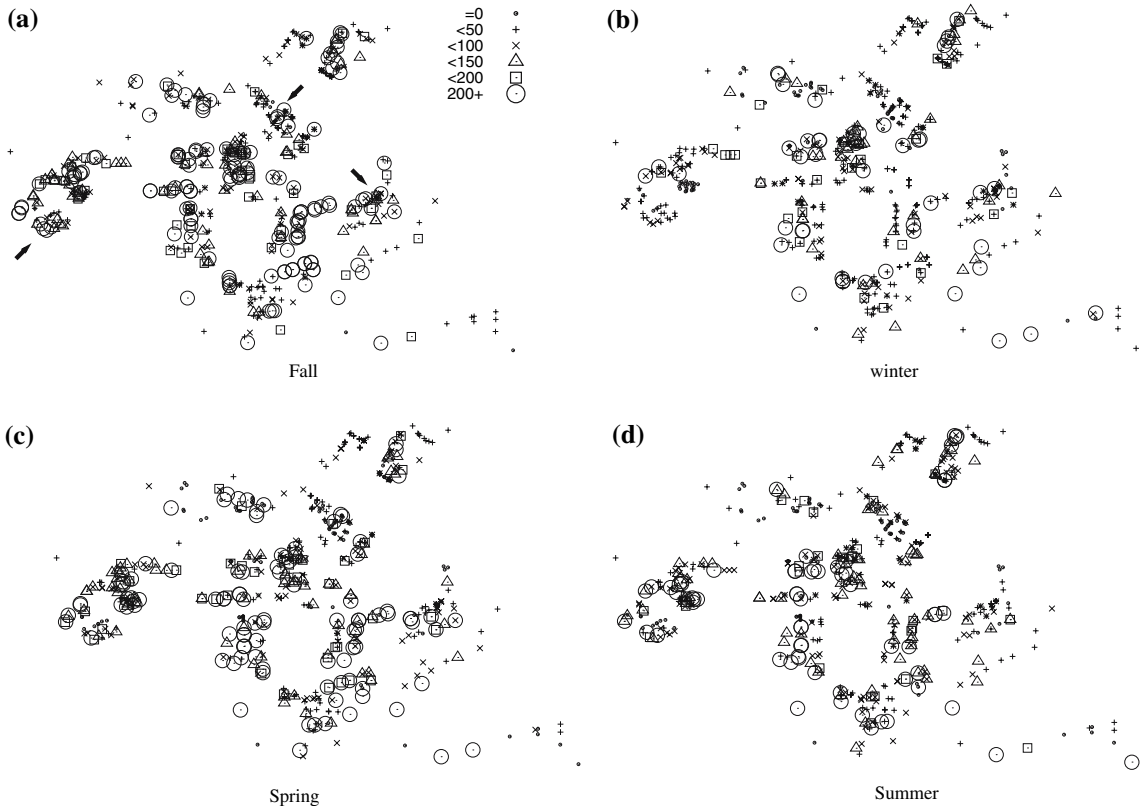

Fig. 16 Geographical shift in AP popularity for four academic terms. *Three arrows* in **a** denote areas whose popularity changes significantly across different academic terms

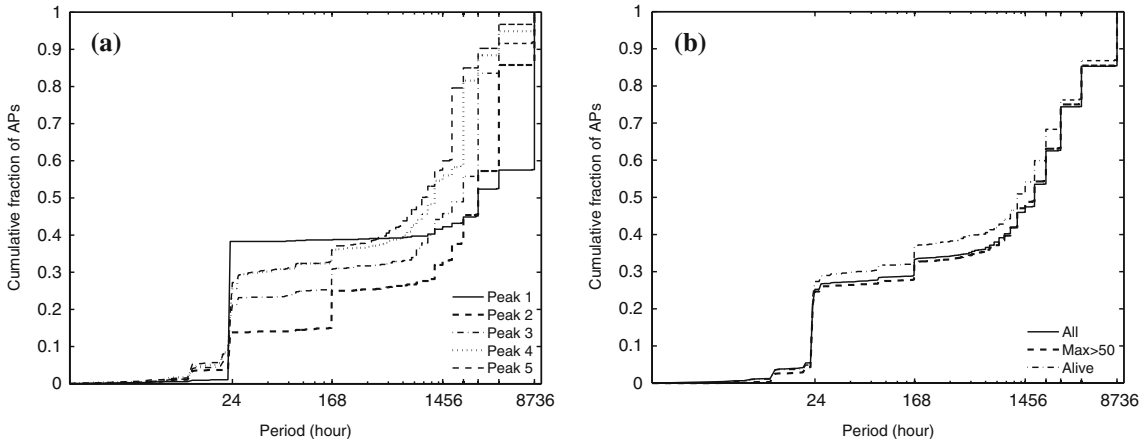

Fig. 17 Periodicity in APs. **a** Shows the cumulative fraction of five strongest periods of hourly visitors over APs. Peaks 1 through 5 are the five strongest periods in descending order of strength. **b** Shows the cumulative fraction of the aggregated data of the five peak periods over APs. ''All'' denotes all 760 APs. ''Max > 50'' shows APs whose maximum visitors is bigger than 50. ''Alive'' shows the APs that were alive during the AP radio trouble in March and April of 2004

tion of strong periods over APs. The *X*-axis shows the period in hours in a log scale. ''All'' denotes all 760 APs. ''Max > 50'' shows 550 APs whose maximum visitors were bigger than 50. ''Alive'' shows 185 APs that were alive during the AP radio trouble in March and April of 2004. All three lines have their biggest jump at 1 day and have smaller jumps at 1 week, 2.5, 3, 4, 6 and 12 months. Compared to ''All'' and ''Max > 50'', ''Alive'' has a relatively *smaller* jumps at 12, 6, 4, 3, and 2.5 months. We expect that this is because ''Alive'' has more APs with regular periodic patterns than the other two groups. Apparently, the gap during March and April affected periodicities of those failed APs.

