Electronic Journal of Statistics
Vol. 4 (2010) 58–116
ISSN: 1935-7524
DOI: 10.1214/09-EJS473

Statistical analysis and modeling of Internet VoIP traffic for network engineering

Bowei Xi*

Department of Statistics, Purdue University, West Lafayette, IN 47907, USA
e-mail: xbw@stat.purdue.edu

Hui Chen

Yahoo! Inc., 701 First Avenue, Sunnyvale, CA 94089, USA
e-mail: chenhui@yahoo-inc.com

William S. Cleveland†

Department of Statistics, Purdue University, West Lafayette, IN 47907, USA
e-mail: wsc@stat.purdue.edu

Thomas Telkamp

Cariden Technologies Inc., 888 Villa Street, Suite 500, Mountain View, CA 94041, USA
e-mail: telkamp@cariden.com

Abstract: Network engineering for quality-of-service (QoS) of Internet voice communication (VoIP) can benefit substantially from simulation study of the VoIP packet traffic on a network of routers. This requires accurate statistical models for the packet arrivals to the network from a gateway. The arrival point process is the superposition, or statistical multiplexing, of the arrival processes of packets of individual calls. The packets of each call form a transient point process with on-intervals of transmission and off-intervals of silence. This article presents the development and validation of models for the multiplexed process based on statistical analyses of VoIP traffic from the Global Crossing (GBLX) international network: 48 hr of VoIP arrival times and headers of 1.315 billion packets from 332018 calls. Statistical models and methods involve point processes and their superposition; time series autocorrelations and power spectra; long-range dependence; random effects and hierarchical modeling; bootstrapping; robust estimation; modeling independence and identical distribution; and visualization methods for model building. The result is two models validated by the analyses that can generate accurate synthetic multiplexed packet traffic. One is a semi-empirical model: empirical data are a part of the model. The second is a mathematical model: the components are parametric statistical models. This is the first comprehensive modeling of VoIP traffic based on data from a service provider carrying a full range of VoIP applications. The models can be used for simulation of any IP network architecture, wireline or wireless, because the modeling is for the IP-inbound traffic to an IP network. This is achieved because the GBLX data, collected on an IP link, are very close to their properties when they entered the GBLX network.

*Communicating author. Supported by NSF DMS-0904548.
†Supported by ARO MURI Award W911NF-08-1-0238.

AMS 2000 subject classifications: Primary 62P30, 62-07; secondary 62M10, 62M15.

Keywords and phrases: Statistical model building, very large datasets, semi-empirical models, long-range dependence.

Received August 2009.

Contents

1. Introduction

1.1. Voice over the Internet (VoIP)

Voice is increasingly being transmitted over the Internet. It is migrating from the Public Switched Telephone Network (PSTN), the traditional voice phone network that 25 years ago carried nearly all voice traffic (Peterson and Davie, 1999). The technology for the Internet is referred to as the Internet Protocol (IP), and voice over the Internet is VoIP. Two widely used sets of VoIP protocols are SIP-RTP (Rosenberg et al., 2002; Schulzrinne et al., 2003), which is the set for the VoIP data reported in this article, and Skype (Baset and Schulzrinne, 2006; Suh et al., 2006).

A VoIP communication starts with a signal, a waveform, that is sampled at rates typically from 5.3 kilobits/sec to 64 kilobits/sec depending on the implementation. An increasing rate produces better signal reproduction. The captured bits are accumulated over intervals of 20 ms or 30 ms. At the end of each interval, the bits become a packet of information that is transmitted over the Internet. If the capture rate is 64 kilobits/sec and the interval is 20 ms, as it is for our data, then the size of the packet information is (64 kilobits/sec)/(0.002 sec) = 1280 bits = 160 bytes. Added to each packet are headers that Internet protocols must read to carry out their tasks of getting the packet to its destination and reforming the signal. The device that carries out this IP packetization is the *IP-inbound gateway*.

Each call has packets in two directions: caller host to callee host and callee host to caller host. The packets in each direction are a *semi-call*. For our data, the headers are 40 bytes, so each packet is 200 bytes = 1600 bits. A packet is released for transmission to the destination as soon as it is formed, so for a semi-call this happens at intervals of 20 ms or 30 ms.

Packets of a semi-call are routed across the Internet. They move from one router to the next over a transmission link. A router has input links and output links, and each link can carry the packets of other Internet communication connections. These can be other VoIP calls. They can be connections in which files are transferred by the Internet protocol TCP; one example is the transfer of a Web page from a server host to a client host. For such data connections, files are also broken up into packets that carry up to 1460 bytes = 11680 bits of the file. Such data traffic using TCP is referred to as *best-effort traffic*.

An output link has a speed in bits/sec at which the equipment can "write" the bits of a packet onto the link. The link from which our measurements come is 100 megabits/sec. (Once on the link, a bit travels down it at the speed of light.) The combined VoIP traffic using the link has a packet arrival rate which we can average over any interval of time such as 5 minutes to get a rate in packets/sec. Each packet is 1600 bits, so the VoIP traffic rate in bits/sec is the packet rate times 1600 bits/packet. For the link from which our measurements come, 5 minute traffic rates vary over about 2-11 megabits/sec.

1.2. Quality of Service (QoS) for VoIP

The link transmission start times of the packets of any semi-call, the times that the first bits of the packets are put on the link, form a transient point process. The start times of the packets of all semi-calls using the link are superposed, or statistically multiplexed, on the link. (Interestingly, statisticians and probabilists use the term *superposition*, but network researchers use the term *statistical multiplexing*.)

Only one packet at a time can be transmitted over the link. If a packet arrives for transmission on an output link that is not in use, transmission begins nearly immediately. Otherwise, the packet is put into a queue to await transmission. If the queue buffer has no space available, the packet is dropped; this happens when the traffic bit-rate exceeds the link speed for a while, the traffic backs up, and the queue fills to capacity. If there is space in the queue, the packet is just delayed. The service time for a packet is the transmission time of the packet, which is the time it takes the router interface to write the packet bits on the output link. The link speed for the link from which we collected our data is 100 megabits/sec. The transmission time of a 1600-bit VoIP packet at 10^8 bits/sec is 16 μs.

Network quality of service (QoS) for a VoIP semi-call, how well the speech sounds at the destination, is determined by the transit of the packets across the Internet from the source host to the destination host. Queueing delays at the routers have a major impact on quality. The quality can be very well characterized by two QoS criteria (Karapantazis and Pavlidou, 2009). The first criterion is packet travel time from source to destination (mouth to ear); the upper bound beyond which degradation begins to appear is about 150 ms. The more time a packet spends in queues, the longer the travel time. A second criterion governs the inter-arrival times of the packets of the semi-call at the destination.

At the source, VoIP packets are sent every 20 ms or 30 ms, according to the accumulation interval. VoIP protocols at the receiving end must feed speech to a listener at a constant rate. Varying delays for the packets of a semi-call in a queue causes jitter: deviations of the destination semi-call packet inter-arrivals from their original 20 ms or 30 ms. The protocols can compensate to some extent by buffering the speech, but in general more jitter results in lower speech quality. A QoS criterion for the mean absolute jitter is 30 ms.

VoIP traffic also contains connections that are machine-to-machine transactions such as faxes, point-of-sale credit card charge verifications, and computer connections to the Internet (Karapantazis and Pavlidou, 2009). These connections occur in VoIP traffic because they are carried out on lines used for voice. Quality of service (QoS) requirements for voice are more stringent than for the other applications. Voice is a real-time application. Speech, to be intelligible, needs special treatment. So QoS standards for VoIP traffic are set for voice.

1.3. Network engineering for efficiency and QoS

VoIP network engineering has two parts: VoIP traffic engineering and VoIP protocol implementation engineering. The goals are to operate the network as efficiently as possible while assuring QoS. Traffic engineering practices determine the topological design of an IP network, its link speeds, and the amount of traffic sent over the links. Implementation engineering involves choosing VoIP implementation factors such as the signal capture bit-rate and the duration of the accumulation interval. Some practices reduce traffic and others reduce queueing delay.

Network traffic engineering

The queueing in a router buffer is affected by the arrival bit-rate of the statistically multiplexed VoIP traffic on the link, and the speed, or bit-rate, of the link. A higher traffic rate increases the queueing distribution, and a higher link speed reduces the service time and therefore decreases the queueing distribution (Kesidis, 2007; Rolls et al., 2005). The link *utilization*, u, is the mean traffic rate divided by the link speed. The goal is to find the QoS utilization: the maximum utilization u_{qos} that satisfies the QoS criteria for delay and jitter. Over-provisioning, which means operating at $u < u_{\text{qos}}$, wastes resources.

Network implementation engineering

One implementation engineering practice, silence suppression, operates on the packets of a semi-call at the IP-inbound gateway. An algorithm listens to what is being sent, determines that there is silence, and stops the sending of packets from the gateway. Transmission starts up again when the algorithm hears a signal. During the silence period, the gateway sends a packet to start the silence

interval, and keep-alive packets every 2 sec to let the destination host know that the connection has not been severed. When silence suppression is used, the semi-call consists of a sequence of alternating transmission and silence intervals whose lengths form an on-off process.

Another practice is priority queueing. If no action is taken, VoIP packets must wait in the queue for the transmission of best-effort traffic packets such as Web page transfers. However, a VoIP semi-call has much more stringent QoS requirements, and typically has much less traffic volume overall. The practice is to move an arriving VoIP packet ahead of all best-effort packets in the queue.

Other implementation engineering factors are the signal capture rate, compression of the packet contents, and the accumulation interval. The first two change the traffic rate, and the third keeps the traffic rate constant but changes the packet size and the time interval between packets of each semi-call.

2. Justification and overview of the work presented here

2.1. Statistics and Internet traffic engineering

Network engineering for VoIP QoS is challenging because the statistical properties of the multiplexed VoIP arrival process and the effect of network engineering factors on them are very complex, as they are for all types of Internet traffic. When the semi-call rate increases, the multivariate distribution of any fixed number of consecutive inter-arrivals can change in complex ways because of the increased multiplexing of packets of different semi-calls (Cox and Isham, 1992). Furthermore, characterizations of the queueing delay distributions are typically not tractable, although there has been much work in approximated tail behavior (Massoulie and Simonian (1999); Choe and Shroff (1999)).

What typically does not apply is *simple network speed-up (SNP)* (Rolls et al., 2005). Suppose the call arrival process is stationary. Suppose the call arrival rate and the link speed are increased by the factor $f > 1$. Then the traffic bit-rate and the expected number of active calls are always increased by the factor f. If SNP applied, the finite sample distributions of the inter-arrival times as a time sequence, and the distributions of the delays as a time sequence would change only by the scale factor $1/f$. In other words, SNP behaves as if we filmed the original system and speeded up the film by the factor f.

SNP would apply if the packet arrivals were Poisson. VoIP arrivals, and Internet traffic arrivals generally, are not Poisson, but rather tend toward Poisson in a certain sense (Cao et al., 2002). This complex change in the arrival process changes the queueing distribution in complex ways, not simply by the factor $1/f$. It is for this reason that finding the QoS utilization, u_{qos}, for traffic engineering is fundamentally a statistical problem (Belottia et al., 2008).

To get comprehensive answers to the effect of changing engineering factors of QoS, it is necessary to run statistical multi-factor simulation experiments that recreate network transmission and queueing for one or more routers, and vary the engineering factors (Shin and Schulzrinne, 2009). The queueing properties

are themselves mathematically complex, typically intractable if the goal is to characterize the entire distributions of delay and jitter. The stochastic properties can be guided by theory, but the simulated queueing process creates another set of data that typically need to be studied empirically using statistical methods. Rolls et al. (2005) discuss this issue in detail and put forward a number of useful tools.

However, to insure the validity of simulation study of queueing, it is critical that the model used to generate the multiplexed packet traffic accurately reproduce the statistical properties of live multiplexed traffic (Fraleigh et al., 2003b). Furthermore, the modeling must be for the IP-inbound packet traffic from a gateway. If we attempt comprehensive simulation, then we must start with traffic as it is when first seen on the network. The models can then be used for simulation of any IP network architecture, wireline or wireless. If the modeling is carried out with measurements of multiplexed traffic that has undergone more than very light network processing, then the statistical properties will have been altered in ways that make their use as IP-inbound traffic invalid.

2.2. The modeling

This paper presents detailed analyses of multiplexed packet-level VoIP traffic that lead to models for the packet arrival process of the inbound IP traffic from an IP gateway. The analyses provide validated models: the statistical properties of the resulting traffic is a good approximation of the live traffic for the purpose of QoS traffic engineering.

The modeling is done to provide generation of multiplexed VoIP traffic for valid network QoS simulation as described above. Because of our approach, the simulation can be for a fully wireline network or one with wireless links. There are two statistical models. One is a mathematical, parametric model. The second is a semi-empirical model, which means that a subset of the measurements are a part of the model.

To carry out the modeling, we collected VoIP packet timestamps and headers for 48 hr in both directions on a link of the Global Crossing international network; there are 332018 calls and 1.315 billion packets. Modeling is based on analyses of the marginal distributions and time dependencies of call-level properties (semi-call arrivals, durations, bit-rates, and transmission-silence intervals), and packet-level properties (timestamp accuracy, jitter, and 20-ms packet-counts).

There are two models. Both use a mathematical parametric model for call arrival times; each arrival time becomes the time of the first packet of a semi-call. The two models then depart in modeling the semi-call packet inter-arrival times. One is a semi-empirical semi-call model, and the other a mathematical model. In the semi-empirical model, the inter-arrivals are those of a semi-call randomly sampled from 277540 semi-calls from the Global Crossing dataset. In the mathematical semi-call model, the semi-call inter-arrivals are generated by the following:

- a call duration is generated

- alternating transmission and silence interval lengths are generated by a model of the transmission-silence interval process until their sum reaches (closely) the call duration
- packet arrivals are inserted every 20 ms into each transmission interval beginning with the start time of the interval
- silence packets are inserted every 2 sec into each silence interval beginning with the start time of the interval.

Then for both models, the packets of the semi-calls are superimposed to form the multiplexed VoIP traffic, just as packets of the semi-calls are multiplexed on the live network. Different multiplexed traffic bit-rates are achieved by different semi-call arrival rates.

This modeling at the semi-call level is far simpler than directly modeling the multiplexed packet traffic arrivals, which would require models that change with the expected semi-call rate. Also, the semi-call modeling allows extensibility to different VoIP implementation factors from those of the measured traffic. For example, we could investigate the effect of silence suppression by eliminating the silence intervals in the semi-call models, using only the durations in the modeling.

One critical task in the semi-call modeling is establishing the validity of certain assumptions about independence and identical distribution of the modeled semi-call random variables for both the semi-empirical model and the mathematical model. These variables model the semi-call durations and the lengths of the alternating transmission-silence intervals.

Another critical task, for reasons discussed above, is to establish that our measured traffic has, to an excellent approximation, the statistical properties of gateway IP-inbound arrivals. Our measured traffic begins as IP-inbound traffic, but does undergo some router queueing before arriving at our monitored link. We demonstrate that the effect of the network processing on the arrival times is negligible.

Setting the goal of the analysis as QoS study affects the modeling results. The impact is aptly characterized by Box (1976): "All models are wrong, but some are useful." Statistical effects in the data that are insignificant to QoS issues do not need to be modeled. In particular, it is not necessary to account for unique properties of a subset of semi-calls that have a negligible packet rate. Because we have so much data, we can clearly see that no assumption in our modeling is true. Model selection is based on judgments about whether obvious deviations are likely to have negligible impact on delay and jitter.

Multiplexed best-effort Internet packet traffic arrivals, discussed in Section 1, are a long-range dependent point process (Leland et al., 1994). Inter-arrivals have an autocorrelation function that has a polynomial decay as the lag increases, which means that dependencies continue for long periods. The outcome is much larger queueing delays than other arrival processes with the same arrival rate, such as Poisson (Erramilli et al., 1996). The long delays are a major factor in QoS for best-effort traffic and have had a substantial impact on the engineering of the Internet (Paxson and Floyd, 1995). Here, after establishing the

models, we investigate the time dependence of multiplexed VoIP traffic, both the modeled traffic and the measured traffic. This provides additional model checking, but also gives qualitative insight into VoIP queueing processes.

