Internet User Access via Dial-up Networks - Traffic Characterization and Statistics

Ron Hutchins Ellen W. Zegura Andrew Liashenko Philip H. Enslow, Jr.

College of Computing
Georgia Institute of Technology
Atlanta, GA 30332-0280
{ron,ewz,liashenk,enslow}@cc.gatech.edu

Abstract— **Understanding network traffic from operational networks is critical to the design and evaluation of network protocols. We present analysis of a data set comprised of eight months of RADIUS authentication data taken from a large national dial-up Internet Service Provider (ISP). We present basic statistics, including session counts based on time-of-day, session length distribution, session inter-arrival times, and growth in the customer base over the measurement period. We investigate more deeply several properties of the data. We use area code information to correlate account location with basic statistics. For example, we find that west coast accounts tend to have earlier-than-average mean session start time. We find that 40% of sampled accounts exhibit concurrent sessions (two or more sessions active at the same time), while 20% show multiple originating phone numbers. Both phenomenon are likely to increase as users become more mobile and sophisticated. Finally, we offer evidence of significant session activity due to hypothesized automated processes, characterized by periodic interarrival times and/or constant session durations. Our results provide important data for the simulation and modeling of access network protocols and applications. They may also form the basis for creating a workload model of access networks.**

I. Introduction

In recent years, considerable research effort has been devoted to collection and analysis of traces taken from operational networks. Such studies provide valuable information about the behavior of users, both individually and in aggregate, and allow trace-driven evaluation of proposed protocols and algorithms. Among past studies, relatively few have focused on collecting and analyzing behavior for the *access networks* that serve as the entry point for many end-users into the Internet[1]. Notable examples include studies of a building-wide wireless network [1], a campus dial-in modem bank [2], an ISP's cable modem population [3], and the AT&T WorldNet modem bank [4]. These past studies largely focus on application-level behavior of access network users (e.g., web access behavior and performance).

In this paper, we present results from a data set comprised of eight months of RADIUS authentication data [5] taken from a large national dial-up Internet Service Provider (ISP). Over the eight months of data collection, the number of users seen by the system grew from 630,135 on May 4, 2000 to 1,776,822 by Dec 19, 2000, and included users from each of the continental United States. We believe this is the largest published study of access network usage characteristics.

The RADIUS records used in this study give us session start times and lengths, on a per account basis. For the majority of sessions, we have access to the *dialed* (access point) phone number for each session and the *dialing* (origin) phone number. We do not have access to application-level information (e.g., TCP port usage) nor to packet-level information, thus our results focus on session-level characteristics [2].

We present basic statistics, including session counts based on time-of-day, session length distribution and session inter-arrival times. We characterize growth in the customer base over the measurement period, observing a linear increase in the number of new accounts per day (approximately 2200 additional per day) and a four-fold slower linear increase in the number of active accounts per day (approximately 500 additional per day). We investigate more deeply several properties of the data. We use area code information to correlate account location with basic statistics. For example, we find that west coast accounts tend to have earlier-than-average mean session start time. We find that 40% of sampled accounts exhibit concurrent session (two or more sessions active at the same time), while 20% show multiple originating phone numbers. Both phenomenon are likely to increase as users become more mobile and sophisticated. Finally, we offer evidence of significant session activity due to hypothesized automated processes, characterized by periodic interarrival times and/or constant session durations.

The next section provides more detail on the RADIUS system and the dataset we collected. Section III presents

[1]In the category of "access networks" we do not include campus or corporate-wide, richly-connected networks, such as Ethernet, and instead refer to more widely-available, lower-bandwidth access technologies such as dial-in and cable modems.

[2]It may be possible to combine our session-level results with prior studies of application-level behavior, especially for users connected via lower bandwidth access technologies, to get a "complete" model of access network user behavior. However, we have not attempted that synthesis of results in this paper.

1092-1658/01 $17.00 © 2001 IEEE

the first part of our analysis. Section IV contains a deeper analysis of customers from different area codes, the use of multiple dialing phone numbers and concurrent use of a single account, and apparent computer generated automatic processes. We describe related work in Section V and conclude with a discussion of future directions in Section VI.

II. Background

In this section, we describe the basic information available via the RADIUS access logs, as well as the initial processing we applied. That processing removes records not immediately applicable to our study, such as login and error records, and anonymizes the account identification information and individual user's phone number.

The data consists of eight months of authentication data from PPP [6] sessions, taken from a national scale ISP, during the months of May through December 2000. These months span the Summer and early Winter period and therefore exhibit seasonal patterns of usage as well as growth in the user population.

The data set was collected from the distributed RADIUS authentication servers across the country used for collecting billing information for the ISP. This data set includes authentication records from dial-in, DSL, and a very few dedicated (low bit rate) service users. These different services are not separable here due to unavailability of defining information. However, all services are controlled by PPPoE [7] for authentication and the idle timeout mechanism controls session lengths for all services. The session based model of dial-up services still holds, so the inclusion of these technologies has no adverse effect on the dial-up access analysis.

Some RADIUS data from employee accounts is included in the captured data. These accounts are not subject to the automated timeouts that control session length of customer accounts and so exhibit session length characteristics very different from normal customer sessions. These sessions are generally of long duration and of very few number. They are included in the analysis for completeness, but extreme values are truncated in many plots for clarity of presentation.

