Mining Frequent and Periodic Association Patterns

Guanling Chen, Heng Huang, and Minkyong Kim

{glchen, hh, minkyong}@cs.dartmouth.edu

Dartmouth College Computer Science Technical Report TR2005-550

July, 2005

Abstract

Profiling the clients' movement behaviors is useful for mobility modeling, anomaly detection, and location predication. In this paper, we study clients' frequent and periodic movement patterns in a campus wireless network. We use offline data-mining algorithms to discover patterns from clients' association history, and analyze the reported patterns using statistical methods. Many of our results reflect the common characteristics of a typical academic campus, though we also observed some unusual association patterns. There are two challenges: one is to remove noise from data for efficient pattern discovery, and the other is to interpret discovered patterns. We address the first challenge using a heuristic-based approach applying domain knowledge. The second issue is harder to address because we do not have the knowledge of people's activities, but nonetheless we could make reasonable interpretation of the common patterns.

1 Introduction

Wireless Local Area Network (WLAN) has been increasingly deployed on enterprise and academic campuses, given WLAN's flexibility, ability for incremental growth, and potential for great cost savings. A large-scale deployment involves many (potentially overlapping) cells, each covered by a wireless access point (AP). A mobile client in such a network may frequently change its associated APs for network access, depending on its location. By observing and studying the sequence of a client's associated APs, we can obtain insights on users' mobility behaviors from seemingly random associations. The applications of profiling clients' movements include mobility modeling, anomaly detection, and location prediction.

Existing mobility models work at micro-level and assign limited semantics to the generated sequences. By profiling user movements, we may be able to adjust the mobility models to produce realistic long-term properties, such as the frequent and periodic patterns discovered in this paper. The profiles can also be used to find anomalies in user movements, which may be a sign of possible intrusion. Location predication algorithm commonly uses Markov model that considers frequencies of previous movements, and its accuracy could be improved with other event context such as periodicity.

In this paper we focus on studying both frequent and periodic patterns by data-mining clients' association sequences. Our results reflect an academic-centric environment, where much of the mobility is driven by students moving between residential, academic (classrooms), and library areas. Some of results we got make sense after simple reasoning. For instance, though we know that many residential buildings had large amount of wireless usage while there were few usage for some other parts of campus, we do not expect to see exponential differences. Thus the popularity of frequently visited locations would not follow power-law distribution. There is, however, no intuitive explanation why the popularity of frequent movements did follow power-law distribution. It is also not surprising for us to discover that the most prominent periodic patterns were weekly patterns. However, we also saw some abnormal behaviors such as a client visited several locations every 12.5 days. To further reveal these anomalies, we may have to gather users' true activities using more active techniques beyond passive observations, such as the Experience Sampling Method proposed by Henderson et al. [3].

In Section 2, we discuss our data collection and how we addressed the challenge of removing noise from clients' association history. We give definitions of frequent and periodic patterns in Section 3. Section 4 and Section 5 present our experimental results. Finally, we discuss related work and conclude in Section 6 and Section 7, respectively.

2 Data Preparation

There are about 550 access points covering around 190 buildings on our campus. These APs send syslog messages to a server. The syslog messages contain clients' association and disassocation events. For the purpose of this paper, we studied a 60-day syslog trace from January 1, 2004 to February 29, 2004 (inclusive). Note that the changing of APs does not necessarily mean the client has physically moved; it may be due to so-called "ping-pong" effect. Namely, a client may decide to associate with a new AP if its perceived signal strength from current AP drops below certain threshold. The degrading of signal strength, however, may be caused by many other reasons than mobility, such as interference and multipath effects. Thus a poorly positioned client, even physically stationary, may continuously associate back and forth with a set of nearby APs. So the challenge is to pre-process the syslog trace to find appropriate clients that have meaningful movement patterns, and to remove the ping-pong noise from the traces.

Many clients only appeared in our 60-day trace for a short time; they may be guests visiting our campus. We say a client is *active* in a particular day if it associated with at least one AP. We found that the median number of active days of all clients was 23. We also observed that some clients associated with only a small set of nearby APs, exhibiting limited mobility. The median total number of APs a client ever associated with is 9 and the average is 15.5. About 20% of clients associated with only one or two APs in their entire trace. To measure clients' mobility, we use *diameter* defined as the maximum distance of any two APs that appeared in its association history, similar to the definition used in [4]. The median of all diameters over 60 days is 228.9 meters and the average is 242.5 meters. Note that the diameter we computed may be smaller than the true value, since there are 55 APs (out of total 537 APs) in the trace that we do not know their coordinates. Also the diameter is two-dimensional, so two APs on different floors may appear to be very close on a two-dimensional plane. Since we are mostly interested in client's frequent and periodic (such as weekly) patterns, we focus our study on the clients who were active for at least 30 days and whose diameter was greater than 50 meters. This reduced the number of clients from 5,979 to 2,259.

