User Transitioning Among Networks - a Measurement and Modeling Study

Technical Report UM-CS-2013-020

Sookhyun Yang, Simon Heimlicher, Jim Kurose, Arun Venkataramani
University of Massachusetts, Amherst, MA, 01003
{shyang, heimlicher, kurose, arun}@cs.umass.edu

***Abstract*—Physical human mobility has played an important role in the design and operation of mobile networks. Physical mobility, however, differs from mobility in a network or network-addressing point of view - a physically mobile user may be stationary (i.e., maintain its network address) from a network point of view, and a physically stationary user may change access networks or move among contemporaneous connections to different networks. We perform a measurement study of user transitioning among networks from a network-level point of view and discuss insights and implications drawn from these measurements. We characterize network transitioning in terms of transition rates, network residency time, degree of multihoming, and more. We find that users typically spend time attached to a small number of access networks, and that a surprisingly large number of users access two networks contemporaneously. We also develop and validate a parsimonious Markov chain model of canonical user transitioning among networks.**

I. Introduction

Physical human mobility has played a central role in the design and operation of mobile networks (including cellular, Wi-Fi, and mobile ad hoc networks) and their protocols for hand-off, routing, location management, and more. Consequently, numerous research studies have developed models of human physical mobility and used these models in the design and evaluation of mobile network protocols.

Physical user mobility, however, is quite different than mobility from a network or network-layer addressing point of view. For example, a user physically moving among access points or base stations within the same subnet retains its IP address. Conversely, a multi-homed stationary user or a stationary user shifting among multiple devices attached via contemporaneous connections to different networks will change access networks and the IP address to which his/her identity was most recently associated. In the former case, the physically mobile user is stationary from a network perspective; in the latter case, the physically stationary user is mobile from a network perspective.

This distinction between physical mobility and mobility among networks (i.e., a changing network address associated with a device or an end user) is an important one, since it is this mobility among networks that is important to location management protocols such as mobile-IP [11], HLV/VLR registration in cellular networks [16], and name/address resolution protocols in current (e.g., LISP [5]) and next generation (e.g., MobilityFirst [15], XIA [8]) network architectures and protocols. The amount of network-level signaling for location management depends on mobility among networks rather than physical mobility; similarly it is mobility among networks (rather than physical mobility) that determines the network or set of networks in which a user is reachable at a given point in time. Recognizing the ambiguity between physical and network mobility, we will refer to a user moving among networks from a network-layer/addressing viewpoint as *transitioning* among networks.

In this paper, we perform a measurement study of user-transitioning among networks and discuss insights and implications drawn from these measurements. Based on these measurements, we also develop and validate a parsimonious Markov chain model of canonical user-transitioning among networks. Our measurement study, conducted using IMAP server logs of a population of approximately 70 users over the course of three months, quantitatively characterizes network transitioning in terms of transition rates among networks, network residency time, degree of contemporaneous connection to multiple networks, and more. We find that users spend the majority of their time attached to a small number of access networks, and that a surprisingly large number of users access two networks contemporaneously. We also show that our Markov chain model of a canonical individual user, in spite of its many simplifying assumptions, can accurately predict aggregate transition rates, the degree of contemporaneous multi-homing, and other key network-transitioning performance metrics for an aggregate population. Our measurements provide quantitative insight into the location management signaling overhead needed by modern and proposed name/address translation and location management protocols; our models provide the ability to design, dimension and analyze such systems. More generally, we believe that while physical mobility and the design of link-layer and intra-subnetwork handoff protocols are relatively well-understood, the behavior, modeling and measurement of users transitioning among networks and the design of protocols for managing that mobility at global scale are much less well-understood. This paper is an important step in deepening that understanding.

The remainder of this paper is structured as follows. In sec-

tion II we present our measurement scenario and methodology. In Section III we then examine our traces of user transitions among networks, quantifying various characteristics of user-network-transitioning; we also discuss insights drawn from these measurements. Section IV presents and validates a parsimonious Markov chain model of canonical user-transitioning. In Section V, we discuss related past research. Section VI concludes this paper.

II. Measurement

In this section we describe our measurement scenario and methodology and the properties of our collected traces.

A. Measurement methodology

Measuring user-transitioning among networks is itself a challenging task. Measuring network connectivity directly at the end user requires a population of users willing to install software on all of their network-connected devices (e.g., laptop, home/office desktops, tablet and/or smartphone), periodically monitoring/logging network connectivity on all interfaces on all devices, and then collecting measurement data. In addition to the difficulty of finding and managing such a user base, the task is technically complicated by concerns regarding battery drain for connectivity monitoring on mobile devices with limited battery capacity. For these reasons, a more centralized, server-based approach might seem preferable. In particular, since a client's connection to a server provides that client's IP address, the (possibly changing) access network used by each of the server's multiple clients can thus be easily logged at a server.

Yet there are also many challenges associated with server-side measurement of user-transitioning. Each server implements a single service/application and each user runs many services and applications. Monitoring all service and application servers is impossible - there are far too many servers, and most commonly-accessed servers (e.g., Google, Amazon) are proprietary. Moreover, a user invoking multiple applications has a different "identity" in each application; correlating a user's identity on one application with his/her identity on another application is a difficult research problem [6]. From a practical viewpoint then, we ideally need a server application that *(i)* is frequently (ideally always) used by an on-line user, *(ii)* can be monitored at a non-proprietary server, and *(iii)* provides both a user "identity," so that the same user can be tracked across multiple sessions, and the network address from which that identified user accesses that server.

Although no single application server meets this ideal, we believe that an IMAP mail server [4] is a compelling choice. Many users frequently check and read email when online; additionally, many mail clients periodically poll (often at short timescale) the IMAP server so that recently arriving mail can be proactively pushed to the client. Email checking, reading, polling and delivery all create entries in the IMAP server's log containing an associated client IP address, as well as an identifier - the email address - for that client; this email address typically remains the same across a user's many devices. A user who accesses the IMAP server from a desktop while at work, and then from a mobile device while commuting, and then from a laptop at home will create IMAP log entries evidencing transitions from office network to cellular provider network to home access network. Of course, not all users periodically access their IMAP server while online. Consequently, using IMAP logs to trace a user's transitioning among access networks may miss a network transition or underestimate the amount of time spent in a network. In this sense, using IMAP logs represents an informal lower bound on user-network-transitioning activity.

