Measurement and Modeling of User Transitioning Among Networks

Sookhyun Yang, Jim Kurose, Simon Heimlicher, Arun Venkataramani
University of Massachusetts, Amherst, MA, 01003
{shyang, kurose, heimlicher, 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 user identity (name) mobility in both traditional mobility management protocols such as Mobile-IP and in new architectures, such as XIA and MobilityFirst, that support identity mobility and location independence as first class objects. A multi-homed stationary user or a stationary user shifting among multiple devices attached to different networks will persistently keep his/her identity but will change access networks and the IP address to which his/her identity is associated. We perform a measurement study of such user transitioning among networks from a network-level point of view, characterizing the sequence of networks to which a user is attached and discuss insights and implications drawn from these measurements. We characterize network transitioning in terms of network residency time, degree of multi-homing, transition rates 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 develop and validate a parsimonious Markov chain model of canonical user transitioning among networks that can be used to provision network services and to analyze mobility protocols.**

I. Introduction

”Mobility” in computer networks takes two distinct forms: physical (human) mobility among a network’s access points and base stations, and *virtual* mobility of a user identity among the many networks that make up the larger Internet. 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, intra-network routing and 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. Virtual mobility – mobility *among networks* – is a more recent concern of protocols such as Mobile-IP and new architectures such as XIA [8] and MobilityFirst [16], which aim to provide location independence (mobility transparency) by separating identifiers (names) from addresses or network locations. Here, the need to map a user’s identity to his/her current network location via mobility registration and lookup/indirection protocols, are central concerns. Thus, a quantitative understanding of how a user identity transitions among access networks – the networks through which that identity is addressed and ultimately reached – is of great interest for mobility architecture and protocol design and analysis.

Recognizing the potential ambiguity between physical and virtual mobility, we will refer to a user identity moving among networks from a network-layer/addressing viewpoint as *transitioning* among networks. To appreciate this distinction, consider an individual, say Alice, often connected to the Internet via numerous different networks during the course of her day. She might begin her day reading email on a tablet, connected to the Internet via a residential DSL or cable network or a wide-area wireless network; she might later work a bit from home using a computer connected via Ethernet to her residential network and then later connect wirelessly via a smartphone to her wide-area wireless network service provider as she bikes or drives to work. At work, Alice connects via the company network, but also uses a smartphone. At the end of the day, her transitioning among networks is repeated in reverse. Together, these networks might be considered Alice’s set of frequently used “home” networks. When traveling, Alice connects via a smartphone’s wireless provider network and via airport, airplane, cafe, hotel and remote institutional networks. Indeed, we see that the identity that is “Alice” connects to the Internet via *many* different networks over time and is sometimes connected using different devices on different networks at the same time.

A user’s transition between networks can occur in a number of different scenarios: *(i)* a user might detach from one network and attach to a new network (e.g., a user explicitly disassociating from one wireless network and then associating with a different wireless network); *(ii)* a user with multiple devices[1] might move his/her activity from a device attached to one network to another device attached to a different network, or use both devices concurrently; we will refer this latter as a user being “contemporaneously connected” to two (or more) networks; *(iii)* a user with one device with multiple NICs may change the interface being used, or use multiple NICs on the single device contemporaneously (which we believe is rare); *(iv)* a user may connect to a VPN, thus changing its network-visible IP address.

In this paper, we perform a measurement study of user-transitioning among networks and discuss insights and implications drawn from these measurements. Our study thus differs from previous mobility studies that have developed models of a single device’s changing MAC or IP addresses while physically moving between access points or base stations. 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 two sets of IMAP server logs (a year-long log of approximately 80 users, and a four-month log of a different population of more than 7,000 users) quantitatively characterizes network transitioning in terms of transition rates among networks, network residency time, degree of contemporaneous connection

[1] The use of multiple devices is increasing rapidly. The Pew Internet Research Project [15] notes that in addition to traditional Internet access via computers, 58% of Americans own a smartphone, with approximately 50% of these users using a smartphone as their primary Internet-connected device. 43% of Americans own a tablet, a thirteen-fold increase in ownership over four years.

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.

II. Measurement Methodology

In this section we first discuss the challenge of measuring user-transitioning at large scale and our decision to use IMAP logs to do so. We then provide details of the IMAP logs themselves and discuss the set of networks visited by users in our logs. We conclude this section with a discussion of how we estimate user session lengths based on log data.

A. Why IMAP mail access logs?

Measuring user mobility between networks is itself a challenging task. Measuring network connectivity directly at the end user requires a population of users willing to install software on *each* 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 monitoring connectivity on mobile devices. 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 among networks. 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 many commonly-accessed servers 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 used by an online 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 server [4] is a compelling choice. Email checking, polling and delivery all create entries in the IMAP server's log containing an associated client IP address, as well as a client's identifier (i.e., the email account); the email account typically remains the same across a user's 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 logs evidencing transitions from office network to cellular provider network to home access network. Although many e-mail clients periodically and automatically access their IMAP server while online (providing a rich source of IMAP data), not all clients do so. 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. And email is indeed but one application (albeit popular one). Thus, we can think of our results here informally as a lower bound on the actual amount of network-transitioning performed.

IMAP logs can be also 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 NICs connected to different networks, the multi-homed IMAP accesses 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 log collection

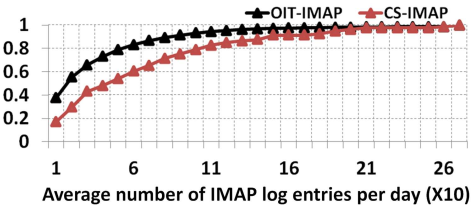


Fig. 1. CDF of the average number of IMAP entries per day over all users.

