Analysis of Wide Area User Mobility Patterns

Kevin Simler Steven E. Czerwinski Anthony D. Joseph
University of California at Berkeley

Abstract

In this paper, we present an analysis of user behavior and mobility patterns based on a trace of accesses to a department e-mail server. In contrast to previous studies, we consider a single service and examine how a user community connects to it while moving across a variety of different service providers' wireless and wired networks. By measuring an e-mail service we were able to monitor a large number of sessions originating from a diverse set of locations because e-mail is one of the few services that users habitually access. Our contributions include: a unique approach to extracting user mobility information from traces of client application interactions; a novel approach to modeling user behavior and mobility; and a demonstration of how such models can be used to generate synthetic traces. Overall, although some users are highly mobile, we find most users have a low degree of mobility – 70% of users access their e-mail from 2 or fewer unique locations. We also find that our observed session times are longer than those reported by previous mobility studies in wireless networks.

1 Introduction

Several recent studies have explored and analyzed the mobility pattern of users of wireless devices in local area networks [2, 4, 7] and a wireless Internet Service Provider (ISP) [8]. In this paper, we supplement these studies by analyzing and modeling user behavior, usage patterns, and mobility across wireless and wired networks in the wide area.

A detailed understanding of the behavior of mobile users in the wide area is important for the design and implementation of many distributed and Internet applications. For a large-scale Internet application, such as an e-mail service or a distributed file system, delivering high performance and easy usability for the service depends on a good understanding of how and from where users interact with the system. Thus, system designers need to account for user behavior and mobility in the architecture of the system, as well as in the feature set and user interface.

As a first step in exploring wide-area user behavior and mobility, we obtained and analyzed a month long trace of the e-mail server in the EECS department at UC Berkeley. The server provides both an Internet Mail Access Protocol (IMAP) [3] and a web front-end, so that users can access their e-mail from anywhere. The user community reflected in the trace is relatively broad and varied, and includes approximately 1,000 active users (graduate students, faculty, administrative/support staff, and others).

As a widely used service, e-mail is particularly well-suited to the task of capturing user behavior and mobility since it is a simple, fairly robust indicator of users' presence (location) on the Internet. The first thing many users do when they have access to the network (through their own device or a web browser) is to check their e-mail. Thus, we believe that a characterization of a users' behavior with respect to e-mail access can be used to draw tentative conclusions about users' more general network behavior.

The study in this paper presents an analysis of an e-mail service trace with an eye toward extracting useful nuggets of understanding about how users access their e-mail, and more broadly, how users migrate across wireless/wired local- and wide-area networks. We present three contributions in this paper: a unique approach to extracting user mobility information from traces of client application interactions; a novel approach to modeling user behavior and mobility; and a demonstration of how such models can be used to generate synthetic traces, which are useful in testing the performance of network applications with large numbers of users.

We found two interesting observations in our analysis of the trace set. First, although some users are highly mobile, the majority are not. On most days, most users tend to access their e-mail from a single location[1]. Given the large number of students in our community population, we expected to see a large number of multiple location accesses.

Second, user sessions are relatively long, typically lasting more than an hour and often more than a

[1]We broadly define a single location as access from a single ISP or Autonomous System(AS).

few hours. Users tend to interact with their e-mail for as long as they need before they move to another location or are idle for a long period of time.

In the next section, we discuss the collected trace dataset and processing techniques we developed. In Section 3, we explore and attempt to understand user behavior and mobility. In Section 4, we show how to build a model of an e-mail user's behavior, use the model to generate a synthetic trace, and compare the model with the original real trace. We discuss related work in Section 5 and present our conclusions and plans for future work in Section 6.

2 Methodology

2.1 Trace data

Our analyses are based on a 31-day trace (May 2003) of the e-mail server at UC Berkeley's EECS department. The server provides access through both the IMAP protocol and a web (HTTP) front-end for users. The user community consists of 1,004 active users (active at least once over the duration of the trace) and an unknown number of inactive users. The users are predominantly professors, graduate students, and administrative and support staff, although a small number of undergraduates and other affiliated persons have accounts as well.

Each entry in the trace includes: timestamp (in seconds), username, request type, and IP address from which the request originated. The request type is typically a login, logout, or select mailbox event, although a few error messages appear as well. For this study, we only examine login and logout events.

Since this study involved the collection and analysis of sensitive personal data (specifically username and IP address), we only present aggregate results. While we provide an initial exploration of the characteristics of this dataset, we expect that other researchers will want to apply their own analyses. To enable future research while protecting our users' privacy, we will make an anonymized version available at `http://www.cs.berkeley.edu/~czerwin/traces/`.

2.2 Preprocessing

The raw trace data is not immediately useful to us, because the trace measures client application behavior rather than user behavior. Specifically, the trace entries reflect every time that the e-mail client (*i.e.,* Outlook, PINE, Mozilla, etc.) interacted with the server, rather than just those times when the user herself interacted with the server. The biggest problem is that a user may leave their e-mail client running on their computer while they are performing other tasks or are away. Meanwhile, the client periodically polls for new e-mail by logging into the server (recorded in our trace), checking for e-mail (not recorded in our trace), and logging out (recorded in our trace), at regular periodic intervals. These polling events do not represent true user accesses.