2.3. Contents of the sections

Section 3 describes the VoIP traffic collection, data processing, and certain bit-rate properties. The collected data are a large, complex dataset. Section 4 describes our approach, methods, and computational environments for the analysis of large datasets, which enables comprehensive, detailed analysis of the data. Section 5 is an overview of previous work. Section 6 discusses two bit-rates of importance for the modeling, the multiplexed traffic bit-rate and the cumulative bit-rate. Section 7 demonstrates that the Global Crossing data are an excellent approximation of IP-inbound traffic. Section 8 sets the stage for the two models developed in the paper: the semi-empirical model and the mathematical parametric model. The basic structure of the models is described. Section 9 begins the study of the semi-call properties with an analysis and modeling of the semi-call arrival times. Section 10 describes the analysis and modeling of the semi-call durations. Section 11 analyzes semi-call transmission and silence intervals and develops models for them. Section 12 analyzes semi-call bit-rate; even though this variable is not modeled directly, it provides important added model checking. Section 13 is an analysis of the time dependence of VoIP multiplexed traffic, both the observed traffic and traffic generated from the semi-empirical model. Section 14 presents and discusses the final model assumptions.

3. Traffic collection and processing

3.1. Collection

We collected the VoIP packet timestamps and headers of multiplexed calls on a link of the Global Crossing (GBLX) network in Newark, New Jersey. The principal application in the traffic is voice calls, but the traffic also contains other applications such as faxes, point-of-sale terminals (e.g., credit card processing), and computer connections to the Internet. Nevertheless we use the term "call" for each connection.

On the GBLX network the caller signal and the callee signal both originate in the Public Switched Telephone Network (PSTN), the traditional telephone network that developed starting in the 19th century. Each signal arrives at a gateway where it is converted to VoIP packets. The signal is sampled at a rate of 64 kilobits/sec, the accumulation interval is 20 ms, and silence suppression is employed. For the caller, packets emerging from the gateway are transmitted over an IP link to a first IP router, travel across the GBLX IP network, and exit to the PSTN through another PSTN-IP gateway that converts the packets back to a PSTN signal. The signal continues on to the callee over the PSTN. The signal from the callee has a similar path but in the other direction, entering the

FIG 1. *Global Crossing location in Newark New Jersey. A monitor with an Endace card collected packet headers and timestamps of both directions of calls using a link between an IP-PSTN gateway and an IP network edge router.*

IP network at the second gateway used by caller, and leaving the IP network through the first gateway to travel over the PSTN to the caller.

Collection was carried out on a 100 megabit/sec link between a gateway in Newark, New Jersey and an edge router of the GBLX IP network. Figure 1 shows the architecture. The data are packet headers and timestamps for both directions of each call using the link — the caller-to-callee semi-call and callee-to-caller semi-call. The caller semi-call conversion at Newark is either PSTN-to-IP or IP-to-PSTN, and the callee semi-call is the other. Data were collected by a 100 megabit/sec Endace DAG 3.5E card installed in a workstation. The hardware timestamps have been reported to be accurate to 100 ns (Arlos and Fiedler, 2003). The header fields used in our analyses are the source and destination IP addresses and port numbers, and a flag that indicates whether a packet has transmission information or is a silence-suppression control packet. We will describe time of day by 24-hour notation, denoting the beginning of a day by 00:00:00 and the beginning of the last second of a day by 23:59:59. Collection began on 12:05:00 on calendar day 1 and lasted 47 hr 55 min 30 sec through 12:00:30 on calendar day 3. All three calendar days were workdays.

We also obtained data for the first 23 hr 55 min of the collection period from a PSTN call-detail database. The variables collected are call start and end times as seen by the PSTN, whether the caller was PSTN-to-IP or IP-to-PSTN on the Newark link, and whether the call was an attempt or a connected call that was answered.

3.2. Processing

Analysis of different statistical properties used different subsets of the data, created by successive deletion of calls. There are 5 such subsets described next whose names are the following: *full*, *processed*, *arrival*, *complete-call*, and *detail-augmented*. Each data subset of the list is a subset of the preceding.

The *full data* are the packet header fields and timestamps of all calls. There are 1.315 billion packets from 332018 calls. A call has packets in each direction, caller-to-callee and callee-to-caller. Each direction is a semi-call, so there are 664036 semi-calls. A silence suppression algorithm is used by GBLX for VoIP; each semi-call has transmission (on) intervals alternating with silence (off) intervals. Start and end times of the on and off intervals are calculated and become a

fundamental part of the data. During silence periods, keep-alive control packets are sent every 2 sec.

The *processed data* eliminate calls that are small, and calls with large gaps. A large fraction of the calls are extremely small: 20 or fewer callee packets and 20 or fewer caller packets, all lasting less than 400 ms. They represent calls that ended almost as soon as setup began either through communication problems or caller actions. Very small calls and their frequency has been discussed in detail by Bolotin (1994) who writes that their number on the PSTN is "radically larger" than the number of calls lasting more than a few sec. In addition, there are 19 gap calls that have very large time gaps with no packets. Packets for a single call are determined by a 4-tuple: the source and destination IP addresses and port numbers of the two participating gateways. It is possible but improbable to have two calls with the same 4-tuple. The small and gap calls add a negligible amount to the traffic bit-rate in each of the 1150 5-min inbound and outbound time intervals of the collection period (not including the last 30 sec). In only 3 intervals did they contribute more than 1% of the traffic, and in only 8 intervals did it exceed 0.7%. We deleted the extremely small and the gap calls because their bit-rate is inconsequential to network QoS, and as we will see in Section 9, they add needless complexity to the call arrival process. The *processed data* consist of the 144185 remaining calls.

The *arrival data* are the calls of the *processed data* that arrived during the measurement period. Eliminated are calls in progress at the beginning of the measurement period. A call arrival is measured by the arrival time of the first packet of the call from either direction. During the measurement period of 2875 min = 172500 sec there are 144046 arrivals, which is a call-rate of 0.835 calls/sec and an inter-arrival time of 1.20 sec.

The *complete-call data* are the calls of the *arrival data* that finish during the measurement period with an estimated probability of 0.9999. These are used to study call-level statistical properties such as call duration and call bit-rates for callers and callees. Call duration is measured by the time of the last packet of a call from either direction minus the call arrival time. With just the packet traces, we cannot determine precisely the end of a call. However, the call-detail records have durations for calls ending during the first 23 hr 55 min of collection. The 0.9999-quantile of these durations is 9111 sec = 2.53 hr. From this we take any call of the *arrival data* to be in the *complete-call data* if its arrival time is more than 2.53 hr before the end of the measurement period of 47 hr 55 min. The number of semi-calls is 277540, and we take the duration of each to be the duration of the call of which it is a part.

The *detail-augmented data* are the 78050 calls of the *complete-call data* for which we have call-detail information. The additional information allows us to determine which semi-call of a call is caller-to-callee and which is callee-to-caller, and whether the call is an attempt or a connect.

4. Divide and recombine for large datasets

Our collected data, which consist of timestamps and headers for 1.315 billion packets, constitute a large, complex dataset. The challenge in analyzing such datasets is to maintain the comprehensiveness of analysis that we can achieve with small datasets, ensuring that important information in the data is not missed. Recent work addressing this goal has resulted in moderate progress, making feasible comprehensive analysis of many datasets that would have been infeasible even a short time ago (Guha et al., 2009). The general approach begins by partitioning the dataset into small subsets in one or more ways, and applying numeric methods or visualization methods to each of the subsets of a sample, analyzing some in great detail. The sampling frame can vary from one analysis method to the next. It is common to have numeric methods applied to more subsets than visualization methods. In some cases, sampling can be exhaustive, all subsets, particularly for numeric methods.

Two non-exhaustive sampling methods are representative and regional. In representative sampling, survey variables are defined that measure properties of the subsets. A subset sampling frame is chosen to encompass the multidimensional space of the survey variables in a uniform way by some definition. In regional sampling, samples have values of the survey variables lying in a region of the multidimensional space that is important to the analysis. For example, in Section 11 we present the analysis of a representative sample of 2000 semi-calls.

Partitioning can be carried out in many different ways. Often, we start with a *core partition* that arises naturally from the structure of the raw data. This is a soft concept but useful nevertheless. The subsets of the core are often further partitioned or combined by variables other than those that define the core. For the VoIP data there is one very natural core partition that we used extensively: a breakup into semi-calls. As we proceed through the analyses in coming sections here, the samples of the subsets analyzed vary.

Partitioning leads to embarrassingly parallel computation. Large amounts of time can be saved by distributed computing environments. One is RHIPE (Guha, S.), a recent merging of the R interactive environment for data analysis (`www.R-project.org`) and the Hadoop distributed file system and compute engine (hadoop.apache.org). It allows work to proceed fully within R, but to distribute computation across many processors in a cluster, which has vastly reduced run time in many of our applications.

Nothing serves comprehensive analysis better than data visualization. This principle has been accepted and practiced for decades (Daniel and Wood, 1971; Anscombe, 1973). For a large, complex dataset this requires making a large number of displays many of which can have a large number of pages and many panels per page. The total number of pages might be measured in thousands or tens of thousands in this *visualization database*, or *VDB* (Guha et al., 2009). For the analysis of the 2000 semi-calls of Section 11, we made a number of displays with 1000 pages each. Our extraordinary human visual system can readily process very large displays by methods of viewing not utilized for the single-page display. One is rapid scanning. It is possible to click at a fast rate

through the pages of a document with the same display method replicated across the panels and pages, one subset per panel. But for this to succeed, it must involve the assessment of a gestalt: a pattern that forms effortlessly without attentive search of basic elements of the display. The pattern "hits you between the eyes." When there are two or more gestalts to assess, it is best to scan through pages assessing one at a time. Attempting simultaneous assessment slows down the process remarkably because a cognitive shift of assessment is time consuming. Using these concepts in the VoIP analysis we were able to design the 1000-page displays so that rapid scanning approaching animation was achieved.

5. Previous work and the approach of this article

5.1. Best-effort vs. VoIP

The Internet research literature contains many analyses of packet traces of multiplexed traffic on operational links carrying the full spectrum of Internet applications, which in the beginning was mostly data traffic (Leland et al., 1994; Paxson and Floyd, 1995; Crovella and Bestavros, 1997; Willinger et al., 1997; Riedi et al., 1999; Fraleigh et al., 2003a; Cao et al., 2002, 2001). Analyses of such live traffic are vital because they convey the realities of traffic; the properties of the traffic are very dependent on (1) the mix of applications and (2) the properties of the transferred information. Laboratory testing of traffic can be very informative, but study of live traffic is essential to building validated traffic models, which are a basis for not just network engineering, but the development of the technology of the Internet. The development of models based on early studies of traffic was a major factor in the development of Internet technology (de Pereira et al., 2002).

However for VoIP, there have been very few studies in the literature that report on the analysis of live VoIP packet traces on operational links carrying a cross-section of the voice and machine-to-machine applications used in VoIP. Such studies are needed if validated models for VoIP engineering study are to be developed. Birke et al. (2007) appear to have been the first to publish such a study: "Traffic monitoring and characterization have always been seen as a key methodology to understand telecommunication technology and operation, and the complexity of the Internet has attracted many researchers to face traffic measurements since the pioneering times. Data traffic has hogged the majority of this effort, while the attention toward VoIP traffic measurements only recently increased. ... In this paper, we present the first extended set of measurement results collected via passive monitoring of real-world VoIP traffic."

5.2. Analyses of VoIP traffic from large operational networks

Birke et al. (2007) and Ciullo et al. (2008) report on packet traces collected on FastWeb, a network of an Italian ISP that provides end-to-end VoIP for customers. The VoIP service is carried on links with data applications such as

HTTP, and with television transmissions. The VoIP capture rate is 64 kilobits/sec and the accumulation interval is 20 ms. There is no priority queueing for VoIP and no silence suppression. Packet traces were collected in both directions on two links. Collection was carried out for approximately 3 weeks on a link with a two-way average load of 100 megabits/sec and peaks up to 380 megabits/sec and for approximately two weeks on a link with a two-way average load of 500 megabits/sec and peaks up to 900 megabits/sec. The VoIP traffic is a small fraction of the total traffic and appears, given the available information, to have multiplexed bit-rates ranging from very small to about 9 megabits/sec, quite similar to the bit-rates of our VoIP packet traces.

Zhang et al. (2007) collected packet traces of aggregate traffic in both directions on 3 links within 3 provinces of China that contain VoIP traffic. They report the details of their analyses of one of the links, which is part of a metropolitan area network of China Telecom, a large ISP. On this link they collected 525×10^9 packets in two directions over a period of about 25 days. A classification algorithm divides the traffic into 7 categories, one of which is VoIP. About 10% of the packets are VoIP so the average VoIP bit-rate is about 0.6 megabits/sec. No information is given about the VoIP implementation. The principal goal is to investigate the long-range dependence of the traffic through estimation of the Hurst parameter. Their conclusion is that the aggregated VoIP traffic is not long-range dependent based on the observation that no estimate was above 0.17.

Dang et al. (2004) obtained call detail records from a corporate VoIP network with nearly 800 VoIP phones from November 2002 to July 2003. The call arrival process was fitted by a Poisson process and the call duration distribution by a generalized Pareto distribution with parameter values indicating finite variances. The silence and transmission durations were fitted by a generalized Pareto distribution as well.

5.3. Direct measurement of QoS criteria

A number of research projects report on direct QoS study of operational networks by sending probe packets across the network that imitate certain properties of VoIP packets. QoS metrics are collected from the probes. Tobagi et al. (2002); Markopoulou et al. (2003) sent constant probes, one set every 100 ms and a second set every 10 ms, over 43 backbone paths for 17 days. Packet loss rate and end-to-end delay were measured from the probes, but not jitter. The results were used to evaluate whether at that time, the backbone networks were ready for VoIP deployment.

In this last study, probe packets did not have interarrivals that are the same as that of common implementations of the VoIP protocol. One can use common accumulation intervals and then add jitter measurements to the QoS metrics. Avaya ExpertNet has a general probing tool that distributes a number of linux machines over strategic points of a network. The machines inject probe packets with different accumulation intervals. One-way delay, jitter and packet loss rate are gathered and used to evaluate whether a network is ready for VoIP.

In one experiment, Agrawal et al. (2006, 2007) set up two host machines in Bangalore, India, and in Murray Hill, New Jersey. The two machines established connections between them using the SIP protocol, and exchanged RTP probe packets with the rate of 50 VoIP calls. During part of the experimental period, background traffic was also sent between the hosts. One-way and round-trip delay, jitter, loss, call set-up time, and call tear-down time were measured from the synthetic traffic. The results showed that VoIP packets mixed with background data traffic experienced service quality degradation.

Toral-Cruz and Torres-Roman (2005) generated synthetic VoIP traffic between two local area networks in Guadalajara, Mexico. The probe traffic had different packet accumulation intervals and imitated silence suppression. One-way delay, jitter, and loss were measured from the probe packets. Large accumulation intervals and silence suppression were recommended for efficient utilization of network resources.

While probe methods can give valuable information about the performance of VoIP on a live network, it is difficult to carry out controlled experiments whose designs systematically study the effect of the engineering factors on QoS for which there are interactions. It can be far more effective to run simulations of the queueing processes, either a single queue or a series of queues. An example is Shin and Schulzrinne (2009), who measured VoIP traffic in a wireless IEEE 802.11 testbed and drew many interesting conclusions just from the empirical study. They went on to run simulations, which provided substantial information about important engineering factors for wireless.

5.4. Call-level properties

A non-homogeneous Poisson arrival process and independent exponential duration were established in the early days of telephony as a model for telephone voice calls (Babu and Hayes, 2004; Brown et al., 2005). But current VoIP traffic contains applications other than voice that can change these properties since machine-to-machine transactions and other factors in modern telephone networks can potentially create bursts of arrivals (Bolotin, 1994), as well as call durations with longer tails than the exponential (Bolotin, 1994; Dang et al., 2004).

The study of talk-spurts and silence-gaps in two-way conversations has a very long history in connection with the design of the PSTN, going back to at least 70 years (Norwine and Murphy, 1938; Brady, 1969; Lee and Un, 1986). This work provides useful background, but for the purpose of modeling VoIP, studies in the context of current VoIP technology are far more relevant. There are two reasons. First, current codecs have configurable parameters that can affect onsets of transmission and silence (Jiang and Schulzrinne, 2000). Second, as the terminology "talk-spurts" and "silence-gaps" indicates, the traditional work was for speech, but today VoIP carries machine-to-machine connections that have the potential to change the statistical properties. One approach to the study of on and off intervals is to record a semi-call, play it through a VoIP codec, and analyze the statistical properties of the resulting on-off duration process. This has

been done in a number of papers with interesting results (Jiang and Schulzrinne, 2000; Casilari et al., 2002; Biernacki, 2006). This approach does miss variability that arises from the many different types of calls that occur in the live traffic on an operational VoIP network, but the results of these investigations are very helpful to the overall validation of our modeling.

5.5. Relationship of past work and this article

The past work in the statistical properties of VoIP traffic collectively provides useful information for our goal of building validated models of the multiplexed packet-level traffic for QoS study. The earliest papers address a much simpler call population, which has become more complex with time due to an increase in the number of applications of telephone communications, and to associated technologies such as automated repeat dialing. No single past work, nor all collectively, have the range of investigations of statistical properties or data to support validated complete modeling.