For this study, we collected two sets of traces from IMAP servers located at the University of Massachusetts Amherst. The CS-IMAP set contains logs from IMAP servers in the Computer Science Department from Apr 14, 2013 to Feb 22, 2014; the CS-IMAP has a population of 81 users mostly consisting of CS faculty and staff members. The OIT-IMAP set contains approximately four months of logs from IMAP servers that supports a mail service for university students (primarily), faculty and staff that is separate from the CS mail service. The OIT-IMAP has a campus-wide user population of 7,137 users; these traces were taken from Dec 1, 2013 to Mar 25, 2014. The total number of CS-IMAP and OIT-IMAP log entries per user over the measurement ranged from 2 to 79,392, and from 1 to 1,490,473, respectively. Fig. 1 plots the CDF of the average number of daily IMAP entries per user and shows that users in CS-IMAP (mostly faculty members) tend to access mail servers more frequently than OIT-IMAP (mostly students).

Each trace consists of a series of individual IMAP log entries stored by *syslog* [5], recording a user's e-mail activities including signing into the mail server, checking the INBOX, deleting messages, and unilateral server decisions to close (idle) connections. We processed only a user's sign-in logs which allowed us to extract the following pieces of information for each entry: *(i)* user's account ID - we consistently anonymized a user's account ID (email address) using SHA2-hashing for privacy purposes, *(ii)* timestamp - the time at which a user signs into the IMAP mail server to poll, check, or retrieve email, and *(iii)* a client-side IP address - this is the user's (client-side) IP address when accessing the IMAP server[2].

Given an IP address, we determined the user's IP prefix network, Autonomous system number (ASN), and network domain ownership via *whois* using *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. The CS-IMAP set contains 1,405 unique IP prefixes and 387 unique ASNs, and the OIT-IMAP set contains 9,016 unique IP prefixes and 1,777 unique ASNs. The network information for two IP addresses in the CS-IMAP and 63 IP addresses in the OIT-IMAP was unknown, but the number of IMAP logs generated from such IP addresses was negligible; these entries were excluded from our analysis.

C. IMAP traces: network information

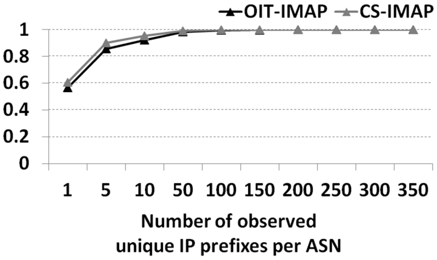


Fig. 2. CDF of the number of observed IP prefixes associated with an ASN over all users.

Fig. 2 shows the CDFs of the number of observed unique IP prefixes associated with an ASN over all users in the CS-IMAP and the OIT-IMAP sets. Fig. 2 shows that approximately 61% and 57% of ASNs had only a single observed IP prefix in the CS-IMAP, the OIT-IMAP, respectively. In the traces, the following ASN and IP prefix information of frequently visited service providers have been observed (we will investigate the length of time a user is resident in an IP prefix or ASN network in Section III). AT&T, Sprint, T-Mobile, and Verizon wireless are mobile access service providers. Comcast, Charter, Cox, Time Warner, and Cablevision are residential wired Internet service providers (e.g., cable and DSL access networks); the Hughes network supports a satellite Internet service used in rural communities lacking wired and cellular broadband service. The UMass Amherst network is part of the Five Colleges (AS1249) network. SAS in the CS-IMAP (a DSL and Wi-Fi service provider in France) and Unicom in the OIT-IMAP (a mobile service provider in China) were used for a non-negligible amount of time in our measurements.

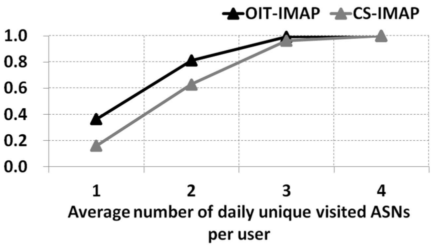


Fig. 3. CDF of the number of unique ASNs visited daily per user over all users.

Fig. 3 plots the CDF of the number of unique ASNs visited daily per user over all users, indicating that users in both OIT-IMAP and CS-IMAP access at most four unique ASNs in a day, but users belonging to CS-IMAP (mostly faculty members) access more ASNs than OIT-IMAP (mostly students).

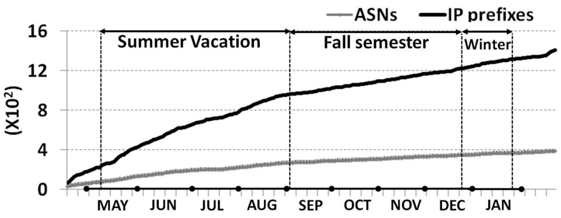


Fig. 4. CS-IMAP. Cumulative number of unique ASNs accessed by all users over time.

Fig. 4 plots the daily cumulative numbers of unique IP prefixes and ASNs accessed by all users over time. These figures indicate that the cumulative number of unique IP prefixes and ASNs each increase roughly linearly over time; the slopes of two curves during vacations (when users would be out of town more frequently) are steeper compared with the slope during the academic term. This constant increase in the daily number of new networks accessed (after the initial startup period) was initially surprising, as we had expected that users would generally access the same set of networks over time. We'll see later that a user typically does indeed 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 roughly constant rate, resulting in the positive slope in Fig. 4.