We developed the following process to eliminate polling events from our trace. First, we divide up the trace by user and, for each user, we group the login and logout events into pairs[2]. See Figure 1a. Some of these pairs overlap, reflecting clients that can open multiple simultaneous connections to the server. This occurs when the user is browsing multiple IMAP folders.

Next, we look for high periodicity in the sequence of login/logout pairs (representing client polling). Essentially, this process consists of taking the Fourier transform of all login events and looking for the frequency or period, p, with the highest 'energy' (see Figure 1b). Typically, p ranges from 1 to 15 minutes and represents how often the user has set the client application to poll for new messages.

Then, we discard all login events (and corresponding logout events) that fall p minutes after another login event. For a long sequence of login/logout pairs that occur every p minutes, we discard all but the first pair (Figure 1c). We keep the first pair because it indicates some form of user behavior (*i.e.,* the user starting their e-mail client).

Finally, we clump together the remaining connections into user sessions, which represent a coherent period of time during which the user is accessing their e-mail. We consider two consecutive connections to be the same user session if there are no more than fifteen minutes between them (Figure 1d).

Although this process substantially mitigates the effect of client polling, it is not perfect. Client-initiated connections are subject to false identification, both positive and negative. A false positive identification occurs when a user-initiated connection happens, coincidentally, to fall exactly p minutes after an earlier connection. In this case, the user-initiated connection is erroneously discarded. Similarly, a false negative occurs when a client-initiated connection happens not to fall exactly p minutes after some other connection. For example, if the user's computer stalls for a few seconds right when the client is about to poll the server, the polling event will be delayed and will look like a user-initiated connection. In this case, the client-initiated connection will be erroneously preserved in the trace. Thus, while the

[2]A logout entry in the trace indicates which login entry initiated the connection that is now being closed by the logout.

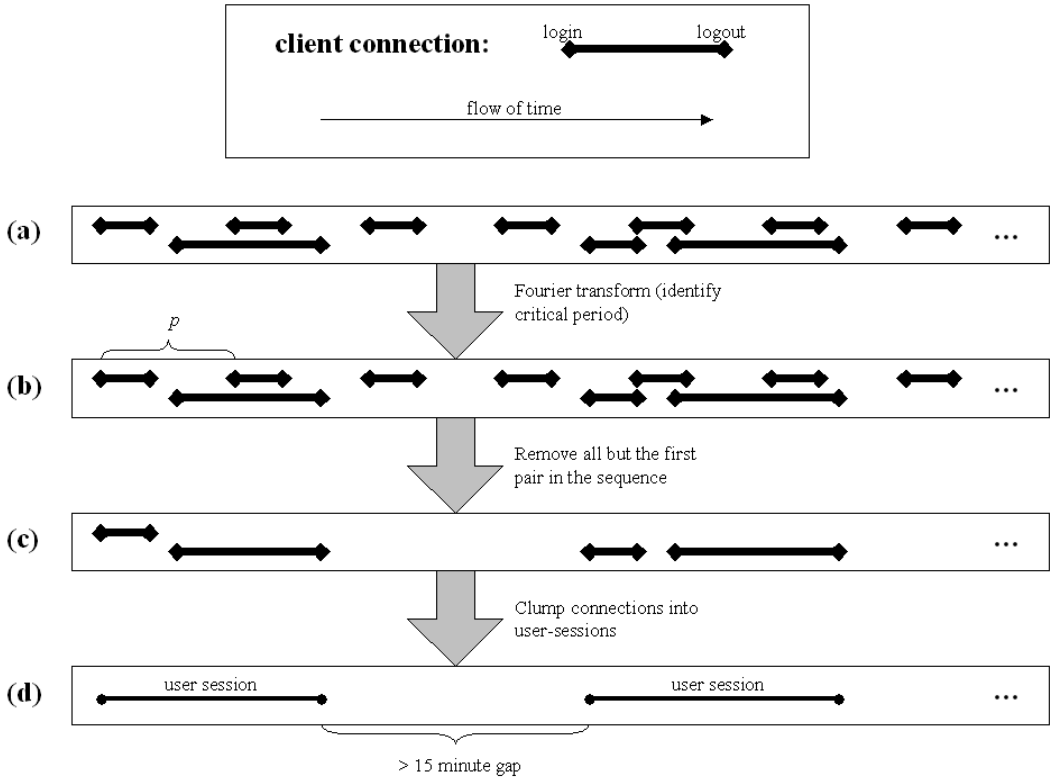

Figure 1: Preprocessing of Traces to Identify True User Sessions

process is fairly successful, it is important to recognize that some of our results may reflect the effects of these false identifications.

Now that we have removed most of the client-initiated connections and grouped connections together into user sessions, we have the data we want to study — a trace of user behavior.