It is difficult to detect ping-pong behavior by passively observing the association sequences of the clients. We took a heuristic-based approach to convert the association sequence to a higher-level sequence that reflects physical movements. As we scan through a client's association sequence, we computed a group of APs that the client was likely to switch back and forth due to the ping-pong effect. When a new AP appears in the sequence, we put it in the existing group if the distance of the new AP to any AP in the group is less than 50 meters, or the client's moving speed to the new AP is greater than 10 meters per second (unlikely for human walking). Otherwise, we choose the *significant* AP from the existing group; the significant AP in a group is the AP on which the client spent most of its time. We then start a new group with the new AP.

After scanning all 2,259 clients' traces, we reduced the total 3,350,443 association records to only 554,566, each of which can be considered as a representative of an AP group identified using previous approach. Of all these AP groups, there are 384,074 groups containing one AP. Of those groups with more than one APs, about 95% of them has less than 5 APs and about 0.1% of them has more than 10 APs. The maximum group contains 17 APs. Figure 1 shows distribution of the ratio of the time spent by a client on the significant AP to the time it spent over the entire AP group. It shows that only 6% of the groups whose ratio is less than 50%, and about 54% of the groups has the ratio greater than 80%. This means that the significant APs we picked are indeed outstanding from the group and could be used to approximate the client's current location.

3 Pattern Definitions

We can view a client's sequence of association sequence as a time series,

$$A_1, A_2, A_3, ..., A_n$$

where A_i is an AP the client had associated with. For any $0 < i < n$, we require $A_i \neq A_{i+1}$. We define (A_i, A_{i+1}) as a single *movement*, and a *pattern* with length k to be another sequence of APs, such as $B_1, B_2, ..., B_k$. We say pattern

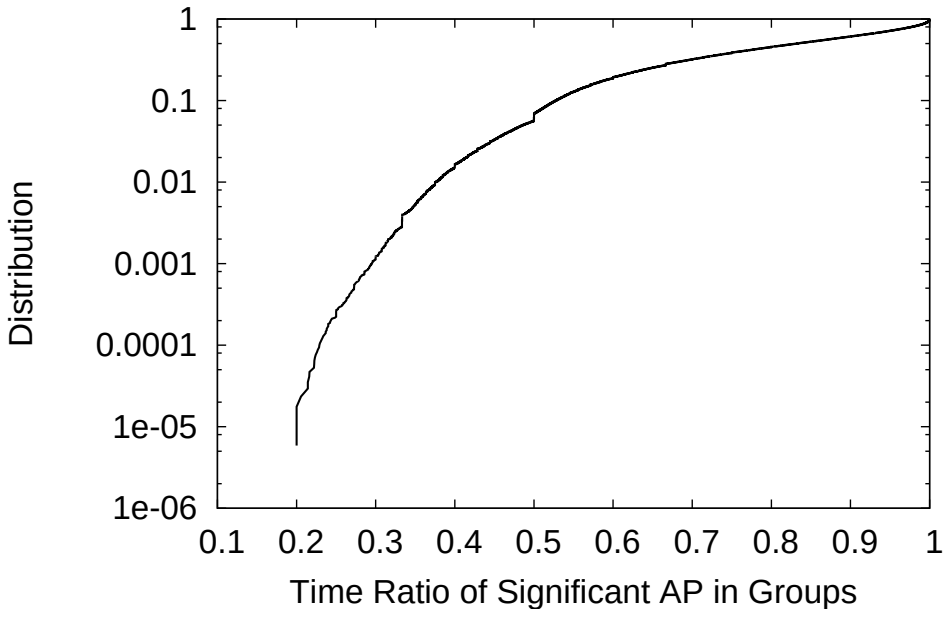

Figure 1: Ping-pong group statistics.

B appears in client's association sequence A, if there exists i ($0 < i < n$) such that

$$A_i = B_1, A_{i+1} = B_2, ..., A_{i+k-1} = B_k.$$

Since we can easily convert the name of AP to the name of the building in which that AP is installed, we can also represent both association sequences and patterns in terms of buildings.

Because there are not many matches of the patterns with length longer than two, we only consider patterns with short lengths (1 and 2) in this paper. We also only considered exact pattern matching. Given a parameter k as the desired pattern length, we can simply scan once through the whole association sequence to count the number of times all patterns with length k appeared in that sequence. It is easy then to find the most frequent patterns for each client and for all clients by aggregating the counts. There are other efficient data-mining algorithms to find patterns with inexact matches [7] and to find infrequent patterns without specifying desired length [5]. We plan to explore these algorithms in the future work.

A periodic pattern is the pattern that appears in a time sequence with the interval p, called *period*. This requirement is generally too strong for practical reasons. First, we need to add an error bound δ to the interval. So as long as the pattern always appears within interval $[p-\delta, p+\delta]$, then we say it is a periodic pattern with period p. Second, we need to consider occasional absence of patterns. For instance, a regular class might be canceled because the instructor is sick. In this case, we still want to consider the occurrence of this class as periodic patterns. This requires an algorithm to partition the whole timeline into on-segments and off-segments, so the pattern appeared periodically with p over all on-segments but did not occur during any off-segments. Such a pattern is called *partial periodic*, and a data-mining algorithm can find all periodic patterns without specifying the period as input parameter [8]. This approach is better than Fast Fourier Transform (FFT), which has large computational complexity and does not cope well with random off-segments.