D. From IMAP log data to sessions

We use the notion of a *time window* to determine intervals of time during which a user is connected to a network.

Definition 1: Time is divided into consecutive *time windows,* each of length Δt. A **session** is defined as a series of consecutive time windows, each of which has one or more IMAP log entries from the *same* network (distinguished by either its IP prefix or ASN).

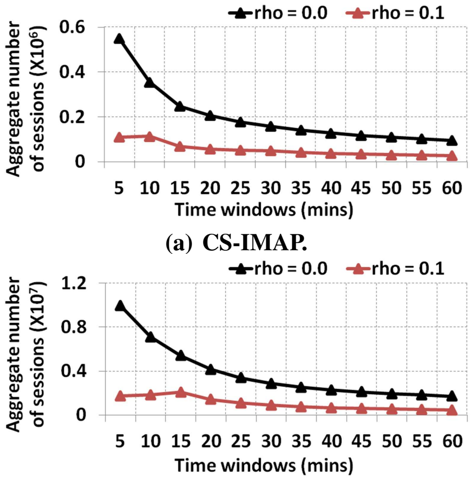
[2] Users in the CS-IMAP set occasionally accessed mail via a departmental web-based server, rather than directly from a client email application. In this case, the user's logged IP address is recorded in the IMAP log as 127.0.0.1; we analyzed the server's web logs to determine the client address of the user browser associated with this IMAP access. Only 1.6% of all IMAP web-based log entries could not be identified due to missing web logs; those entries were excluded from our analysis. 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.

By Definition 1, two IMAP log entries in the same time window that have different IP addresses but the same IP prefix (or the same ASN) would be regarded as belonging to the same session. Our measurements indicate that a user may be also connected to *more than one network* during a window of time.

Definition 2: Given time window of length Δt, a **multi-sessioned** time window for a user is one in which that user has IMAP entries from two or more different networks.

Choosing a value for Δt for session identification via Definition 1. If we choose a small time window value, this would break a user's single session into multiple distinct sessions separated by empty Δts having no IMAP logs entries. If a user was indeed connected during these empty Δt intervals, then we would overestimate the amount of user network-transitioning. Conversely, if the time window is too large, intervals of time during which the user disconnects and then reconnects to that same network would be coalesced into a single session, thus underestimating the amount of user transitioning. This dilemma is often faced when reconstructing user session behavior from discrete log entries [3, 13]. We choose the length of the time window Δt by observing the number of sessions as a function of Δt, as discussed below.

Definition 3: Given time window Δt, define ρ as the fraction of time windows that (i) contain no entries; (ii) fall between two time windows that contain IMAP entries, and (iii) in the ground truth case, the user remains connected to the network (even while producing no IMAP entries).

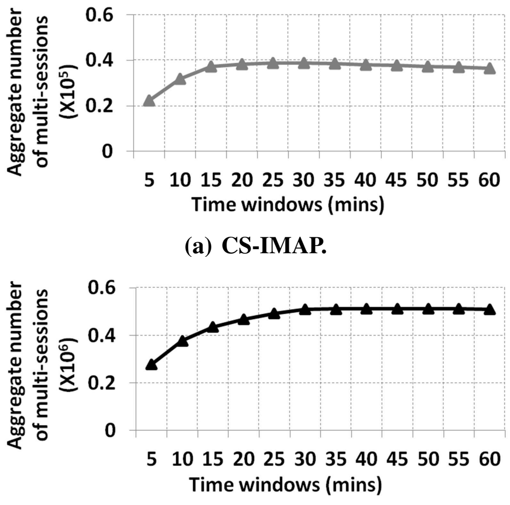


(a) CS-IMAP.

(b) OIT-IMAP.

Fig. 5. Aggregate number of sessions over all users.

Fig. 5 plots the total number of all users' ASN-based sessions[3] as a function of a time-window length for different values of ρ. The black curve in Fig. 5a shows that the number of sessions with $\rho = 0$ in CS-IMAP initially decreases sharply with increasing values of Δt, and then, at around a time-window length of 15 minutes, begins decreasing more slowly. Fig. 5a's red curve plots that the hypothetical number of sessions with $\rho = 0.1$ for different time-window sizes in CS-IMAP. The red curve is significantly lower than the black curve in the inital region, and then shows a knee of the curve at 15 minutes; this pattern was also found for different values of ρ. Similarly, the knees of the curves in OIT-IMAP appears at approximately 20 minutes as shown in Fig. 5b. We also noted that approximately 97% of the time intervals between a user's two consecutive IMAP log entries in CS-IMAP were less than or equal to 15 minutes, and approximately 82% of the time intervals between a user's two consecutive IMAP log entries in OIT-IMAP were less than or equal to 20 minutes.



(a) CS-IMAP.

(b) OIT-IMAP.

Fig. 6. Aggregate number of multi-sessioned time-slots over all users.

A similar analysis can be applied to the case of a user being contemporaneously connected to multiple networks. Fig. 6 plots the total number of all users' ASN-based multi-sessioned time windows for different time-window sizes. Fig. 6 shows that the number of multi-sessioned time windows in CS-IMAP increases until a window length of 15 minutes and then flattens out and the knee of the curve appears at 20 minutes, the same knee location found in the Fig. 5. Thus, a user who has been connected to multiple networks is likely to be completely offline for an amount of time greater than the time interval length at the knee. We will thus choose **15 minutes** in CS-IMAP and **20 minutes** in OIT-IMAP to be the length of the time window and identify user sessions accordingly via Definition 1. We will only show the results with $\rho = 0$ in our subsequent discussion.