The users were generically aware that the ISP has the right to perform monitoring, however they were unaware of this specific study. The data thus represents typical user activity, with no change in behavior due to non-standard monitoring.

The original data consisted of LOGIN records, START records, STOP records and ERROR records, totaling about 1.25 Gb/day. The data was filtered to include only STOP records, since all information of interest to us was contained in these records. STOP records include the session end time, the session length, and the originating and called phone numbers for the session. ERROR records contain information about problems in the authentication process, which is not a focus of our study. No specific information was available on call blocking in the dial-up system.

Some data was missed over the eight month collection period, typically due to problems transferring from the RADIUS servers. Specifically, we are missing data on May 1-3, 20-28; June 9-12, 19-21, 29-30; August 4,30; September 20-26; October 6, 10-17; November 30; and December 20-31, 2000. Of the records collected, some contained anomolies. Specifically, we observed the following:[3]

- STOP records showing a zero length session. (3319 of 2052594 records, or .16%, from May 4 data)
- Extremely long reported session lengths (1 of 2052594 records longer than 30 days from May 4).

Zero length sessions were included in the analysis for completeness. Long sessions were truncated in the graphs but generally included for completeness. Some RADIUS files were missing some sections of the day, especially late night data, but if the data files transferred properly they were processed.

Some STOP records did not contain originating or dialed telephone numbers due to different technologies used at the access points (T1 lines do not pass calling line id, PRI lines do). We note, for example, that 363463 of 2052594, or 17%, of the RADIUS records from the May 4 data have no originating phone number. These records were included in the analysis except for the area code study.

A condition of our use of this data was the removal of any ability to associate usage characteristics with a particular real user. To provide this anonymity, each individual account was assigned a sequential unique identification number (UID for this analysis) the first time a RADIUS record for the account appeared. Subsequent activity on that account was recorded using the assigned UID, thus preserving per-account statistics. The originating (dialing) phone number was anonymized similarly, with the exception that the area code was preserved. Because there are many users in each area code, preservation of this information does not identify individual users, yet allows us to examine coarse geographic density of the user population. We are, of course, unable to identify more fine-grained (neighborhood) geographic information from the anonymized dataset.

III. Data Analysis: Part One

This section presents the first results of our analysis, including aggregate traffic statistics as well as individual customer characteristics. We present time-of-day data for typical weekday and weekend traffic, and outliers such as the Thanksgiving holiday. We show characteristics of session lengths and counts for all traffic and for individual customers.

A. Overall Characteristics of User Activity

We begin by examining general characteristics of the overall data set, including number of sessions based on time-of-day and session length distribution. When time-of-day results are reported, all times are the **local** time of the dialed (access) point. Because the customers are distributed throughout the three time zones of the United

[3]The percentages offered here are taken from a single day's RADIUS data as examples.

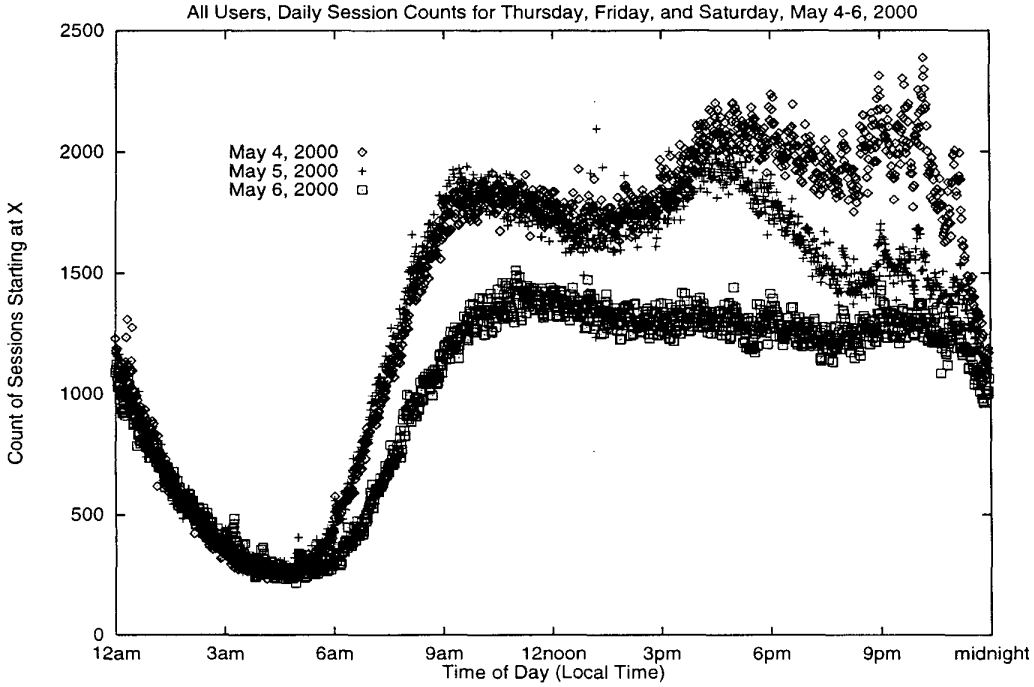

Fig. 1. Count of Sessions Starting During the Day, May 4-6, 2000

States, this essentially merges results that occur at the same relative, but not absolute time.

