Activity Based Session Generation for Personal Communication

June 10, 2005

Contents

1 Introduction

1.1 Motivation

1.2 Previous Work

1.3 Contributions

- activity type has a large influence on communication intensity
- unified tempral behavior model (inter-arrival time) for voice and data communication

2 Model Overview

We focus on modeling the communication behavior of a *single* user (person), as opposed to model combined traffic as it appears on communication network elements (switched, routers, etc.) To do that, we need to find answers to the following questions:

- when communication occurs,
- who communicates with whom,
- and what is being communicated.

We measure the time when communication occurs as time between two communication sessions (inter-session time), finding who communicates with whom is selecting source and destination of a session, and we refer to what is being communicated as the content of the session.

Each of the three components is obtained using a stochastic process. Parameters to such processes in general depend on *activity type* (such as work, staying home etc.) of a person whose session is being generated. General description of the processes, including their dependence on activity types, is given bellow and details for each session type are discussed in section 3.

2.1 Inter-session Time

The inter-session time can be measured either as time between end of one session and beginning of the next (I_i), or as time between two consecutive session starts (A_i). Which of the two definitions of an inter-session time is used depends on the type of sessions. Sessions with considerable length and that cannot be made in parallel (e.g. phone conversations) are modeled using I_i, while instantaneous session and sessions that can be performed in parallel are modeled using A_i (e.g. emails or Web browsing).

Let s_i, e_i be start and end times of the ith session, respectively. Then then the inter-sessions times are defined as follows: $A_i = s_i - s_{i-1}$ and $I_i = s_i - e_{i-1}$.

We used Weibull distribution to model inter-sessions times (both A_i and I_i) of all session types we considered. It is defined with a cumulative distribution function of:

$$F(x) = \begin{cases} 1 - e^{-(x/\beta)^\alpha} & \text{if } x > 0 \\ 0 & \text{otherwise} \end{cases}$$

with *shape parameter* $\alpha > 0$ and *scale parameter* $\beta > 0$ [4].

Even though the distributions of inter-session times look qualitatively very different for different types of sessions, they can be modeled with Weibull distribution using suitable shape parameter α. This is shown in Figure **??**. Moreover, Weibull distribution arrises as an inter-arrival time of Power-law Nonhomogeneous Poisson Process, whose relevance and nice properties are discussed in section 4

For data sessions (emails or Web browsing), probability of having short inter-session times is high. This results in natural clustering of sessions in time, with relatively long periods between the clusters. It corresponds to Weibull distribution with $\alpha < 1$. On the other hand, phone call sessions have low probablility of very short inter-session times. Thus, calls are spread in time without a tendency to form clusters. This situaton arrises for Weibull distribution with $\alpha > 1$.

In other words, using the shape parameter α allows us to model the fact that people have a tendency to send emails in bulks, while they wait longer time between calls.

2.2 Source and Destination

A source of a session is choosen implicitly as the person who received a session-begin event. Destination is chosen explicitely at the source's side. Since the nature of a destination is different for each session type (another person for phone calls, a web or email server for data sessions), the way of choosing a destination will depened on the session type.

There is, however, a common underlying structure for destination choosing. Each source (person) has a *destination list* for each session type. The destination list consists of pairs of destination identifiers and weights for choosing that identifier. A destination identifier is then interpreted either as the destination itself (e.g. person or server, this will depend on session type), or it may be a special value singnifying that extra operations need to be done to find the destination (e.g. choose randomly among all people at work etc).

2.3 Session Content

The "what" is being communicated will again depend heavily on the particular session type. It ranges from finding a duration of a phone call, to determining number and sizes of various objects and requests for an HTTP session.

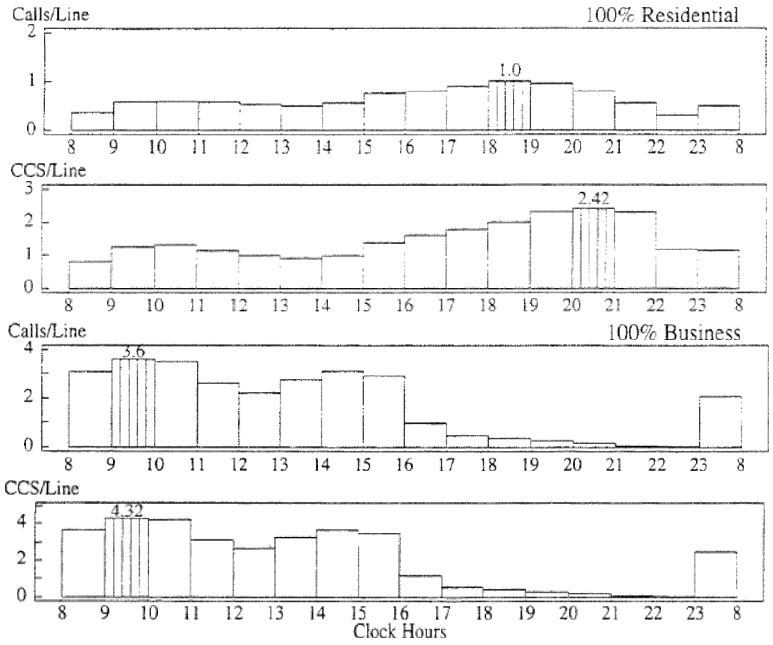

Figure 1: Call intensity for residential (top panels) and business (bottom panels) wire-line phones with respect to the time of day, taken from [5].