IMAP logs can also be used to indicate a multi-homed user, or a user contemporaneously belonging to multiple networks via multiple devices. In the former case, if the user with a single device accesses the IMAP server using multiple device interfaces connected to different networks, the multi-homed IMAP access via these different client IP addresses (and networks) will be evidenced in the IMAP log. In the latter case, a user accessing the IMAP server from multiple devices (e.g., working and reading email on laptop or PC, while also having email pushed to a smartphone) within the same period of time will have IMAP accesses via multiple contemporaneous connections during this period of time evidenced in the IMAP logs.

B. IMAP logging data set

For this study, we collected logs produced by two load-balanced IMAP servers in the Computer Science Department at the University of Massachusetts Amherst from April 14th to July 5th 2013. The traces consist of a series of individual IMAP log entries. We combined and processed these logs, extracting the following pieces of information for each entry:

- A **user's account ID.** We consistently anonymize a user's account ID (email address) using MD5-hashing for privacy purposes.
- A **timestamp**. The time at which a user accesses the IMAP mail server to poll, check, or retrieve email.
- A client-side **IP address.** This is the user's (client-side) IP address when the user accesses the IMAP server. Occasionally, users accessed mail via a departmental web-based server, rather than directly from a client email application. In this case, the user's logged IP address is recorded in the IMAP log as 127.0.0.1; in this case, we analyzed the server's web logs to determine the client address of user browser associated with this IMAP access. Only 0.04 percent of all IMAP web-based log entries could not be identified due to missing web logs; those entries were excluded from our analysis.

Given an IP address, we then identified its IP-prefix range, AS number, and network domain ownership information using UNIX's *whois* command with *whois.cymru.com* [1]. Information at whois.cymru.com is updated every 4 hours from the regional registries including ARIN, RIPE, AFRINIC, APNIC, and LACNIC.

Our traces include 70 users consisting mostly of UMass Amherst CS faculty and staff members. Fig. 1 shows the distribution of the average number of a user's IMAP log entries

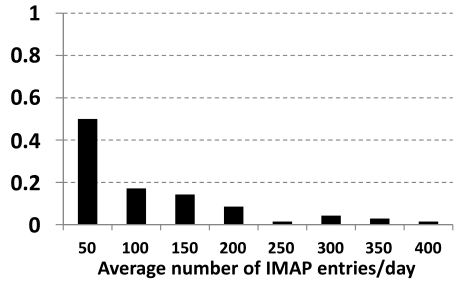


Fig. 1. Distribution over all users: average (per-user) number of IMAP entries/day.

per day. The total number of IMAP log entries per user ranges 26 to 28,584. In this and subsequent bar graphs, the y-value indicates the count for the interval ending at an upper limit given by the corresponding x-value.

# IP prefixes	AS descriptions
1	127 ASes including Verizon wireless (AS6167, AS22394), Free Mobile SAS (AS51207)
2	14 ASes including AT&T wireless (AS20057), and Hughes Network Systems (AS6621)
3	14 ASes
4	7 ASes
5	11 ASes including AT&T wireless (AS7132), and Five college (AS1249)
6	3 ASes incl. AT&T Internet (AS7018)
7	Free SAS (AS12322)
8	Sprint wireless (AS3651)
9	British telecom (AS2856)
11	Charter communications (AS20115), and Verizon Online (AS19262)
16	France telecom (AS3215)
28	Comcast cable network (AS7922)

TABLE I
NUMBER OF OBSERVED IP PREFIXES PER AS.

The traces contain 398 unique IP prefixes and 183 unique ASes. The AS numbers associated with 11 of these IP prefixes were unknown. 97% of IMAP log entries have client addresses in networks registered in the United States.[1]

Table I shows a list of selected ASes with their owner and the number of observed IP prefixes per AS in our measurements. 127 of 183 ASes had only a single IP prefix in the trace. The UMass campus network, part of the Five College AS (AS1249) network, consists of two IP prefixes. The largest number of observed IP-prefixes per AS was 28 from Comcast cable network (AS7922). Comcast (AS7922), Verizon Online (AS19262), and Charter (AS20115) are primarily residential wired Internet service providers (e.g., cable and ADSL access networks); the Hughes network (AS6621) supports a satellite Internet service used in rural communities lacking wired and cellular broadband service. Among mobile access service providers, we find Verizon wireless (AS6167, AS22394), AT&T wireless (AS7132, AS20057), and Sprint wireless (AS3651). AS51207 and AS12322 owned by Free SAS, an ADSL and Wi-Fi service provider in France, were used for a non-negligible amount of time in our measurements.

[1]VPN access to the IMAP servers is not required. Anecdotally, we believe VPN access is used primarily for accessing library and other restricted campus resources. VPN IMAP access t would be logged as a client access from within cs.umass.edu, perhaps appropriately so given that server-to-client IMAP replies would then be addresses and delivered to the cs.umass.edu network (and then tunneled to the remote client).

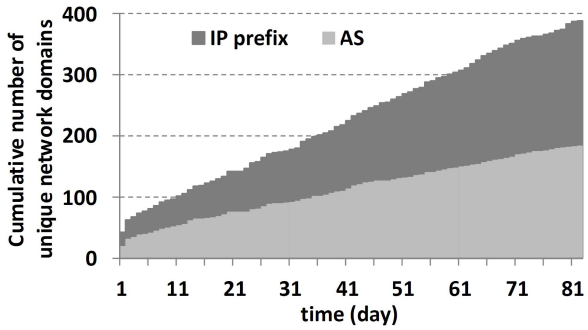


Fig. 2. Cumulative number of unique network domains accessed by all users over time.

Fig. 2 plots the cumulative number of unique network domains accessed by all users over time. As shown in Fig. 2, both the cumulative number of unique IP prefixes and the cumulative number of unique ASes linearly increase over time. The slopes indicate that approximately five new IP prefixes and two new ASes appear in a day. This constant increase in the number of new networks accessed per day (after the initial startup period) was initially surprising, as we had expected that users would generally would access the same set of networks over time. We'll see later that a user typically does spend most of the time in the same (relatively small) number of networks over time, but does visit new networks outside of this set of common networks at a constant rate, resulting in the constant slope in Fig. 2.