Count of Sessions Across a 24 hour Period. Figure 1 shows the number of sessions that **begin** in each 1 minute interval throughout the three 24 hour periods Thursday, Friday, and Saturday May 4-6, 2000.

The weekday activity follows an expected diurnal cycle, with much less activity between 3am and 6am, and a quick increase in activity beginning at 6am. There are several peaks that appear on Thursday and Friday, including one around 4pm and another closer to 9pm. The 4pm peak seems reasonable for work customers who may dial in shortly before leaving work, and the 9pm for home users who dial in shortly before bed.

Thursday appears in this analysis as a "typical" weekday. Activity on Friday closely matches that of Thursday until the afternoon where an expected drop in activity happens. The weekend (Saturday) activity follows the same diurnal cycle, but without the evening and late night peaks present in the weekday data. The overall amount of customer activity is also lower on the weekends, with a peak at midday of about 1500 sessions, as compared to a peak of 2300 between 9pm and 11pm during the week. The count of sessions in the early morning hours is relatively constant across all days.

We may ask whether data access networks show a characteristic decline or increase during holiday periods. Observations of several days around the Thanksgiving holiday, November 23, 2000, show an expected slowdown of sessions starting during the afternoon on Wednesday. The afternoon on Thursday is noticeably lower than even a weekend, returning to normal weekend activity later in the day. The Friday after Thanksgiving models weekend activity and the Monday following shows a robust rebound to expected session activity, with a much earlier start to the rise in sessions (around 5am) compared to a normal work day.

Session length distribution. We now look at session length distributions. Figure 2 shows data for each of four Thursdays across four months of the collection. Session lengths, quantized into one minute buckets, are shown on the x-axis, with the counts of sessions of length x shown on the y-axis. Spikes are clearly visible at each 30 minute increment, at the 4 hour mark and at 9 hours. Due to the very sharp peaks at the 30 minute and 4 hour times we hypothesize that these spikes are due to preset timers within client application software such as network dial-up software as provided by the ISP, or other software applications. The 9 hour spike shows a quick rise but a more gradual decrease in counts over about the next hour. We hypothesize this peak is explained by an 8 hour work day with an additional hour lunch period for customers who utilize their ISP account in their work. The sharp 12 hour spike on all 4 days identifies the ISP's enforced timeout, verified by discussions with engineers from the ISP. The sessions that last beyond 12 hours are those not subjected to the enforced timeout. We also see growth in population, with higher counts for the later collection days (eg. August 3).

Sessions longer than 24 hours have been truncated from the graphs for presentation. For example, approximately 10 sessions were seen per day with duration of 24 hours, about 4 per day of 36 hours length, and about 2 of 48 hours length. A very small number of other sessions appear, ex-

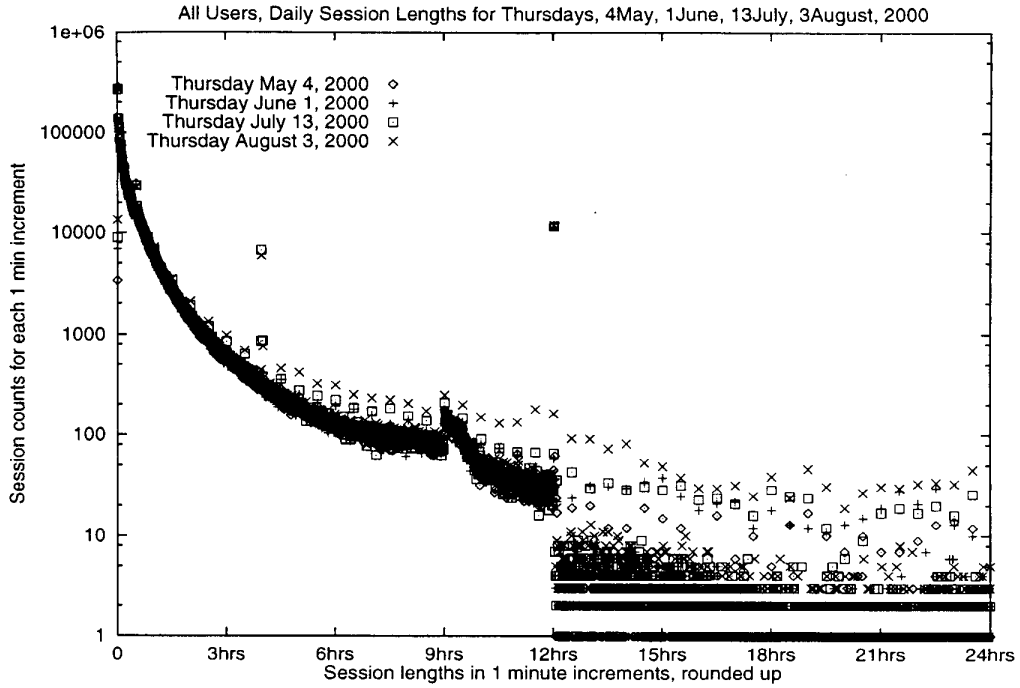

Fig. 2. Session Lengths for Typical Thursdays: May 4, June 1, July 13, August 3, 00

hibiting long extremes [4].