2.4 Activity Dependence

All of the above processes that determine properties of communicaton sessions may vary with time. We identify one important variation using which we are able to reporoduce realistic session intensity curve during a 24 hour period, and that is *activity*. We use three basic categories of activities (activity types): work, sleep, and default (all other activities). Using different parameters for the inter-session call time process for each actity type, complicated session-intensity curves at Figure 1 can be reproduced.

The call intensity curves in Figure 1 are strongly related to work and non-work activities in a simulated population in Figure 2. This qualitatively justifies the approach of varying inter-session times with activity types. While data often shows that the intensity curve for a *single* user over *many* days itself resembles the curves shown, we reproduce the shapes by generating sessions for *many* users during a *single* day.

Other parts of the session-generation process can be varied likewise. So the destination list could vary, resulting in workers calling other workers more likely than people at home.

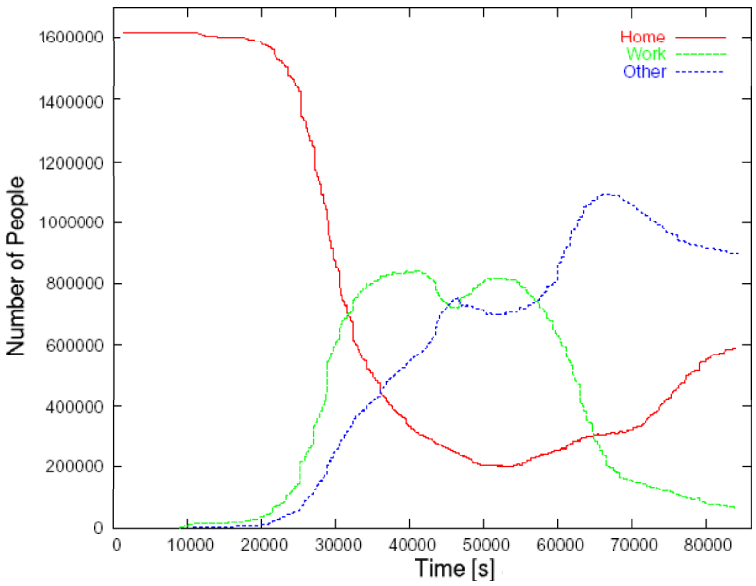

Figure 2: Activities of the generated synthetic population in the Portland study (in 15 minute bins).

3 Individual Session Models and Empirical Validation

In this section, models that were introduced before are made concrete for each of the three session types we model: phone calls, http traffic and emails. Results from using the models are also compared to real-world data that is available to us, or found in the literature. Many of the validation steps are performed by observig the *emergent* behavior of the whole population, because that is what empirical data is available for.

3.1 Phone Calls

The phone call model captures user behavior in making both wireline and wireless calls. The distinction between the two is not described in this paper, since it is part of device usage modeling, which is not dealt with here. If data was available that would justify making the distinction at the session level, the wireline and wireless call would use the same model, but with different parameters. As it stands now, we combine bits and pieces of information we find from both worlds to find out parameter values for the unified model.

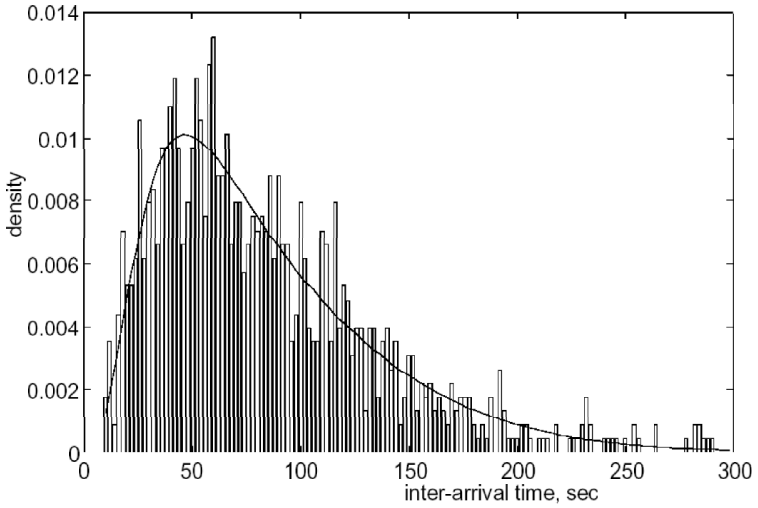

Figure 4. Inter-arrival time distribution: empirical vs. theoretical for ρ=0.6.

Figure 3: Inter-arrival time for cellular base station with a fit of Erlang-3,8 distribution, taken from [8].

3.1.1 Inter-session Time

The inter-session time in case of phone calls is modelled as time between end of one call and beginning of the next (using the I_i notation). The shape parameter α for the Weibull distribution is chosen to be greater than 1, according to empirically observed distributions. In Figure 3, a distribution of inter-arrival times of calls at a cellular basestation is plotted.

Note that data in Figure 3 differs from our modeling approach in two important ways: it provides a *combined* inter-arrival time (for all users of the basestation), and it plots time between *two consecutive session starts* (so the A_i times, using our inter-session time notation). But we can still use it to draw conclusions about how a distribution of inter-session time should look like in our model. In particular, we see that the distribution is *not* exponential. The fact of having combined data for the empirical distribution does not matter, because superposition of exponential distributions would again yield an exponential distribution. The error obtained by going from A_i to I_i is small, because of the very short length of a phone call compared to a time between calls for a single person. Fitting a Weibull distribution to the Erlang-3,8 data yields the shape parameter $\alpha = 1.8$ for our model.