III. MEASUREMENT RESULTS, ANALYSIS, AND DISCUSSION

In this section, we present our measurement results regarding user occupancy within networks and transitioning among these networks. We also discuss insights and implications drawn from these results.

A. A user's online time: sessions, residence time, and multi-session time

Let us define the following terminology to describe and characterize user behavior, using the notion of a time-interval of length Δt, as shown in Fig. 3. Time is divide into intervals of length Δt. A period of time during which the user is "online" (a user's **online-time**) is a series of *consecutive* time intervals during which the user has one or more IMAP log entries. Similar to online-time, a **session** consists of a series of consecutive time intervals during which a user has one or more IMAP log entries with a client-side IP address in the *same* network domain (distinguished by either an IP-prefix or

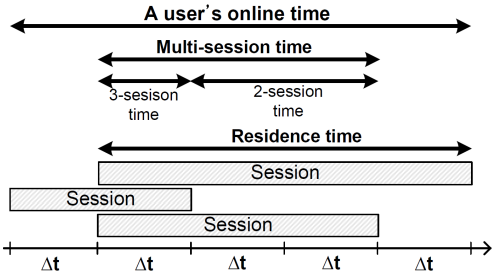


Fig. 3. An illustrative example on a user's online time

an AS).[2] The length of a session is the user's **residence-time** in that network; this is time that the user is considered to be continuously attached to that network.

Our measurements indicate that a user may also have IMAP log entries from *more than one network domain* during a time interval. As discussed earlier, this could result from a multi-homed user (i.e., a user with a single device connected to two or more networks), or a user accessing the IMAP server from multiple devices that are contemporaneously connected to different network domains. In either case, the user is reachable from the IMAP server in two different networks during that interval. Thus, we define a **multi-session** as a series of consecutive time windows, each of which has IMAP log entries for that user, with client addresses in two or more different network domains.

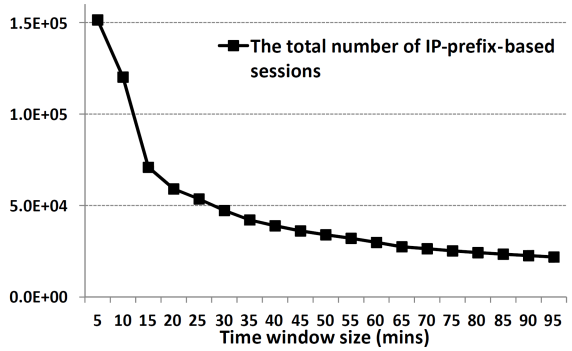


Fig. 4. Total number of all users' IP-prefix-based sessions for different time-window sizes.

But what value should we choose for Δt? If we choose a session time that is too small, a series of consecutive time intervals for a user who is online for that period of time (i.e, has a device actively connected to a network during that period of time) may contain time intervals with no IMAP entries for that user. Thus, a small time window could break that user's single online session into multiple distinct online sessions separated by intervals during which the user is considered offline (i.e., not connected to that network). Conversely, if the time interval is too large, two intervals of time during which the user is truly online and separated by a period of time during which the user is truly offline could be coalesced into a single online session. This dilemma is often faced when reconstructing user session behavior from discrete log entries [13,3]. Fig. 4 plots the total number of all users' IP-prefix-based sessions for different time-window sizes. We see that the number of sessions initially decreases sharply with increasing values of Δt, and then, at around a time interval length of 15 minutes, begins decreasing more slowly. We will thus choose the rough "knee" of this curve at 15 minutes to be the length of the time window in our subsequent discussion, unless otherwise noted. We also found that the percentage difference in the number of sessions observed in our IMAP logs (for different window sizes) based on whether we use IP prefixes or AS numbers reported by *whois* to distinguish among different "networks." That difference was always less that 1.5% for different window sizes, differing by only 0.05% for a 15-minute time window, giving us further confidence in our choice of 15 minutes as the standard time window value.

[2]Thus, two IMAP log entries in the same interval and having different IP addresses but the same IP-prefix (or the same AS number) would be regarded as belonging to the same session.

B. Residency time

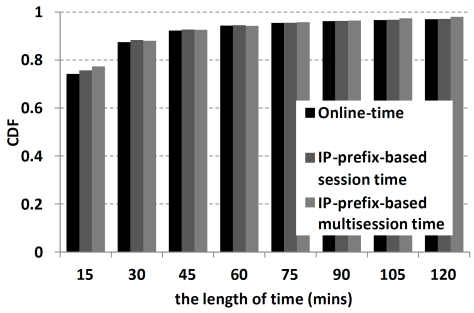
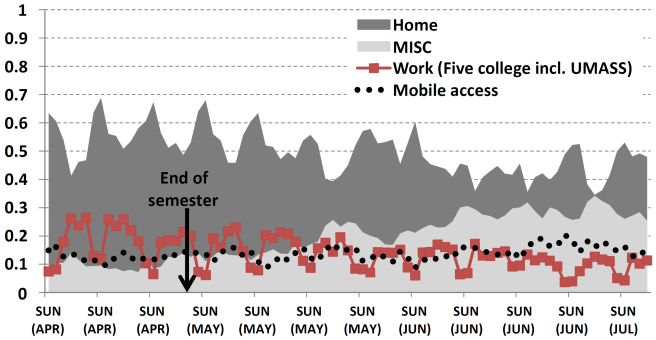


Fig. 5. The CDF of the length of all users' online-time, IP-prefix-based session time, and IP-prefix-based multi-session time.

Fig. 5 plots the CDF of the length of all users' online-time, residence-time, and multi-session time using IP-prefix distinction. Fig. 5 shows that approximately 80% of online-times, residence-times, and multi-session times are 15 minutes or less in length. The CDFs of the online-time is slightly less than the CDF of residence-time, which in turn is slightly less than the CDF of multi-session time, but the differences are not significant. A comparison of IP-prefix and AS distinctions in the CDF of the length of all users' online-time, residence-time and multi-session time also indicates that there is not a significant distinction between the number of IP-prefix-based and AS-based sessions (or multi-sessions). Thus, our results indicate that a user is rarely connected using different IP prefixes within the same AS within a single session.