3 Trace Analysis

We divide the analysis of the trace into three categories: (1) user mobility where we investigate how users move among locations (Section 3.1); (2) session characteristics where we analyze the small-scale time component of user behavior (Section 3.2); and (3) mail server load where we track the aggregate behavior over time of all users in the system, as perceived by the mail server (Section 3.3).

For the analysis, we have found it suggestive (but by no means definitive) to think in terms of the following classes of users. Staff are users employed by the university for administrative and support functions, typically working 7:30AM to 4:30PM, Monday to Friday, and rarely (if ever) accessing their e-mail from off-campus locations. Graduate students and professors, on the other hand, typically work longer, more eclectic hours, often bringing their work home with them and/or using their EECS e-mail account as their personal account. Finally, travelers are those graduate students and professors who travel frequently and access their e-mail 'on the road', from many different locations.

3.1 User mobility

In this section, we analyze user mobility. The most basic user mobility is from one IP address to another. However, we use two methods to group IP addresses into meaningful and more useful clusters. The first clustering method is based on IP subnet — two IP addresses are in the same subnet cluster if their IP addresses have the same three byte prefix. We assume subnets are /24 addresses, however the method would work equally well for variable size subnets. The second method uses Autonomous System (AS) numbers — two IP addresses are in the same AS cluster if they map to the same AS number, using IP address to AS

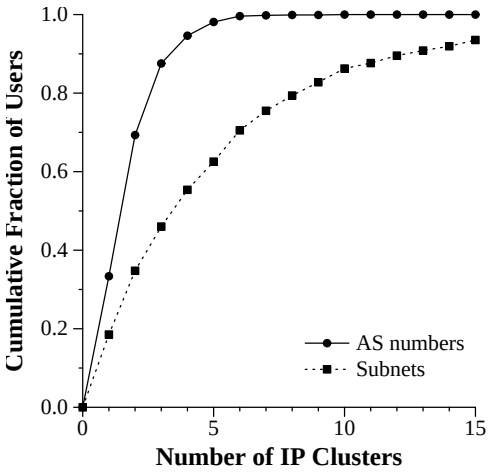

Figure 2: Number of IP clusters Per User

number mapping information from Routeviews [6].

Figure 2 shows the results of applying the two methods. As we expected, the subnet line is pushed out farther to the right than the AS number line. This difference indicates that most users login from a greater number of different subnets than AS numbers over the course of the trace. An ISP owns very few AS numbers (typically one), but usually owns many subnets (*e.g.,* UC Berkeley owns a single AS number, but has nearly 100 subnets). So, a subnet typically clusters together fewer IP addresses than an AS number does. Thus, the difference reflects mobility *within* an ISP (*e.g.,* users moving around floors within a building or across campus)[3]. Overall, we observe that 70% of users log in from 2 or fewer AS numbers during the trace, compared to 6 or fewer subnets. Similarly, 10% of users log in from more than 12 subnets throughout the 31 days of the trace, whereas virtually no one uses more than 6 AS numbers in the same period.

Using the IP address to 'location' method we develop in Section 4.1.1, we measured the number of locations visited by a user each day (see Figure 3). We roughly define location as a way to connect to the Internet (*e.g.,* a dial-up ISP or a campus wireless connection), and calculate the distribution of locations per user-day (a user-day equals one user on one day of the trace). Surprisingly, we found that users are not as mobile as expected, with slightly over 50% of user-days showing user e-mail access from only one location. Contributions to this 50% include staff members during the work week, graduate students and professors during the weekends, and travelers when they are staying in one place. We also saw that for 30% of the user-days, the user did not access the e-mail server, as indicted by zero locations visited. We observe relatively few users logging in from more than one location on any given day.