The scale parameter β determines how frequently calls will be made. We do not have a good source for this type of information. We use 2002 Yankee Group

Survey[1] [10] to estimate the number of wireless calls per person per day using the number of monthly minutes (165) and average reported wireless call length (3.7 min). This means cca 1.5 wireless calls/person/day. Moreover, from the same survey, we learn that the average percentage of wireline calls replaced by wireless is 28%, so there are about twice as many wireline calls as wireless. This brings us to about 4.5 calls/person/day. From this, we can compute a mean inter-session time of about 320 minutes, which corresponds to scale parameter $\beta = 360$ mins.

The β parameter varies with activity type, so that the resulting number of calls per person will be different than 4.5, and the default β value must be tuned accordingly. We assign (possibly different) value of the $\beta_{default}$ parameter to each person in our model (scale parameter for default activity). To do that, we use a *social network* of our synthetic population [11], which is a graph in which nodes correspond to people and an edge is present between two nodes if the corresponding people somehow know each other. Now following an intuition (we have no data to show it) that people with more social contacts tend to call more often, we compute the individual parameters as

$$\beta_{default} = \frac{\text{average number of social contacts}}{\text{particular number of social contacts}} \cdot \beta$$

(smaller $\beta_{default}$ value means more frequent calls).

From Figure 1, we learn that a peak intensity of business calls is approximately 3.6 times larger than home calls. So the β parameter for work activity is $\beta_{work} = \frac{1}{3.6}\beta_{default}$.[2] We have no data to set the night intensity, so we estimate $\beta_{sleep} = 10 \cdot \beta_{default}$. The shape parameter α does not change with activity type (we have no data to support the change).

Even though we concentrate on modeling exactly one day of communication sessions, it is worthwhile to mention that [5] suggests that day-to-day call rate distribution can be very well modeled using normal distribution.

3.1.2 Source and Destination

Destination lists for phone calls contain individual people that a particular person might call, plus special entries for calling random destination and random people that are currently at work.

The entries of individual contacts are obtained from the social network mentioned before. The weights are proportional to a strength of the social contact (a weight on the edges in the network, e.g. time duration of the contact). We have no data to set weights of the special entries, so we used the following values: 10% of random calls, and 20% of calls to random workers during work activity (no calls to random workers during other activities).

The difference in calling random workers mentioned above is the only place where destination lists differ with activity.

[1] 2004 survey does not have the required fields.

[2] We can do this because the mean of Weibull distribution is linear in its β parameter

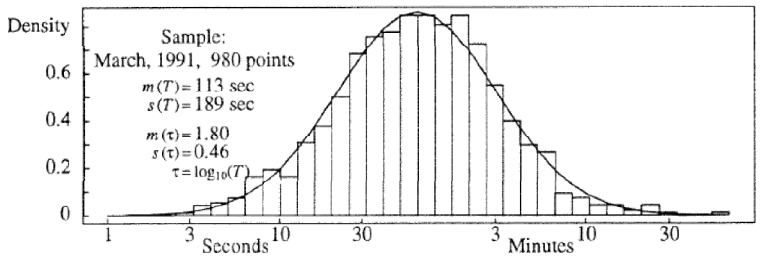

Figure 4: Call-length distribution with a normal fit on logarithmic scale, taken from [6].

3.1.3 Session Content

Content of a phone call session is defined only by its length. Figure 4 from [6] suggests that a log-normal distribution represents a very nice fit for wireline call lengths. The log-normal fit (or combination of log-normals) for session lengths is also suggested for other scenarios [7, 9]. The mean and standard deviation for the call lengths is taken from Figure 4, $m = 113$s and $s = 189$s, respectively.

The activity dependence can be derived from Figure 1 by dividing CCS[3] by then number of Calls in corresponding time bin. We find that the mean of 113 seconds corresponds best with work activity, while default activity (e.g staying home) has calls of approximately twice the length, 226 seconds. This nicely agrees with the information about wireless calls obtained from the Yankee Group Survey [10] discussed above. Parameters for to the log-normal distribution are then $meanlog = 4.07$, $sdlog = 1.15$ for work activity[4], and $meanlog = 4.76$, $sdlog = 1.15$ for default activities. In lack of any data to validate it, we do not vary the *sdlog* parameter with activity, so the standard deviation of the resulting call lengths will increase with increased mean.

Length of home phone calls also depends on time, as noted in [5]. This is also apparent from Figure 1 by observing that the ratio between CCS and Calls is not constant (it is more or less constant in case of business calls). This time dependence is not captured by our model.

3.2 Http Traffic

The HTTP sessions that our model describes are sessions generated by people using the World Wide Web service of the Internet. Other applications that may operate using the HTTP protocol (such as streaming [12], web crawlers or other automated HTTP usage) are not captured. The model is based on, and is very simlar to, the one presented in [13, 14].

[3]Hundred Call Seconds, a measure of call intensity.

[4]values given in Figure 4 are for logarithm-10 base log-normal distribution.

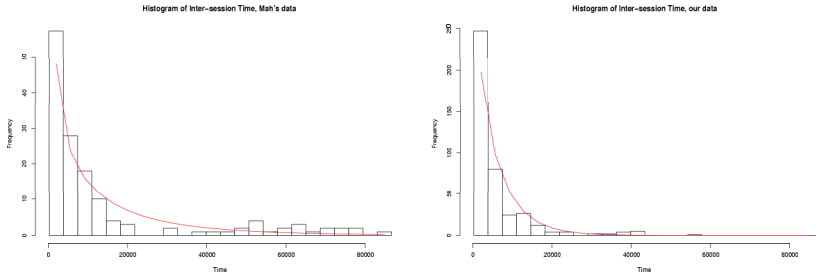

Figure 5: Distribution of time between two consecutine WWW sessions. Histogram with Weibull distribution fit, left panel data taken from [15], right panel our data.