We implemented this algorithm to find all periodic patterns with length one, and chose the parameter δ to be one hour. Note that this algorithm may choose arbitrarily long off-segments and lead to undesirable output. For instance, if a pattern appeared twice both on the first day and the last day with half an hour interval, the algorithm would report period p to be half an hour with all intermediate 58 days considered an off-segment. To avoid this problem, we filter out the output as follows. For a given period p, we computed the maximum possible number of occurrences during 60 days to be M. We then only consider a periodic pattern valid if it occurred at least $M * 0.75$ times.

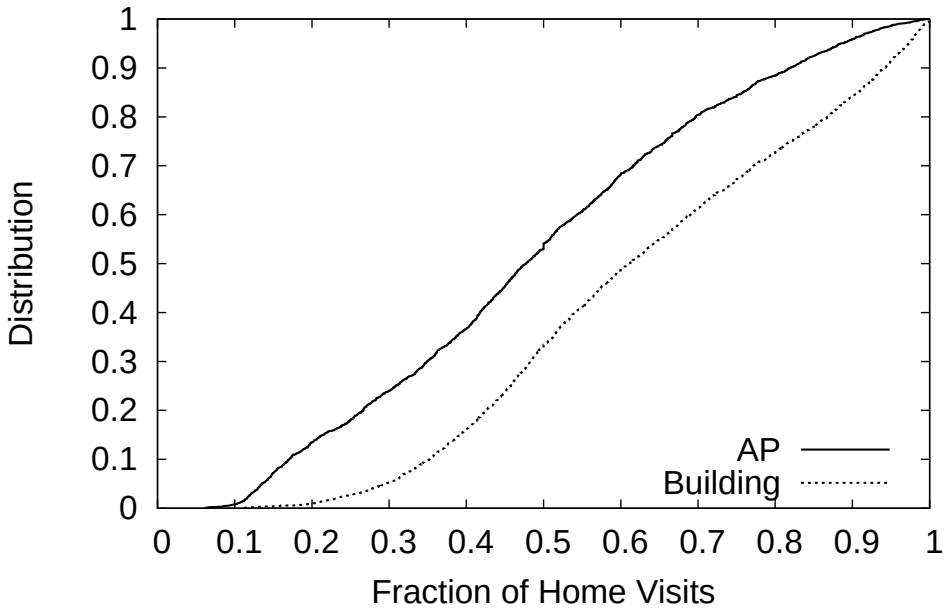

Figure 2: Fraction of clients' home visits.

4 Analysis of Frequent Patterns

We scanned each client's association sequence and computed the frequencies of all patterns occured. The pattern with length one is the AP or building a client visited, depending on the representation of the association sequences. The pattern with length two is a client movement. We define the "home" location of a client as its most frequently visited AP or building. Note that we use the number of visits to define home location, while others use the duration of visits [1, 4]. We call the mostly frequently occurred movement as client's *top movement*.

The most popular home APs and buildings are in academic and residential areas. Figure 2 shows the fraction of a client's visits to its home location, either a particular AP or building. As expected, a client visited its home building more often than its home AP since a building has a coarser granularity. The median visiting fraction to home AP is 48% while the median visiting fraction to home buildings is 61%. We aggregated the home APs and buildings across all clients, and found that the most popular home APs and buildings are in academic and residential areas. Among the top 10 most popular home APs (out of 339 in total), 4 are from academic buildings and 6 are from residential buildings. Among the top 10 most popular home buildings (out of 129 in total), the top two are academic buildings while the rest are residential. This divides clients into two groups: one likely to be the students living on campus and the other likely to be the students taking classes or staffs working in academic buildings.

Distributions of the popularity of home APs and buildings are not power-law. We now look at the popularity of APs and buildings serving as home locations to clients. We computed the number of clients having same home locations and plot them in Figure 3 (both axis in log scale). The figure reads as there are fraction Y of home locations who are shared by more than X clients. Neither curve follows a straight line, indicating that the popularity of home locations does not follow a power-law distribution. The median number of clients having same home AP is 4 and the maximum number is 49. On the other hand, the median number of clients having same home building is 8 and the maximum number is 177. The most popular home AP is also in the most popular home building, which turns out to be an academic building where the business school holds regular classes. This building has always been a " hotspot" on our campus since most of the business-school students own a laptop (often with wireless enabled).