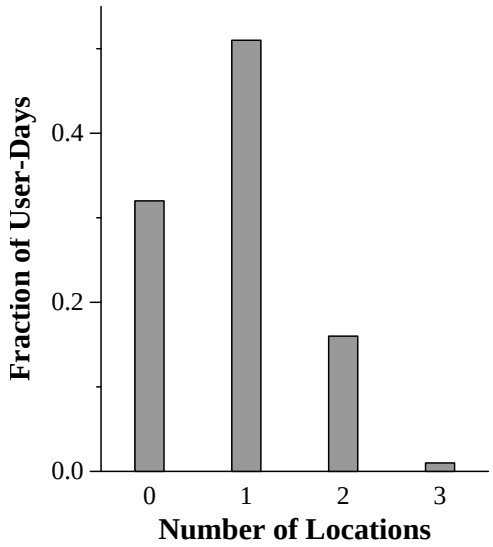

Figure 3: Locations Visited Per Day, Per User

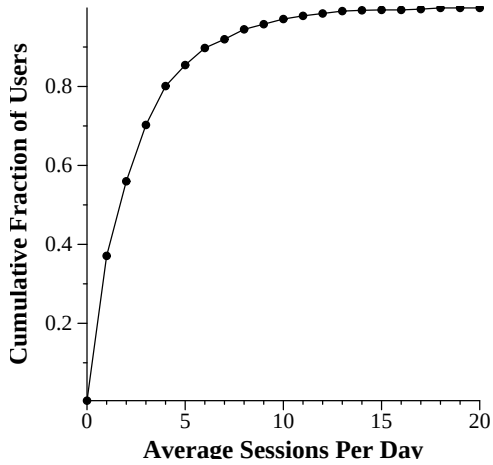

Figure 4: Sessions Per Day, Per User

3.2 User session characteristics

A user session, or just session, is a series of connections between a client application and server that were initiated by the user and separated by no more than 15 minutes of idle time. A session has a start time and an IP address, and represents a coherent, consistent period during which the user is actively accessing e-mail.

Figure 4 shows a Cumulative Distribution Function (CDF) of the average number of sessions per day for each active user[4]. Nearly 40% of users average one

[3]Note that some of the changes may be due to changes in dynamically assigned IP addresses.

[4]For a given user, this is the number of sessions in the trace divided by 31 days.

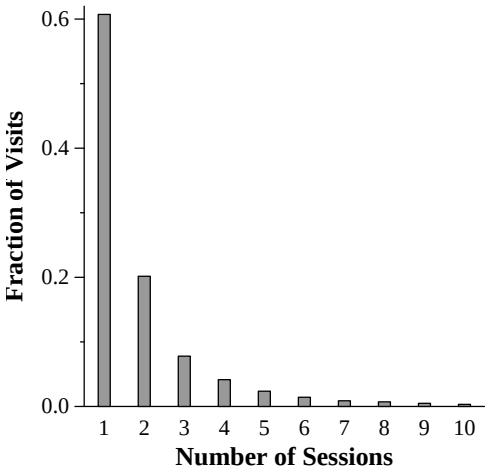

Figure 5: Sessions Per Visit, Per User

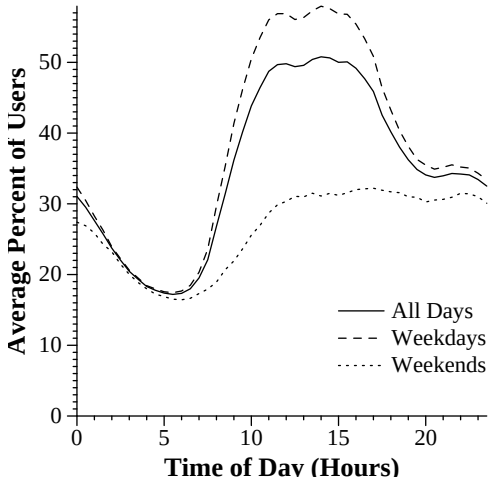

Figure 6: Hourly Server Load

or fewer sessions per day. A large component of this 40% is likely staff, who check their e-mail only at work and typically access it frequently throughout the day (*i.e.,* they have only one or two very long sessions each day). On the other hand, 20% of users average more than 5 or more sessions every day (*i.e.,* shorter, more sporadic sessions).

Another way to look at sessions is to group them together into visits, a series of consecutive sessions at the same location separated by no more than two hours, and representing the period of time that a user spends at a given location. For example, a graduate student might be in and around her office, talking with colleagues, going to meetings, and periodically accessing her e-mail. When she goes home and starts checking her e-mail there, it will mark the beginning of a new visit. Alternatively, a staff member might access her e-mail all during the workday but never at home. When she comes into work the next day she will begin a new visit, because more than two hours have elapsed since she left the office the day before.

In Figure 5 we see a distribution of the number of sessions in each visit. The majority of visits (61%) comprise only a single session. A share of these visits come from staff members accessing their e-mail in one long session throughout the day, but, additionally, travelers (when they are traveling) are likely to generate a lot of single-session visits (*e.g.,* when they check their e-mail at an airport kiosk). Indeed, any user who accesses his e-mail only periodically will contribute heavily to this category. Interestingly, the distribution quickly falls off as the number of sessions per visit increases. We infer that most of the time a user accesses his e-mail, he engages with it for as long as he needs, but then he either leaves for another location or waits a while before checking again.

3.3 Mail server load

In this section, we will analyze the trace's impact on the server. By observing the load on the server, as measured by the number of simultaneous users, we get a picture of user behavior in aggregate across different periods of time. Two particularly fruitful ways of viewing server load are by time of the day, and by day of the week.

Figure 6 shows the server load at each time of the day, and we can see a pronounced diurnal cycle. Users are most active starting around 8 or 9AM and tend, in aggregate, to drift off to sleep around 1AM. This pattern, as expected, is consistent whether we look at weekends or weekdays, since sleep is not an activity specific to any particular day.

Other patterns, however, surface only during the work week. For example, 8AM to 5PM shows the highest user activity, representing not only staff but also graduate students and professors who do the bulk of their work during the day. The busiest time of the day is early afternoon, where on average nearly 60% of users are active. We also record a small blip around lunchtime. These patterns are either absent or significantly muted during the weekend.