3.2.1 Inter-session Time

In WWW sessions, inter-session time is measured as time between two consecutive session starts (A_i). This approach differs from the one used in [13, 14], where individual sessions alternate between on/off states. Scheduling a next session right after one starts (as oppose to when it ends) allows for having multiple simultaneous sessions for the same person, which we believe corresponds better to reality.

Fitting Weibull distribution to data from [15] and to data we collected ourselves (see Appendix A) is shown in Figure 5. We use 900 seconds as threshold to distinguish http requests that belong to different sessions (shorter inter-request times are considered to belong to the same session). This technique was suggested in [13]. The fitted parameters are $\alpha = 0.76$ and $\beta = 12000$ for [15] data, and $\alpha = 0.95$ and $\beta = 5600$ for our dataset, in seconds (K-S and χ^2 goodness-of-fit tests both reject it utterly, though). The shape parameters α are similar, while β values differ more significantly. This suggests that the model for WWW inter-session time is similar to the phone calls model. Shape parameter α is help constant for all users, while scale parameter β varies from user to user. Due to lack of data to explore the user dependence, we keep both parameters constant for all users at $\alpha = 0.81$ and $\beta = 6900$ (fit for combined data).

Figure 6 shows Internet usage intensity by time of day. Assuming that the usage is dominated by WWW browsing, the pattern looks remarkably similar to that of phone calls: combination of default/work/sleep activity types with different session intensities for each type will recreate the pattern. Seeing that work intensity is about twice the default one, we set $\beta_{work} = \frac{2}{3} \cdot \beta$ and $beta_{default} = \frac{4}{3} \cdot \beta$. We set again $\beta_{sleep} = 10 \cdot \beta_{default}$ as in the case of phone calls.

3.2.2 Source and Destination

The destination list in case of HTTP traffic contains HTTP servers, along with weights coresponding to their usage. An HTTP server can either be specified

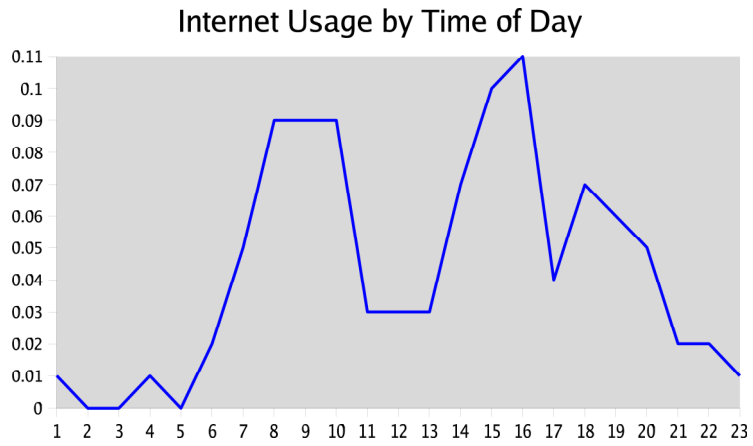

Figure 6: Internet usage by time of day, modified data from [16]

in terms of URL, or as an id of a particular server. Note only that while most surveys show raning of URLs, many URLS may be mapped to go to the same server (e.g. Akamai servers).

The weights for individual web servers should follow the Zip's Law [15] (weight of the ith most popular site is proportional to $1/i$). The destination list may also be different for work and default activity, but we have no data to show that.

3.2.3 Session Content

Each WWW session consists of several *requests*, and each requests has a primary request and reply, plus several secondary request-reply pairs. The primary request-reply correspond to the main HTML file, and the secondary request-replies pairs correspond to inlined object within that web page. The process of constructing a WWW session content, along with the random distributions that are used, is outlined bellow:

1. Number of requests in the session: X=lognormal(meanlog=1.8, sdlog=1.68) (from [13])
2. For each request:
   (a) Choose another destination server with prob 0.3 (conforms to consecutive document retrievals measure from [15])
   (b) Primary request-reply pair:
      i. Request size [kB]: S=lognormal(meanlog=0, sdlog=0.29) (combining information from [13, 15])
      ii. Reply size [kB]: M=lognormal(meanlog=1.31, sdmean=1.41) (from [13])

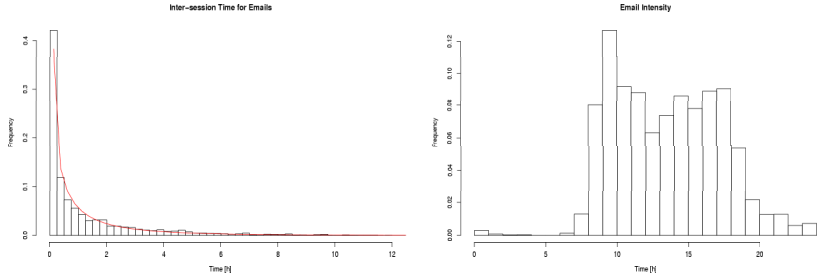

Figure 7: Distribution of inter-session time for emails and its Weibull fit (left panel) and email send intensity during a day (right panel). Obtained from our data (work-related emails).

(c) Number of inlined objects in the request: N=gamma(shape=0.24, scale=23.42) (from [13])

(d) For eech inlined object (secondary request-replies)

i. Request size [kB]: R=lognormal(meanlog=-1.4, sdlog=0.29) (combining information from [13, 15])

ii. Reply size: O=lognormal(meanlog=-0.75, sdlog=2.36) (from [13])

3. Time between requests [s]: V=weibull(shape=0.51, scale=21) (data from [15] fitted to median and stddev mentioned in [14])

The number of request that actually will need to be fetched from the original server will depend on the browser's and local network's caching mechanism. Data from [14] suggest that there is only a relatively small fraction (less than 10%) of locally cached data. Local network caches (such as corporate proxies) might possibly have a larger amount of prestored data from popular sites. Due to unavailability of data, we do not include any caching mechanism in our model.