The top movements are among libaray, academic, and residential areas. The number of unique movements between APs made by clients is generally small: the median is 6 but the maximum is 430. The median and maximum number of unique movements between buildings is 5 and 285, respectively. Figure 4 shows the distribution of the ratio: the number of occurrence of top movement divided by the total number of movements. The median of AP-level movements is 22% and the median of building-level movements is 32%. Both are much lower than the counterparts of distributions for home locations (Figure 2). We observed that the top movement by most clients were made between

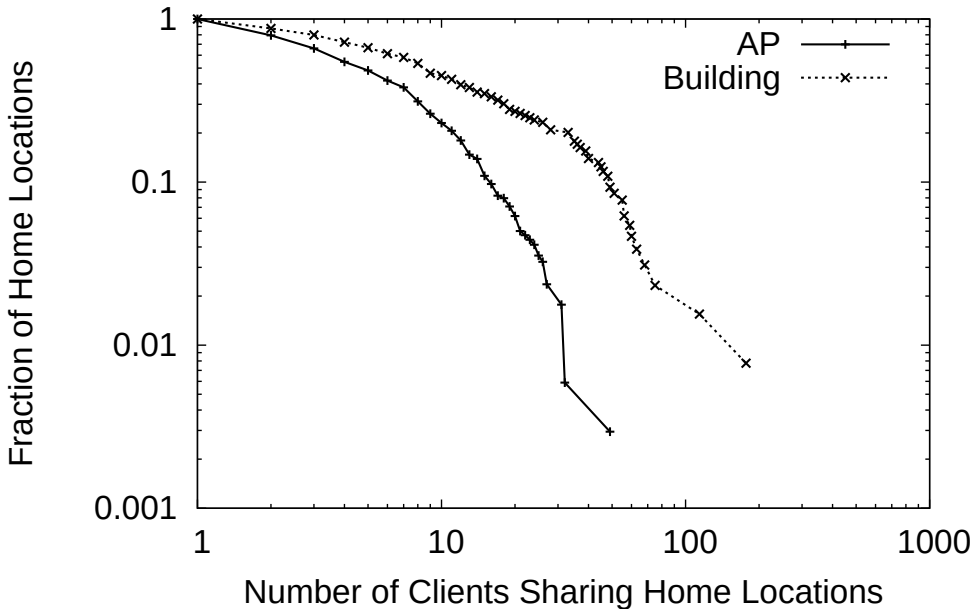

Figure 3: Popularity of home locations.

residential buildings and academic or library buildings, which makes sense for the students who live on campus taking classes and studying in libraries. The top movements of about 60% of clients were made either from or to a residential building. Figure 5 shows the distribution of the distance of clients' top movements. We note that about 13% movement distances is less than 50 meters, or less than the threshold we used for defining ping-pong groups (Section 2). One possible explanation is that the client associated with a new AP and created a new group, but the significant APs of the new and old groups were still close by. It is hard to tell whether changing of that AP was due to client moving away and back, or simply a ping-pong with longer distance. However, the distance of over an half (54%) of the top movements is longer than 100 meters.

Distributions of the popularity of top movements are almost power-law. Previously we observed that as many as 177 clients shared the same home building. Here, we note that less people shared the same top movement. Only 11 clients shared same AP-level top movements and 30 shared same building-level top movements. Figure 6 shows that the popularity of the shared top movements follows a power-law distribution with some deviation at the tail. At this point, we do not have a clear explanation of why the popularity distribution of home locations does not follow power-law while the distribution of top movements does.

5 Analysis of Periodic Patterns

We ran the pattern recognition algorithm described in Section 3 over every client's association sequences. We discarded the reported periods longer than 15 days since it only occurs at most 4 times in our 60-day trace and it is hard to interpret them as real regular visits.

The number of clients with regular periods is small, but half of the APs and buildings are visited periodically. We found that only 102 clients had periodic patterns at AP-level and only 24 clients had periodic building-level patterns. However we also found 322 APs and 97 buildings had regular periods, indicating about half of the APs and buildings on our campus were visited periodically at least by some clients. This indicates that if a client ever had periodic patterns, it tended to visit more than one location periodically. Figure 7 shows the distribution of the number of location a client visited periodically. The median and maximum numbers are 6 and 89 for AP sequences, and 8 and 48 for building sequences, respectively.

APs in academic buildings have most periodic visiting clients. Figure 7 shows the distribution of the number of periodic visiting clients to the same location. The median number is 2 for both APs and buildings sequences, while the maximum values are 27 and 10, respectively. Note that the number of periodic clients to the same AP is