The least active period of the day is at night, but it is still surprisingly busy: Around 17% of the userbase is active throughout the night. Of course, no single group of users is responsible for this 17%. This is merely the aggregrate load placed on the server by everyone working late or getting up early. Some of the 17% may represent client-initiated polling events that were not correctly identified and discarded from the trace (see Section 2.2).

Figure 7 shows the server load[5], during each day

[5]This analysis is not intended to measure the amount of per

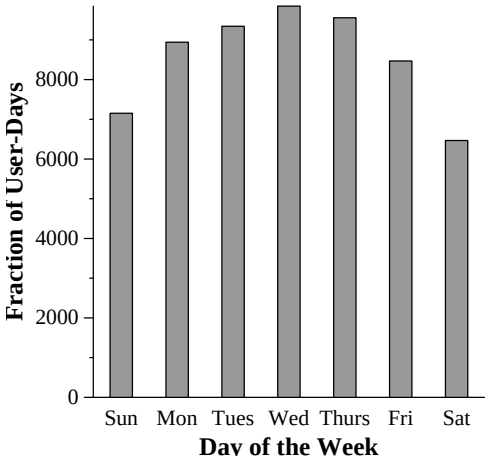

Figure 7: Server Load Per Day of Week

of the week. Here, server load is measured in user-hours, where one user active for one hour equals one user-hour of load. As before, our results are not surprising. The server sees more activity on weekdays than during the weekend, and both Monday and Friday see less activity than the middle of week. In this regard, our results agree with other user studies. However, unlike the results in [2], our user community shows substantially more weekend activity relative to weekday activity. One factor that explains this difference is that [2] measured a corporate wireless LAN, which users cannot access over the weekend without physically being at work. Our study, on the other hand, looks at an e-mail server, which users can access from virtually anywhere, especially given the web front-end. Additionally, some of the weekend activity (as well as some of the weekday activity) may be explained by polling events that were not successfully removed from the trace. Removing these events may increase the relative disparity between weekday and weekend activity.

4 User Modeling

In this section we examine one approach for generating user behavior models. Our purpose in creating the model is two-fold. First, we would like to generate synthetic traces which can be used to emulate the load on a global-scale Internet service of an arbitrary number of users over an arbitrary period of time with realistic access patterns. Second, we believe our models can help predict the behavior of real users in real time, so that an adaptive network service could adapt to users' needs in changing conditions. We first discuss the basic modeling approach and structure of our model, and then how we train and test it.

day 'work' a server actually has. Rather, our goal is to uncover the cyclical, time-dependent component of user behavior.

4.1 Approach

Our approach to creating user models has three major steps. First, we group the users in our trace into categories based on similar mobility patterns. Second, within each category, we create a model of how users transition from location to location. Finally, within each location, we model the series of login and logout events (*i.e.,* the user sessions).

Before we take any of these steps, however, we must first provide a definition of 'location.'

4.1.1 Location

Users access their e-mail from different locations. For this paper, we define a user's network location as the connection she uses to access the Internet. For example, a graduate student might check e-mail from her office desktop, where she is connected by Ethernet to the department network, or she might use her laptop with a connection to the university-wide wireless network. These two connections represent two different locations, but a location need not be specific to a single machine. Two machines may share a connection (*e.g.,* in a home network with a single DSL connection), or one machine may have more than one connection (*e.g.,* dial-up access to two different ISPs).

Given this rough definition of what a location is, now our problem is how to uniquely identify a user's location based only on their IP address. The first thing to note is that the IP address itself is at least a fair approximation of the user's location. Different locations, for instance, will always have different IP addresses. The converse, however, is not true: different IP addresses are not always associated with different locations, due to such things as dynamic addresses assigned by the Dynamic Host Configuration Protocol (DHCP). For example, every time a subscriber dials into an ISP that uses DHCP, he is assigned a new IP address, but the different IP addresses represent the same location.

So we need a way of clustering IP addresses in order to identify locations more accurately. We briefly discussed two such techniques in Section 3.1: clustering by AS number and clustering by subnet. To this set we add a third clustering technique: group IPs with the same authoritative Domain Name Server (authoritative DNS). To find the authoritative DNS associated with a given IP address, we perform a reverse DNS lookup and record the DNS server responsible for the reverse lookup. However, we found that this

method is quite unreliable. Two-thirds of the reverse IP lookups in our data set fail to return an authoritative DNS.

Thus, we chose to identify locations as follows. For a given IP address, if the reverse DNS lookup succeeds, we use the authoritative DNS as a unique identifier for the location. If the reverse DNS lookup fails, we use the AS number as a unique identifier for the location. Finally, if we cannot find the AS number associated with the IP address (which happens on very rare occasions), we use the subnet as a unique identifier for the location. We tested this method of identifying locations against the previously mentioned methods and found that this was the best way to compute location. We describe the testing methodology in detail in Section 4.2.