3.3 Emails

There is not much email traffic analysis in the literature. A nice overview of what is available is in [13]. We also obtained some data ourselves, as described in Appendix A, which we use to estimate parameter values for our model.

Unlike other session types, sending emails is a three-stage process: its sending to a local email server (e.g. SMTP server), transmission to the recipient's email server (e.g. POP3 server), and its download to the recipient. Both first and last stage include modeling of user behavior, but we focus on the first stage. Another popular way of working with emails is via a WWW interface, which we hope to capture in our WWW session model.

3.3.1 Inter-session Time

The inter-session time is again measured by tim between two consecutive email sends (A_i). We used our dataset to estimate the parameters for Weibull distribution, as shown in Figure 7 (left panel). The Weibull fit yields parameters $\alpha = 0.6$ and $\beta = 3000$ (in units of seconds). These values correspond to an average of the three users we analyzed, and to work-related emails only. Moreover, we expect that the values obtained will be higly above average due to the nature of their work. We therefore set the average work-related scale parameter to $\beta_{work} = 15000$. We keep the shape parameter $\alpha_{work} = 0.6$ the same, which is also supported by our data (the relative difference in the α values is small in our sample compared to βs). Even though we observe large varience in the scale parameter among different users (1.5K, 3K and 5.2K), we keep it fixed in our model and do not vary it by user. This is due to lack of data to explore the dependence properly.

The right panel of Figure 7 shows that email sending intensity follows again the curve of work activity type. Similrarly to WWW sessions, we define the scale parameters for other activity types as follows: $\beta_{default} = 2 \cdot \beta_{work}$ and $\beta_{sleep} = 10 \cdot \beta_{default}$ (Figure 6 can be used again).

3.3.2 Source and Destination

The destination model is essectially the same as in case of WWW sessions. The destination list consits of email *server* entries (i.e. final recipients full email address is not necessary) along with weights for choosing each.

Emails can possibly have multiple recipients. There is no source in the literature to statistics about this, and our limited sample analysis is inconlusive as to which distribution should be use for this purpose. Therefore, only single recipients are included in our current model.

3.3.3 Session Content

Email session content is fully defined in our model by the email size. Following the model in [17], we fit our data to a trimmed Chauchy distribution (because there is a minimal email size of about 0.4KB, all values bellow this are discarded). The email size distribution along with the trimmed Cauchy fit is in Figure 8. The fitted parameter values for the Cauchy distribution are location = 0.8 and scale = 1.4. As noted in [13], the Cauchy model tends to underestimate fraction of long emails. This is somewhat true also in our case, althought our location parameter is larger than the one considered in [13].

It is hard to regenerate the empirical data in terms of the same mean and standard deviation. This is due to the fact that neither of the measures is defined in the case of Cauchy distribution, so when a sample is drawn from such distibution, the average and sample standard deviation vary significantly. A median (a measure that *is* defined for Cauchy distribution) of the trimmed distribution matches well with the observed data, though.

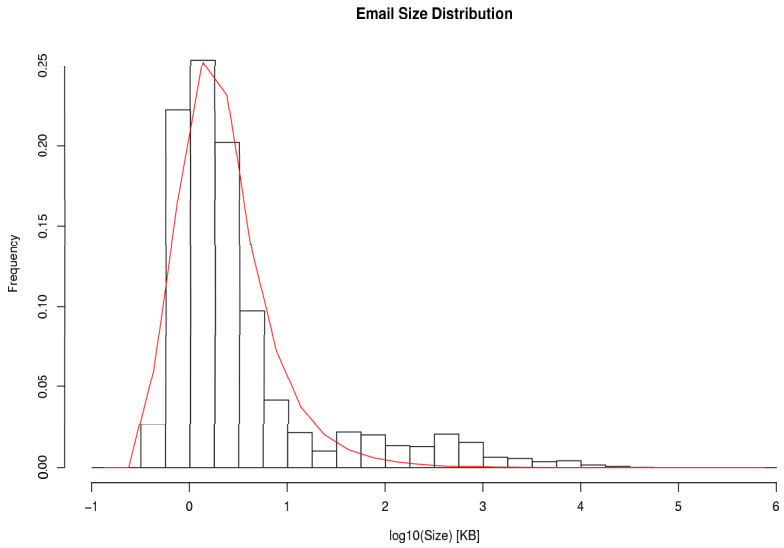

Figure 8: Email size distribution (on log scale) along with its (trimmed) Cauchy fit, from our data.

3.4 Other

This section contains information about other possible session types which generation could benefit from activity information. Models for these are not developed in detail, though.

3.4.1 Streaming

From Figure 9, it can be seen that intensity of streaming applications corresponds faily well with work activity (in case of live streams), or combination of default/work/sleep activity similar to phone calls (for on-demand streaming of pre-recorded clips). But there is a *very* large variability of the intensity from day to day ([12] documents a six-fold increase from one day to the next). This suggests that more data need to be explored before a model for such highly-varying process is developed.

4 Theoretical Underpinnings

The inter-arrival times A_i are modeled as a renewal process [1]. This means that A_i are independently identically distributed according to some distribution. While the distribution is different for different session types, it can be very well modeled as Weibull distribution with different shape parameters (see section 2.1).