The range of the statistical investigations and the VoIP traffic measurements presented here support complete modeling and validation. Because the measured traffic underwent little network processing, the models apply to IP-inbound traffic exiting a gateway that converts call signals to IP packets.

The models have a number of new statistical components since standard model components cannot describe the complexity of the live multiplexed VoIP packet traffic. For example, simple, standard distributions such as Weibull, gamma, and log normal are inadequate for many aspects of the modeling; the data require more complex models with random-effects and piece-wise linear components. Still, observed traffic properties of past work described in coming sections are consistent with our results when increased complexity is taken into consideration; this both supports the modeling and suggests that it can be extended to links other than the one we measured.

6. Multiplexed traffic, cumulative, and semi-call bit-rates

There are three types of bit-rate of importance to VoIP traffic engineering, traffic analyses, and traffic model building. In this section we introduce all, and discuss the first two. In Section 12, we discuss the third.

The first is the bit-rate of the VoIP multiplexed packet traffic on a link. The multiplexed traffic has a packet arrival rate. To measure it, we can average over any interval of time to get a packet-rate in packets/sec over the interval. Each packet is 1600 bits, so the bit-rate in bits/sec is the packet rate times 1600 bits/packet. As discussed in Section 1, the major goal of traffic engineering is to determine the QoS utilization, u_{qos}: the largest multiplexed traffic bit-rate that satisfies QoS delay and jitter criteria for the link, divided by the link speed. In addition, certain of the statistical analyses must take account of the changing rate.

The second bit-rate is the cumulative sum of the bits of all semi-calls less than or equal to a duration, from the smallest duration to the largest. This gives us a sense of the durations that are important for traffic engineering. Semi-calls with short durations that carry a very small fraction of the traffic are not critical for traffic engineering. This guides the modeling; deviations from model assumptions for such short semi-calls are less important than for the longer semi-calls.

The third is the bit-rates of the two semi-calls of each call: the numbers of bits of each semi-call divided by the duration. The purpose of silence suppression is to reduce the semi-call bit-rate. This then reduces the collective multiplexed packet bit-rate/semi-call, resulting in greater efficiency. We do not model this semi-call bit-rate directly, but its analysis does constitute an important part of the model validation.

6.1. Multiplexed traffic bit-rate

Bit-rates of the multiplexed traffic of the *processed data* in 1150 5-min intervals of the collection period, 575 inbound intervals and 575 outbound, are displayed in Figure 2. There are strong diurnal patterns in the bit-rate resulting from diurnal patterns in the call-arrival rate. The bit-rates range from about 2 to 11 megabits/sec (1250 to 6875 packets/sec). In Section 12 we show that the average bit-rate of a semi-call is 0.04465 megabits/sec. Thus for 11 megabits/sec, the expected number of active calls is $11/0.04465 = 246$ calls, and for 2 megabits/sec, the number is 45 calls.

6.2. Cumulative traffic bit-rate

It is instructive to study cumulative sums of semi-call bits as a function of duration for the *complete-call data*. We order the semi-calls by the durations. At each value of duration we compute the sum of the bits of all calls whose durations are less than or equal to the selected duration. We divide the cumulative sums by the total bits for all semi-calls. Figure 3 is a graph of the fractional cumulative bit sum against log base 2 of duration. The vertical lines show 5 quantiles of duration — 0.05, 0.25, 0.5, 0.75, and 0.95. The plot is quite striking. We can see that about 50% of the calls, those above the median of 39.35 sec, account for 97.15% of the bits; and those above 128 sec account for only 16.50% of the calls and 87.90% of the bits. Because our focus is on QoS, we could have deleted the smallest 50% of the calls as we did for the extremely small calls. We did not do so because the information about them provides subsidiary information about VoIP call properties of general interest. However, it does allow us to tolerate more readily deviations of the data from model assumptions for calls with durations less than 30-128 sec. This forms an important part of our assessment of the i-id properties of the semi-calls.

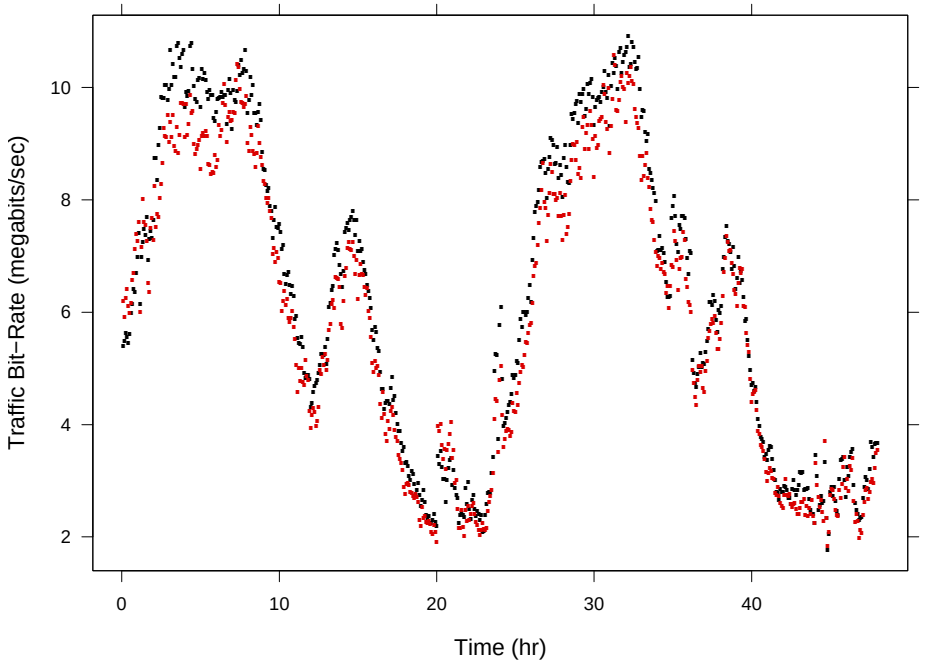

FIG 2. *Multiplexed packet traffic bit-rates for 1150 5-min intervals. IP-to-PSTN (black) and PSTN-to-IP (red).*

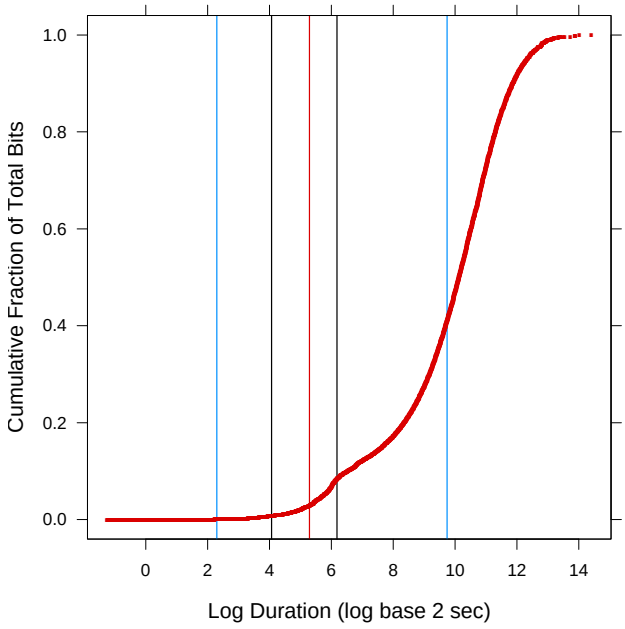

FIG 3. *Cumulative fraction of bits by duration vs. log base 2 duration.*

7. Verification of semi-call measurements as IP-inbound traffic

7.1. General issues

The overall goal of this work is packet-level modeling of the VoIP traffic emerging from a source that packetizes many in-progress semi-call signals for forwarding to a first IP link. The packets coming to an IP link from the gateway, and the times when the packets have finished their conversion, are the *IP-inbound* traffic. This offered load has not been altered by the many delay mechanisms that can occur during transmission across the network. To build a model that accurately reflects the IP-inbound traffic, it is important that the network processing for the measured traffic has resulted in negligible change from the traffic in its offered state.

Our measured link connects a gateway to an IP edge router of the GBLX network. The PSTN-to-IP semi-call packets coming from the gateway onto the IP link are subjected to only one queue, that for the link itself. The IP-to-PSTN semi-call packets coming from the edge router onto the link arrive from worldwide sources, and many semi-calls have experienced numerous network mechanisms. However, the network traffic engineering on the GBLX network at the time of measurement provided both priority queueing for VoIP packets and small VoIP traffic loads on links. For example, we have seen in Figure 2 that the maximum 5-min traffic rate in each direction of the 100 megabits/sec measured link is about 11 megabits/sec, a utilization of only 11%. This leads to the hypothesis that the measured load is close to the offered. In the next two sub-sections, methods of statistical analysis provide a check of this hypothesis.

7.2. Time stamp accuracy

The first critical matter is the accuracy of a packet timestamp written by the collection hardware that records when the first bit of the packet is put on the link. For the measurements to reflect the offered load, the timestamps must be accurate. Hardware lab tests for the Endace card used in the collection show timestamp accuracy to be $\pm 0.1\mu$s (Arlos and Fiedler, 2003), far better than our accuracy requirements. However, we can check accuracy to a level needed for our QoS study: $\pm 0.5\mu$s. A VoIP packet is 200 bytes plus another 38 bytes of Ethernet encapsulation, so the number of bits is 1904. The link speed of the GBLX measured link is 100 megabits/sec, so the transmission, or service, time is 19.04 μs. Suppose while packet k is in service, packet $k+1$ arrives. The arrival time of k is the time when k begins transmission. Packet $k+1$ begins transmission at the moment packet k has finished transmission, so its arrival time is 19.04 μs after that of packet k. The inter-arrival time between k and $k+1$ is 19.04 μs, and no true inter-arrival time can be less than this value.

To verify the timestamp accuracy using our data we rounded the arrival times of all packets in the *full data* to the nearest μs, our accuracy requirement, and computed the inter-arrival times. The minimum is 19 μs, which verifies the requisite accuracy.

7.3. Jitter

A powerful way to study the validity of using the measured traffic as the offered load is to analyze jitter. The packets of a semi-call are divided into transmission intervals separated by silence intervals. During a transmission interval there is a sequence of arrivals whose initial inter-arrivals at the time of packetization are intended to be 20 ms with a very small allowable error. The measured jitter of semi-call packets on a link are the absolute values of the measured inter-arrivals of the packets inside the transmission intervals minus 20 ms. Network processing in the form of queueing delay creates jitter because packets of a semi-call are delayed by varying amounts.

As discussed above, the PSTN-to-IP traffic coming out of the local GBLX gateway onto the measured link encounters much less network processing than the IP-to-PSTN traffic. For this reason one expects, a priori, to see more jitter for the IP-to-PSTN semi-call traffic than the PSTN-to-IP. In addition, one expects the jitter to increase with the multiplexed traffic bit-rate on the measured link because an increase creates more delay, and therefore jitter, in the link queues.

We studied jitter for the 288370 semi-calls of the *processed data*. We broke the measurement period into 575 5-min intervals as in Figure 2, which graphs 1150 5-min multiplexed traffic bit-rates against time, 575 for the IP-to-PSTN semi-calls and 575 for the PSTN-to-IP semi-calls. Each semi-call has a collection of values of jitter. Each 5-min interval is assigned the values of jitter for packet pairs whose first packet arrives in the interval. The 5-min intervals are broken up into six groups according to the 1150 multiplexed traffic bit-rates. The six groups are formed from six equally-spaced intervals from the minimum of the 1150 5-min multiplexed bit-rates to the maximum. Then 12 distributions of jitter are formed, grouped by the 12 combinations of values of two variables: the two values of PSTN-to-IP and IP-to-PSTN, and the six groups of intervals.

Figure 4 graphs six quantiles of the 12 jitter distributions against the mid-points of the six bit-rate intervals, with "+" for the PSTN-to-IP distributions and "o" for the IP-to-PSTN. The frequencies of the quantiles are shown in the strip labels. Figure 4 shows that overall, the jitter is small, with all but the very highest quantile less than 0.25 ms. (The end-to-end standard is 30 ms.) Each jitter quantile tends to increase with the multiplexed traffic bit-rate, as expected. The quantiles for IP-to-PSTN are either somewhat smaller or about the same as those for PSTN-to-IP except for quantile 0.999 where the IP-to-PSTN values are larger. A priori, as explained above, one expects a consistently higher distribution of jitter for IP-to-PSTN. We addressed this anomaly, and found an explanation, described next.

We studied jitter for each semi-call as a sequence in time order, and discovered a periodic component with a period of 1 cycle per 9 jitter values, which is 180 ms. Figure 5 graphs jitter for a portion of one transmission interval of one call; the cycle is very apparent. The statistical properties of the cycle were studied over sufficiently long transmission intervals to get good estimates. We selected calls both of whose two semi-calls had at least 25 transmission intervals, each with at least 180 jitter measurements, which is at least 20 cycles of the periodic

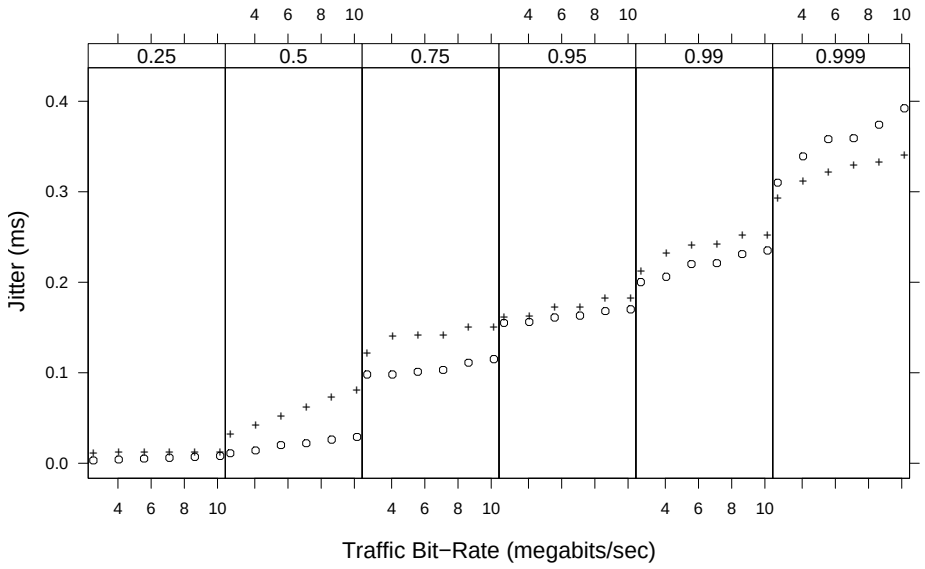

FIG 4. *Six quantiles of jitter are graphed against the midpoints of six intervals of traffic bit-rate for IP-to-PSTN (∘) and PSTN-to-IP (+) semi-calls.*

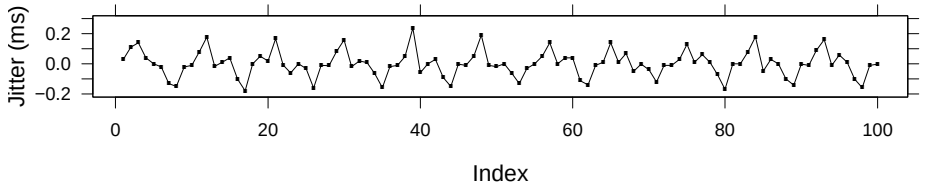

FIG 5. *Jitter for a portion of one transmission interval of one semi-call is graphed against the index (time order).*

component. The result was 7278 calls, with 532968 transmission intervals for the PSTN-to-IP semi-calls and 546723 transmission intervals for the IP-to-PSTN semi-calls.

Analysis began by fitting a regression model to the jitter of each semi-call as a waveform function of the index, the order of occurrence k of the jitter values. The 9 explanatory variables of the waveform are cosinusoids: $\cos(2\pi kj/9), j = 0, \ldots, 4$, and $\sin(2\pi kj/9), j = 1, \ldots, 4$. Initial fitting to each transmission interval of a semi-call independently together with regression diagnostics within and across intervals, revealed that all terms contribute to the fitted waveform, the phase of the cycle is not maintained from one transmission interval to the next for a single semi-call, and the errors have a symmetric and long-tailed error distribution. Because of the error distribution, each transmission interval regression was carried out using the bisquare robust estimator with a constant of 6 (Andrews et al., 1972).