Having considered session length characteristics, let us next examine the networks in which user sessions occur. Fig. 6 plots (over all users) the fraction residence time spent in home, work, mobile-access, and MISC (Miscellaneous) networks, which are defined as follows:

- The home category includes the Comcast cable network (AS7922), Verizon Online (AS19262), Charter



AS description	percent
Comcast cable (AS7922)	37%
Five college (AS1249)	14%
Charter communications (AS20115)	7%
AT&T wireless (AS7132)	6%
Verizon online (AS19262)	6%

Fig. 6. The fraction of all users' network-residence time.

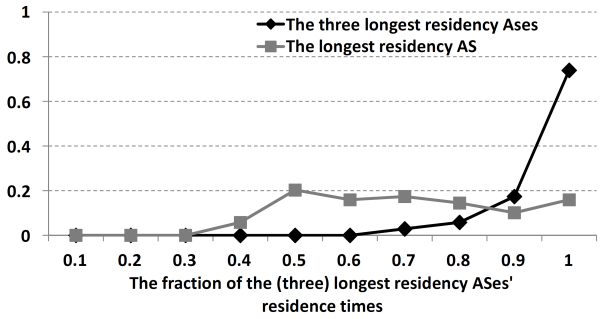
communications (AS20115), and Hughes network systems (AS6621).

- The work category includes the Five College AS (AS1249), including the UMass campus network.
- The mobile-access category includes Verizon wireless (AS6167 and AS22394), AT&T wireless (AS7132, AS20057), and Sprint wireless (AS3651).
- The MISC category includes all other network domains observed in our logs.

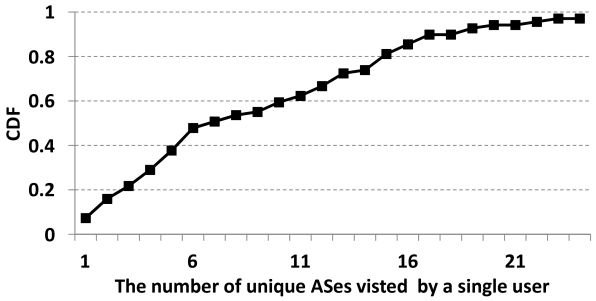
Fig. 6 plots the fraction of residence time spent (daily) in various types of networks (home, work, mobile-access and MISC) as a function of time. Perhaps not unexpectedly, the residence time in home and work networks shows a weekly pattern, with percentage of time in work networks higher on workdays and less on weekend days, and the percentage of time in home networks higher on weekend days and less during workdays. Fig. 6 also shows a decrease in home and work occupancy and a concomitant increase in MISC in particular after the beginning of May, which corresponds to the end of classes for the Spring 2013 semester.

Given that the home, work and mobile-access networks are collectively constituted by only 10 ASes (out of the 183 ASes observed in our traces), Fig. 6 also shows that users spend the majority of their time (more than 80% through early May and approximately 70% after early May) resident in only a small number of networks. The table in Figure 6 lists the ASes for which the overall user residency time (over all users' time) is greater than 5%, also confirming the observation that the lion share of user time is spent in a relatively small number of networks. Note that just two networks (Comcast and Five College) account for more than half of the overall residency time and that the five most common networks collectively account for 70% of the overall residency time.

Having observed that users *in aggregate* spend most of their time in a small number of home, work, and mobile-access networks, it is natural to examine the extent to which this is



(a) PDF of the fraction of the (three) longest residency ASes' residence times to the total residence times.



(b) CDF of the number of unique ASes visited by a user.

Fig. 7. Individual user network domain residency, networks visited.

true for *individual* users as well. For a given user, what is the fraction of its residence time spent in the singe network in which it is most often resident, or in the three networks in which together it is most often resident? Fig. 7(a) and (b) plot the empirical distribution (over all users) of the fraction of time that a user spends resident in the network in which it is most often resident (grey line with box points), and in the in the three networks in which together it is most often resident (black line, diamond point). The solid black curve in Fig. 7(a) indicates, for example, that approximately 75% of the users spend between 90% and 100% of their time in their top three networks, and that nearly 20% of the users spend between 80% and 90% of their time in their top three networks. Thus we see that individual users generally also spend the lion share of the residency time in just a few (e.g., three) networks. A much smaller fraction of the users spend their time in just one network - the gray curve in Fig. 7(a) in dictates that less than 20% of the users spend 90% to 100% of their time in their most commonly resident network. Fig. 7(b) plots the CDF of the number of networks visited by a user (over all users) during our study. Approximately 90% of the users were resident in 15 or fewer networks during the three month measurement period.

C. Multi-sessioned users

Having considered a user's connectivity to individual networks, let us next examine the fraction of time a user spends contemporaneously connected to two or more networks. Fig. 8 plots the fraction of users (y-axis) who spend a given fraction of their time (x-axis) contemporaneously connected to multiple networks. Fig. 8 indicate, for example, that 20% of the users

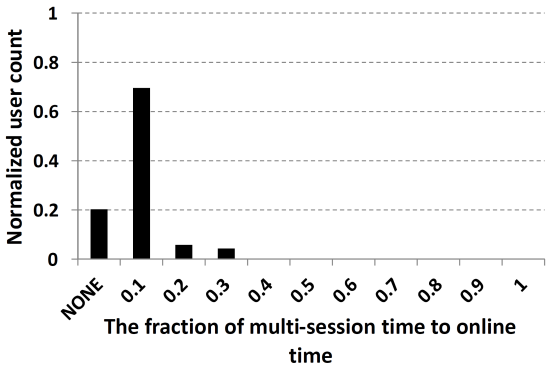


Fig. 8. The distribution of the fraction of total IP-based multi-session times to total online-times per user with 15 mins-time-window.

were always connected to a single network (when online). Approximately 70% of the user spent less than 10% (but greater than 0%) of their time multi-sessioned (i.e., contemporaneously connected to more than one network), and approximately 7% of users were multi-sessioned between 10 and 20% of their time online. We also observed that 98% of multi-sessions consist of only two network domains, but observed instances where a user would be contemporaneously connected to four network domains in a 15 minute interval, from both an IP-prefix and AS distinction.