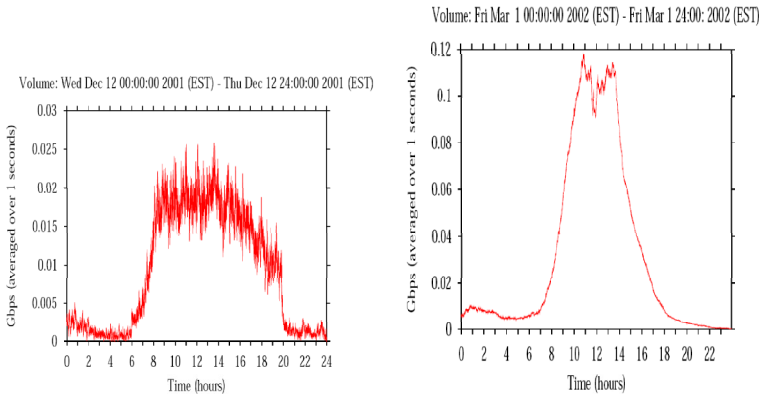

Figure 9: Intensity of streaming usage for on-demand clips (left panel) and live streams (right panel), taken from [12].

The Weibull distribution has some nice properties, which are discussed bellow. It is a distribution of inter-arrival times arrising from a non-homohenegous Poisson Process, with an intuitive explanation of the difference in the distribution shapes. This in turn allows us to reschedule already-scheduled events without afecting the overall inter-arrival time distribution. This is important because we have to reschedule future session events whenever an activity of a person changes.

4.1 Power-law Non-homogeneous Poisson Process

To review basic definitions and properties of Poisson process and to see its relevance to session generation, see Appendix B. A non-homogeneous Poisson process (NHPP) is a Poisson process where the rate λ is a function of time, $\lambda(t)$ [2]. The inter-arrival time cumulative distribution function (CDF) is given by:

$$P\left[A_i \leq t\right] = 1 - e^{-\int_0^t \lambda(\tau)\mathrm{d}\tau} \tag{1}$$

(see Appendix C for derivation)

For $\lambda(t) = \alpha \cdot t^{-\beta}$ ($\alpha > 0$, $\beta < 1$), we obtain a Power-law Non-homogeneous Poisson Process [3]. Using this in Equation 1, we obtain:

$$P\left[A_i \leq t\right] = 1 - e^{-\frac{\alpha}{1-\beta} t^{1-\beta}} \tag{2}$$

which is a CDF of Weibull distribution with a shape parameter $1-\beta$ and scale parameter $\left(\frac{1-\beta}{\alpha}\right)^{\frac{1}{1-\beta}}$

In session simulation, the time t always measures time since the last event. The α parameter sets the initial rate, and β governs how the rate chages in time. When an event occurs, the rate is reset to α (by setting t to 0). For

$\beta = 0$, the rate is constant, and the process is therefore a homogeneous Poisson process with exponentially distributed inter-arrival times. For $0 < \beta < 1$, the rate *decreases* in time, and events are therefore more likely to happen soon after a previous event, naturally creating clusters in time (typical for data sessions). For $\beta < 0$, the rate *increases* in time, and consecutive events are therefore separated by longer intervals (typical for call sessions).

4.2 Rate Parameter Change with Activities

When a change in activity of a person occurs, different α_{new} and β_{new} parameters may have to be supplied to the rate function $\lambda(t)$. A new event has to be scheduled with the new rate function, replacing the old event. Let T denote the time elapsed since the last event, $T \geq 0$. Since T may now be non-zero, we have to alter Equation 1 appropriately:

$$P_T\left[A_i \leq t\right] = 1 - e^{-\int_T^{T+t} \lambda(\tau)\mathrm{d}\tau} = 1 - e^{-\int_0^t \lambda(T+\tau)\mathrm{d}\tau} \tag{3}$$

where t measures time since the rate change.

Plugging α_{new} and β_{new} into the rate function $\lambda(t) = \alpha \cdot t^{\beta}$, we obtain a generalized version of Equation 2:

$$P_T\left[A_i \leq t\right] = 1 - e^{-\frac{\alpha}{1-\beta}\left((T+t)^{1-\beta} - T^{1-\beta}\right)} \tag{4}$$

From Equation 4, it follows that the time to the next event after the rate paramater change does *not* depend on the previous rate paramaters, but only on the new parameters values and the time of the last event occurance.[5]

There is one important property that we need to preserve when rescheduling events. We need to make sure that the probability of an event occurring in a certain interval $(t, t+\Delta]$ does not change by rescheduling the event with the *same* rate parameters. This is captured by the following equality:

$$P_{T_1}\left[A_i \leq t + \Delta | A_i > t\right] = P_{T_2}\left[A_i \leq t + \Delta | A_i > t\right] \tag{5}$$

for $x \geq T_1 \geq T_2$, where $T_{1,2}$ are two times after the last event occurance when we are rescheduling the next one. If $x < T$, then this would mean that the event had already happened at time T and we would not be rescheduling it. To show that Equatily 5 holds, let $INT(T, t_1, t_2) := -\int_{t_1}^{t_2} \lambda(T+\tau)\mathrm{d}\tau$. Then using a definition of conditional probability and Equation 3, we have:

$$\begin{aligned} P_T\left[A_i \leq t + \Delta | A_i > t\right] &= \frac{1 - e^{INT(T,0,t+\Delta-T)} - (1 - e^{INT(T,0,t-T)})}{1 - (1 - e^{INT(T,0,t-T)})} \\ &= 1 - e^{INT(T,0,t+\Delta-T) - INT(T,0,t-T)} \\ &= 1 - e^{INT(T,t-T,t+\Delta-T)} \\ &= 1 - e^{INT(0,t,t+\Delta)} \end{aligned}$$

Since $P_T\left[A_i \leq t + \Delta | A_i > t\right]$ does not depend on T, it follows that Equality 5 holds.