Of the 2,724 locations visited by our users over the duration of the trace, 897 are identified by authoritative DNS, 1823 are identified by AS number, and 4 are identified by subnet.

4.1.2 Categorization of users

For the purpose of predicting what a user will do next, it would be ideal to model that user individually. Instead, we choose to model whole classes or categories of users. There are two reasons for this choice. First, we have comparatively little data on each individual user, whereas when we pool users together, we have a much larger dataset with which to train our model. Second, we want to not only predict an individual user's behavior, but also to generate synthetic traces which mimic the original trace in relevant ways. If we model each user separately, we run the risk of over-fitting our model to the data, in which case the synthetic trace would be too similar to the original.

For the synthetic traces we created, we decided the most important distinction between users was how mobile they were (*e.g.,* how many unique locations they visited on average and how frequently). As shown earlier, some users do move around much more often than others. By using distinct models to represent each different mobility type, we feel we can more accurately model the population as a whole.

We thus categorize users based on the number of primary locations they have. A primary location is a location where a user spends an amount of time large enough to call for special attention in our model. Our threshold is 5%: if a user spends more than 5% of his time (of all the time he spends accessing his e-mail) at a given location, then that location is a primary for him. Table 1 shows the breakdown of the 1,004 active users in our trace. We throw out 114 users because they don't have enough sessions (< 10) to accurately determine a primary location. Also, in the third category, there are 51 users with four or more primary locations, but we decided to group them together with the 141 users with exactly three primary locations.

Primary Locations	#(Users)
1	350
2	348
≥ 3	192

Table 1: Categorization of users by number of primary locations

Users in the first category only access their e-mail from one location, which is most likely their work location. The second category represents users who have two distinct locations, which are most likely their work location and a home computer. Finally, users in the third category have a third primary location in addition to home and work (*e.g.,* the home of a significant other or close friend).

4.1.3 Model structure

For a user in a given category, we model his movement using a Markov model with states as locations. We refer to this as the location Markov Model (location-MM). A user in category two, for example, is represented by a 3-state Markov model. One state represents the user accessing e-mail from his most primary location; another state represents him at his second most primary location; and a third state represents him traveling to other locations (*i.e.,* when the user visits locations other than his two primary ones). The traveling state behaves differently from the other states in that whenever the model transitions into that state during trace generation, the user is assigned a new location at random from the set of all locations seen in the trace.

We also associate the transitions between locations with an average idle time, t_i, that is unique to each transition type (*e.g.,* a transition between work and home might take an hour on average, whereas the home to work transition might take eight hours on average, if the user typically checks his e-mail at home before going to sleep and then waits until he gets to work to check it in the morning).

Within each location, we model a user's session behavior at that location by a separate Markov model. The session Markov Model (session-MM) has only two states: Logged-In and Logged-Out (see Figure 8).

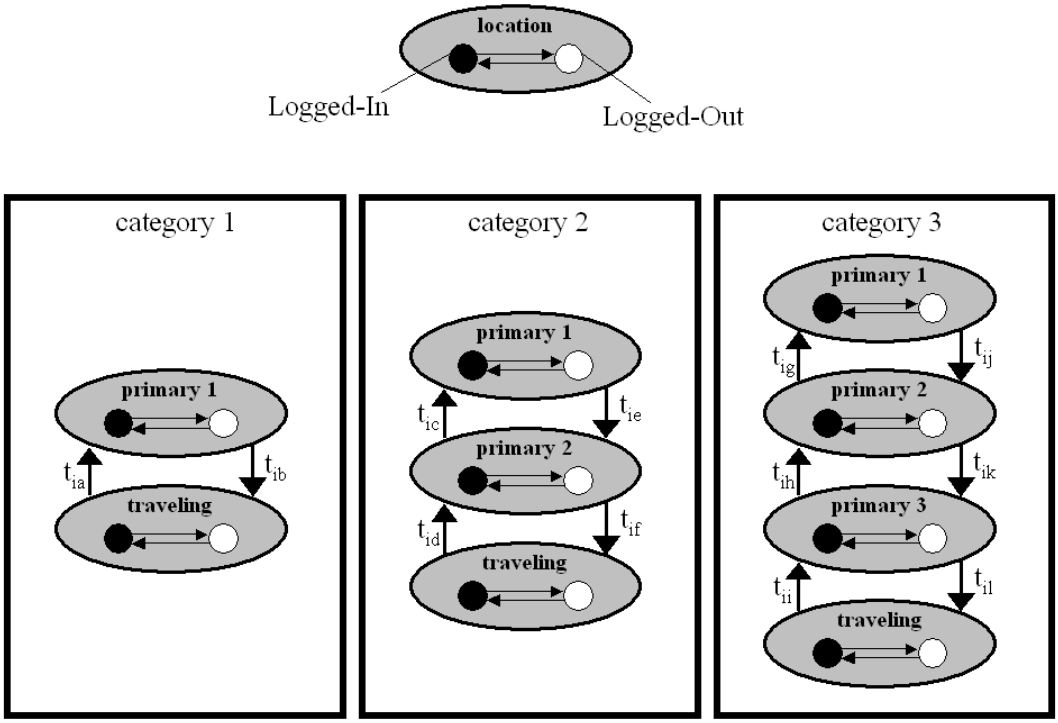

Figure 8: Structure of the Model: Some transitions between locations have been omitted for clarity; in fact, it should be a complete graph with transitions between every location pair, including self-transitions