	ASes used as a multi-session	
1)	Comcast (AS7922), Five college (AS1249)	20%
2)	Comcast (AS7922), Verizon wireless (AS22394)	15%
3)	Five college (AS1249), AT&T Internet (AS7132)	14%
4)	Free Mobile SAS (AS12322, AS51207)	8%
5)	Charter (AS20115), AT&T Internet (AS7132)	7%
6)	Comcast (AS7922), AT&T Internet (AS7132)	4%
7)	Five college (AS1249), Verizon wireless (AS6167)	4%
8)	Comcast (AS7922), Five college (AS1249), Verizon wireless (AS22394)	2%

TABLE II
A LIST OF THE MOST FREQUENTLY OCCURRING AS-COMBINATIONS IN A MULTI-SESSION.

TABLE II examines multi-session behavior in more detail, showing the most commonly observed AS-combinations constituting a multi-session, together with the fraction of the amount of time for that combination to the total amount of multi-sessioned time. We make the following observations:

- TABLE II rows (1) and (8) corresponds to the case of two networks (Comcast and the Five College network) with little overlap in their physical footprints - the Five College network is generally confined to campus locations, and Comcast is a residential network. Contemporaneous access to these two networks in a 15-minute interval could result from a user physically moving from one network to another (e.g., office to home or vice versa). This would be consistent with our observation (discussed shortly) that by far the most common transition between networks (i.e., moving from residency in one network in one interval to a different network in the subsequent interval) is between the Five College and the Comcast network. Contemporaneous residency in the Comcast and Five College network could also result from VPN access to the Five College network via the Comcast residential network, as discussed earlier.
- 46 percent of multi-sessions in TABLE II, corresponding to rows 2), 3), 5), 6), 7), and 8), consist of a fixed (residential or Five College) and a mobile-access network. These scenarios could correspond to cases of a user with multiple devices contemporaneously connected to different networks (e.g., a laptop connected to a wired network and a smartphone connected to a cellular data network), or to a single device with multiple NICs connected to different networks.
- *Transitions within the same provider.* 8 percent of multi-sessions in TABLE II, corresponding to row 4), shows contemporaneous access from two ASes owned by a single mobile service ISP (Free Mobile SAS). This may correspond to the case of a user who is either physically moving and connecting to different Free 802.11 base stations while in motion, or a stationary user connecting to different Free base stations within the 15-minute interval.

D. User network transitions

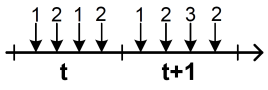


Fig. 9. An example of user network transitioning.

Having considered network residency and multi-session characteristics, let us next examine user transition among networks. We will analyze our measurement results using 15-minute discrete time intervals (as before) and in a more fine-grained continuous time manner. Suppose that a series of IMAP entries from a user from network domains 1, 2, and 3[3] in intervals t and $t+1$ are observed as shown in Fig. 9. In discrete time, the user is contemporaneously resident in networks (1,2) in interval t and in networks (1,2,3) in interval $t = 1$. We thus have one discrete time transition in the example.[4] In continuous time, we consider the two consecutive IMAP entries by a user from different domains as a network transition. In this example, 7 continuous time transitions occur. A discrete analysis is useful for defining and analyzing session characteristics, while a continuous time analysis is useful for analyzing the amount of signaling that might be associated with a user joining a new network.

Table III lists the most common transitions between network pairs (in both directions) from a continuous time perspective. Not surprisingly, since we have seen that a user is typically resident in a small number of networks, Table III indicates that

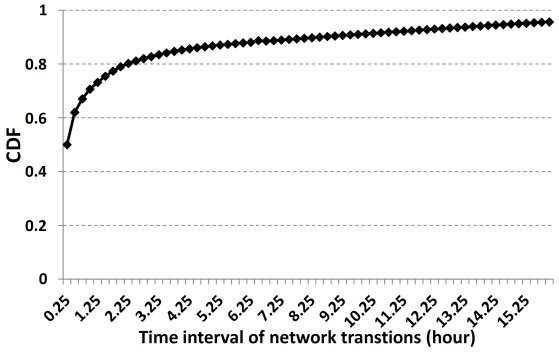
[3]AS and IP prefix distinctions do not produce significantly user-transitioning results; thus we present data for transitions among ASes only.

[4]We consider a transition to occur in the discrete time case whenever interval $t+1$ contains an IMAP entry from that user in at least one network not seen in interval t.

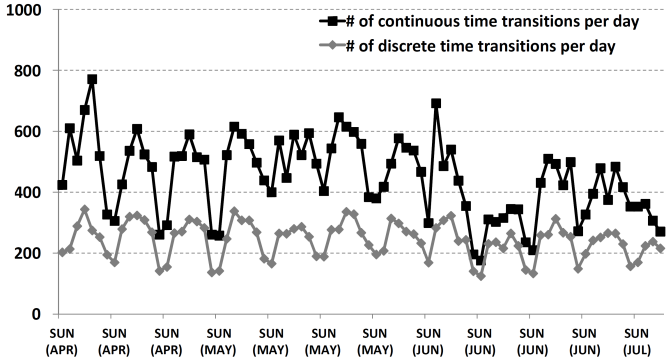
Network transition	percent
Comcast (AS7922) ⇌ Five college (AS1249)	11.30%, 11.17%
Comcast (AS7922) ⇌ Verizon wireless (AS22394)	6.13%, 6.10%
Five college (AS1249) ⇌ AT&T wireless (AS7132)	5.60%, 5.40%
Comcast (AS7922) ⇌ AT&T wireless (AS7132)	3.13%, 2.99%
Free SAS (AS12322) ⇌ Free mobile SAS (AS51207)	2.86%, 2.86%
Five college (AS1249) ⇌ Verizon wireless (AS6167)	2.66%, 2.62%
Charter (AS20115) ⇌ AT&T wireless (AS7132)	2.23%, 2.21%
Five college (AS1249) ⇌ Verizon wireless (AS22394)	1.10%, 1.10%
Comcast (AS7922) ⇌ Verizon wireless (AS6167)	0.86%, 0.83%

TABLE III
MOST FREQUENTLY OCCURRING AS TRANSITIONS, AND PERCENTAGE OF OCCURRENCES OF THESE TRANSITIONS

most of the transitions occur between a relatively small number of networks. Specifically, seven networks appear in Table III, and approximately 71% of the 34,610 network transitions in our logs take place among these seven networks.



(a) CDF of the time interval between AS transitions.