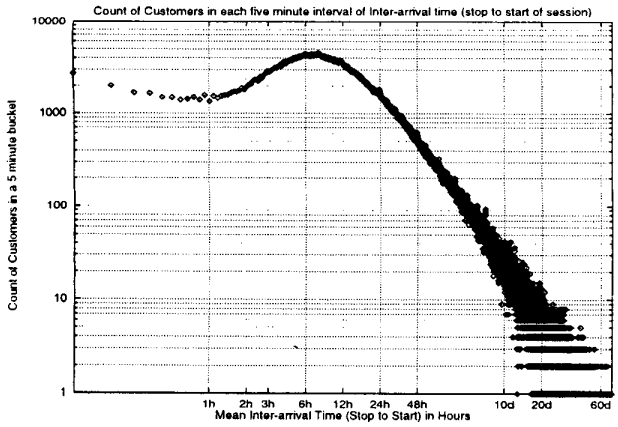

Fig. 3. Interarrivals for All Customers over the Collection Period

Session Interarrival times. Figure 3 shows the distribution of the interarrival (stop to start) times for all sessions over the entire period of May to December. A problem exists when calculating interarrival times for an account that exhibits concurrent sessions (discussed in Section IV-B). For the analysis, when a session record showed a start time before the end time of the previous session, the new record was discarded. Sessions with length of zero were discarded, as were sessions of length longer than the ISP's 12 hour enforced timeout. A bucket size of 5 minutes was selected to group the data for presentation.

[4]Employees of the ISP were exempt from the enforced 12 hour session timeout. This accounts for the longer sessions.

The figure shows the count of customers whose mean interarrival time falls within each 5 minute bucket. A very short interarrival time (less than 5 minutes) is expected for those customers who continually reconnect when timed-out. The peak is at the 6 to 8 hour mean interarrival time. We do not have a definitive explanation for the peak, however we note that it does correspond to approximately the time between the morning (9am) and afternoon (3-4pm) peaks in session starts and the afternoon (3-4pm) and evening (9-10pm) peaks in session starts observed for the normal workdays that make up the majority of the collection period. The very short interarrival means could be the direct cause of the ISP's enforced 12 hour timeout and an immediate reconnect from some users. We have verified with the data that this behavior exists in some customers. The values between 10 and 60 days represent those very infrequent customers who have long times between connecting. There are far fewer of these, but they are represented in this data and in the growth plots to follow.

Overall Session Counts. We now move from session length data to an examination of counts of sessions. Observation of session count data shows the the high prominance of customers with low session counts. We observe a knee in the curve at around 200 sessions, indicating those customers with around one session per day average. The data begins spreading at around 1000 sessions, indicating another change in the nature of the customers. We hypothesize that the session length characteristics may be dominated by the group of users with high counts of very short sessions.

Growth of Customers Over the Sample Time period. During the time of this data collection, the count

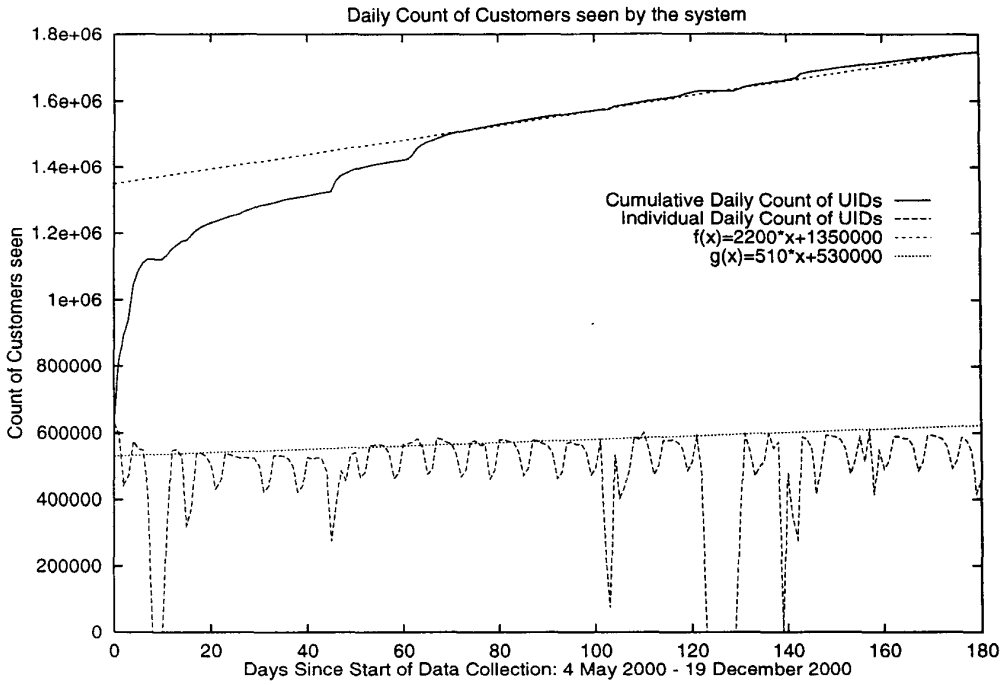

Fig. 4. Growth in Customers Seen by the System

of customers seen grew from 631,135 on May 4, 2000 to 1,776,822 by December 19, 2000. The access network exhibits two areas of growth: the total number of customers seen to date, and the number of unique customers seen on a given day. Figure 4 shows the two growth patterns. For both curves, the x-axis marks days of the sample period with May 4, 2000 as Day 0, and the y-axis shows the count of customers. The data points of both customer count curves consist of a single point at each day. The points on the graph have been connected with lines to more clearly show the growth.