TABLE 1
0.05, 0.5, and 0.95 quantiles of the estimated jitter-cycle magnitude for 6 categories: IP-to-PSTN by PSTN-to-IP by three intervals of sample sizes

Sample Size	Gateway In-Out	Number of Intervals	0.05 (ms)	0.5 (ms)	0.95(ms)
All	IP-to-PSTN	546723	0.184	0.275	0.315
≥ 825	IP-to-PSTN	25836	0.185	0.275	0.308
≤ 185	IP-to-PSTN	25304	0.183	0.273	0.319
All	PSTN-to-IP	532968	0.262	0.298	0.319
≥ 825	PSTN-to-IP	21196	0.270	0.297	0.311
≤ 185	PSTN-to-IP	24558	0.259	0.298	0.323

We cannot readily draw statistical inferences about the variability of the amplitude of the waveform fit for each transmission interval individually because the error terms have complex correlations due to the jitter time process. However, we can assess variability by studying amplitude estimates as the sample size changes, that is, the number of jitter values in an interval. Table 1 shows this. For both IP-to-PSTN and PSTN-to-IP there are three collections of transmission intervals: all intervals, intervals with 825 or more observations, and intervals with 185 or fewer observations. For the resulting 6 categories, the last four columns show the number of intervals and three quantiles of the amplitudes: 0.05-quantile, median = 0.5-quantile, and 0.95-quantile.

For IP-to-PSTN and PSTN-to-IP separately, the quantiles do not depend on sample size. This suggests that sampling variability is small and different transmission intervals have different amplitudes. There are differences between the IP-to-PSTN and PSTN-to-IP quantiles, but they are relatively small except for the 0.05-quantile. However, they are close enough to suggest a common mechanism producing the jitter cycling.

The only plausible explanation for the cycle is the gateway packetization algorithm. All gateway devices on the GBLX network at the time of measurement were the same model from one company, and presumably configuration was the same; this explains the similar magnitudes in both the IP-to-PSTN and PSTN-to-IP directions. The offered load is itself not pristine, but rather just very-nearly-pristine.

The fitted waveforms of the regressions are the variation due to the cycle in the offered load, and the residuals are the variation due to network induced jitter. Figure 6 is the same display method as in Figure 4, but uses the residuals of the regressions. The network-induced jitter distribution is substantially smaller than the total jitter variation of Figure 4 due to the removal of the offered load cycle, and the anomaly of greater PSTN-to-IP jitter is no longer present; IP-to-PSTN and PSTN-to-IP jitter are now very close for all but the 0.999-quantile where the IP-to-PSTN is greater.

These results and the high timestamp accuracy show that it is entirely reasonable to consider our measured multiplexed packet traffic as an excellent approximation of the IP-inbound traffic.

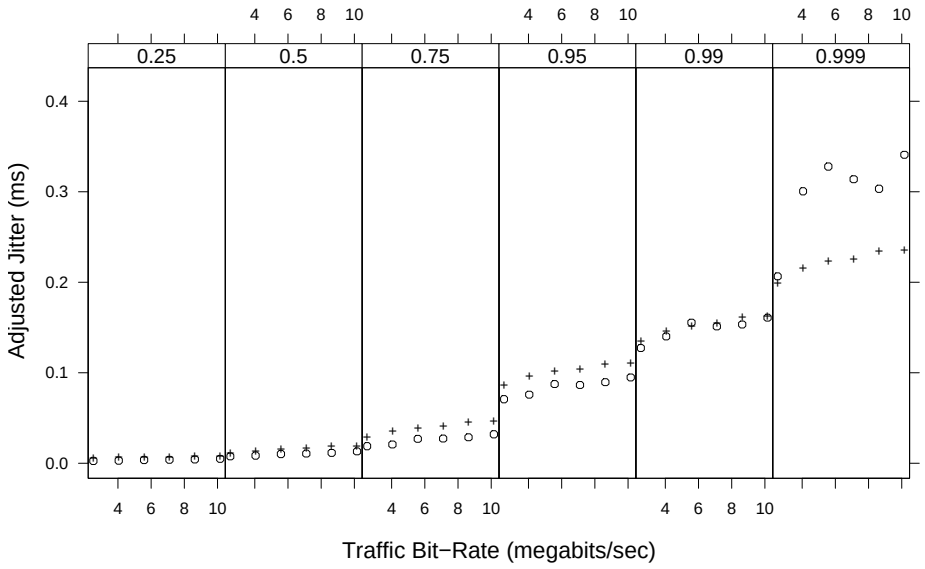

FIG 6. *Jitter Residual Quantiles.*

8. Specification of basic model structures

This section sets the stage for the two models developed in the paper: the semi-empirical model and the mathematical model. The semi-empirical model uses empirical observations as part of its modeling, and the mathematical model consists of parametric statistical specifications of mathematical components. Next, the basic model structures are described through showing how each model generates multiplexed packet traffic, and then specifying the assumption of independence and identical distribution of the models.

The semi-empirical model generates traffic in the following way.

- A non-homogeneous Poisson semi-call arrival model generates the semi-call arrival times for a sequence of generated semi-calls. In applications the rate is likely to be stable on the order of 5-15 minutes and can be modeled as a deterministic or slowly varying process. We do not specify the changing rate in our modeling since it depends on the specific network being studied. For many network studies, a simulation interval with constant rate suffices.
- Semi-calls are randomly sampled from the 277540 semi-calls of the *complete-call data*, with one semi-call assigned to each semi-call arrival.
- The generated arrival time for the first packet for the generated semi-call is the generated semi-call arrival time.
- The arrival times for the subsequent packets of the generated semi-call are the generated time of the first packet plus the empirical packet inter-arrival times of the sampled semi-call.
- The generated multiplexed packet traffic arrival times are the superposition of the packet arrival times of the generated semi-calls.

The mathematical parametric model generates traffic in the following way.

- The same Poisson semi-call arrival process for the semi-empirical model generates the times of the first packets of a sequence of generated semi-calls.
- Each semi-call duration is generated from a parametric semi-call duration distribution that is piece-wise Weibull.
- The marginal distribution of the on-off (transmission-silence) interval lengths of each generated semi-call is a square-root gamma with a bivariate parameter: log shape and log scale.
- The bivariate parameter for the transmission lengths of a generated semi-call is generated from a transmission-interval bivariate normal. The transmission lengths are generated from the square-root gamma with this generated bivariate parameter. Each semi-call has its own generated parameter.
- The bivariate parameter for the silence lengths of a generated semi-call are generated from a silence-interval bivariate normal. Silence lengths are generated from the square-root gamma with this generated bivariate parameter. Each semi-call has its own generated parameter.
- Transmission and silence lengths are generated until their sum equals, approximately, the generated duration.
- Packets for a transmission interval begin at the start of the interval, occur every 20 ms, and end when there is less than 20 ms remaining.
- Packets for a silence interval begin at the start of the interval, occur every 2 sec, and end when there is less than 2 sec remaining.
- The generated multiplexed traffic packet arrival times are the superposition of the packet arrival times of the generated semi-calls.

The following assumptions of independence and identical distribution (i-id) form parts of the two above models. For clarity, the list entries below are redundant and overlap with some of the assumptions for the two models above.

- The bivariate parameters for the transmission lengths of the semi-calls are i.i.d., the bivariate parameters for the silence intervals are i.i.d., and the bivariate silence parameters are independent of the bivariate transmission parameters.
- Durations are i.i.d. and are independent of the bivariate transmission and silence parameters.
- Given the bivariate parameter of the transmission intervals of a semi-call, the transmission lengths are i.i.d.
- Given the bivariate parameter of the silence lengths of a semi-call, the silence lengths are i.i.d.
- Given the bivariate parameters of the on-off processes of all semi-calls, there is joint independence of all transmission-silence lengths and semi-call durations.

9. Modeling the call arrival process

9.1. The data and their adjustment

A call arrival is measured by the arrival time of the first packet of the call from either semi-call. During the measurement period of 2875 min = 172500 sec there are 144046 arrivals, so the call-rate is 0.835 calls/sec, which corresponds to a mean inter-arrival time of 1.20 sec. Let a_k be the time of the k-th call arrival for positive integer k and and let $t_k = a_{k+1} - a_k$ be the k-th inter-arrival time. We partitioned the measurement period into 191 consecutive 15-min intervals and a final interval of 10 min, and analyzed each interval separately. The number of arrivals ranges from 142 to 1847 across the 192 intervals.

The call-rate is stable over 15 min for most intervals, but for a small number, there can be a smooth drift in the rate. We corrected for drift in each interval by the following method. Let $\ell_k = \log_2(t_k)$, where $\log_2$ is log base 2. ℓ_k is smoothed as a function of the index k using the loess method with a span of 2/3 and local quadratic fitting (Cleveland and Devlin, 1988). This results in fitted values $\hat{\ell}_k$ that estimate the local log rate. We then compute the residuals $r_k = \ell_k - \hat{\ell}_k$ and study their statistical properties. The equivalent degrees of freedom of the fit is 1.8, very small compared with the sample sizes, so the use of the residuals for model building is reasonable.

Weibull quantile plots of the log-scale r_k for the 192 intervals show the empirical distributions are well-approximated by the Weibull. Figure 7 shows the plot for 6 intervals. On each panel, the ith largest r_k in the interval is graphed against the log base 2 of the quantile of the unit exponential with probability $(i - 0.5)/n$. The oblique line on the plot is drawn through the upper and lower quartile points of the empirical distribution and the theoretical distribution. The horizontal lines show the 0.01, 0.05, 0.25, 0.75, 0.95 and 0.99 empirical quantiles. The 6 intervals shown in Figure 7 were chosen to summarize how well the Weibull fits across the 192 intervals. For each of the 192 plots, we computed the mean absolute deviation of the points of each plot from the oblique line on the plot, and ordered from smallest to largest to form empirical quantiles of the deviations of orders $(k - 0.5)/n$ for $k = 1$ to 192. The 6 panels of Figure 7 show intervals whose deviations are closest to the quantiles of orders 0.25, 0.40, 0.55, 0.70, 0.85, and 0.99, respectively, which are shown in the strip labels. The first quantile of order 0.25 is that for which the plotted points are very close to the oblique line for intervals with smaller deviations, but where the fit starts to have an increasing systematic departure for larger deviations. So for about 25% of the intervals, the Weibull fit is very close. As the deviation measure increases beyond the 0.25 quantile, a departure occurs in the lower tail of the distribution. Overall the departure from Weibull is minor; where it occurs on the log base 2 ms scale, the differences are quite small on the ms scale compared with the overall variation in the empirical values in ms.

Independence of the call inter-arrivals was assessed by computing and plotting autocorrelations of the r_k at lags 1 to 50 for each interval. There was no

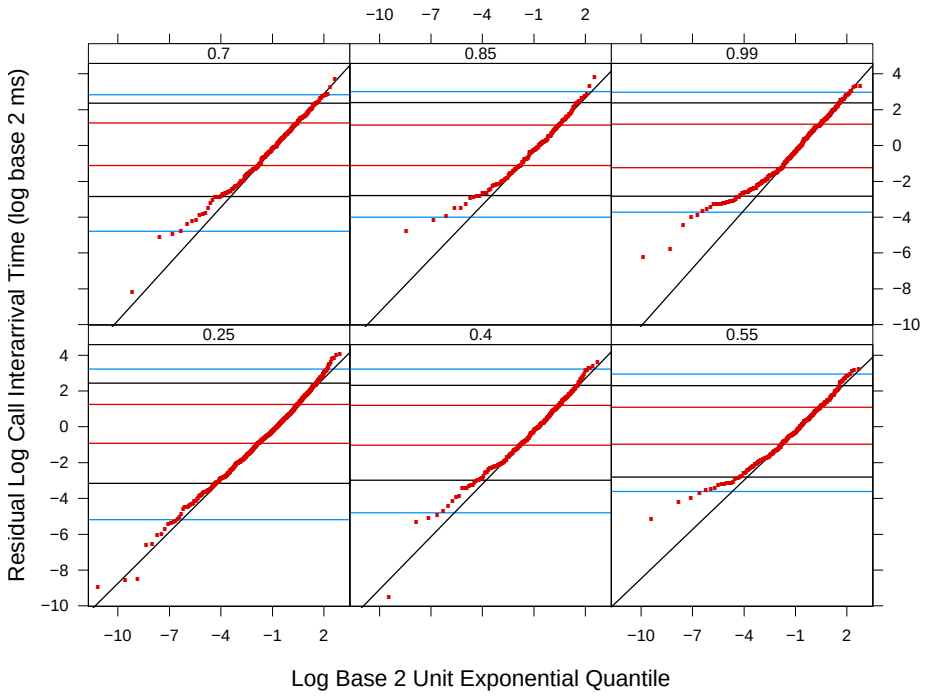

FIG 7. *Log Weibull quantile plot for 6 intervals of residual semi-call inter-arrivals.*

indication of significant correlation across the intervals. The lack of autocorrelation is a reasonable indicator of independence since the r_k, while not normal, do not depart grossly from normal.

The shape parameter of the Weibull was estimated by maximum likelihood from the values of 2^{r_k} for each interval. Figure 8 is a cumulative frequency quantile plot of the 192 estimates. The k-th largest estimate is graphed against its empirical cumulative frequency, $(k - 0.5)/192$. The horizontal lines show quantiles of frequencies 0.01, 0.05, 0.25, 0.75, 0.95, and 0.99. The results suggest a true shape slightly bigger than the shape of 1 of the exponential, but so close that the exponential is a good approximation of the true distribution.

9.2. Previous work

While the Poisson result is both simple and traditional, its background is more complex. A non-homogeneous Poisson process was established in the early days of telephony as a model for telephone voice call arrivals (Babu and Hayes, 2004; Brown et al., 2005). But as discussed in Section 3, networks carrying voice have gotten more complex due to rapid dialing and other technology advances. Our Poisson property holds for the *arrival data*, which result from deleting, among other calls, the very small calls of the *full data*. In fact, with the very small

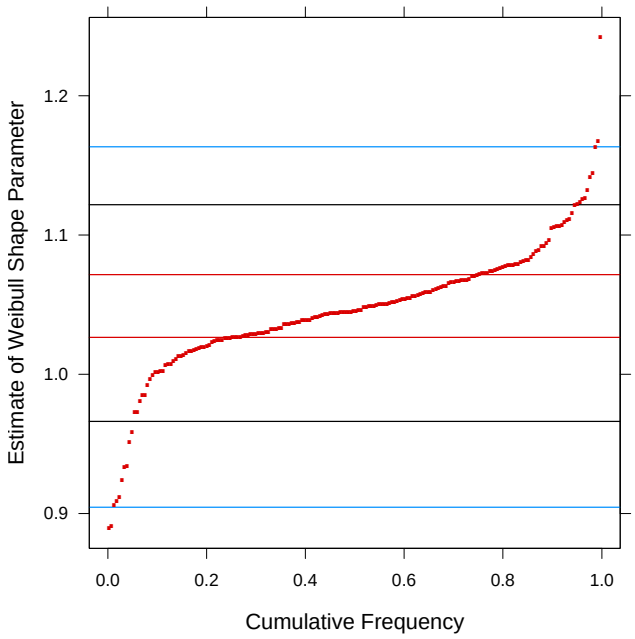

FIG 8. *Quantile plot of 192 maximum likelihood estimates of the Weibull shape parameter.*

calls present, the process is not Poisson, but rather is bursty: positively auto-correlated inter-arrivals, which has also been observed by Bolotin (1994). Thus removing the small calls, unnecessary for QoS studies because of their very small multiplexed traffic bit-rate, leads to a much simpler semi-call arrival model.

9.3. Modeling results

Our analysis has shown that the inter-arrivals adjusted for a changing rate are independent and have a marginal distribution that is well approximated by the exponential. This means that locally, the semi-call arrivals are a Poisson process, and globally are a non-homogeneous Poisson process. This then is the first step in our modeling of the multiplexed VoIP packet traffic.

10. Modeling semi-call duration

This section describes analyses that lead to a parametric model for the marginal distribution of call duration that is part of the multiplexed traffic mathematical model. The analyses also address the i.i.d. assumption for call duration that is part of the overall set of i-id assumptions about independence and identical distribution. The i-id assumptions are critical to both the mathematical model and the semi-empirical model.

10.1. Marginal distribution

Call duration is measured by the time of the last packet of a call from either direction minus the time of the call arrival, which is the time of the first packet from either direction. The durations of the two semi-calls of the call are taken to be the same as the call duration. To study the statistical properties of the durations, we analyzed the durations and start times of the 138770 calls of the *complete-call data*, as well as the 78050 calls of the *detail-augmented data* when we needed information about caller-to-callee and callee-to-caller or attempt and connect.