III. Measurement Analysis and Findings

In this section, we present and discuss our measurement results regarding user residence time in various networks and multi-sessioned behavior.

A. Network residence time

House	Comcast (AS7015, AS7922, AS33651, AS33668), Charter (AS20115), Cox (AS22773), Hughes (AS6621), Time Warner (AS11351), Cablevision (AS6128)
Work	Five colleges AS (AS1249) (including UMass Amherst)
Mobile	Verizon (AS22394, AS701, AS6167), AT&T (AS20057, AS7018), T-Mobile (AS21928), Sprint (AS3651)

TABLE I. House, work, and mobile categorization.

[3] A comparison of using IP prefix vs. ASN distinctions to identify the number and length of sessions indicates that there is not a significant difference between IP prefix-based and ASN-based session lengths. Thus we only show ASN results.

Let us first consider the *aggregate* network residence time over all users spent in various networks. Table I defines house, work, and mobile networks whose constituent ASNs or ISPs are accessed by users for more than 0.5% of aggregate network residence time. The MISC category, which includes all other network domains observed in our logs, may thus include rarely-used residential wired service provider or mobile access provider ASNs that account for negligible fractions of network residence time. Broadly, we may consider the house/work/mobile networks as a user's "home" networks and the remaining MISC networks as a user's "visited" networks.

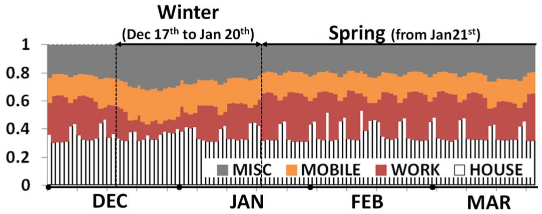
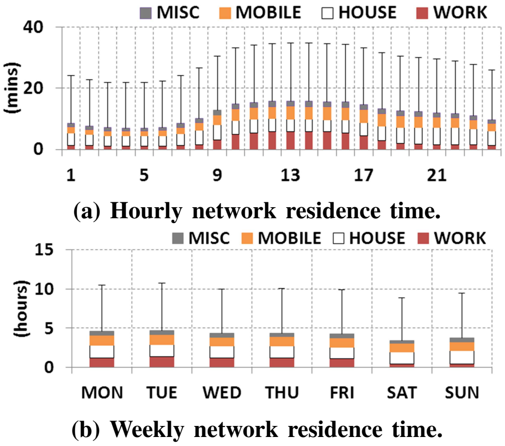


Fig. 7. OIT-IMAP. Time series plot of daily fractions of network residence times over all users.

Fig. 7, 8 plot the daily fraction of *aggregate* residence time spent in house, work, mobile and MISC ASNs over all users for OIT-IMAP, CS-IMAP respectively. Given that the house, work and mobile networks are collectively constituted by only 17 out of the 1,858 ASNs observed in CS-IMAP and OIT-IMAP, Fig. 7, 8 show that users spend the majority of their time (approximately 80% through a measurement period, and in particular more than 90% during fall semester in CS-IMAP) resident in only a small number of networks. We also observed that just two ASNs (Comcast AS7015, and Five colleges As1249) account for more than half of the overall residency time in OIT-IMAP and CS-IMAP, and that the ten most common ASNs collectively account for approximately 85% (for OIT-IMAP) and 90% (for CS-IMAP) of the overall residency time, confirming the observation that the lion share of aggregate user time is spent in a relatively small number of networks.

Fig. 7, 8 also show seasonality corresponding to the academic calendar; a decrease in work network occupancy and a concomitant increase in MISC network occupancy during vacations; conversely, an increase in house network occupancy and work network occupancy but a decrease in MISC network occupancy during semesters. Not surprisingly, Fig. 7, 8 also show per-week periodicity for house and work network residence times, with the percentage of time in work networks higher on workdays and less on weekend days, and the percentage of time in house networks higher on weekend days and less during workdays.

We also observe hourly and weekly patterns in the aggregate average and maximum for hourly and weekly network residence times (shown as box plots with whiskers in Fig. 9a and 9b over all users in OIT-IMAP. Fig. 9a shows that users tend to be connected approximately 10 minutes on average and up to 35 minutes per hour. Network residence time during daytime is longer than during nighttime, with an increase of residence time in work networks during the day. Fig. 9b shows that users are connected approximately 5 hours a day on average up to 10 hours per day. Network residence time during workdays is longer than during weekend days, with an increase of residence time in work networks during the week. Similar hourly and weekly results are also found in CS-IMAP.



(a) Hourly network residence time.

(b) Weekly network residence time.

Fig. 9. OIT-IMAP. Box plot with whiskers with average and maximum for hourly and weekly network residence time over all users.

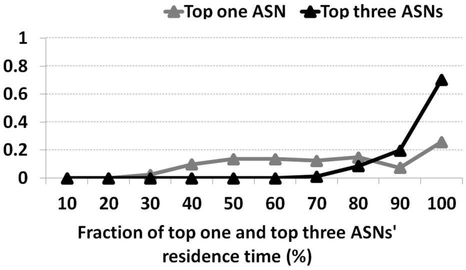


Fig. 10. CS-IMAP. pdf of the fraction of the (three) longest residency ASNs' residence times to the total residence times.