[5] For $T > 0$, Equation 4 no longer corresponds to the Weibull distribution.

5 Conclusions

A Obtaining Data

We were able to obtain a limited sample of WWW behavior of one user and email statistics of few. In hope that somebody may someday be able to do a more complete study, a description of was was done is included here. All scripts are assumed to be run from `bash`.

A.1 Http Traffic

Mozilla has a logging facility that can be used to gather data about requests and responses that a user generates. It does not have any negative consequences (like speed) as long as the logging level is kept low. To enable logging, do the following before running Mozilla (e.g. put the follofing lines into `mozilla` script, with a correct path for the log files):

```
export NSPR_LOG_MODULES=nsHttp:3
export NSPR_LOG_FILE=/home/kroc/HttpLogs/‘whoami‘-‘date +"%s"‘.txt
```

This logs all request and reply headers for the HTTP traffic. From this, the number of seconds in a day when a request was issued and time between requests can be obtained by:

```
echo -e "Seconds\tInterSeconds"
grep -E "http request|Date: " |
grep "http request" -A1 |
grep "Date:" |
cut -d: -f3- |
while read DATE
do
    date -d"$DATE" +"%H %M %S"
done |
gawk ’
    function getSec(h,m,s) { return (h*60 + m)*60 + s }
    NR==1 {
        Last=getSec($1,$2,$3);
        print Last,"NA";
        next }

    {
        This=getSec($1,$2,$3)
        if( This >= Last ) { print This,This - Last }
        else { print This,"NA" }
        Last = This
    } ’
```

The above is only an approximation, it takes a "Date:" field from the first responce to the request, because the request header does not have the date and time information in it. It only outputs inter-requests times for requests issued in the same day, we do not want "over-night" intervals (an approximation again, different days are recognized by decrease in time, which may not always be the case).

A.2 Email

It is possible to extract information about emails sent by Mozilla (or Nescape) by examining a file that contains the email Sent folder (usually something like ∼/.mozilla/default/*/Mail/*/Sent). The samples we used (Jim, Stephan and Lukas) were all work emails with not necessarily an average work schedule :-) . The following script does that and outputs [Time, Size, Recipients] fields for every email sent:

```
gawk -v FS="" -v OFS="" '
    BEGIN {
        Time="Time"
        Sum="Size"
        Recipients="Rcps"
        Email=0
        Rcps=0 }

    # beginning of a new email
    /^From - / && Email==0 {
        Email=1
        print Time,"\t",Sum,"\t",Recipients
        Time=""
        Sum=0
        Recipients="" }

    # date section
    /^Date: / && Email==1 {
        sub("^Date: ","",$0)
        Time=$0 }

    # turn recipient info off
    /^[[:graph:]]+: / {
        Rcps=0 }

    # turn info about recipients on
    (/^[tT][oO]:/ || /^[cC][cC]:/ || /^[bB][cC][cC]:/) && Email==1 {
        Rcps=1 }

    # looking for an empty line at the of an email header
```

```
    /^[[:space:]]*$/ && Email==1 {
        Email=0
        Rcps=0 }

    # remember recipients
    Rcps==1 {
        Recipients = Recipients " " $0 }

    # sum up size of content
    { Sum += NF }

    END {
        print Time,"\t",Sum,"\t",Recipients }'
```

Then the following script takes the above output and turns it into [Seconds in a day, Size, Number of recipients, Seconds between sends] fields (as with the WWW sessions, the Seconds between sends may not be correct):

```
echo -e "Seconds\tSize\tRcpsNum\tInterSeconds"
IFS="#"
tail +2 |
tr "\t" "#" |
while read Time Size Rcps;
do
    Time2=`date -d"$Time" +"%H %M %S"`
    Rcps2=`echo $Rcps | sed -e 's/[^@]//g' | wc -c | gawk '{print $1-1}'`

    echo -e "$Time2\t$Size\t$Rcps2"
done |
gawk -v OFS="\t" '
    function getSec(h,m,s) { return (h*60 + m)*60 + s }
    NR==1 {
        Last=getSec($1,$2,$3);
        print Last,$4,$5,"NA";
        next }

    {
        This=getSec($1,$2,$3)
        if( This >= Last ) { print This,$4,$5,This - Last }
        else { print This,$4,$5,"NA" }
        Last = This
    } '
```

B Poisson Process

A Poisson process can be characterized as a renewal process whose inter-arrival times $\{A_i\}$ are exponentially distributed with rate parameter λ:

$$P[A_i \leq t] = 1 - e^{-\lambda t}$$

Equivalently, it is a counting process satisfying:

$$P[N(t) = n] = \frac{e^{-\lambda t}(\lambda t)^n}{n!}$$

(the number of arrivals in a given interval $(0, t\rangle$ has a Poisson distribution with a parameter λt, where λ is the event arrival rate and t is the length of the interval).

The derivation of the above distributions and the process's relevance to the problem of session generation can be viewed using a simple thought experiment. Suppose we have i discrete time instances of length Δt. In each of these intervals, an event may occur with a constant probability $p = \lambda.\Delta t$ independent of any other event occurrences. This corresponds to a situation where a person decides each second whether to make a call or not, independent of his/her decision history. Now the question is what is the probability that an event occurs exactly n times. This can be modeled using binomial distribution:

$$P[\mathrm{Bi}(i, p) = n] = \binom{i}{n} p^n (1-p)^{i-n}$$

For $i \to \infty$, $\Delta t \to 0+$, and $i.\Delta t \to t$ where t is a constant (overall time), the above expression becomes $\frac{e^{-\lambda t}(\lambda t)^n}{n!}$. This means that it has a Poisson distribution with a parameter λt, corresponding to the Counting process definition of the Poisson process stated above.