(b) AS transition rates per day.

Fig. 10. Characteristics of empirically-observed AS transitions.

Fig. 10(a) and (b) plot the CDF of the time interval between a user's AS transitions, and the aggregate number of (daily) discrete and continuous time transitions over time, respectively. Here we define a transition to occur when the network(s) in which a user is newly resident differs from the previous network(s) in which the user was last resident, regardless of the amount of time that has passed since residency in the previous network(s). Fig. 10(a) shows that users change their network domains at least once a day and 50% of transitions happen in less than 15 minutes. Since a large number of transitions happen in less than 15 minutes, we also observe in Fig. 10(b) that the aggregate network transition rate in discrete time (where at most one single transition can occur from one 15 minute interval to the next) is less than the continuous time transition rate. Fig. 10(b) also shows that network transitions occur less frequently during weekend - not surprising since 48% of transitions involve a work network (AS1249) and our user population is primarily faculty and staff.

IV. MARKOV MODEL OF USER TRANSITIONING AMONG NETWORKS

In this section, we develop a parsimonious discrete-time Markov chain model of an individual user's transitions among networks and validate how well performance measures for the aggregate population (in particular, signaling overhead due to transitioning between networks) predicted by the model match empirical measures determined from the traces.

A. Parsimonious discrete-time Markov chain

Since a primary goal of our model is to characterize signaling overhead (e.g., signaling to the location management components of a mobility-aware architecture), our discrete-time Markov chain model encodes enough state information to compute the rate at which a user generates signaling messages as it goes online/offline and transitions among networks. Our unit of discrete time is the time window discussed in section III; see Figure 9.

Our Markov chain model has two dimensions. Let the random variable $X = X_t$ be the number of networks in which a user is resident at time t. X_t may take values $\{0, 1, *\}$, where $*$ denotes the case that a user has two or more contemporaneous sessions; see Figure 3. For simplicity, we do not distinguish the case of more than two contemporaneous sessions from the case of exactly two such sessions, since 98% of multi-sessions consist of only two network domains, as discussed in Section III. Our model can be easily extended to cover the more general case. The first dimension of the Markov chain tracks the value of X. Let $Y = Y_t$ be the number of *new* network domains in which the user is resident at time t, with respect to time *t-1*. The second dimension of the Markov chain tracks the value of Y. A value of $Y_t = *$ encodes the case that the number of new networks domains is two or more. This second state variable will be used to quantitatively compute signaling overhead using our model.

Our Markov model thus consists of six states, $\{(0,0), (1,0), (1,1), (*,0), (*,1), (*,*)\}$. With these six states, we have a stochastic transition probability matrix $P = [p_{ij}]$ where $p_{ij} = Pr\{(X_t, Y_t) = j | (X_{t-1}, Y_{t-1}) = i\}$ and $\sum_j p_{ij} = 1$. These transition probabilities will be

determined empirically from our traces. A signaling cost matrix $C = [c_{ij}]$ is also associated with these states, where c_{ij} denotes the number of signaling messages generated when a user makes a transition from state i to state j:

$$C = \begin{array}{c} \\ (0,0) \\ (1,0) \\ (1,1) \\ (*,0) \\ (*,1) \\ (*,*) \end{array} \begin{array}{c} \begin{array}{cccccc} (0,0) & (1,0) & (1,1) & (*,0) & (*,1) & (*,*) \end{array} \\ \begin{bmatrix} 0 & 0 & 1 & 0 & 0 & 2 \\ 0 & 0 & 1 & 0 & 1 & 2 \\ 0 & 0 & 1 & 0 & 1 & 2 \\ 0 & 0 & 1 & 0 & 0 & 2 \\ 0 & 0 & 1 & 0 & 0 & 2 \\ 0 & 0 & 1 & 0 & 0 & 2 \end{bmatrix} \end{array}$$

The costs of infeasible transitions in C is 0. For example, the transition from $(0, 0)$ to $(*, 1)$, i.e., a user is newly connected to more than one domain but only generates one signaling message, can not occur and thus has cost 0.

B. Trace stationarity

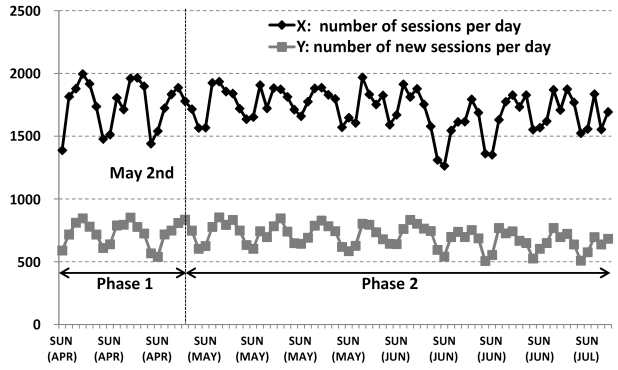


Fig. 11. The number of all users' sessions and new sessions per day.

Since our Markov chain model models stationary behavior, we check the traces themselves for stationarity using the *runs test*. The runs test examines underlying trends or variations of data using null hypothesis tests. As shown in Fig. 11, we divide our traces into two subtraces: *phase-1* and *phase-2*. Phase-1 consists of 19 days ending May 2nd, 2013 (roughly corresponding to the end of classes on the UMass Amherst campus); phase-2 consists of 63 days from May 3rd to July 5th 2013. Fig. 11 plots the time series of the number of all users' IP-prefix-based sessions per day and the number of all users' *new* IP-prefix-based sessions ("new" in the sense of Y_t) per day. Given a trace, we apply the runs test as follows.

1) Using 15-minute time windows, we derive a time series of the number of all users' IP-prefix-based sessions (corresponding to X) and a time series of the number of all users' new IP-prefix-based sessions (corresponding to Y).
2) We then divide the time series into time intervals of one-day and apply the runs test.

We separately apply the runs test to the entire trace, phase-1 trace, and phase-2 trace. At the 5% significance level, the phase-1 and phase 2 traces separately pass the stationarity test, but the entire trace as a single time series does not. We found similar results using AS-based sessions. Thus, we work with phase-1 and phase-2 traces separately, focusing here on the longer phase-2 trace.