Let c_k be the duration of the kth call, let D be its cumulative distribution function, and let $d(c)$ be its density. We studied the marginal distribution of our measurements, c_k, using quantile plots. Figure 9 is a Pareto quantile plot. The i-th largest of the $n = 138770$ values of $\log_2(c_k)$ is the empirical quantile of the log durations of frequency $(i - 0.5)/n$. It is graphed against the quantile of the unit exponential with probability $(i-0.5)/n$. The horizontal lines show the 0.01, 0.05, 0.25, 0.75, 0.95, and 0.99 quantiles. The oblique line is drawn through the upper and lower quartile points of the empirical distribution and the theoretical distribution. If the pattern of the points were linear, the distribution would be Pareto but it clearly is not, showing a faster rate of decay in the upper tail.

Figure 10 is a Weibull quantile plot. It is the same as Figure 9 except that the horizontal scale is now the log of the exponential quantile. The pattern of the

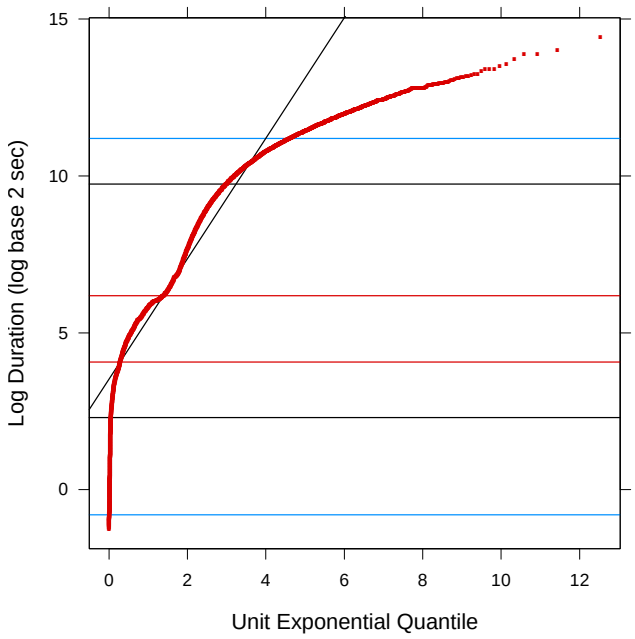

FIG 9. *Pareto quantile plot of 138770 call durations of the complete-call data.*

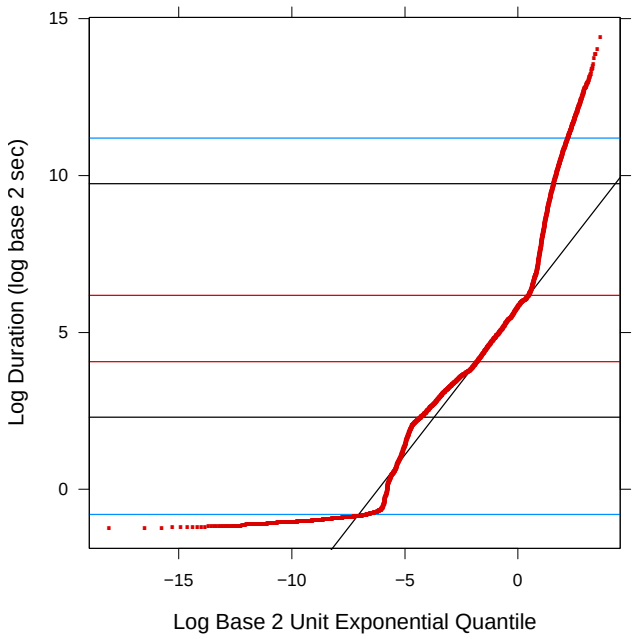

FIG 10. *Weibull quantile plot of 138770 call durations of complete-call data.*

points in Figure 10 is a piecewise linear continuous function with about 6 pieces. The tail is heavier than that of a Weibull in the sense that the slope is steeper than the Weibull distribution as defined by the pattern of the points fitted by the line through the lower and upper quartiles. It is highly likely that these different pieces result from a mixture of different types of calls. We are not able to differentiate different applications, voice and the various machine-to-machine transactions, because we do not collect the packet payload. But the 78050 calls of the *detail-augmented data* do provide information about call type: attempt or connect. There are 33361 attempts and 44689 connects; the frequency of connects is 0.573, an estimate of p.

Figure 11 has Weibull quantile plots (red dots) of the attempts (top panel) and the connects (bottom panel); most data points of the plot are obscured by a close fit (blue lines) to the distributions that will be described shortly. The connect distribution has two distinct duration regions. The lower likely has many machine-to-machine connections, and the upper is likely mostly voice calls and computer connections. Figure 12 shows nonparametric density estimates of attempts and connects using the ed method with 20-gaps smoothed by local cubic fitting and a span chosen to optimize the fits (Hafen and Cleveland, 2009). The connect and attempt distributions are complex; no simple, standard parametric family will provide an adequate fit.

We will model duration by a mixture of attempts and connects. $\log_2(c)$ is modeled separately for attempts and for connects by a piecewise linear contin-

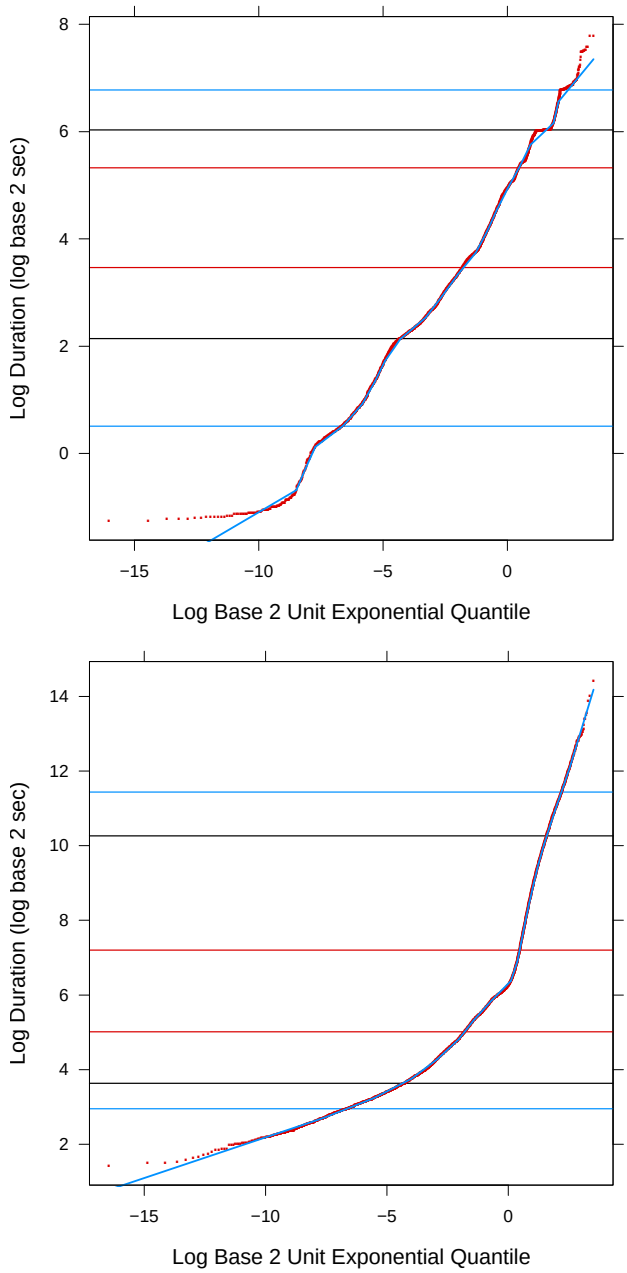

FIG 11. *Weibull quantile plots (red) of 33361 attempt call durations (top panel) and 44689 connect durations (bottom panel) from the detail-augmented data. The line segments on each panel (blue) are a piecewise linear continuous fit with 20 intervals.*

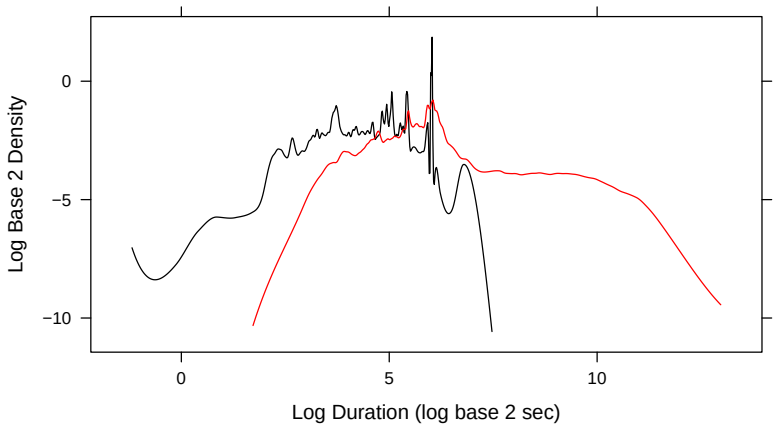

FIG 12. *Ed nonparametric density estimates of 33361 attempt durations (black curve) and 44689 connect durations (red curve) from the detail-augmented data.*

uous function of log Weibull quantiles with 18 interior line segments, and two outer rays to make the fit defined for all real values. The result is shown by the blue oblique line segments in Figure 11. Log duration is divided into 20 equally spaced intervals from the 0.001-quantile to the 0.999-quantile of $\log_2(c_k)$. Let υ_j for $j = 1, \ldots, 21$ be the endpoints. Let f_j for $j = 1, \ldots, 21$ be the fraction of $\log_2(c_k)$ less than or equal to υ_j. The f_j-quantile of the unit exponential is $-\log(1 - f_j)$ where log is the natural log. Let $\chi_j = \log_2(-\log(1 - f_j))$. Let ψ_j be the slope and ϕ_j the intercept of the line through (χ_j, υ_j) and $(\chi_{j+1}, \upsilon_{j+1})$, for $j = 1, \ldots 20$. The fitted function has the following: a line segment with slope ψ_j and intercept ϕ_j from (χ_j, υ_j) to $(\chi_{j+1}, \upsilon_{j+1})$ for $j = 2, \ldots, 19$; a ray to the right from $(\chi_{20}, \upsilon_{20})$ with slope ψ_{20} and intercept ϕ_{20}; and a ray to the left from (χ_2, υ_2) with slope ψ_1 and intercept ϕ_1.

The generation of a $\log_2$ connect duration, υ, uses the information in the left panel of Table 2. A uniform random variable f on $(0, 1)$ is generated and $\chi = \log_2(-\log(1 - f))$ is computed. Then $\upsilon = \alpha_j \chi + \beta_j$ where the slope and intercept are those in the row for which f falls in the interval of column 1. The generation of a $\log_2$ attempt duration is similar and uses the right panel of Table 2. To generate υ unconditional on connect or attempt, a Bernoulli variable is generated with probability p for the value `connect` and $1 - p$ for `attempt`, determining which log call distribution to sample. p can be altered to reflect network conditions. For the calls of the *detail-augmented data*, the fraction of connects is 0.573.

10.2. i-id assumptions

The basic assumptions of the multiplexed traffic modeling are discussed in Section 8. A critical part of the assumptions prescribe independence and identical distribution the modeled semi-call random variables for both the semi-empirical

TABLE 2
Piecewise Weibull distributions for connect and attempt calls

Connect

Quantile Interval	Intercept	Slope
(0.0000 , 0.0052)	4.3697	0.2186
(0.0052 , 0.0212)	4.6630	0.2573
(0.0212 , 0.0618)	5.0811	0.3328
(0.0618 , 0.1238)	5.7342	0.4972
(0.1238 , 0.2142)	6.0426	0.6029
(0.2142 , 0.3129)	6.4860	0.8189
(0.3129 , 0.4617)	6.3511	0.7235
(0.4617 , 0.6591)	6.3044	0.6559
(0.6591 , 0.7253)	6.1642	1.9809
(0.7253 , 0.7658)	5.7470	3.1095
(0.7658 , 0.8018)	5.6278	3.3312
(0.8018 , 0.8372)	5.7434	3.1647
(0.8372 , 0.8720)	5.9646	2.9075
(0.8720 , 0.9066)	6.3477	2.5391
(0.9066 , 0.9382)	6.6932	2.2617
(0.9382 , 0.9631)	6.8805	2.1349
(0.9631 , 0.9816)	7.3085	1.8863
(0.9816 , 0.9922)	7.3326	1.8742
(0.9922 , 0.9970)	7.0965	1.9779
(0.9970 , 1.0000)	6.7227	2.1249

Attempt

Quantile Interval	Intercept	Slope
(0.0000 , 0.0028)	1.6218	0.2721
(0.0028 , 0.0036)	8.5912	1.0935
(0.0036 , 0.0047)	8.1294	1.0366
(0.0047 , 0.0103)	2.8594	0.3546
(0.0103 , 0.0174)	3.9976	0.5273
(0.0174 , 0.0244)	5.7412	0.8264
(0.0244 , 0.0326)	6.4035	0.9504
(0.0326 , 0.0492)	4.9776	0.6605
(0.0492 , 0.0977)	3.8246	0.3928
(0.0977 , 0.1510)	4.5052	0.6001
(0.1510 , 0.2240)	4.6033	0.6377
(0.2240 , 0.3391)	4.4672	0.5690
(0.3391 , 0.4362)	4.8370	0.8597
(0.4362 , 0.5276)	4.9798	1.0375
(0.5276 , 0.6380)	4.9306	0.9190
(0.6380 , 0.7588)	4.9326	0.8304
(0.7588 , 0.8523)	4.8758	0.9422
(0.8523 , 0.9734)	5.3489	0.4366
(0.9734 , 0.9849)	2.6056	1.9131
(0.9849 , 1.0000)	5.4053	0.5596

model and the mathematical model. Next we consider the i-id assumptions for duration: durations of semi-calls appearing on one direction of a link are are independent and identically distributed.

By our definitions, the two semi-calls of a single call, caller-to-callee and callee-to-caller, have the same duration: the time between the first packet from either direction and the last packet from either direction. Our analyses show that if we change the semi-call durations to make each semi-call have a duration defined by its own packets, the resulting durations for the two semi-calls change by a negligible amount. This means, obviously, that for the *complete call data*, the assumption of identical duration distributions for the caller-to-callee semi-calls and the callee-to-caller semi-calls is an excellent approximation.

Obviously the durations of the two semi-calls of a single call are not independent because we can (by definition) predict one from the other exactly. However, both directions are almost never multiplexed on a single direction of a link. Note that in Section 8, the semi-call sampling frame does not select both semi-calls of single call.

We also investigated the assumptions of independence and identical distribution of call duration through time as a function of the start time of a call. Violations of these assumptions can come through callers starting connections with certain patterns. Callers can have a tendency to use certain applications more at one time of day and less at others, which would change distributions. It is also possible for individual callers to have a burst of applications of the same type such as sending many short faxes, which would create autocorrelation. Because the most likely cause of deviations from assumptions is caller behavior,

we study the duration assumptions using two subsets of the *detail-augmented data*. The first subset is the calls in which the caller packets travel PSTN-to-IP, and the second subset is those in which the caller packets travel IP-to-PSTN.

We checked for a changing duration distribution through time by computing quantiles of the call durations of frequencies 0.01, 0.05, 0.25, 0.75, 0.95, and 0.99 for each of the 192 15 min intervals of the *detail-augmented data*, one set of quantiles for PSTN-to-IP and one set for IP-to-PSTN. We did the same for the mean duration. We did find minor changes in the distribution with time of day, and further investigations showed that most of the change was due to a decrease in frequency of short calls from about 10 p.m. to 6 a.m., which is a time of day when multiplexed bit-rate tends to be small. The change involves just calls with small duration whose cumulative multiplexed bit-rates are small, so we ignore this effect in the modeling and assume duration to be identically distributed.

The change in distribution also induces autocorrelations in each of the two log duration sequences in each 15 minute interval — positive values at lags up to 500, but always less in absolute value than 0.018 for the PSTN-to-IP direction and 0.06 in the IP-to-PSTN direction. The different magnitudes reflect the different sample sizes of the two link directions. A loess fit with a span of 0.2 for each 15 minute interval is quite smooth, is small in magnitude, and results in residuals with positive and negative autocorrelations whose absolute values at lags up to 500 are not more than about 0.01 for PSTN-to-IP and 0.04 for IP-to-PSTN. This confirmed that the effect of the distribution change is minor, and that there does not appear to be short-term burstiness in the durations. The latter is also commensurate with the non-homogeneous call arrival process of Section 9, which suggests that the calls arise from many different unrelated sources.

10.3. Modeling results

The work of this section results in a parametric distribution for call duration — the mixture of piecewise Weibull distributions described in Table 2. This distribution is used in the mathematical model for semi-calls. The work also verified the assumption that the call durations are i.i.d. This is important for both the mathematical and semi-empirical models. In particular it implies that we do not need to consider semi-call direction, callee-to-caller and caller-to-callee, in the modeling and therefore in simulation studies. The same is true of time-of-day effects.