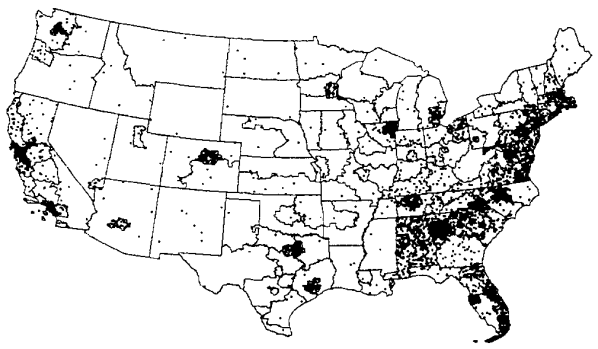
Fig. 5. Area codes represented in the data

The Cumulative Daily Count curve shows a significant "start-up" time for our system to record each current customer's appearance. This start-up period is related to data in the interarrival plots (Figure 3) which shows some customers with up to 80 days between sessions. We assume that by some point (after the 80 days) nearly all current customers have been seen and growth after that time represents new customers. The Individual Daily Count curve is a day by day count of unique customers seen by the system during that 24 hour period. It exhibits fluctuations due to day-of-the-week effects as well as occasional data collection errors.

The linear equation f(x) is fit to the Cumulative Daily Count, showing a steady state growth of about 2200 new customers per day. g(x) is fit along the weekday peaks of the Individual Daily Count, showing a slope of about 510 per day growth in active users ignoring data drop-outs. It is interesting to note that for this ISP, the growth of new customers is linear and not exponential, as in the early days of Internet adoption. It is also interesting that, although we have no information on the rate of sales of new accounts by this ISP, using first appearance of a customer (past the start-up period) as time of new customer account, the growth in new customers is about four times the growth in active users. This is an important factor for provisioning services.

Summarizing, our customer population of over 1.7M customers exhibits an expected diurnal cycle in activity, but with significant differences in weekday and weekend behavior. Weekday activity contains peaks in mid and late evening, while weekend activity is more smooth. Some sessions last far longer than average. Session interarrival times (stop to start) vary widely with peaks in the five minute and seven hour times but with a tail extending out to 80 days between sessions for a small number of customers. We

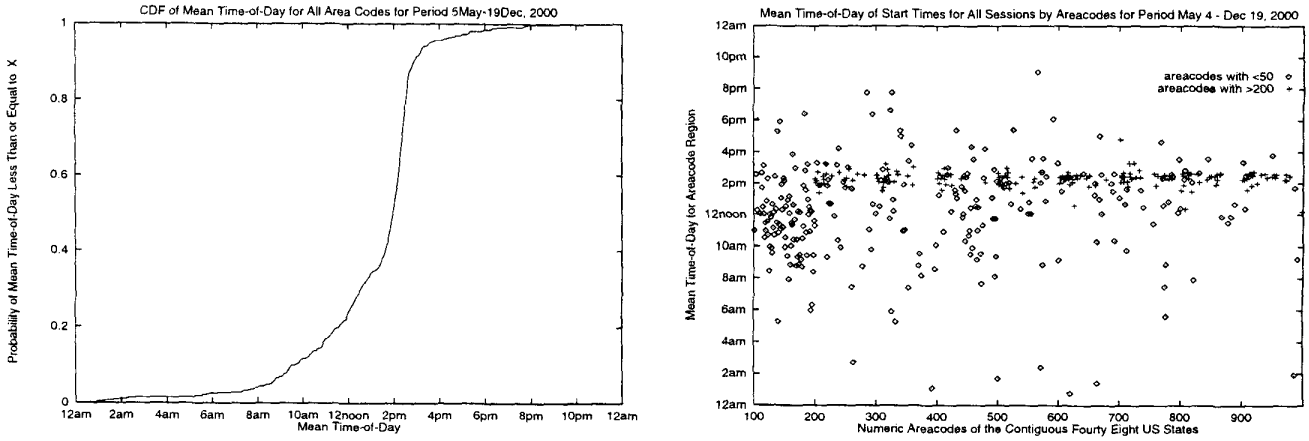

Fig. 6. Area Code Data: CDF and Distribution of Users for Mean Session Starting Time-of-Day

note that the cumulative growth of customers is about four times the growth rate of unique customers seen on a daily basis.

IV. Data Analysis: Part Two

In this section we examine details of the data. Specifically we present answers to questions such as:

- How are customers distributed across the ISP's serving area?
- How do customers from different area codes differ in their characteristic usage patterns? Are metropolitan areas different from non-metro areas?
- How much do computers initiate the sessions on the access network as opposed to humans?
- Are access network customer accounts overloaded with multiple customers? Can we differentiate multiple customers on a single account?

A. User Location: Area Code Analysis.

We examine in detail the characteristics of the customer population associated with the geographical location of the originating phone number. We present the distribution of customers across each area code, the mean time-of-day of customers for each area code using a CDF, and show that the mean time-of-day of start of a session tends toward 2pm, except on the west coast where mean session start time is about an hour earlier.