Now a question arises of what is the waiting time until the first event arrival. It is a random variable, which cumulative distribution function can be derived as follows:

$$\begin{aligned}
F(t) &= P[\text{the first event's time} \leq t] \\
&= 1 - P[\text{the first event's time} > t] \\
&= 1 - P[\text{no event in interval of length } t] \\
&= 1 - P[\mathrm{Po}(\lambda t) = 0] \\
&= 1 - \frac{e^{-\lambda t}(\lambda t)^0}{0!} \\
&= 1 - e^{-\lambda t}
\end{aligned}$$

It is hence an exponentially distributed random variable with rate parameter λ, as stated in the Inter-arrival time process characterization of the Poisson process.

C Inter-arrival time in Non-homogeneous Poisson Process

See [2] for the full story. Let $N(t)$ denote the number of times an event occurs in time interval $(0, t]$. Let $p(n, t) = P[N(t) = n]$ be the probability that the event occured exactly n times. From the definition of NHPP, it follows that

$$\begin{aligned} p(0, t + \Delta t) &= p(0, t)\,(1 - \lambda(t) \cdot \Delta t - o(\Delta t)) \\ &= p(0, t) - \lambda(t) \cdot p(0, t) \cdot \Delta t - p(0, t) \cdot o(\Delta t) \end{aligned}$$

(because probability that an event happens in time interval of length Δt must be $\lambda(t)\Delta T + o(\Delta t)$, so that at maximum one event occurs in a sufficiently small interval)

Therefore,

$$\frac{p(0, t + \Delta t) - p(0, t)}{\Delta t} = -\lambda(t) \cdot p(0, t) - \frac{p(0, t) \cdot o(\Delta t}{\Delta t}$$

Now assuming that $p(0, t)$ is differentiable with respect to t and letting $\Delta t \to 0$, we obtain:

$$\frac{\mathrm{d}}{\mathrm{d}t} p(0, t) = -\lambda(t) \cdot p(0, t)$$

We solve the above equation with an initial condition of $p(0, 0) = 1$, and get:

$$p(0, t) = e^{-\int_0^t \lambda(\tau)\mathrm{d}\tau}$$

Now following the same logic as in Appendix B, we conclude that

$$\begin{aligned} F(t) &= P\left[\text{the first event's time} \leq t\right] \\ &= 1 - P\left[\text{the first event's time} > t\right] \\ &= 1 - P\left[\text{no event in interval } \textit{(0,t]}\right] \\ &= 1 - p(0, t) \\ &= 1 - e^{-\int_0^t \lambda(\tau)\mathrm{d}\tau} \end{aligned}$$

References

[1] V.S.Frost, B.Melamed: *Traffic Modeling For Telecommunications Networks,* IEEE Communications Magazine, March 1994

[2] A.Høyland, M.Rausand: *System Reliability Theory, Models and Statistical Methods*; Wiley, 1994

[3] NIST/SEMATECH e-Handbook of Statistical Methods, http://www.itl.nist.gov/div898/handbook/, 2005.

[4] A.M.Law, W.D.Kelton: *Simulation Modeling and Analysis, 3rd ed.* McGraw Hill, 2000.

[5] V.A.Bolotin: *New Subscriber Traffic Variability Patterns for Network Traffic Engineering,* 15th International Teletraffic Congress ITC 15, 1997.

[6] V.A.Bolotin: *Telephone Circuit Holding Time Distributions,* 14th International Teletraffic Congress ITC 14, 1994.

[7] F.Barcelo, J.Jordan: *Channel Holding Time Distribution in Cellular Telephony,* 9th International Conf. On Wireless Communications (Wireless '97), Calgary, 1997.

[8] F.Barcelo, J.I.Sanchez: *Probability Distribution of the Inter-arrival Time to Cellular Telephony Channels,* Proc. 49th Vehicular Technology Conference, Houston, 1999.

[9] J.Jordan, F.Barcelo: *Statistical Modeling of Transmission Holding Time in PAMR Systems,* IEEE Globecom 97, Phoenix AZ, November 1997.

[10] The Yakee Group: *Technologically Advanced Family Survey,* 2002

[11] Anybody: *Anything on Episim,* Anytime.

[12] J.Merwe, S.Sen, C.Kalmanek: *Streaming Video Traffic: Characterization and Network Impact,* Proceedings of the Seventh International Web Content Caching and Distribution Workshop, Boulder, CO, August 2002

[13] P. Tran-Gia, D. Staehle, K. Leibnitz: *Source Traffic Modeling of Wireless Applications,* International Journal of Electronics and Communications (AE), 55, 2001.

[14] Hyoung-Kee Choi, John O. Limb: *A Behavioral Model of Web Traffic,* Seventh Annual International Conference on Network Protocols, Toronto, Canada, 1999.

[15] B.Mah: *An Empirical Model of HTTP Network Traffic,* Proceedings of the IEEE Infocom 97, Kobe, April 1997.

[16] *Internet Usage By Time of Day,* http://www.zonalatina.com/Zldata386.htm

[17] G.Brasche, B.Walke: *Concepts, services, and protocols of the new GSM phase 2+ generalpacket radio service,* IEEE Communications Magazine, August 1997.