Let us now turn our analysis from the aggregate to the individual, and investigate the fraction of an *individual* user's residence time spent in the single network in which it is most often resident, as well as in the three networks in which together it is most often resident? Fig. 10 plots the distribution (over all users) of the fraction of time that a user in CS-IMAP spends resident in the network in which it is most often resident (grey line with triangle points), and in the three networks in which together it is most often resident (black line with triangle points). The black curve 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 their 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 indicates that roughly 25% of the users spend 90% to 100% of their time in their most commonly resident network. Similar results are also found in OIT-IMAP.

B. User's multi-sessioned behavior

Having considered a user's connectivity to individual networks, let us next examine a user's contemporaneous connection to two or more networks. In our measurements, we observe that 99% of the ASN-based multi-sessioned time

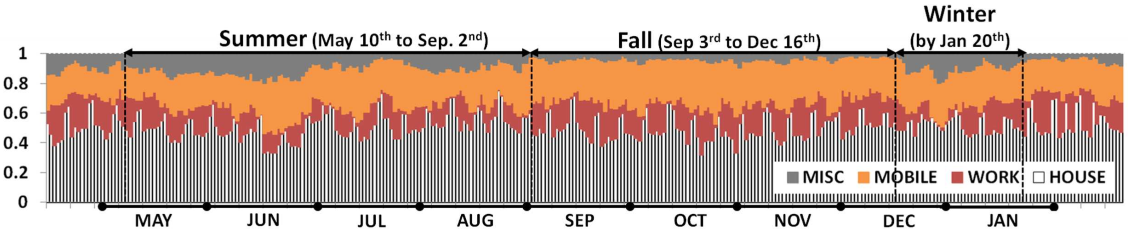


Fig. 8. CS-IMAP. Time series plot of daily fractions of network residence times over all users.

windows in OIT-IMAP and 99.5% of the ASN-based multi-sessioned time windows in CS-IMAP consist of only two ASNs.

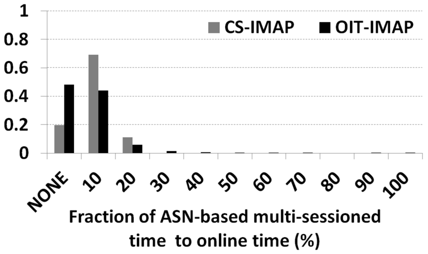


Fig. 11. pdf of ASN-based multi-session time per user.

Fig. 11 plots the fraction of users (y-axis) who spend a given fraction of their time (x-axis) connected to multiple networks in CS-IMAP and OIT-IMAP. Fig. 11's gray bar indicates, for example, that 20% of the users in CS-IMAP were always connected to a single network (when online). Approximately 70% of the users spent less than 10% (but greater than 0%) of their time multi-sessioned and approximately 7% of users were multi-sessioned between 10 and 20% of their time online. Fig. 11's black bar shows that approximately 50% of the users in OIT-IMAP were always connected to just a single network. Overall, however, we found the amount of multi-sessioned time to be much higher than we would have expected, suggesting that contemporaneous connectivity to multiple networks should not be considered "outlier" behavior.

A deeper investigation in the multi-sessioned time windows revealed three common scenarios, with the following potential causes of multi-sessions:

Fixed and mobile networks. 55% of multi-sessioned time windows in OIT-IMAP and 51% in CS-IMAP consisted of a fixed (residential or Five colleges) and a mobile network (as defined in Table I's mobile category). *(i)* These scenarios could correspond to the cases of a user carrying multiple devices or a single device with multiple NICs being contemporaneously connected to different networks (e.g., a laptop connected to a wired network and a smartphone connected to a cellular data network). *(ii)* Network transitions between fixed and mobile networks within a time window could also have resulted from a user's switching his/her devices.

Fixed networks across different ISPs. 17% of multi-sessioned time-slots in OIT-IMAP and 27% in CS-IMAP consisted of two fixed networks (residential and Five colleges) with little overlap in their physical footprints - the Five colleges network is generally confined to campus locations. *(i)* Contemporaneous access to these two networks in the same time window could have resulted from a user physically moving from one network to another (e.g., office to home or vice versa) or *(ii)* could also have resulted from emails being automatically by a user device in a different physical location that the user him/herself, or from VPN access to the Five colleges network via the residential network.

Network transitions within the same ISP. 6% of multi-sessions in OIT-IMAP and 4% in CS-IMAP show multiple networks access from two ASNs owned by a single service provider such as SAS, Verizon, AT&T and Comcast. This may correspond to the case of a user who is either physically moving and connecting to different 3G/4G or 802.11 base stations while in motion, or a stationary user connecting to different base stations within a time window.

Let us conclude this section by further dissecting the cases above to determine which multi-sessioned time windows might result from a user's transition between networks (e.g., as indicated by a series of IMAP log entries from one network followed by a series of IMAP log entries from another network during a time window) versus a user switching back and forth between networks in that time window. Let $S_{t_1}^{t_2}$ be a sequence of networks to which a user is connected from t_1 to t_2. For instance, if a user at t generates three consecutive IMAP log entries via network B followed by one IMAP log entry via network A, then $S_t^t = \{B, A\}$. We determine whether a user performs a network transition or is contemporaneously connected to multiple networks at multi-sessioned time window t based on the following proposition.