During the initial data reduction and anonymization process, the area code of a customers' originating phone number was maintained and statistics were kept for count of customers calling from each area code in the continental United States. This data is shown in Figure 5, with each dot representing 250 accounts, randomly placed in the geographic region represented by that area code. Where multiple area codes serve an overlapping geographic region, the data points overlap.

The plot illustrates that the customers are distributed across the continental United States, but are concentrated in the east and west. Additional regions of heavy density appear around larger cities (e.g., Denver, Phoenix, Dallas).

A cumulative distribution graph and a plot of mean session start time by area code are presented in Figure 6. The CDF shows that the mean starting time of sessions strongly tends toward 2pm. We notice, however, that there is greater likelihood of an area code exhibiting mean starting time earlier than 2pm rather than later.

The right plot of Figure 6 shows mean time of day of use for selected area code, differentiating the area codes by number of customers. We can see that for area codes with very small customer counts (feweer than 50), the mean start time can be either later or earlier than the overall 2pm mean, though it is more likely to be earlier. For area codes with larger customer counts (200 or more), the mean tends to be closer to 2pm.

In Figure 7, we have pulled out particular area codes and analyzed their geographic location. We include for reference for each areacode: total session count, mean session length, mean time-of-day of use, and count of customers who are primarily seen in this area code. We see no significant difference in metropolitan areas as compared to non-metro areas. We do see differences from east coast to west coast. As seen in Figure 7, west coast locations such as Seattle, San Francisco, and San Diego exhibit an earlier mean. We hypothesize that this earlier than expected mean is due to the importance of Eastern Time in financial transactions and related businesses.

B. Concurrent Usage and Multiple Originating Numbers

Basic trends in mobile computing suggest that over time customers will tend to use an account from an increasing number of locations. We utilize phone number data to examine the use of multiple originating phone numbers from a single account. We also present data on *concurrent sessions* using one account and discuss implications of these two characteristics. For part of this analysis, where using the entire data set was not practical, we sampled 3213 accounts from the 1.7M total customer accounts.

Concurrent Use of a Single Account. Concurrent usage – two or more dial-up sessions active concurrently on one account – may be explained by multiple scenarios. For

Selected Area Codes from Metro and Non-Metro Areas, East and West Coast Focus

acode	sesn cnt	mean seslen (s)	mean tod (s)	usrcnt	city-state
401	526831	2088	51529	4101	RI (all)
212	2607607	2296	50264	19917	NYC, NY
646	51968	2735	51742	414	NYC, NY
516	880000	2361	51796	7675	Nassau County, NY
914	1466460	2312	52342	11447	White Plains, NY
631	1010432	2438	51672	8799	Suffolk County, NY
718	2677458	2671	52043	21521	Bronx, Brooklyn, NY
203	41040	2371	51041	1266	Stamford, CT
302	426866	3328	52230	3714	DE (all)
202	980394	2594	50535	7398	Washington, DC
919	10766541	2108	53142	82605	Raleigh, NC
336	682083	2460	52508	4745	Greensboro, NC
404	8982931	2110	50473	49236	Atlanta, GA
770	29587731	2233	51820	157606	Atlanta suburbs, GA
321	489169	2505	52358	2866	Orlando, Fl
352	509132	2964	52071	3893	Gainesville, Fl
305	2507899	2702	50736	16394	Miami, Fl
205	3916618	2376	53027	26362	Birmingham, AL
270	1354144	2305	52528	8241	Bowling Green, Ky
281	1717445	2889	52382	12004	Houston, TX
206	1373213	2335	49874	12741	Seattle, WA
509	14868	2739	49802	321	Spokane, WA
425	1147985	2883	49905	9892	Everett, Redmond, WA
360	74037	2781	49943	1589	Bellingham, Olympia, WA
253	350691	4415	49628	3321	Tacoma, WA
541	7650	1277	48169	245	Eugene, OR
503	33459	2101	50166	956	Portland, OR
415	1007834	2510	50223	11383	San Francisco, CA
661	5693	2203	50016	255	Bakersfield, CA
209	324913	2730	49976	2456	Modesto, CA
831	361675	14196	50718	3682	Monterrey, CA
858	23113	2232	50159	571	La Jolla, CA

Fig. 7. Selected Area Codes from Metro and Non-Metro Areas, East and West Coast Focus

example:

- A customer initiates a new session without disconnecting a previous session.
- Multiple customers share one account concurrently.
- A customer configures a computer to automatically place calls while using the same account concurrently, for example, accessing web pages while a computer separately downloads email.

We find that 1282 of the 3213 sampled accounts, or 40%, exhibit concurrent sessions during the collection period. For the 1282 accounts, the average of concurrent time is around 1000 seconds, not significant in this context. One account exhibited 33120 minutes worth of concurrent use over 69120 minutes online. This online time encompassed 1362 separate sessions over 148 calendar days. Two different originating phone numbers were noted for this account.

Multiple Originating Phone Numbers. We find that 978 of the 3213 sampled accounts (30%) show multiple originating phone numbers. The highest count seen is 55 unique numbers on a single account. Interestingly, this account shows only 1260 minutes of concurrent use out of 14400 minutes of online time. There were 1015 sessions over the 137 calendar days of monitoring. A count of 2 unique numbers is the most prevalent at 16% of the sample set, but 62 of the 3213 (2%) show 10 or more originating phone numbers. Of the 3213 sampled accounts, 351 (11%) show both concurrent usage and multiple originating phone numbers.