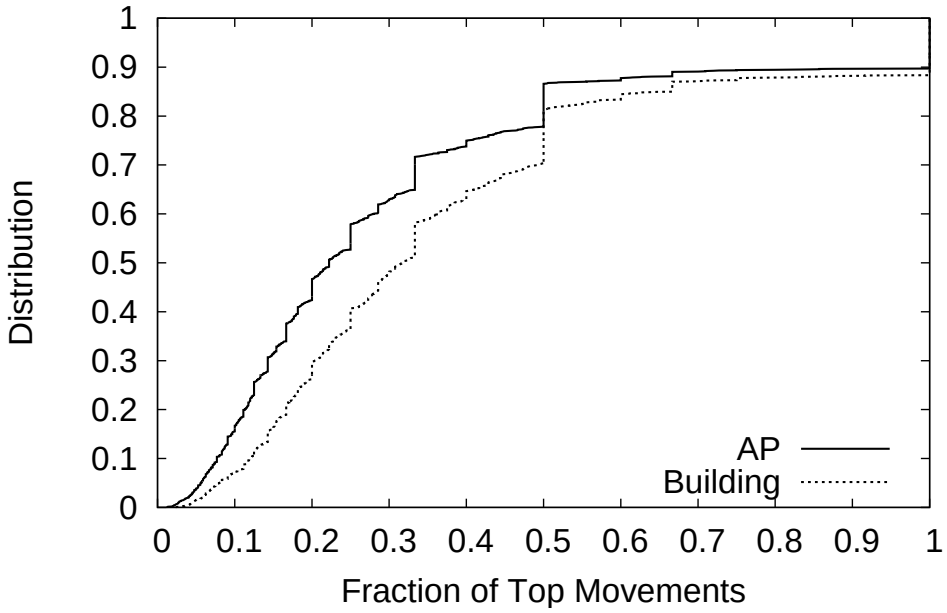

Figure 4: Fraction of clients' top movements.

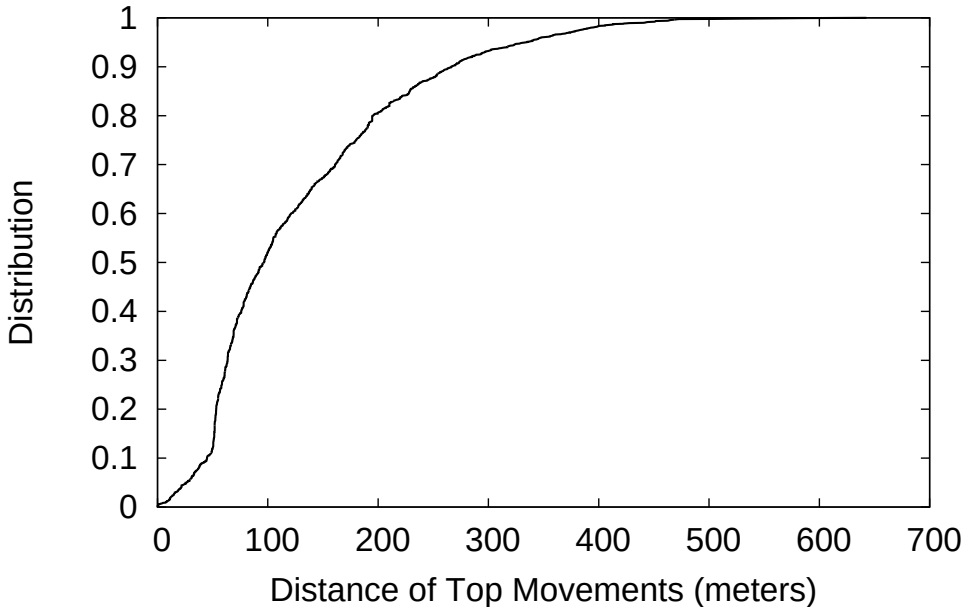

Figure 5: Distances of top movements.

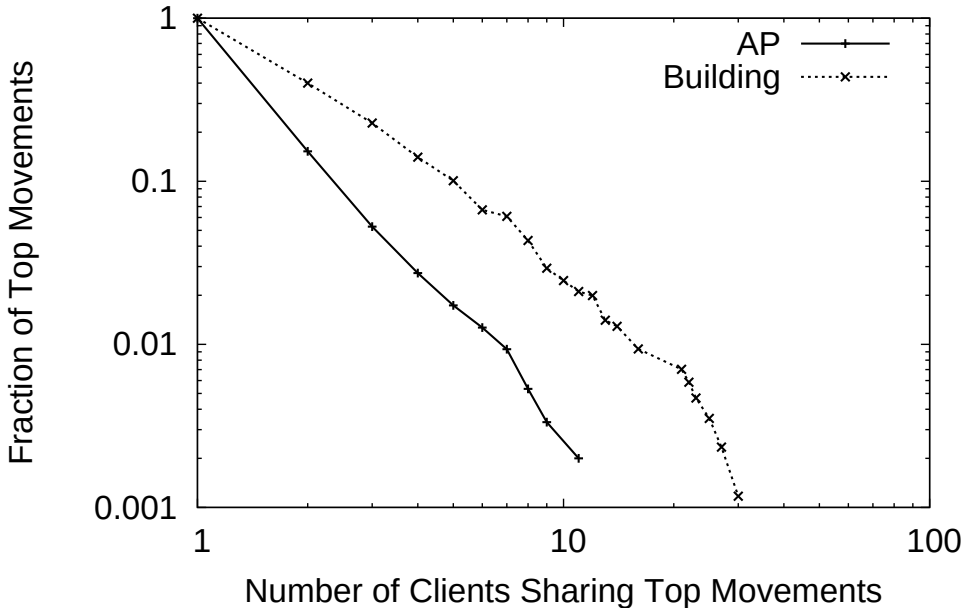

Figure 6: Popularity of top movements.