11. Modeling the silence suppression on-off process

This section describes analyses of the statistical properties of the lengths of the on-off, or transmission-silence, intervals for the 156100 semi-calls of the *detail-augmented data*. A parametric model for the on-off process is developed that is part of the multiplexed traffic mathematical model. We verify certain of the i-id assumptions made in both the semi-empirical and mathematical models.

We use the term "on-off" for convenience, but in fact the off intervals are not fully off, sending an initial silence packet 5 ms after the last transmission packet, which signals the start of the silence interval, and sending connection keep-alive packets every 2 sec. We altered the on-off intervals slightly by imposing a holdover of transmission when a silence interval was less than 25 ms. This removed only a small fraction of silence intervals which otherwise could be as small as 5 ms.

Let $\dot{w}_k$ be the lengths of intervals of a semi-call where k indexes the order of occurrence. The intervals alternate between transmission and silence starting with a transmission interval. Let $\dot{t}_r = \dot{w}_{2r-1}$ be the transmission lengths, and $\dot{s}_r = \dot{w}_{2r}$ be the silence lengths. The minimum length is 25 ms for $\dot{s}_r$ and 80 ms for $\dot{t}_r$; this means that the origin for each random variable is not 0, which makes parametric modeling more complex. One remedy is to subtract the minima, but it is convenient to take logs for certain purposes, and a 0 origin would lead to distortions at the bottom ends of the distributions. Instead, we subtract 60 ms from the $\dot{t}_r$, taking $t_r = \dot{t}_r - 60$, and 15 ms from the $\dot{s}_r$, taking $s_r = \dot{s}_r - 15$. This is sufficient for parametric modeling with a 0 origin, and allows the taking of the logarithm.

We study t_r and s_r for individual semi-calls as a bivariate time series in r, considering the marginal distributions of each, the autocorrelations of each, and the cross-correlations of the two. We do not model jointly the intervals of the two semi-calls of a single call; this is not needed for QoS study because the two semi-calls are not multiplexed with one another on an operational network. Understanding the silence suppression mechanism as a whole, independent of QoS issues, would require joint modeling.

11.1. 1000 semi-call samples: Introduction, trend, and independence

We analyzed semi-calls with larger durations so that sample sizes of the transmission intervals and silence intervals would be larger. This is important because random effects modeling is used, necessitating identification of random-effects distributions, which is greatly enhanced by reduced sampling variability.

Here we report on one subset analysis: 1000 samples of both directions, caller-to-callee and callee-to-caller, of the 44689 connected calls in the *detail-augmented data*. Each call has four sets of interval lengths: t_r and s_r, each for caller-to-callee and callee-to-caller, so there are 4000 sets of lengths. Sampled calls had the following properties: call durations from 2^8 sec to 2^{12} sec; call durations as nearly equally spaced as possible on a log scale; and each of the four sets of interval lengths for a call have at least 80 values. The medians and the maxima of the numbers of intervals for the four sets are the following: t_r caller-to-callee (331, 1861); t_r callee-to-caller (370, 2341); s_r caller-to-callee (330, 1859); s_r callee-to-caller (369, 2341).

t_r and s_r have low-frequency trends for many calls. Figure 13 shows one example. $\log(t_r)$ for a caller-to-callee semi-call is graphed against the index r; the curve is a loess fit with locally quadratic fitting and a span of 2/3

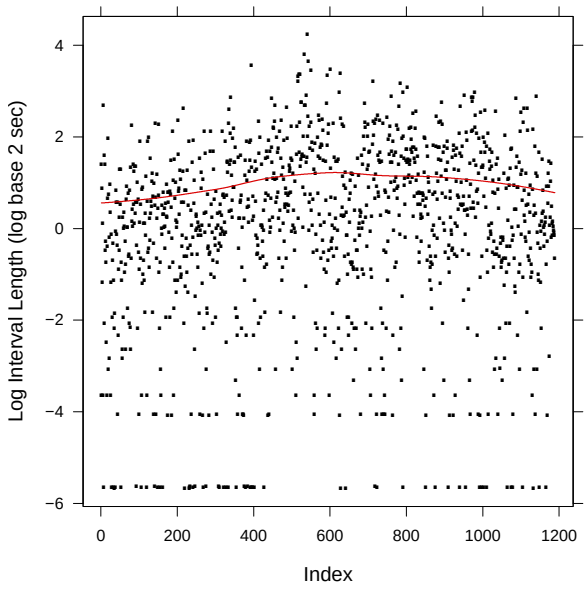

FIG 13. *1189 log base 2 transmission interval lengths are plotted against order of occurrence.*

(Cleveland and Devlin, 1988). With a very small number of exceptions, the magnitudes of the trends are small compared with the overall variability in log lengths, but result in a number of small positive autocorrelations that is much larger than expected. However, the estimated autocorrelations from the residuals of the loess fit are small, and vary as expected for true autocorrelations of 0. The cross-correlations of the residuals also are consistent with values of 0. In our analysis below we study interval lengths on the original scale corrected for the minor trends. The corrected intervals are 2 to the power of the mean of the loess fitted values plus the residuals. With an abuse of notation we will let t_r and s_r be the corrected values. The end result is t_r and s_r that are uncorrelated bivariate time series. We take this to imply independence as well since the log lengths, while not exactly a Gaussian process, are not too far off.

11.2. 1000 semi-call samples: Marginal distributions

The t_r and s_r to be modeled consist of 4000 sets of interval lengths of varying sample sizes. These come from 1000 calls each with 4 categories of semi-calls: transmission and silence by caller-to-callee and callee-to-caller. To study the marginal distributions, we made 4000 quantile plots for each parametric distribution whose ability to fit the empirical distributions was checked. We found that the square-root lengths, $\sqrt{t_r}$ and $\sqrt{s_r}$, are well approximated by the gamma distribution, so t_r and s_r are a square-root gamma distribution. t_r and s_r on the original scale have tails too long for the Weibull and gamma distributions, and too short for the log normal. This is illustrated in Figure 14, which has log normal, square-root gamma, and Weibull quantile plots, for callee off lengths. For

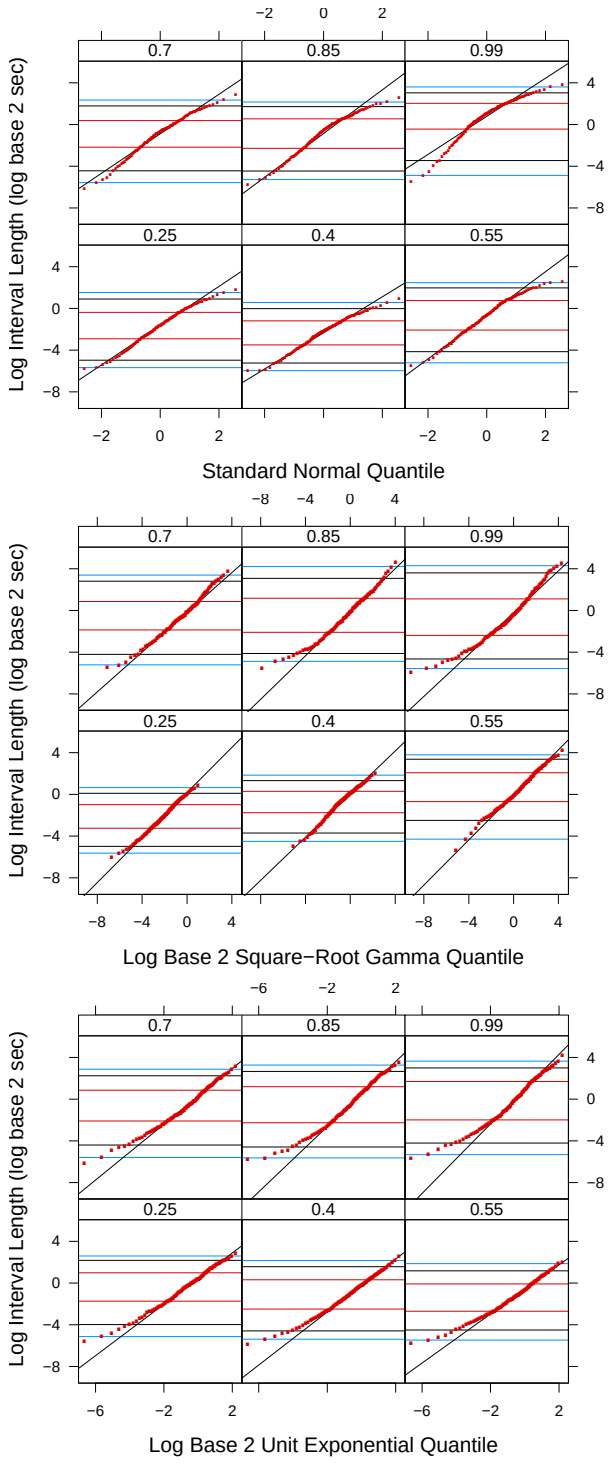

FIG 14. *Three sets of quantile plots on a log scale investigate approximations of empirical quantiles for callee off lengths by the log normal (top), square-root gamma (middle), and Weibull (bottom) distributions.*

each parametric distribution there are log quantile plots for 6 semi-calls. The six chosen semi-calls for each parametric distribution summarize the behavior of a large number of displays as described next.

On each plot of Figure 14, the empirical log quantiles of orders $i/100$ for $i = 1$ to 99 are graphed against the quantiles of orders $i/100$ of a parametric distribution; from the top set of 6 to the bottom set, the distributions are normal, log square-root gamma, and log exponential. Parameters for the square-root gamma plot are estimated by maximum likelihood. The log normal and Weibull do not need parameter estimation because on the log scale, the quantiles are linear in the two unknown parameters. The horizontal lines show 6 empirical quantiles of log length — 0.01, 0.05, 0.25, 0.75, 0.95, and 0.99. The oblique lines go through the 0.25 and 0.75 quartiles. The plots were made from the 460 callee semi-calls out of 1000 that had at least 400 off intervals. For each of the 460 plots and each parametric distribution, we computed the mean absolute deviation of the points of the plot from the oblique line on the plot. We ordered the resulting 460 deviations for each parametric distribution from smallest to largest to form empirical quantiles of orders $(k - 0.5)/460$ for $k = 1$ to 460. We computed deviation quantiles of orders 0.25, 0.40, 0.55, 0.70, 0.85, and 0.99 from the 460 values using linear interpolation. Then we found the 6 semi-calls of the 460 whose deviations were the closest to the 6 interpolated quantiles. Each set of 6 quantile plots in Figure 7 for a parametric distribution display the 6 semi-calls chosen for the distribution by this process.

For $\sqrt{t_r}$ and $\sqrt{s_r}$ there are 4000 pairs of maximum likelihood estimates of the gamma shape and scale. Let α denote the log base 2 of the shape of a gamma distribution and let β be the log base 2 of the scale. Let $\hat{\alpha}$ and $\hat{\beta}$ be the logs of the maximum likelihood estimates. The 1000 pairs of values of these estimates for the 4 categories are shown by boxplots in Figure 15. The four marginal distributions of the shape are centered at very similar values, but the variability is much smaller for silence than transmission. The scale distributions for silence are shifted toward lower values. This is to be expected because we reported that the overall bit-rate of the semi-calls of the *detail-augmented data* is 45.17 kilobits/sec, more than half the full bit-rate of 80 kilobits/sec, which means that transmission intervals are longer, taking into account that the bit-rate of a transmission interval is 80 kilobits/sec and that of a silence interval is about 0.8 kilobits/sec. The marginal distributions for caller-to-callee and callee-to-caller are very close; this provides strong support for the on-off process marginal distributions being identically distributed across call direction, part of the semi-call i-id assumptions.

Variability in the $\hat{\alpha}$ and $\hat{\beta}$ comes from two sources. One is the sampling variability arising from finite samples, and the other is potential changes in the value of a parameter from one semi-call to the next. We used the bootstrap to assess the sampling variability. For each of the 4000 estimations of the two parameters, we took 1000 bootstrap samples, resulting in 1000 bootstrapped estimates of the pair of parameters. For each of the 4000 bootstrap simulations, we made a normal quantile plot of the 1000 log shape bootstrap estimates, a normal quantile plot of the 1000 log scale estimates, and a scatterplot of the 1000

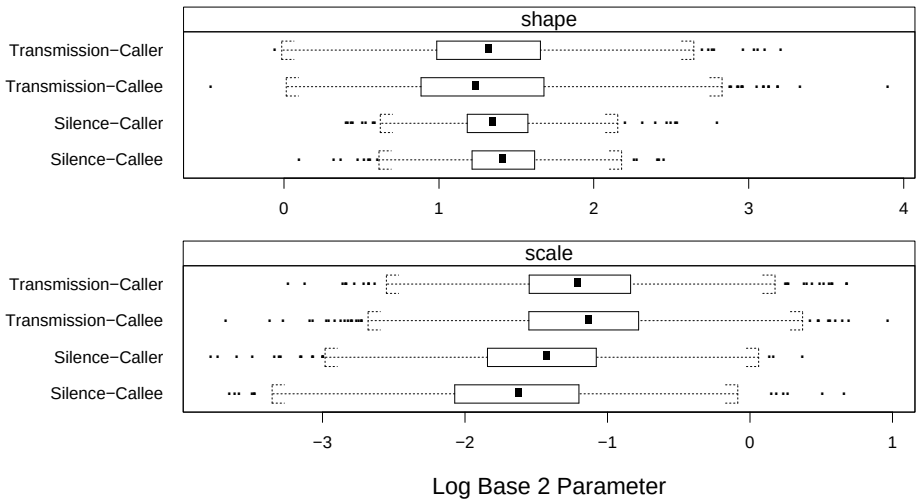

FIG 15. *Boxplots of estimates of log gamma shape and scale estimates, 1000 per boxplot.*

log shape estimates against the 1000 log scale estimates. The displays show that the bivariate bootstrap distribution is well approximated by the bivariate normal except for a small number of the smaller samples where there was slight skewness in the marginals. They also show substantial correlation of the estimates. For each estimate we computed the bootstrap variances and covariance.

Because the distributions of the estimates of $\hat{\alpha}$ and $\hat{\beta}$ are nearly the same for callee-to-caller and caller-to-callee, we proceed next by merging the two so that we now have two categories, transmission and silence, with 2000 pairs of estimates for each category. Let $\hat{\alpha}_k$ and $\hat{\beta}_k$ for $k = 1$ to 2000 be estimates for either category. The estimates can be decomposed into two components:

$$\begin{aligned} \hat{\alpha}_k &= \alpha_k + \epsilon_k(\alpha) \\ \hat{\beta}_k &= \beta_k + \epsilon_k(\beta) \end{aligned}$$

α_k and β_k are independent random samples from a bivariate distribution, $F_{(\alpha,\beta)}$, with means $\mu(\alpha)$ and $\mu(\beta)$, variances $\sigma^2(\alpha)$ and $\sigma^2(\beta)$, covariance $c(\alpha, \beta)$, and correlation $\rho(\alpha, \beta)$. α_k and β_k model true changes in the gamma parameters. $\epsilon_k(\alpha)$ and $\epsilon_k(\beta)$ are bivariate normal random variables with mean 0, and variances and covariance $\sigma^2_\epsilon(\alpha)_k$, $\sigma^2_\epsilon(\beta)_k$, and $c_\epsilon(\alpha, \beta)_k$ from the bootstrap sample for call k; the variances and covariance are assumed fixed and known. $\epsilon_k(\alpha)$ and $\epsilon_k(\beta)$ model sampling variability in the estimates with assumptions based on our analysis of the bootstrap samples.