4.2 Testing

We have now defined three different models for three different classes of e-mail users. We are now ready to train each model: we can simply compute the probabilities for each transition in all the Markov models using the trace data. Note that training the model for each category is no different, in principle, from training a model for an individual user. If we had enough data to model an individual user, we would determine how many primary locations that user has, create a state for each location as well as one extra for the traveling state, and then train the Markov model on the user's traces.

Once the models are fully trained, we can use them to generate synthetic traces. The first step is to create a synthetic user community. To match the original trace, we created 890 synthetic users, representing the same number of users minus the ones with too few sessions. Each of our 890 synthetic users is randomly assigned to one of the three different categories according to the distributions observed in the trace: with a probability of 350/890 the user will be assigned to category 1; with 348/890, to category 2; and with 192/890, to category 3. After creating the user community, we simply generate events using our Markov models for each user for 31 days. This action will tell us when a user is logging in and logging out, and from which locations. Working backwards, we also assign one or more IP address to each location. This combination results in a full synthetic trace.

To test how well our synthetic trace stacks up against the real trace—or, in other words, how well our model represents true user behavior—we extract five representative metrics from both the real and the synthetic traces, and then we compare them. The five metrics are as follows:

1. How many sessions does a user generate between visits to his most primary location?
2. How long does a user take between visits to his most primary location?
3. What are the network latencies (ping times) for each session?
4. How long is each session?

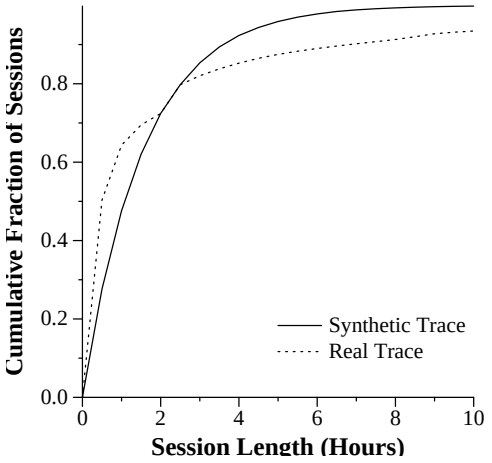

Figure 9: Distribution of Session Lengths

5. How long is the idle time between sessions?

Metrics (1) and (2) measure how well we have modeled user mobility. Metric (3) measures how well we have approximated location (see Section 4.1.1)[6]. And metrics (4) and (5) measure how well we have modeled session characteristics.

Each metric generates a CDF that, in some sense, captures the characteristics of the entire trace with respect to that metric. For example, metric (4) will generate a CDF of session length over all sessions in the trace. We then compute the disparity between the CDF from the synthetic trace and the CDF from the real trace by measuring the area between the curves and dividing it by the area underneath the real CDF. Figure 9 shows the CDFs for metric (4).

We find that the model is reasonably accurate, but that there is significant room for improvement. The disparities for each of the metrics are: (1) 0.04 (2) 0.21 (3) 0.19 (4) 0.09 (5) 0.28. For some of the statistics, we come close, and for others, we perform reasonably.

There are other serious limitations to our model. For example, it does not recreate the temporal characteristics of the load on the e-mail server across the times of the day and the days of the week (see Section 3.3). Our model is time-independent in the sense that the modeled user has an equal chance of being connected to the server at any time. Similarly, a user has an equal chance of accessing his e-mail from any location at any time. These are both unrealistic assumptions for a model of true user behavior, but they reflect an accuracy versus simplicity tradeoff.

This provides a simple first-cut at the problem, but, we are working on solutions to these limitations. For example, a multi-level model could combine a low frequency B-spline-based model to capture long-term temporal trends (*e.g.,* diurnal events), with our Markov models for capturing higher frequency transient events.

[6]If we do a poor job of approximating location, then our synthetic trace will show the wrong distribution of network latencies. For example, if we compute location too broadly, agglomerating what should be a number of different locations into a single represented location, then the network latencies in the synthetic trace will not be sufficiently distributed.

5 Related Work

We believe this paper is the first to analyze wide-area user mobility and behavior patterns. However, over the last several years, others have analyzed user mobility patterns in networks of varying scales, usually focusing on wireless devices [1, 2, 4, 5, 7, 8].

Tang and Baker [7] first studied the mobility patterns of wireless devices in a single college building. They found that many of the wireless devices were stationary (only used one access point), while some were somewhat mobile (used a few access points), and a small number were very mobile (used a large number of access points). Other studies that followed, including ours, show that this distribution of mobility holds for larger networks and communities of users.