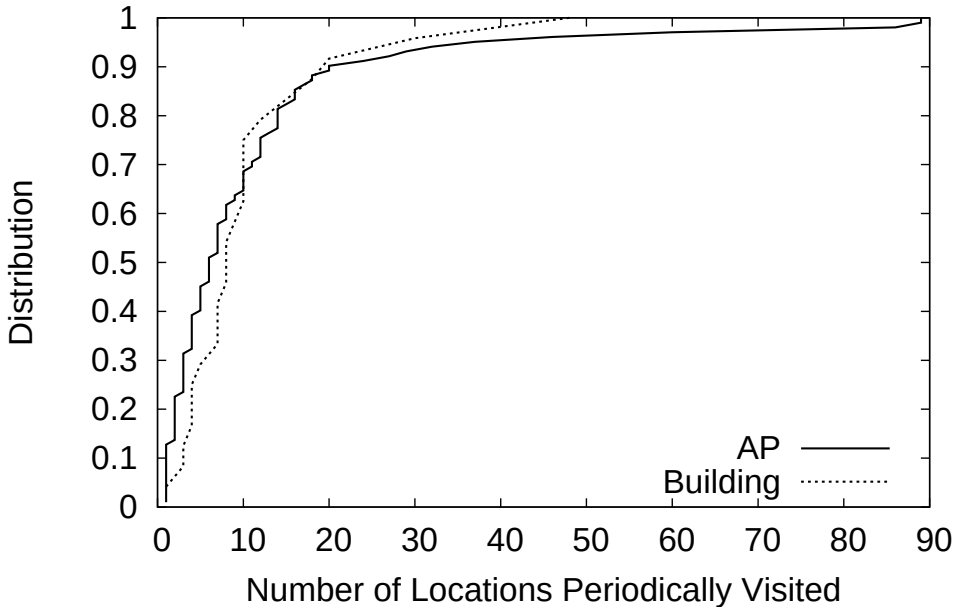

Figure 7: Number of periodic visiting clients.

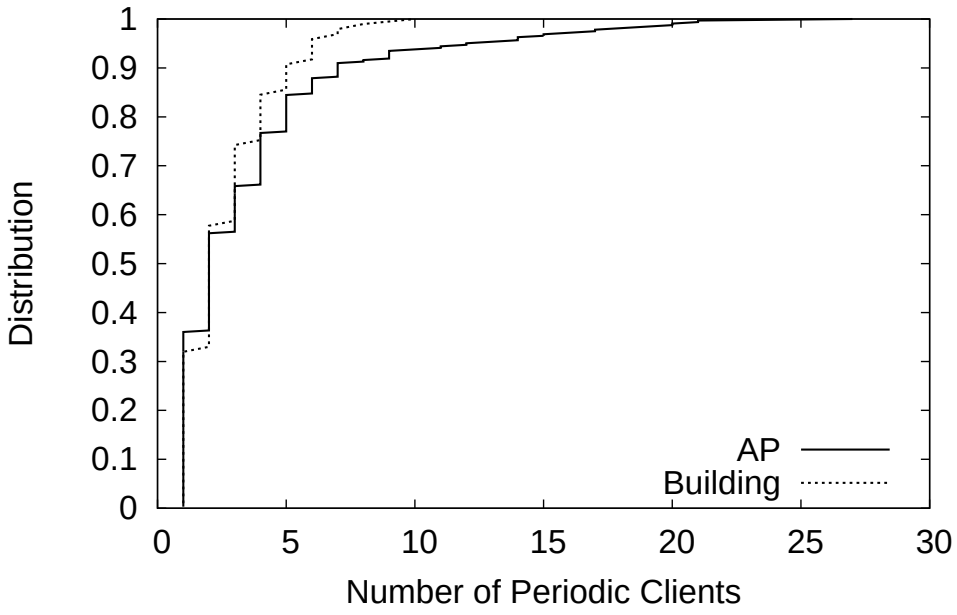

Figure 8: Number of periodic visiting clients.

generally larger than the number to the same building. One possible explanation is that though a client may visit one AP periodically, it may also visit other APs in the same building at other times so that its visiting interval to the whole building becomes non-periodic. Most of the clients with periodic patterns visited the APs in academic buildings; they are likely to be the students attending regular classes. Some library and social buildings also had periodic visitors, though the number of visitors was small (less than 10).

Most periods are integral days in length. Figure 8 shows the distribution of deviation of period length (in days) from closest integers. Almost all of the periods discovered using building sequences lasted in N days, while N is an integer. There are, however, several periods from AP sequences appear to contain half days. We found that a single client was visiting 25 APs, scattered in different buildings, every 12.5 days and this period has occurred at least 4 times. We are not sure what caused this unusual periodic pattern, whether an coincidence or random noises.