4.4 Summary of seasonal effects

We looked into user mobility and AP popularity patterns in a 1-year trace. Maximum values provided insights into a overall population makeup and DFT revealed periodicities. Our findings can be summarized as following:

- About 31% of all users never visited more than one AP during an hour.
- Of the users who moved, 24% had random mobility; on the other hand, mobility behavior of less than 3% of all APs was random.
- User mobility and AP popularity changed based on the academic calendar.
- User mobility changed from one academic term to another, possibly due to different weather; mobility increased from Winter to Fall, Fall to Spring, and Spring to Summer.
- Always-on devices had bigger maximum diameters than the whole population of wireless devices.
- Thirty percent of non-random users had a strong period of 1 day, while only five had a strong period of 1 week. 32% also showed a strong period of 3 months, which corresponds to the Dartmouth's quarter academic calendar.
- Sixty-five percent and 21% of APs had a strong period of 1 day and 1 week, respectively.

5 Conclusions and future work

In this article, we present a method to extract information from real wireless network traces by transforming the time series to the frequency domain using the Fourier transform. We extracted the two most significant periods from 1-month trace and clustered instances using a Bayesian classification tool. We also looked into long-term seasonal effects by analyzing a 1-year trace. Our study is unique in using the Fourier transform and Bayes' theory to provide insights into user mobility and the behavior of access points. We were able to understand the periodic nature of users on the Dartmouth's wireless network, and we expect that our method would be useful on similar traces at other locations.

We hope that our findings provide a base for modeling user mobility. One approach for modeling is to perform the inverse DFT to obtain the time series that represents each class as mentioned earlier in Sect. 1. Another approach is adapting autoregressive moving average models [4]. Generally, a time series may consist of following components [6]:

$$\text{series} = \text{seasonal cycles} + \text{trend} + \text{regression term} + \text{irregular effects}. \quad (4)$$

In this article, we focused on the first component, seasonal cycles. We divided traces into short-term and long-term traces, and analyzed periodicities. We found that while a daily pattern is common among both users and APs, a weekly pattern is common only for APs. For users, we also found a strong period of 3 months, which corresponds to the Dartmouth's quarter academic calendar. We plan to analyze other components in Eq. 4 in the future. For example, we will look into whether user mobility has a trend such as a constant increase from 1 year to the next.

In the future, we plan to build generalized models for user mobility. We believe that our method will help us build models by identifying some of the significant characteristics, by clustering users into groups that need different models or different parameters, and by abstracting traces.

Acknowledgments This project was supported by Cisco Systems, NSF Infrastructure 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.

References

1. Agrawal R, Faloutsos C, Swami AN (1993) Efficient similarity search in sequence databases. In: Lomet D (ed) Proceedings of the 4th international conference of foundations of data organization and algorithms (FODO), pp 69–84. Springer, Chicago
2. Balazinska M, Castro P (2003) Characterizing mobility and network usage in a corporate wireless local-area network. In: Proceedings of the first international conference on mobile systems, applications, and services (MobiSys). San Francisco, May 2003, pp 303–316
3. Beker H, Piper F (1985) Secure speech communications. Academic, New York
4. Box G, Jenkins G (1970) Time series analysis: forecasting and control. Holden-Day
5. Cheeseman P, Stutz J (1996) Bayesian classification (AutoClass): Theory and results. In: Fayyad UM, Piatetsky-Shapiro G, Smyth P, Uthurusamy R (eds) Advances in knowledge discovery and data mining. AAAI /MIT, Philadelphia
6. Congdon P (2001) Bayesian statistical modelling. Wiley, UK
7. Henderson T, Kotz D, Abyzov I (2004) The changing usage of a mature campus-wide wireless network. In: Proceedings of the 10th annual international conference on mobile computing and networking (MobiCom). ACM Press, Philadelphia, pp 187–201
8. Jain R, Lelescu D, Balakrishnan M (2005) Model T: an empirical model for user registration patterns in a campus wireless LAN. In: Proceedings of the eleventh annual international conference on mobile computing and networking (MobiCom). ACM Press, Cologne, pp 170–184
9. Paxson V (1995) Fast approximation of self similar network traffic. Technical Report LBL-36750
10. Press WH, Teukolsky SA, Vetterling WT, Flannery BP (1992) Numerical recipes in C: the art of scientific computing. Cambridge University Press, Cambridge
11. Tang D, Baker M (2000) Analysis of a local-area wireless network. In: Proceedings of the sixth annual international conference on mobile computing and networking (MobiCom). ACM Press, New York, pp 1–10
12. Tang D, Baker M (2002) Analysis of a metropolitan-area wireless network. Wireless Netw 8(2-3):107–120