Kotz and Essien [4] performed a similar study but for an entire campus. They observed mobility patterns across buildings and measured the length of user sessions at each location. They found that many users (18%) only used one building, while most (50%) visited at least 5 buildings over the two month trace. This result corroborates the observation that few users are very mobile. Similar to our study, they analyzed session times (how long each user accesses the network at a location), and found the median session length to be 16 minutes, whereas we observed it to be 30 minutes with a much longer tail. This difference is not surprising, since our trace includes stationary home and work machines on which users tend to work much longer than a wireless device.

Balazinska and Castro [2] studied wireless LAN roaming across several corporate campus buildings. They again found a wide distribution of mobility patterns, but much less roaming than observed by [4], a difference they attributed to the more rigid interaction patterns of a corporate versus a academic environment. This study had several interesting contributions that influenced our work. First, they differentiated between *persistence* (how long a user stayed at one location) and *prevalence* (how much time a user spent overall at a location), which is similar to our distinction between unique visits and session lengths. Second, they used the notion of home and guest lo-

cations in their analysis. We also found that many users have dominant locations that should be considered primary and secondary home locations.

Finally, Tang and Baker studied mobility patterns in a metropolitan-area wireless network [8]. They analyzed roaming across a much larger geographical space, but were limited to one specific ISP and user population. They also showed how to classify users into different mobility classes based on how many different locations they visited, and used various clustering techniques for classification. In the future, we plan to combine their techniques with the broad user models that we have created.

In contrast to previous studies, we consider a single service and examine how a user community connects to that service from a variety of wireless and wired networks owned by many different service providers. As with previous studies, ours reflects the biases of our academic environment, however we believe that the methodologies we have developed can be usefully applied to users in other environments.

6 Conclusion

In this paper, we have presented an analysis of user behavior and mobility patterns in a month long trace of an e-mail server. Because of its importance and popularity, e-mail is a good measure of how frequently people access the Internet and where from.

We provide several contributions to the analysis and modeling of user behavior: unique Fourier transform-based trace preprocessing to remove the effects of periodic client polling; a novel, simple, but accurate approach to modeling user behavior that splits users into categories based on their degree of mobility and uses Markov models to represent a user's movement between locations and their access pattern; and a demonstration that synthetic traces generated using our model have characteristics similar to a real trace.

Overall, we observed that some users are highly mobile over the duration of the trace, logging in from several different AS's/subnets. Each day, however, most users log in from one location only.

Similarly, we found that users access their e-mail fairly infrequently, but for long periods of time. 70% of active users average 3 or fewer sessions every day. When they do access their e-mail, they tend to spend as much time as they need before moving on to another location or another activity.

Finally, as expected, the aggregate trends in user behavior exhibit a strong diurnal cycle. Over the course of a weekday, there are three levels of activity: highest during normal working hours, dropping off to a medium plateau after work, and settling down significantly at night. Still, there are a fair number of users logged in after midnight.

We plan several directions for future work. First, our analyses could be applied to similar, but larger and more diverse user population traces. Second, traces with more details about user behavior (*e.g.,* the selection of individual messages), would make filtering out the effects of client polling easier and more accurate. Third, there is significant room for improvement of the user model. Specifically, we plan to use a high-order Markov model and include a time component to model the effects of time and day-of-the-week on user behavior.

Overall, we believe the results we have presented shed further light on user access and mobility patterns. Many of our observations and results corroborate the findings in similar studies. Finally, we believe that our study is unique in that it captures wide-area mobility patterns that span multiple ISPs. We expect that the results will be useful to application developers and that they will foster future research into network spanning user mobility.

References

[1] Balachandran, A., Voelker, G. M., Bahl, P., and Rangan, P. V. Characterizing user behavior and network performance in a public wireless LAN. In *Proc. of ACM SIGMETRICS '02* (June 2002).

[2] Balazinska, M., and Castro, P. Characterizing mobility and network usage in a corporate wireless local-area network. In *The First International Conference on Mobile Systems, Applications, and Services* (San Francisco, CA, May 2003).

[3] Crispin, M. Internet Message Access Protocol - Version 4rev1. RFC 2060, December 1996.

[4] Kotz, D., and Essien, K. Analysis of a campus-wide wireless network. In *Proc. of the Eight Annual International Conference on Mobile Computing and Networking (MobiCom)* (September 2002).

[5] Lai, K., Roussopoulos, M., Tang, D., Zhao, X., and Baker, M. Experiences with a mobile testbed. In *Proc. of the Second International Conference on Worldwide Computing and its Applications (WWWCA '98)* (March 1998).

[6] Routeviews. http://www.routeviews.org.

[7] Tang, D., and Baker, M. Analysis of a local-area wireless network. In *Proc. of the Sixth Annual International Conference on Mobile Computing and Networking (MobiCom)* (August 2000).

[8] Tang, D., and Baker, M. Analysis of a metropolitan-area wireless network. *Wireless Networks* (2002), 107–120.