Weekly patterns are most prominent in the trace. Figure 9 shows that the distribution of the length of all discovered periods. 30% of all patterns has the period of 7 days, for both AP-level and building-level association sequences. It can be easily interpreted that those clients visited some APs or buildings regularly every week, such as for scheduled classes and meetings. However what does other periods tell us? For smaller periods less than a week, we believe that it is the artifact of our pattern-recognition algorithm. For instance, if an AP is visited every Monday, Wednesday, and Friday every week, the algorithm will report a period of 2 days. For a pattern occurred every Tuesday and Thursday, the algorithm will report both 2-day and 5-day periods. It is a little harder to explain the period longer than a week, which might be caused by events happened bi-weekly. We note that there were several clients who regularly visited some buildings (mostly residential, academic, and administrative) for at least 4 times with 12-day period. Again without more information on people activities, we could not elaborate what caused this periodicity. It is also interesting to note that we did not see many daily patterns: If a client visited a location daily, it would most likely to visit that location several times a day. Because the mining algorithm searches for the shortest period, it is likely that the algorithm will report an interval, if it is periodic, which is shorter than one day.

Most periodic clients' home locations are in residential and academic areas. Almost half of the clients with periodic behaviors had residential buildings as their home location and the other half homed in academic buildings. Few clients homed in other types of buildings had periodic behaviors. Such patterns make sense for a typical academic campus with many students living on site, where much of the mobility is driven by students moving for classes.

There was no significant temporal relationship between periodic patterns. We were interested in finding if there was any temporal relationship between periodic patterns. Namely, if a client visited location A with period p, did it also visit another location B with the same period and the corresponding visiting timestamps within a predefined window δ? To our disappointment, we did not find any such temporal relationship between periodic patterns with δ

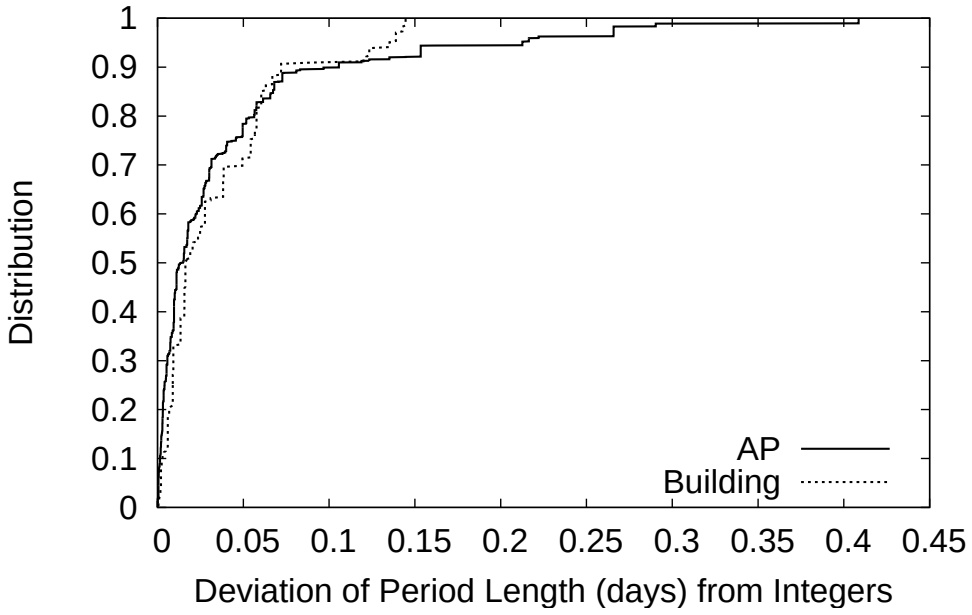

Figure 9: Deviation of of period length (days).

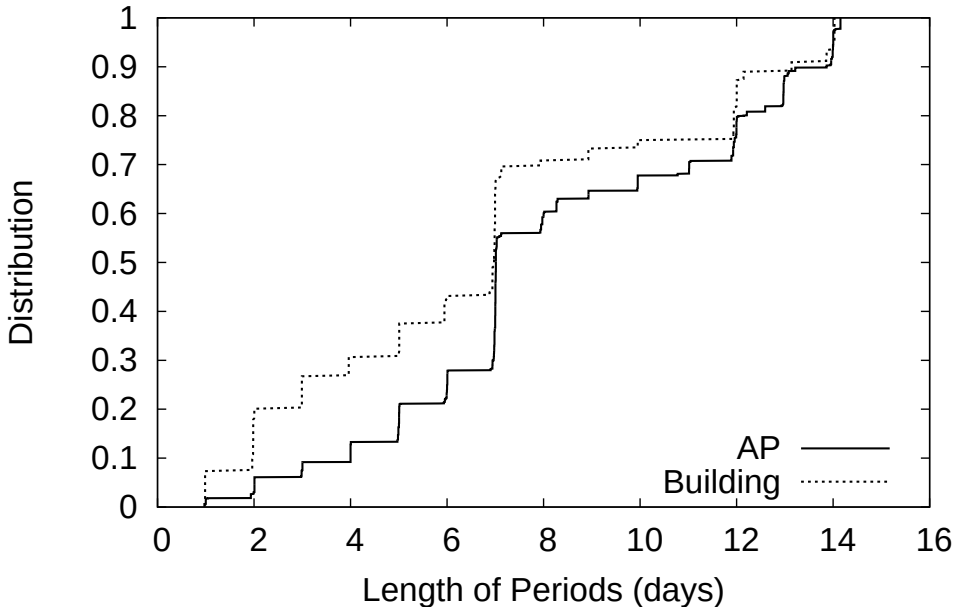

Figure 10: Length of periods (days).