Intuitively, the concurrent use of a single account appears to require multiple originating phone numbers. However, business PBX telephone systems may send the same calling line ID for any call on the system. This may account for concurrent usage without multiple originating numbers.

C. Automated Processes

We also offer evidence of significant session activity due to what we hypothesize are automated processes. These sessions are characterized by regular, periodic interarrival times and/or session durations. For example, Figure 8 shows plots of session lengths from sequential RADIUS

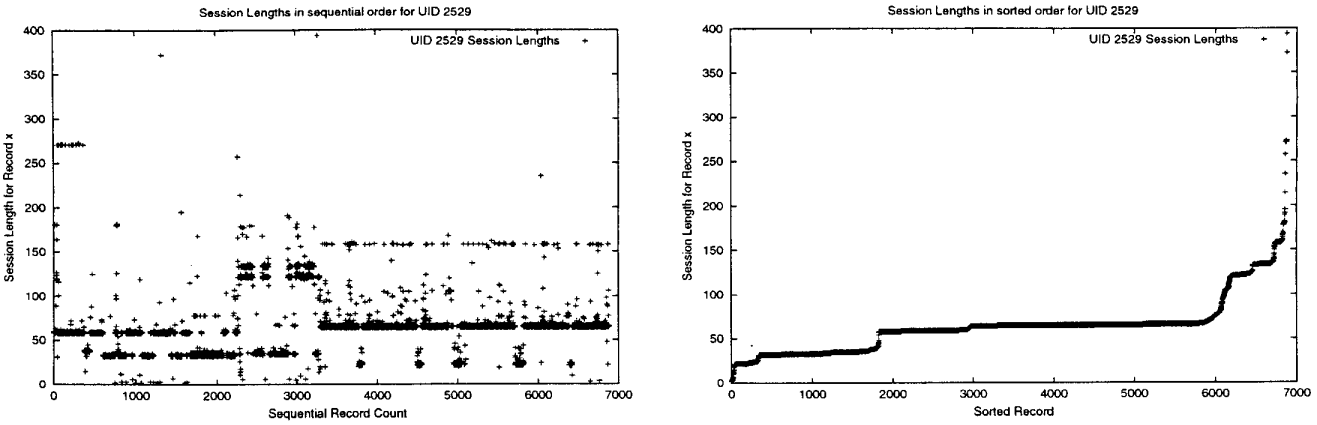

Fig. 8. Session Lengths for UID 2529, Sequential and Sorted

records (left figure) and sorted by length (right figure). It is clear from these graphs that there is regularity in this account unlikely to be explained by human activity.

We discuss scenarios of automatic startup and termination in relation to "human" customers and present statistics on interarrivals and session lengths. We hypothesize four types of sessions for this discussion:

- Sessions both initiated and terminated by humans, exhibiting random session lengths and interarrival times.
- Sessions initiated by humans but terminated by automated means, such as application idle timeouts or the enforced timeout by the ISP, exhibiting random interarrival times, but regular session lengths.
- Sessions that are initiated by a computer agent, but have variable processing time, such as a mail server, exhibiting regular interarrival times, but random session lengths.
- Sessions that are initiated and terminated by a computer agent, such as a POP mail query for new mail, exhibiting regular interarrival times and regular session lengths.

We aim to identify sessions of the last three types specifically.

Due to variability in technical systems, times for events are rarely exactly equal. In order to find correlations in session length or interarrival data, the appropriate data elements were processed to find elements that were "close" to the same value [5].

A second mechanism which utilized second derivative characteristics of the sorted data proved more useful. The data were sorted into non-decreasing order and the original iterative partitioning algorithm was applied to generate the subsets to process. The second deritive was calculated across this window of data elements to identify subsets showing a change from negative slope to positive slope for values within the cluster .

Experimental checking of the processing was necessary to ensure validity. As the algorithm identified accounts with clusters, the data were hand processed and visually checked. We found that clustering of less than about ten percent of the total data points was in general suspect. We attempted to use variance as a measure of accuracy but found that the count of data points proved more accurate on visual inspection of the data. We utilized the same process to examine interarrival times for clusters as well.

For the sampled data, of the 3213 accounts we find that 342 accounts (10.5%) show clusters around session length, 166 (5.2%) show clusters around interarrival time, and 80 (2.5%) show evidence of both session length and interarrival time correlations.

Clusters in session length may be related to automatic timeouts, both at the customer application and at the ISP (as the enforced timeout). Interarrival clusters appear more related to scripted actions taken by a computer. Conservative stopping criteria were used in processing data to identify the clusters. We believe that the numbers presented are at least minimally representative of the characteristics of the data. We observe that many of the accounts showing interarrival clusters contain a large session count. Accounts which show evidence of interarrival clusters may have a pronounced effect on the overall traffic on access networks while only accounting for a small percentage of total users.