Proposition 1: Given a user's IMAP log entries over three consecutive time-slots from $t-1$ to $t+1$, a user is regarded as performing a network transition at multi-sessioned time-slot t if $S_t^t = S_{t-1}^{t+1}$.

For example, suppose that $S_{t-1}^{t-1} = \{A\}$, $S_t^t = \{A, B\}$, and $S_{t+1}^{t+1} = \{B\}$. Then we derive $S_{t-1}^{t+1} = \{A, B\}$, and thus $S_t^t = S_{t-1}^{t+1}$, implying a network transition during the time window. On the other hand, suppose that $S_{t-1}^{t-1} = \{A\}$, $S_t^t = \{A, B\}$, and $S_{t+1}^{t+1} = \{A\}$. In this case, $S_{t-1}^{t+1} = \{A, B, A\}$, and thus $S_t^t \neq S_{t-1}^{t+1}$, indicating the user does not perform a network transition at t; instead we interpret this as there being one session associated with network A from $t-1$ to $t+1$, contemporaneously existing with another session associated with network B during time window t.

Using Proposition 1, we observed that users performed network transitions in 12% of multi-sessioned time windows in OIT-IMAP and CS-IMAP, suggesting that a user is more likely to be using multiple networks contemporaneously during a multi-sessioned time window rather than being the process of transitioning between networks in a single time window.

IV. Empirical investigation of the Markov model

In this section, we develop a parsimonious discrete-time Markov chain model of individual user transitioning among networks. This model can be used to design, analyze and provision protocols and services that support mobility (e.g., Mobile-IP home and foreign agents, or next generation services such as MobilityFirst's GNS [16]). A model of individual user behavior is particularly valuable, as it can be easily used to scale up evaluation workloads. After presenting our model, we validate how well performance measures determined via the aggregation of individual user-level models (in particular, signaling overhead due to user-transitioning between networks) match those determined from the traces.

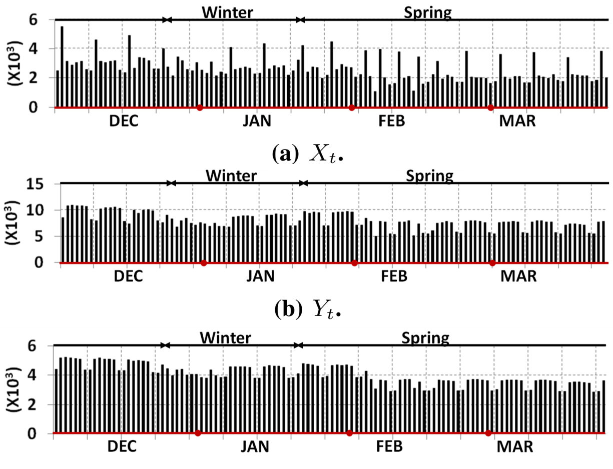
A. Markov Chain Model of User-Centric Network Transitioning

We develop a parsimonious discrete-time Markov chain model of individual user network-transitioning. Our unit of discrete time is the time window discussed in Section II. The Markov chain states encode enough state information to compute the cost of a user's signaling at each time-step. Let X_t be the number of *new* networks to which a user is attached at time t, with respect to time $t-1$. The first dimension of the Markov chain tracks the value of X_t, which will be used to quantitatively compute signaling overhead induced as a user transitions among networks, as we will discuss below. Let Y_t be the number of networks to which a user is attached at time t. The second dimension of the Markov chain tracks the value of Y_t, which will be used to quantitatively compute signaling overhead induced when a user detaches from a network, as we will discuss below.

X_t and Y_t may take value $\{0, 1, *\}$, where $*$ denotes two or more networks contemporaneously connected at t; for simplicity, we do not distinguish the case of more than two contemporaneous sessions from the case of exactly two such sessions, since approximately 99% of multi-sessioned time windows consist of only two network domains in our traces, as discussed in Section III. Our model can be easily extended to cover the more general case. Our Markov model thus consists of six states, $\{(0,0), (0,1), (1,1), (0,*), (1,*), (*,*)\}$. The model has 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.

The overall signaling cost from the user to a network-wide mobility management service (e.g., a Mobile-IP home agent, or the MobilityFirst GNS) on a state transition at $t-1$ to t, is computed as follows. Let $\mathcal{A}$ be the signaling cost generated when a user joins a new network, and let $\mathcal{D}$ be the signaling cost generated when a user departs from a network. (For simplicity, we will not consider signaling costs in the reverse direction from the management service to the user, although these can be easily included in the model.) In the case that network detachment is explicitly signaled, CO_t is computed by $CO_t = \mathcal{A} \cdot X_t + \mathcal{D} \cdot (Y_{t-1} - (Y_t - X_t))$. In the case that network detachment is implicitly signaled by attachment to a new network, CO_t is computed by $CO_t = \mathcal{A} \cdot X_t$.

B. Trace properties



(a) X_t**.**

(b) Y_t**.**

(c) user population.

Fig. 12. OIT-IMAP. Time series plots of "daily" aggregate cost of X, Y and daily user population over all users (using IP prefix distinction).

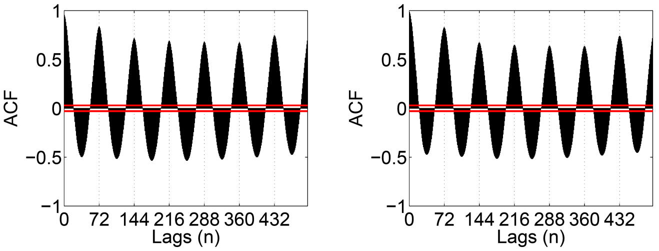
We investigate the properties of our CS-IMAP and OIT-IMAP traces. We first extract subtraces from the CS-IMAP and the OIT-IMAP traces and bisect each subtrace into the *training phase (also called phase 1)* and the *validation phase (also called phase 2),* which will be used in model parameter estimation and model validation, respectively.