as half an day. Intuitively, students might always visit two locations periodically, such as to take two adjacent classes. For our algorithm to detect this relationship, however, they had to use their devices in both classes throughout 60 days, and not to visit the class locations at other times (which will generate noises). We plan to improve the algorithm to be more resilient to noises to study both temporal and spatial pattern relationships.

6 Related Work

There have been several studies of wireless networks on college or university campuses. Schwab and Bunt examined 134 users over one week in January 2003 on the University of Saskatchewan campus [10]. Chincilla et al. analyzed traces of 7,694 wireless clients collected at the University of North Carolina (UNC) at Chapel Hill during 11 weeks in 2003 [2]. They developed a model of the associations using a Markov chain and used the model to predict the next AP. Papadopouli used the same set of data collected at the UNC and studied it with focuses on session and visit durations [9]. Henderson et al. analyzed the characteristics of wireless network usages on the Dartmouth campus using traces collected during the Fall 2003 and Winter 2004 terms [4]. Song et al. performed a study of next-AP predictors using wireless network traces collected at Dartmouth [11]. Kim and Kotz also used Dartmouth traces to find periodic behavior of users and usages of APs [6]. Using a Discrete Fourier Transform, they discovered that both users and APs show a strong repetition period of 24 hours. Our work complements their approach using a general pattern-recognition framework. We focused on the period longer than one day and the data-mining algorithm helped us to discover the patterns that contributed to the overall periodicity.

7 Conclusion and Future Work

The goal of this paper is to discover and understand both frequent and periodic association patterns. Many of our results reflect an academic-centric environment where much of the mobility is driven by students moving on a campus. Some of our results may help define parameters to mobility models so it can reproduce these patterns in some way. Other results may serve as constraints to the output of any realistic models. For instance, the popularity distribution of frequently visited locations does not follow a power-law distribution. The periodic-pattern mining algorithm can be used offline to discover and analyze periodic behaviors with abnormalities. In this paper, we focus on the most frequent patterns across clients. In the future, we plan to study distribution of patterns for individual clients. A limitation of our current approach is that we used the coordinates of an AP to approximate client's geographic location; these two may be different. We plan to investigate ways of accurately estimating clients' location from a sequence of AP associations.

References

[1] Magdalena Balazinska and Paul Castro. Characterizing Mobility and Network Usage in a Corporate Wireless Local-Area Network. In *International Conference on Mobile Systems, Applications, and Services*, May 2003.

[2] F. Chinchilla, M. Lindsey, and M. Papadopouli. Analysis of wireless information locality and association patterns in a campus. In *Joint Conference of the IEEE Computer and Communications Societies*, March 2004.

[3] Tristan Henderson, Denise Anthony, and David Kotz. Measuring wireless network usage with the experience sampling method. In *Workshop on Wireless Network Measurements*, April 2005.

[4] Tristan Henderson, David Kotz, and Ilya Abyzov. The changing usage of a mature campus-wide wireless network. In *International Conference on Mobile Computing and Networking*, September 2004.

[5] Eamonn Keogh, Stefano Lonardi, and Bill 'Yuan chi' Chiu. Finding surprising patterns in a time series database in linear time and space. In *International Conference on Knowledge Discovery and Data Mining*, July 2002.

[6] Minkyong Kim and David Kotz. Classifying the mobility of users and the popularity of access points. In *Workshop on Location- and Context-Awareness*, May 2005.

[7] Jessica Lin, Eamonn Keogh, Stefano Lonardi, and Pranav Patel. Finding motifs in time series. In *Workshop on Temporal Data Mining*, July 2002.

[8] Sheng Ma and Joseph L. Hellerstein. Mining Partially Periodic Event Patterns with Unknown Periods. In *International Conference on Data Engineering*, April 2001.

[9] Maria Papadopouli, Haipeng Shen, and Manolis Spanakis. Characterizing the duration and association patterns of wireless access in a campus. In *European Wireless Conference*, April 2005.

[10] David Schwab and Rick Bunt. Characterising the use of a campus wireless network. In *Joint Conference of the IEEE Computer and Communications Societies*, March 2004.

[11] Libo Song, David Kotz, Ravi Jain, and Xiaoning He. Evaluating location predictors with extensive Wi-Fi mobility data. In *Joint Conference of the IEEE Computer and Communications Societies*, March 2004.