C. Validation

We test how well our model predicts the number of signaling messages generated per time-step by 70 users in aggregate and the expected number of online users per time-step. We bisect the phase-2 trace into two subtraces: *subtrace1* and *subtrace2,* consisting of data from May 3rd to June 3rd, and from June 4th to July 5th 2013, respectively. We will use *subtrace1* to derive the transition rates for our Markov chain model of a canonical user; we will use *subtrace2* to validate how well our model (with parameters empirically derived from *subtrace1*) predicts the signaling overhead found in *subtrace2*. We proceed as follows:

1) **Canonical user model.** Using the data from 70 users in *subtrace1,* consisting of 218,592 user transitions, we derive the transition probabilities for our Markov chain model of an individual canonical user by counting the number of times that all users move from state i to state j, and then normalize these counts so that the sum of the transition counts out of each state equals 1. The empirical transition probability matrix, $\hat{P} = [\hat{P}_{ij}]$ is as follows:

$$\hat{P} = \begin{array}{c} \\ (0,0) \\ (1,0) \\ (1,1) \\ (*,0) \\ (*,1) \\ (*,*) \end{array} \begin{array}{c} \begin{array}{cccccc} (0,0) & (1,0) & (1,1) & (*,0) & (*,1) & (*,*) \end{array} \\ \begin{bmatrix} 0.89 & 0 & 0.11 & 0 & 0 & 0 \\ 0.16 & 0.75 & 0.03 & 0 & 0.06 & 0 \\ 0.67 & 0.25 & 0.05 & 0 & 0.02 & 0 \\ 0.03 & 0.26 & 0.01 & 0.64 & 0.07 & 0 \\ 0.12 & 0.52 & 0.02 & 0.23 & 0.10 & 0 \\ 0.38 & 0.37 & 0.07 & 0.08 & 0.09 & 0 \end{bmatrix} \end{array}$$

2) **Generating synthetic transitions for a population of canonical users.** Using the canonical user model $\hat{P}$ generated from *subtrace1* we synthetically generate 70 users' state transitions for 10^5 time-steps, and then compute the number of aggregate signaling messages generated per time-step using the matrix C. We also determine the number of online users per time-step in the synthetic data.
3) **Validation.** To validate our model, we compare the model-predicted values (whose state transition probabilities were derived from *subtrace1*) with the empirical distribution found in *subtrace2*.

state	$(0,0)$	$(1,0)$	$(1,1)$	$(*,0)$	$(*,1)$	$(*,*)$
Model-based	76%	13%	9%	1%	1%	0%
Observed	78%	12%	8%	1%	1%	0%

TABLE IV
MODEL-BASED AND EMPIRICALLY OBSERVED STATE OCCUPANCIES.

As shown in Fig. 12, the model-predicted distribution of the number signaling messages generated by 70 users shows a good match to the empirically observed distribution in *subtrace2*. This can be confirmed by applying the Chi-Square goodness-of-fit test to compare the equality of two discrete distributions. Our hypothesized distribution passes the test with a 5% significance level. Fig. 12 also shows that the model-based CDF matches

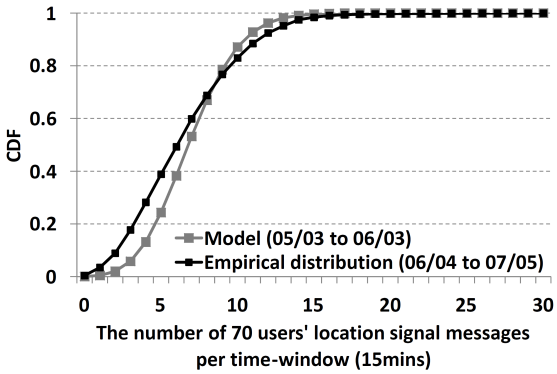


Fig. 12. Model-predicted and empirically observed CDFs for the total number of location signaling messages generated by 70 users.

the empirical distribution best at high signaling rates, precisely when signaling overhead become most critical.

Table IV compares model-based and empirically-observed state occupancies, again showing good agreement. For example, the model predicts that a user is offline 76% of the time, while we empirically observe in *subtrace2* that a user is offline 78% of the time. Additional model validation can be found in Appendix.

V. Related Work

Numerous studies have characterized physical human movement using empirical datasets and discussed the impact of physical user mobility patterns on network performance and design. Human mobility traces have been collected from diverse access networks such as WLAN [12, 9, 2], Bluetooth networks [2], and cellular networks [7, 14, 10]. Research using Wi-Fi access datasets has been done in a single, physically-scoped network domain, such as a campus or enterprise, thus focusing on user mobility within that limited physical domain. In this sense, cellular network data might more fully model human mobility (since users typically carry their cellular phones); such cellular data, however, is typically proprietary. But individual WiFi and cellular traces by definition only include data from an individual type of network, and have not considered contemporaneous residence within multiple networks nor transitions among networks. More generally, we believe there is an important distinction to be made between physical mobility and mobility among networks, as discussed in Section I; our work is the first to characterize and model mobility among networks (which we have referred to as network transitioning).

[7, 14, 3] have related human mobility patterns to network resource use in Wi-Fi access points or cellular network base stations. [7, 14] have found that the extent of users' physical mobility is low and concentrated among a small number base stations, with infrequent visits to other base stations in that network; we have similar, complementary findings for the extent of user transitioning among networks.

VI. Conclusion

In this paper, we performed a measurement study of user transitioning among networks and discussed insights and implications from the measurements. Our measurement study, conducted using IMAP server logs of a population of approximately 70 users for three months, characterizes user network transitioning in terms of transition rates, network residency time, and degree of contemporaneously resident network domains. Based on these measurements, we also developed and validated a parsimonious discrete time Markov chain model of canonical user transitioning among networks. Our measurements and models provide quantitative insight into the location management signaling overhead needed by modern and proposed name/address translation and location management protocols; our models provide the ability to design, dimension and analyze such systems.

Our future work is aimed at extending the scope of our study (both over time and numbers of users), and instrumenting and measuring client devices, and comparing client side measurements with server-side measurements.

Acknowledgments. This research is supported in part by the US National Science Foundation, under NSF Award CNS-104078.

References

[1] Team cymru research nfp, ip to asn mapping, http://www.team-cymru.org/Services/ip-to-asn.html, 2013.