CS-Fall subtrace. The CS-Fall subtrace has 79 users during the Fall 2013 semester (using IP prefix distinction) and its *phase 1* and *phase 2* consist of data from Sep 3rd to Oct 25th and from Oct 26th to Dec 16th, respectively.

OIT-Spring subtrace. Fig. 12 shows the time series plots of daily aggregate values of X_t, Y_t, and daily population of users producing IMAP logs over 7,137 users in OIT-IMAP (using IP prefix distinction). Unlike the CS-Fall subtrace whose daily aggregate values of X_t and Y_t are almost stable over a semester from its time series plot. Fig. 12a, b show a downward drift, particularly during the first half of the trace, likely resulting from the change in user population previously observed in Fig. 12c. Since our goal is to model the system in steady state, we thus only consider the subtrace during Feb and Mar for modeling, with the *phase 1* and *phase 2* consisting of data from Feb and Mar, respectively. This subtrace has 5,793 users generating IMAP logs.

For each subtrace, we derive one set of aggregate values of X_t over all users, and another set of aggregate values of Y_t over all users (using IP prefix distinction), sampled at 15 minutes (for CS-Fall) or at 20 minutes (for OIT-Spring).

Patterns of ACFs. The sample autocorrelation function (ACF) measures the degree of correlation between data at varying time lags (denoted by n), detects any trends and periodicity in a data series, and is also used to check the randomness of data. If random, the autocorrelation should be near zero for any and all time-lag separations. Fig. 13 plots the ACFs of values of X_t, Y_t for the OIT-Spring subtrace. Fig. 13a, b demonstrate that X_t and Y_t in the OIT-Spring subtrace have daily ($n = 72$) and weekly ($n = 504$) periodicity, and drop to near zero correlation at lag 20 so that X_t and Y_t are considered independent at around every seven hours (20·20 minutes).



(a) X_t.

(b) Y_t.

Fig. 13. OIT-Spring. Autocorrelation function for X_t and Y_t.

Similar periodicity and seven-hour independence results were also encountered in CS-Fall trace, but with lower amplitudes.

Testing for Stationarity. We check the subtraces themselves for stationarity using the *KPSS test*[12]. The KPSS assesses the null hypothesis that data is stationary over a range of time lags. The tests at the 1% significance level suggest that X_t and Y_t data in OIT-Spring are stationary for $n > 0$, but X_t data in CS-Fall is stationary for $n > 1$ and Y_t data in CS-Fall is stationary for $n > 4$.

C. Model estimation and validation procedure

We use the observed relative transition rates during the *training phase* to estimate the transition probabilities of our Markov chain model. To determine how well our Markov chain model predicts user behavior we will compare signaling costs determined by the model with those found in data from the *validation phase*. We proceed as follows:

1) Transition probabilities for the Markov Chain Model. Using the *training phase* data, we derive the transition probabilities for our Markov Chain model of a canonical user by counting the number of times that U users move from state i to state j per time-step and then normalize these counts so that the sum of the transition counts out of each state equals 1. This gives us our empirical transition probability matrix, $\hat{P} = [\hat{P}_{ij}]$.

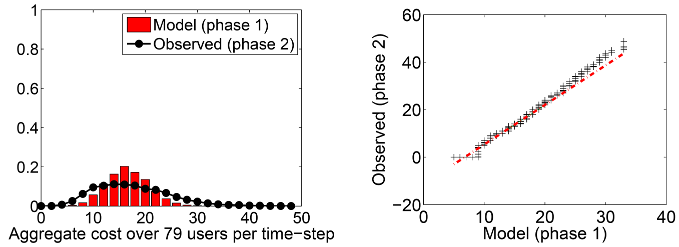
2) Generating a sequence of synthetic transitions between states for a population of U users. For each of the U users, we start from state (0, 0) and generate a next state using the transition probabilities $\hat{P}$. We repeat this process for ϕ time-steps (5,000) and then generate a sequence of length ϕ of state transitions made by the U users.

3) Determining the signaling cost for U users. For each time-step, we compute the aggregate signaling cost of the U users, using CO_t as in the previous subsection; for simplicity, we assume that users explicitly signal network detachment, with $A = D = 1$. Then we compute the distribution of signaling cost for the U users.

4) Model validation. Once the baseline distribution is built, we test how well our model predicts the number of signaling messages generated per time-step for the U users. To validate our model, we compare the model-predicted values (whose state transition probabilities were derived from *training phase* data) with the empirical distribution found in *validation phase*.

D. Prediction with aggregate user population

CS-Fall. Fig. 14a plots the pdf of the model-predicted and the observed aggregate cost over all users for the CS-Fall data set. Fig. 14b shows the Q-Q plot of the model cost data on x-axis versus the observed cost data on y-axis; a data point (x,y) on the Q-Q plot corresponds to one of the quantiles of the distribution plotted on the y-axis against the same quantile of the distribution on the x-axis; the plot has a red reference line through the origin with slope 1; points denoted as + should lie roughly on this line if the x-axis and y-axis data come from the same distribution. Fig. 14 confirms that the model cost and the observed cost datasets come from a Gaussian distribution[4] and the model fits the observed data well based on the linearity evidenced in Fig. 14b, while passing the chi-square goodness-of-fit test with 5% significance level.