The sample mean, $\hat{\mu}(\alpha)$, of the $\hat{\alpha}_k$ is an estimate of $\mu(\alpha)$; we define $\hat{\mu}(\beta)$ similarly. Let $\hat{\sigma}^2(\hat{\alpha})$ be the sample variance of the $\hat{\alpha}_k$. Let $\bar{\sigma}^2_\epsilon(\alpha)$ be the sample mean of the $\sigma^2_\epsilon(\alpha)_k$. An estimate of $\sigma^2(\alpha)$ is

$$\hat{\sigma}^2(\alpha) = \hat{\sigma}^2(\hat{\alpha}) - \bar{\sigma}^2_\epsilon(\alpha).$$

TABLE 3
Statistics for the random effects modeling of log shape and log scale for $\sqrt{s_r}$ and $\sqrt{t_r}$

statistic	notation	transmission	silence
sample mean of shape estimates	$\bar{\alpha}$	1.3307	1.3938
sample mean of scale estimates	β	-1.1940	-1.5675
sample variance of shape estimates	$\hat{\sigma}^2(\hat{\alpha})$	0.3144	0.0973
sample variance of scale estimates	$\hat{\sigma}^2(\hat{\beta})$	0.3546	0.4078
sample covariance of shape & scale estimates	$\hat{c}(\hat{\alpha}, \hat{\beta})$	-0.2928	-0.1649
mean of bootstrap variances of shape estimates	$\bar{\sigma}^2_\epsilon(\alpha)$	0.0151	0.0131
mean of bootstrap variances of scale estimates	$\bar{\sigma}^2_\epsilon(\beta)$	0.0176	0.0173
mean of bootstrap covariances of shape & scale estimates	$\bar{c}_\epsilon(\alpha, \beta)$	-0.0147	-0.0135
estimate of variance of shape random effect	$\hat{\sigma}^2(\alpha)$	0.2993	0.0842
estimate of variance of scale random effect	$\hat{\sigma}^2(\beta)$	0.3370	0.3905
estimate of covariance of scale & shape random effects	$\hat{c}(\alpha, \beta)$	-0.2781	-0.1514
estimate of correlation of scale & shape random effects	$\hat{\rho}(\alpha, \beta)$	-0.8756	-0.8347

Similar notation and formulas hold for a $\hat{\sigma}^2(\beta)$. Finally, let $\hat{c}(\hat{\alpha}, \hat{\beta})$ be the sample covariance of $\hat{\alpha}_k$ and $\hat{\beta}_k$, and let $\bar{c}_\epsilon(\alpha, \beta)$ be the sample mean of the $c_\epsilon(\alpha, \beta)_k$. Then an estimate of $c(\alpha, \beta)$ is

$$\hat{c}(\alpha, \beta) = \hat{c}(\hat{\alpha}, \hat{\beta}) - \bar{c}_\epsilon(\alpha, \beta).$$

The estimates $\hat{\sigma}^2(\alpha)$, $\hat{\sigma}^2(\beta)$, and $\hat{c}(\alpha, \beta)$ provide an estimate $\hat{\rho}(\alpha, \beta)$ of $\rho(\alpha, \beta)$, the correlation of α and β. We carried out diagnostic checking for the bivariate distribution of α_k and β_k and found that the bivariate distribution is well approximated by the normal. The estimates of the parameters of this distribution and other information are given in Table 3.

The bootstrap statistics in the table have much smaller values than the corresponding random effects distributions, which means the variability of the estimates is mostly due to the random effects. Still, we used a method to check the bivariate normality of the random-effect distributions that takes the sampling error into account. We generated 1000 independent bivariate normals with variances and a covariance equal to the random effects estimates in the table. We added to these values a sample of 1000 normal random variables with mean 0 and different variances equal to the bootstrap variances. We sorted the sums from smallest to largest to create order statistics. This was done 10000 times, and the order-statistic means computed. These values were compared with the order statistics of the estimated values by a variety of visualization tools such as scatterplots and probability plots. The displays showed that the bivariate normal is a reasonable approximation to the random effect bivariate distribution. We have chosen to characterize the distributions of $\log_2$ shape, α, and $\log_2$ scale, β, because the logarithms improve the approximation by bivariate normality.

To generate the transmission intervals for a call, values of α and β are generated by the bivariate normal distribution described in Table 3 for transmission. The shape and scale for the gamma distribution used to generate the $\sqrt{t_u}$ are $\dot{\alpha} = 2^\alpha$ and $\dot{\beta} = 2^\beta$, respectively. These generated values are squared and 60 ms added, the amount subtracted at the outset of the analysis. Their expected

value is 60 ms plus the second moment of the generating gamma, $\dot{\alpha}\dot{\beta}^2(1+\dot{\alpha})$. Silence intervals are generated in a similar manner.

11.3. Previous results

Past studies of transmission and silence interval lengths are for the most part consistent with the modeling shown here when we take the increased complexity of the intervals as we observe them into account. This complexity arises because the interval properties change from one call to the next. This has not been fully appreciated in the past, since study has not been carried out on live traffic from an operational network carrying a rich set of VoIP applications.

A very early study, about 80 years ago, reported on the empirical cumulative distribution and density of *talk spurts* (transmission) and *silence gaps* (silence) of speech (Norwine and Murphy, 1938). Statistical modeling began later. Brady (1968, 1969) developed a model for the talk spurts and silence gaps based on analysis of both directions of a call that can be applied to a single direction, a semi-call. The model has independent exponentials for all interval lengths. For talk spurts, the origin is 0 ms, and the mean is 0.902, 1.125, or 1.311 sec depending on the detection algorithm parameters. For silence gaps, the origin is 200 ms, carried out by preventing silence gaps less than that amount in the algorithm; and the mean is 1.664, 1.695, or 1.721 sec. Gruber (1982) and Lee and Un (1986) continued work along these lines, with exponential talk spurts, and a single exponential or a mixture of exponentials to model silence gaps and accommodate small gaps.

When studies for VoIP call properties began, it became clear that the independent exponential model was inadequate. Jiang and Schulzrinne (2000) write: "Previous studies on the performance of voice traffic multiplexers assume that the length of spurts and gaps follow an exponential distribution. Since most of these speech measurements are based on either analog or simple digital silence detectors, we suspected that the spurts/gaps produced by modern voice codecs and silence detectors will no longer fit well to the exponential model, which may in turn affect the packet loss rate at the same utilization." Their findings verified this; they found that the distributions of transmission and silence interval lengths for six monitored calls were consistently heavier-tailed than the exponential. Following up on this work were Casilari et al. (2002) who modeled transmission and silence by the log normal distributions using 10 hours of point-to-point video conferencing, and Biernacki (2006) who modeled by mixtures of exponentials based on 10 hours of VoIP semi-calls. Casilari et al. (2002) also found the small but consistently positive autocorrelation described above; the likely cause, as in our case, is the low-frequency trend that we observed.

These analyses of VoIP traffic are overall consistent with our results. We found the log normal provided a fit that is not an exceedingly poor approximation, but rather is not as good a fit as the square-root gamma. However, our work departs in showing that the wide diversity of calls of our *detail-augmented data* requires varying distributions, modeled by changing parameters of the square-root gamma in a random effects model. In other words, individual semi-calls

admit of a description by a simple parametric distribution, but with changing parameters.

11.4. i-id assumptions

One critical task in the semi-call modeling is establishing the validity of the i-id assumptions described in Section 8. We now discuss assumptions related to the on-off process.

Earlier in this section we showed that the bivariate distributions of the log shape and log scale square-root gamma distribution for transmission interval lengths are nearly the same for caller-to-callee semi-calls and callee-to-caller semi-calls. This means that the identical-distribution assumption holds across call direction. The same is true for the silence intervals. We have also verified that within a semi-call, the on-off interval lengths can be taken as independent of time to a good approximation.

We assume in our modeling that the bivariate random variable of the two log parameters of the on intervals are independent and identically distributed across semi-calls. We investigated this assumption in the same manner as the call duration in Section 10. For each log parameter we made four plots of the 5 quantiles in 15 minute intervals. The four plots are the four combinations of caller-to-callee and callee-to-caller crossed with IP-to-PSTN and PSTN-to-IP. We did the same with autocorrelations. No significant effects were found. We carried this out for the 15220 semi-calls of the 7610 calls with the same properties as those that defined the 1000 sampled calls studied in this section, but not limiting to 1000 calls.

In our modeling we assume that the on-off process random variables are independent of the call duration. We checked this also using the above 15220 semi-calls. For each log parameter we made a four scatterplots against log call duration. The four plots are the four combinations of caller-to-callee and callee-to-caller crossed with on and off. No relationship was detected. Further checking of this assumption is given in Section 12.

It is important to emphasize that this checking can only be carried with the longer calls because short calls have too few observations to get reliable estimates of the two log parameters. However, as discussed in Section 6, the short calls have little effect on QoS.

11.5. Modeling results

The results of this section include further development of the mathematical model. The on-off process of alternating transmission and silence intervals lengths of a semi-call is modeled. All interval lengths of a semi-call are independent. The marginal distribution of the transmission lengths is a square root gamma distribution. The log shape and log scale of the gamma for the transmission intervals of a single semi-call are a random sample from a bivariate normal distribution. The same is true of the silence intervals but with different parameters for

the bivariate normal. For each semi-call, a call duration is generated that is independent of the on-off process random variables; then interval lengths of a realization of the on-off process are generated until the sum of the lengths is nearly the same as the call duration.

12. Semi-call bit-rate

The bit-rate of a semi-call is the sum of the total bits of all packets in the semi-call divided by the call duration. The mean bit-rate across the 277540 semi-calls of the *complete-call data*, which is the sum of the bits of all semi-calls divided by the sum of their durations, is 44.65 kilobits/sec. Without silence suppression, the bit-rate would be 80 kilobits/sec, so suppression reduces the multiplexed call bit-rate to 55.8% of the rate without silence suppression.

The marginal distribution of the semi-call bit-rates is shown in the cumulative frequency quantile plot of Figure 16. On the plot, a semi-call bit-rate value is graphed against its empirical cumulative frequency, which is the fraction of semi-calls whose bit-rates are less than or equal to the value. The horizontal lines show quantiles of frequencies 0.01, 0.05, 0.25, 0.75, 0.95, and 0.99. The oblique line is drawn through the upper and lower quartile points. The bit-rates range from nearly 0 kilobits/sec to 80 kilobits/sec, and are close to uniformly distributed from the 0.2 to the 0.8 quantiles.

We do not model bit-rate directly, but in this section we analyze the semi-call bit-rates to further assess the i-id assumptions of the modeling. Specifically, we

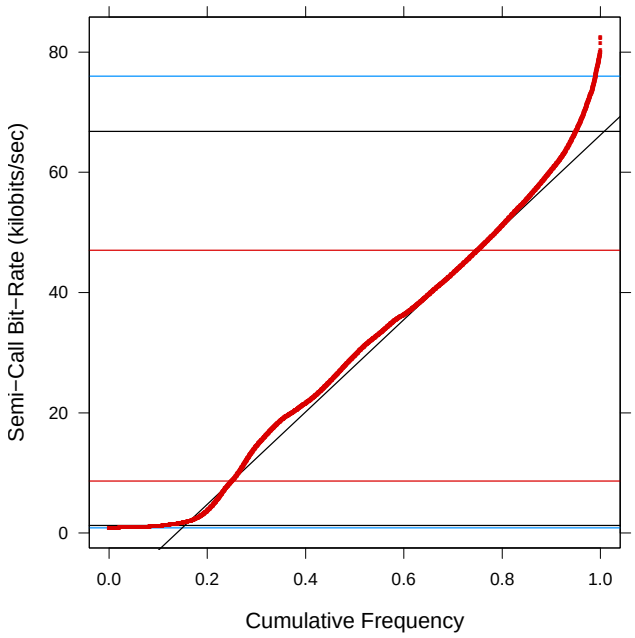

FIG 16. *Cumulative frequency quantile plot of semi-call bit-rates.*

check the 277540 semi-calls of the *complete-call data* to see if we can detect either a difference of the semi-call bit-rate marginal distribution or a correlation in semi-call bit-rate for the following: (1) across the call direction (caller-to-callee and callee-to-caller), or (2) across time (time of day).

Each semi-call bit-rate is a summary statistic for the on-off process of a semi-call. The transmission intervals have 1600-bit packets every 20 ms, which is 80 kilobits/sec. The silence intervals have 1600-bit packets every 2000 ms, which is 0.8 kilobits/sec, 100 times less. Thus the semi-call bit rate is very nearly 80 kilobits/sec times the sum of the transmission interval lengths of the semi-call divided by the semi-call duration.

12.1. i-id assumptions

For each call, we subtracted the bit-rate of the callee-to-caller semi-call from the bit-rate of the caller-to-callee semi-call. We formed three groups of differences based on three intervals of the call duration: 0-32 sec, 32-128 sec, and 128- sec. Figure 17 has three cumulative frequency quantile plots for the three groups, each drawn using the same visualization method as Figure 16. For the 128- sec interval, the largest durations, the distribution of the differences is centered on 0, and the differences are very close to a symmetric distribution; the callee-to-caller and caller-to-callee bit-rate distributions are nearly the same. For the 32-128 sec interval, there are small differences, but overall the callee-to-caller and caller-to-callee bit-rate distributions are similar. For interval 0-32 sec, slightly more that 75% of the callee-to-caller bit-rates are larger, so the distributions differ.

More insight is provided by Figure 18 which shows 10 nonparametric density estimates using the ed method with 20-gaps smoothed by local cubic fitting and

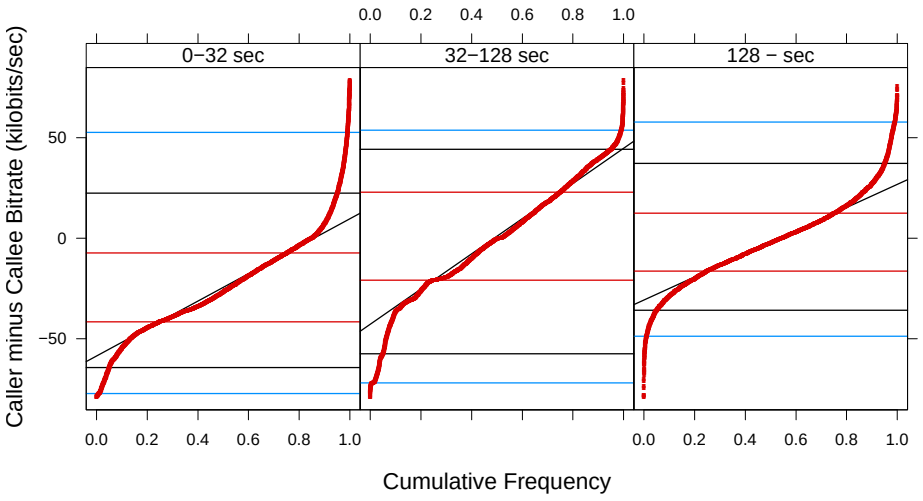

FIG 17. *Cumulative frequency quantile plots of call direction bit-rate differences, caller-to-callee - callee-to-caller, for 3 intervals of duration.*

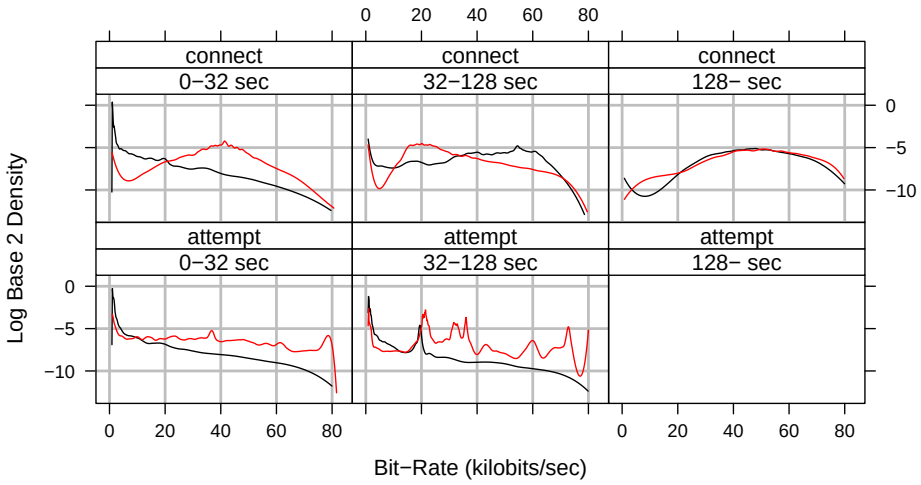

FIG 18. *Ed semi-call bit-rate density estimates for attempts and connects, caller-to-callee (black curves) and callee-to-caller (red curves), and 3 intervals of call duration.*

a span chosen to optimize the fits (Hafen and Cleveland, 2009). There are 12 categories of semi-call bit-rates — 2 directions (caller-to-callee and callee-to-caller) by 2 semi-call types (attempt and connect) by the 3 duration intervals. The figure shows estimates for 10 categories, omitting the two estimates for 128-sec and attempt, which have very few calls. For attempts, the callee-to-caller density is lower than the caller-to-callee at low bit-rates and higher at high bit-rates, and overall callee-to-caller has more bits. The reason is that during the ring period, the VoIP transmission protocol sends more packets callee-to-caller than caller-to-callee. For connects in interval 0-32sec, callee-to-caller also has more bits; the ring period makes up a significant fraction of the call time, and many calls are likely point-of-sale calls for which the callee-to-caller also sends more bits. For connects in interval 32-128 sec, the caller-to-callee density is lower than the callee-to-caller at low bit-rates and higher at high bit-rates, and the caller-to-callee sends more bits; many calls are likely faxes for which the caller-to-callee sends more bits. This effect, however, is minor. For connects in interval 128-sec, the densities are nearly the same; most calls are likely voice, and across calls, caller-to-callee and callee-to-caller have similar patterns.