V. Related Work

Most closely related to our work are studies of access network user populations. Gribble and Brewer conducted an early study in this area, gathering traces from the University of California at Berkeley's dial-in modem banks [2]. They collected HTTP traces for 45 days in the Fall of 1996, observing 8,000 unique clients. They collected information on a per-HTTP-request basis, discarding traffic not destined to port 80. Most comparable to our results are their measurements of requests per minute as a function of time of day. They observe a strong dependence, with a diurnal cycle that somewhat resembles ours. We see a more dramatic difference between weekday and weekend, and a

[5]The algorithm used is loosely based on the description of the CART product in white papers on the Salford Systems web page [8]. Other mechanisms, such as the "autoclass", "k-means", "hierarchical agglomerative clustering", and "expectation-maximization" algorithms have been suggested as well.

weekday late evening peak that is not noticeably present in their data. Both observations are consistent with our population, which includes a strong component of at-home users, while their population is at-work users.

In the area of commercial access network populations, Arlitt et al. collect data at an ISP accessed via a cable modem bank [3]. Their collection took place over 5 months in Spring of 1997, observing "several thousand" subscribers. As in the Gribble data, the collected information is recorded on a per-HTTP-request basis. The Arlitt data also exhibits a strong dependence on time of day, with a general shape that closely matches the Gribble data. Most of the rest of the data analysis focuses on application-level characteristics, rather than session information.

Feldmann et al. also collect a trace in a commercial access network, specifically the AT&T WorldNet modem bank [4]. Data was collected over 12 days in mid-August 1997, observing nearly 80,000 unique users with over 150,000 dial-up sessions. They used the trace to drive a simulation of a web proxy, in order to assess the performance of web proxy caching. The session-level characteristics of the trace are not analyzed in the referenced paper, and therefore there is no basis for direct comparison to our results.

Tang and Baker analyze a 12-week trace of the local-area wireless network located in the Gates Computer Science building at Stanford University [1]. Their user community comprises 74 users; at least half are graduate students, and the remainder are faculty, staff and three robots. They were able to capture packet-level activity (which allows identification of the application via port number), as well as access point information. Their results on overall user behavior are most directly comparable to ours. They observe a different trend in number of active users versus time-of-day, with a much more pronounced mid-afternoon peak on weekdays. This is expected since this network is located in a workplace, while dial-in ISP use includes a large number of at-home users. In looking at how often users are active, their results generally agree with ours in that more users are active on fewer days, while a few users are active on many days, though the limited size of their user population makes the trend less clear.

VI. Conclusions and Future Work

This analysis presents data from a large set of users over a significant time period. The results support basic expectations about time-of-day of network use and session length distribution.

- Session counts are light in early morning hours, showing peaks during specific times in the afternoon and early evening relating to "uptime" of users during the waking hours. This models the diurnal period showing that user activity is closely related to time of day during that period. During the noon to midnight hours, session counts do not follow the diurnal pattern closely but still exhibit some time related behavior.
- Session lengths show a high predominance of short sessions with an extended tail in longer sessions. Distributions of session lengths do not show good fit to an exponential distribution.
- A significant percentage of sampled accounts show concurrent usage, multiple originating phone numbers, or both. This indicates use of the accounts by more than one person, or by an automated process and a person. Multiple originating numbers could be due to customer mobility, but we cannot validate this possibility.
- Significant counts of highly correlated patterns of consistent session interarrival times and session lengths show possible automated processes. These processes do not exhibit the same characteristics of use as humans and again can impact planning for network capacity and upgrade schedules.

In future research, we plan to investigate call blocking in the system. We will analyze data from a wireless network and compare the results to the dial-up network statistics. The demand for mobile networks is growing rapidly and data on current characteristics across the national landscape are needed for designing future protocols and infrastructure.

Acknowledgements

The data and technical expertise were provided by Mindspring/Earthlink, Inc. with invaluable engineering support from Toby Reed and Allen Thomas. Invaluable statistical expertise for automated process discovery was provided by Dr. Russell Heikes and Dr. T. Govindaraj of the School of Industrial and Systems Engineering at Georgia Tech. Dr. Steve French and Claudia Martin of the Georgia Tech GIS Center provided the time and tools to produce the area code map. Many thanks also go to Carl Rigney of Livingston Corporation for help in understanding RADIUS accounting data in this context.

References

[1] Diane Tang and Mary Baker, "Analysis of a local-area wireless network," in *Proceedings of MOBICOM 2000.* August 2000, pp. 1–10, ACM Press.

[2] Steven D. Gribble and Eric A. Brewer, "System design issues for internet middleware services: Deductions from a large client trace," in *USENIX:Proceedings of the Smposium on Internet Technologies and Systems 99*, December 1997, pp. 207–218.

[3] Martin Arlitt, Rich Friedrich, and Tai Jin, "Workload characterization of a web proxy in a cable modem environment," *Performance Evaluation Review 99*, pp. 1–12, August 1999.

[4] Anja Feldmann, Ramon Caceres, Fred Douglis, Gideon Glass, and Michael Rabinovich, "Performance of web proxy caching in heterogeneous bandwidth environments," in *Proceedings of Infocom 99*, 1999.

[5] C. Rigney, A. Rubens, W. Simpson, and S. Willens, "Remote authentication dial in user service (RADIUS)," Internet Request for Comments 2138, April 1997.

[6] W. Simpson, "The point-to-point protocol (PPP)," Internet Request for Comments 1661, July 1994.

[7] Mamakos et. al., "Transmitting ppp over ethernet," Internet Request for Comments 2516, February 1999.

[8] "Salford systems, inc., cart," http://www.salford-systems.com.