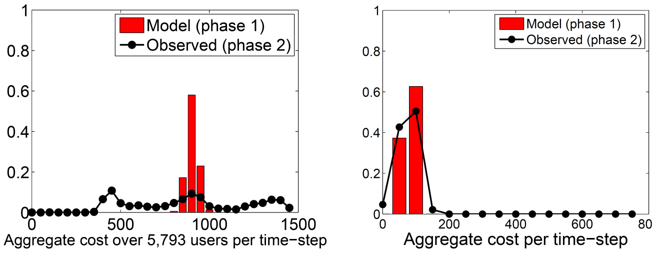
(a) pdf with bin size = 2

(b) Model vs. Observed

Fig. 14. CS-Fall. Aggregate cost over all users.

Recall that our model for U users aggregates the results from U independent user-level models. Since the ACFs of empirical values of X_t, Y_t show both positive and negative correlation at different time lags in Fig. 13, it is not surprising that signaling costs match the least well at the lower and upper extremes of the distributions in Fig. 14a. If the tail distribution is of interest (e.g., for provisioning system resources at the 95% workload maximum), interesting future work would be to develop a model that more accurately matches this tail behavior.

E. Prediction with user clusters



(a) all users

(b) heavy user cluster

Fig. 15. pdfs of aggregate cost over all users vs. over a heavy user cluster

OIT-Spring. Fig. 15a, b plot the pdfs of the model-predicted and the observed aggregate cost over all users and over a heavy user cluster for the OIT-Spring subtrace, respectively. Fig. 15a shows that the Gaussian distribution of cost predicted by the aggregation of individual user models does not fit the observed multi-modal data, which shows three distinct peaks. Visually, Fig. 15a suggests that costs might better be modeled as a *mixture* of Gaussian distributions.

[4] also validated using the Q-Q plot of the model vs. the randomly generated, independent standard normal data

Then what might each component of the mixture correspond to, and how many distributions should be mixed? To answer this question, we performed a clustering analysis, since a user's affiliation is not known in our OIT-IMAP traces. We partitioned the 5,793 users in OIT-Spring subtrace into K clusters based on their signaling cost, using Expectation Maximization (EM) clustering. We estimate the number of clusters that best represents the subtrace via WEKA EM clustering's 10-fold cross-validation [17]. With regard to an average daily cost of a user, the curve of the log likelihood of the cross-validation data as a function of the number of clusters suggests four clusters.

Fig. 15b only plots the pdf of the model-predicted and the observed aggregate cost over the heavy user cluster consisting of 721 users (12% of user population) with the highest cost (a mean cost of 13.62)[5]. Fig. 15b shows that the cost distribution for the resultant four-cluster model is closer to its empirically observed distributions when compared with the single cluster (Fig. 15a) case, although the clustered models do not pass the chi-square goodness-of-fit test. On the other hand, our handpicked heavy user cluster consisting of 100 users having the highest cost (a mean cost of 41) shows a good fit while passing the chi-square goodness-of-fit test with the 5% significance level. These results suggest that proper clustering can improve model performance in predicting signaling costs, a topic we plan to pursue in future research. In the course of our research, we also compared model-based and empirically-observed state occupancies of OIT-Spring, showing good agreement for both the aggregate population of users and for clustered users.

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 [2, 9, 11], Bluetooth networks [2], and cellular networks [7, 10, 14]. 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 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). On the other hand, [3, 7, 14] 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. Those conclusions, however, are based on physical mobility within a single network.

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 two sets of IMAP server logs of populations of approximately 80 users and more than 7,000 users, 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.

Acknowledgments. This research was funded in part by grants from the US NSF awards CNS-1040781 and CNS-1345300.

References

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

[2] A. Chaintreau and et al. 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, Mar. 2003.

[5] R. Gerhards. The Syslog Protocol. RFC 5424, March 2009.

[6] L. H. Goga, Oana and, S. H. K. Parthasarathi, G. Friedland, R. Sommer, and R. Teixeira. Exploiting innocuous activity for correlating users across sites. In *WWW '13*.

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

[8] D. Han and et al. XIA: Efficient support for evolvable internetworking. In *Proc. 9th USENIX NSDI*, 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] M. Kim, D. Kotz, and S. Kim. Extracting a mobility model from real user traces. In *IEEE NFOCOM 2006*.

[12] D. Kwiatkowski and et al. Testing the null hypothesis of stationarity against the alternative of a unit root : How sure are we that economic time series have a unit root? *Journal of Econometrics*, 54(1-3):159–178, 00 1992.

[13] J. Padhye and J. F. Kurose. Continuous-media courseware server: A study of client interactions. *IEEE Internet Computing*, 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] P. I. R. Project. Pew internet research project.

[16] A. Venkataramani, J. Kurose, D. Raychaudhuri, K. Nagaraja, M. Mao, and S. Banerjee. Mobilityfirst: A mobility-centric and trustworthy internet architecture. *ACM CCR*, 2014.

[17] WEKA. Weka em clustering.

[18] S. Yang, J. Kurose, S. Heimlicher, and A. Venkataramani. Measurement and modeling of user transitioning among networks. *Tech. Report UM-CS-2014-023, https://web.cs.umass.edu/publication/details.php?id=2380.*

[5] The detailed results on EM clustering are found in [18]