[2] A. Chaintreau, P. Hui, J. Crowcroft, C. Diot, R. Gass, and J. Scott. Impact of human mobility on opportunistic forwarding algorithms. *IEEE Trans. Mobile Computing*, 6(6):606–620, 2007.

[3] Y.-C. Chen, J. Kurose, and D. Towsley. A mixed queueing network model of mobility in a campus wireless network. In *IEEE INFOCOM*, pages 2656–2660, 2012.

[4] M. Crispin. Internet Message Access Protocol - v4rev1. RFC 3501 (Proposed Standard), Mar. 2003.

[5] D. Farinacci, V. Fuller, D. Meyer, and D. Lewis. RFC 6830: The Locator/ID Separation Protocol (LISP), Jan. 2013.

[6] O. Goga, H. Lei, S. H. K. Parthasarathi, G. Friedland, R. Sommer, and R. Teixeira. Exploiting innocuous activity for correlating users across sites. In *WWW '13*, WWW '13, pages 447–458, 2013.

[7] E. Halepovic and C. Williamson. Characterizing and modeling user mobility in a cellular data network. In *in ACM PE-WASUN*, 2005.

[8] D. Han, A. Anand, F. Dogar, B. Li, H. Lim, M. Machado, A. Mukundan, W. Wu, A. Akella, D. G. Andersen, J. W. Byers, S. Seshan, and P. Steenkiste. XIA: Efficient support for evolvable internetworking. In *Proc. 9th USENIX NSDI*, San Jose, CA, Apr. 2012.

[9] W.-j. Hsu, D. Dutta, and A. Helmy. Structural analysis of user association patterns in university campus wireless lans. *IEEE Trans. Mobile Computing*, 11(11):1734–1748, Nov. 2012.

[10] S. Isaacman, R. Becker, R. Cáceres, M. Martonosi, J. Rowland, A. Varshavsky, and W. Willinger. Human mobility modeling at metropolitan scales. In *ACM Mobisys '12*, pages 239–252, 2012.

[11] D. Johnson, C. Perkins, and J. Arkko. RFC 3775: Mobility Support in IPv6, June 2004.

[12] M. Kim, D. Kotz, and S. Kim. Extracting a mobility model from real user traces. In *IIEEE NFOCOM*, Barcelona, Spain, April 2006. IEEE Computer Society Press.

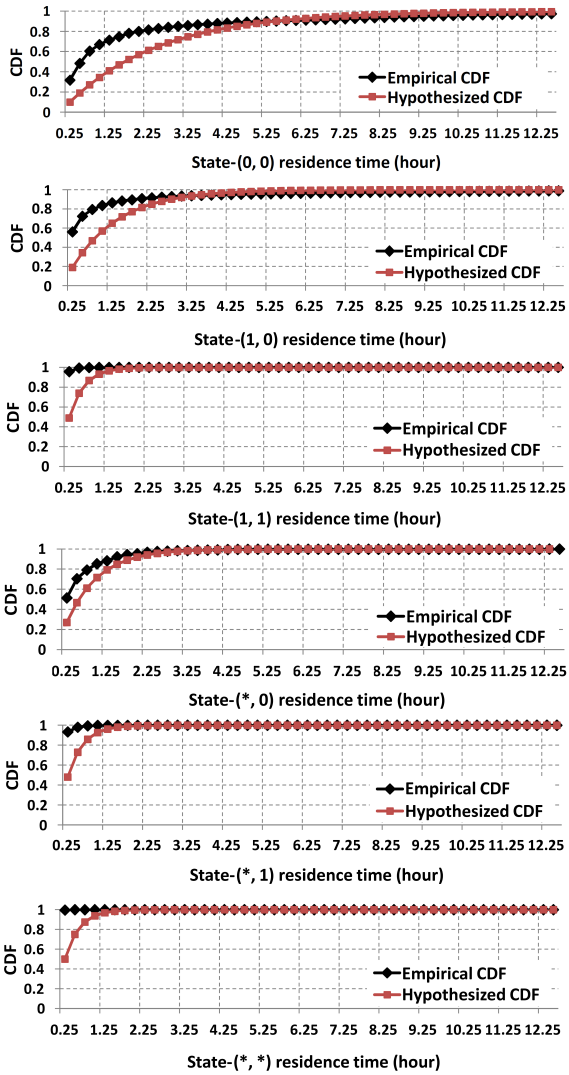
[13] J. Padhye and J. F. Kurose. Continuous-media courseware server: A study of client interactions. *IEEE Internet Computing*, 3(2):65–73, 1999.

[14] U. Paul, A. Subramanian, M. Buddhikot, and S. Das. Understanding traffic dynamics in cellular data networks. In *IEEE INFOCOM*, pages 882–890, 2011.

[15] A. Venkataramani, A. Sharma, X. Tie, H. Uppal, D. Westbrook, J. Kurose, and D. Raychaudhuri. Design requirements of a global name service for a mobility-centric, trustworthy internetwork. In *IEEE COMSNETS*, 2013.

[16] www.3gpp.org.

APPENDIX



state	$(0,0)$	$(1,0)$	$(1,1)$	$(*,0)$	$(*,1)$	$(*,*)$
n	34,300	12,636	34,473	1,041	3,704	1,609
θ_i	0.10	0.19	0.49	0.27	0.48	0.50

Fig. 13. State-(*, 0)'s state residence time CDF and all states' estimated parameters.

We investigate that our trace has goodness-of-fit in each state's state residence time distribution. Using the phase-2 subtrace, we get a sequence of states, determined in time spaced at 15 mins, e.g., $\{(0,0),(1,1),(1,1),(2,1),\cdots\}$ and get a set of the number of time-steps of a user to stay in each state. Hypothesizing that the set follows a Geometric distribution, we estimate parameter θ_i for state i using the *maximum likelihood estimation (MLE)*. Then we the evaluate the quality of fit using Chi-Square goodness-of-fit. Fig. 13 plots the CDF of state-(*, 0)'s empirical distribution and hypothesized Geometric distribution and gives a list of all states' estimated parameters for their hypothesized Geometric distributions. Fig. 13 shows that the empirical distribution seems close to the hypothesized distribution for state-(*, 0) but the test shows that all the states' hypothesized Geometric distribution do not have goodness-of-fit to their empirical results.