We have detected caller-to-callee and callee-to-caller bit-rate distributional differences, but with more than very minor effects only for the interval 0-32sec, which, as shown in Figure 3 has a very small fraction of the total traffic bits. For the remaining calls, the bit-rate distributions for callee-to-caller and caller-to-callee are very nearly the same. This is reflected in the overall bit-rates. For all semi-calls in the *detail-augmented data*, the combined caller-to-callee and callee-to-caller bit-rate is 43.96 kilobits/sec. For callee-to-caller the overall bit-rate is 45.17 kilobits/sec, and for caller-to-callee it is 42.74, a small difference. Thus the detected differences in semi-call bit-rates are not significant for our modeling because of its QoS goal.

We also studied whether the semi-call bit-rate, separately for caller-to-callee and for callee-to-caller, was identically distributed through time, and whether there was correlation in bit-rate as a time sequence. Our methods were the same as that for study of time characteristics of the two log parameters of the on and off intervals of caller-to-callee and callee-to-caller in Section 11. In no case did we find more than minor efforts for the changing quantiles through time, and there was no significant time autocorrelation.

12.2. Modeling results

i-id assumptions, critical to the mathematical model and the semi-empirical model, have been further verified in the sense that departures appear only for semi-calls with short durations that have overall low multiplexed traffic bit-rates. Specifically the results are consistent with the following assumptions: the semi-call on-off process random variables are independent given the bivariate parameters and are independent of call duration; the length random variables of individual semi-calls are independent; the off lengths of individual semi-calls are identically distributed; and the on lengths of individual semi-calls are identically distributed.

13. Statistical properties of the multiplexed packet traffic

13.1. Long-range dependence

The decay of the autocovariance function depends on the tail of the distribution D. If the density is $d(c) = e^{-c}$, a unit exponential distribution, then $\rho(\tau) = C(\int_{\tau}^{\infty} ce^{-c}dc - \tau e^{-\tau})$, so there is an exponential decay of the autocovariance. If the density is $d(c) = \frac{\nu}{s}(\frac{c}{s})^{-\nu-1}$ where $\nu > 1$ and $c \geq s$, a Pareto distribution, then $\rho(\tau) = C\frac{s^{\nu}\tau^{1-\nu}}{\nu-1}$ for $\nu > 1$, so there is a polynomial decay. If $1 < \nu \leq 2$, then the traffic is long-range dependent because $\sum_{\tau} \rho(\tau) = \infty$ (Cox, 1984).

Best-effort Internet traffic — the data traffic on the Internet such as Web page downloads (protocol HTTP), email (protocol SMTP), and interactive encrypted logins (protocol SSH) — is long-range dependent because transferred file sizes have a distribution whose upper tail is well approximated by the Pareto (Paxson and Floyd, 1995; Barford and Crovella, 1998). Queueing delays can be substantially greater for long-range dependent traffic than for traffic for which $\rho(\tau)$ decays exponentially which has had a major impact in the engineering of the Internet (Erramilli et al., 1996). Historically, telephone call duration was modeled by an exponential in an era when telephone calls were voice only (Bolotin, 1994; Brown et al., 2005).

The analysis and modeling of D in Section 10 leads to the conclusion that the VoIP traffic is not formally long-range dependent since the duration distribution does not have a tail that would lead to this. Later in this section, the autocorrelation properties of the multiplexed packet arrivals are analyzed directly to provide further verification. Other duration modeling work has also resulted

in specifications that also do not lead to long-range dependence. Bolotin (1994) models duration by a statistical mixture of log normal distributions, rather than isolating attempts and connects, but the motivation is the same — that the duration distribution is too complex for a single, simple parametric family. Our duration distribution appears even more complex, the reason for our mixture of piecewise Weibull distributions. Dang et al. (2004) are led to the conclusion that the call distribution for VoIP on a local network of 800 VoIP phones is a generalized Pareto distribution with a shape parameter, using our notation, of $\nu = 2.56$, which means the traffic is not long-range dependent because $\nu \geq 2$.

13.2. 20 ms packet counts: Power spectrum and variance-time plot

To understand the statistical properties of the multiplexed packet process of the *processed data*, we analyzed IP-to-PSTN and PSTN-to-IP packet arrival counts in 20-ms intervals, two time series of length 8.625×10^6. The process is non-stationary because of the change in the packet arrival rate, shown in Figure 2. While the process is overall non-stationary, the IP-to-PSTN and the PSTN-to-IP 20-ms counts are stationary within almost all of the 575 5-min intervals that make up the measurement period of 47 hr 55 min. There are significant trends in only a small fraction of the intervals.

Guided by theory, we used the near-stationarity within the 5-min intervals to explore certain statistical properties of the counts. In each interval the 20-ms counts can be thought of, theoretically, as the multiplexing of the packets of v independent and identical source processes, each with mean μ and variance σ^2, because the traffic is made up of many independent calls. Therefore, the multiplexed process has mean $v\mu$ and variance $v\sigma^2$, so the variance is proportional to the mean as v changes. In addition, as v increases, the counts tend to normality by the central limit theorem. If v remains constant, the 20-ms counts are a stationary time series with an autocorrelation function. However, the autocorrelation does not depend on v.

We made 575 normal quantile plots of the PSTN-to-IP 20-ms counts in each 5-min interval, and 575 normal quantile plots for the IP-to-PSTN counts. The empirical marginal distributions are well approximated by the normal. We computed 1150 means and variances of the 20-ms counts for the PSTN-to-IP and IP-to-PSTN counts in the 5-min intervals. Figure 19 graphs the log variance against the log mean. The oblique line is the best least-squares fit with slope 1. All but about 5% of the points follow the line, strongly supporting the above theoretical proportionality. The deviations, large variances, are those intervals with trends in the rates that while small in magnitude, inflate the variance, rendering the theory inapplicable.

The proportionality suggests a method of adjusting the counts for mean and variance nonstationarity that allows for a continuous instantaneous change in the traffic rate with time. The changing $v\mu$ is estimated by a loess fit to the 20-ms counts with a span of 0.0644, which is 180 minutes, and locally quadratic fitting (Cleveland and Devlin, 1988). The packet counts are adjusted by subtracting the loess fit and dividing by its square root. This does not necessarily

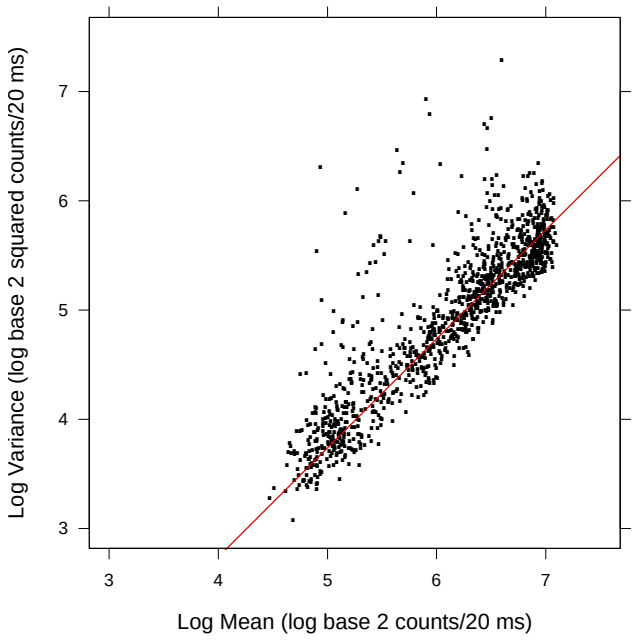

FIG 19. *Log variance is graphed against log mean for 20-ms packet counts in 1150 5-min intervals. The oblique line is the best least-squares fit with slope 1.*

eliminate covariance nonstationarity for nonzero lags, but does provide insight into dependence properties, and appears, as we shall see, to provide an excellent estimate of the above stationary autocorrelation function.

We also analyzed a stationary VoIP process, a synthetic multiplexed packet series generated by the semi-empirical model whose specifications are described in Sections 8 and 14. The call arrival rate was set to make the expected traffic bit-rate 3 mbps, which lies in the range of the observed 5-min traffic bit-rates of the live data. The duration of the synthetic run was 96 hr, about twice the duration of the *processed data* of 47 hr 55 min. The time series of synthetic 20-ms packets counts has 17.28×10^6 values.

We estimated the power spectra of three series: the loess adjusted IP-to-PSTN and PSTN-to-IP 20-ms counts, and the synthetic 20-ms counts. The details of the estimation procedure are given at the end of this section. The three spectra are nearly identical, so the loess adjustment of the counts has resulted in a good estimate of the spectrum of the stationary process. The red curve in Figure 20 shows the synthetic spectrum estimate (red) in decibels graphed against frequency f. The plotted points (black) are the logs of means of the periodogram values in blocks of 6.

Two sets of narrow band peaks appear in the spectrum estimate. The first set is at a fundamental frequency of 0.111 (period = 180 ms) and its three harmonics. This is the cycle in the gateway packetization algorithm described in Section 7. The second is a periodic effect at a fundamental frequency of 0.01

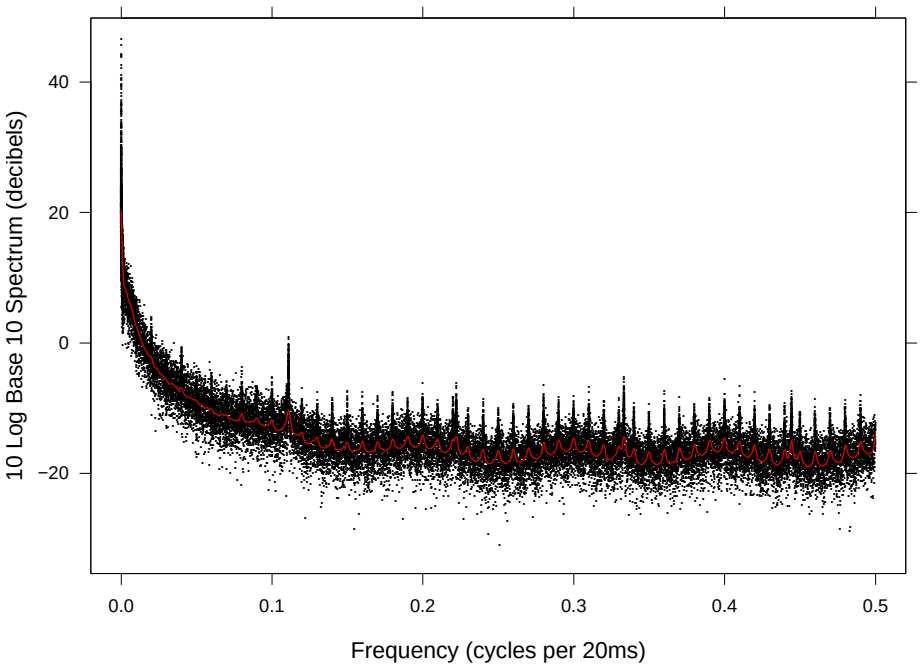

FIG 20. *The estimate of the log power spectrum is graphed against frequency by the curve (red). The plotted points (black) are the logs of averages of 6 periodogram values.*

cycles per 20 ms (period = 2 sec). The effect is caused by the sending of packets every 2 sec during silence periods. Interestingly, the peak at the fundamental does not appear, presumably because of leakage due to the strong rise near the origin, but all harmonics do appear.

The salient feature of the log spectrum estimate is the substantial rise as the frequency decreases, starting at about $f = 0.01$ (period = 2 sec). The rise reflects the high positive persistence in the calls due to the highly skewed distribution of the call durations demonstrated in Section 10. More information about the rise is shown in Figure 21, which graphs the log power spectrum against $\ell = 10\log_{10}(f)$, where $\log_{10}$ is log base 10. The vertical and horizontal scales have the same number of units per cm so the physical slopes are same as the slopes in data units. The vertical lines are drawn at frequencies ℓ with periods 1 hr, 1 min, and 1 sec. For $\ell \le -10$ the pattern is piecewise linear, shown by the four line segments (red). The breakpoints are at values of ℓ that have periods of 2.52 sec/cycle, 25.2 sec/cycle, and 75 min/cycle. The four slopes, for decreasing ℓ, are -1.44, -0.67, -1.36, and 0 decibels/$(10\log_{10}$(cycles per 20 ms)). If the 20-ms counts were formally long-range dependent, the slope near the origin would be constant and equal to $-2d$, where $0 < d < 0.5$ is the Hosking fractional difference power (Hosking, 1981). This value would be greater than -1. Two of the estimated slopes are less than -1, but very close to the origin

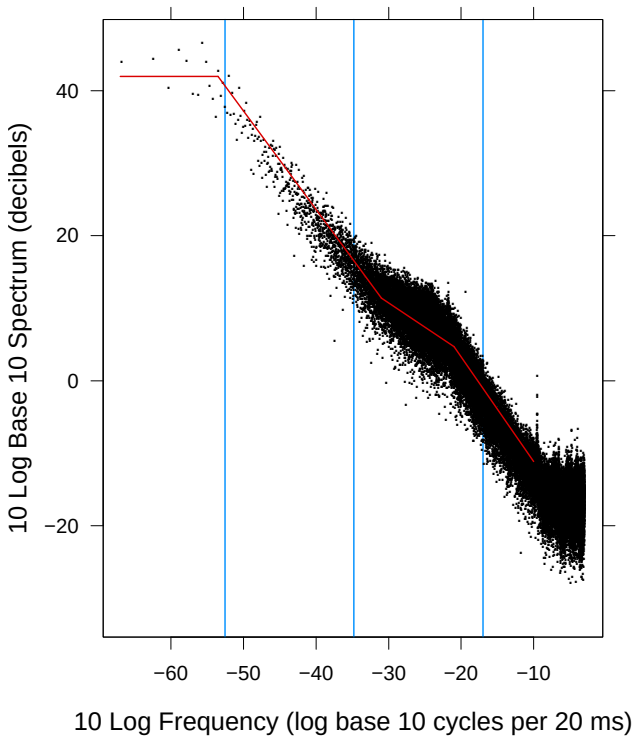

FIG 21. *The plotted points (black) are the logs of averages of 6 periodogram values plotted against log frequency. The oblique line segments (red) describe the pattern of the log spectrum.*

the slope is 0.

Figures 22 and 23 explore the persistence further. The first is a variance-time plot (Leland et al., 1994): for $m = 1, \ldots, 20$, the 2^{24} counts are partitioned into 2^{24-m+1} intervals each with 2^{m-1} observations, and the log base 2 of the variance of the interval means, or m-aggregates, is graphed against log base 2 of m. The second is the variance-time slope plot: each of the 19 slopes of the line segments in the variance-time plot is graphed against the midpoints of the horizontal coordinates of the segment endpoints. If the counts were independent, the slope of the variance-time plot would be -1. If the data were long-range dependent, the slope would tend to $2d - 1$ as m increases, where d is the above Hosking parameter (Beran, 1994). Best-effort Internet traffic tends to have a variance-time plot that is decreasing and convex, and with a slope that increases to $2d - 1$ as m increases. d is typically close to 0.45; furthermore the slope for the smallest m can be close to -1 for higher traffic bit-rates indicating near independence at small time scales and a growing dependence as the time scale increases. Figures 22 and 23 show a very different behavior. The pattern is concave with small absolute slopes that persist through about 10 min and then decrease to -1 indicating an eventual independence.

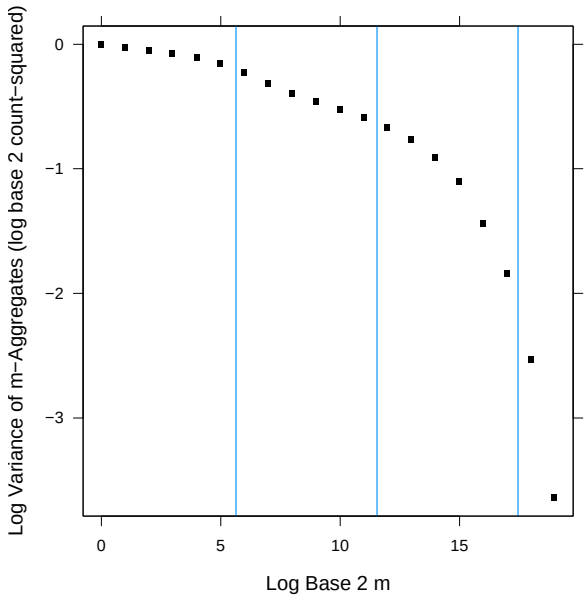

FIG 22. *The variance-time plot graphs the log variance of m-aggregates against log m. The vertical lines are at values of log m corresponding to 1 hr, 1 min, and 1 sec.*

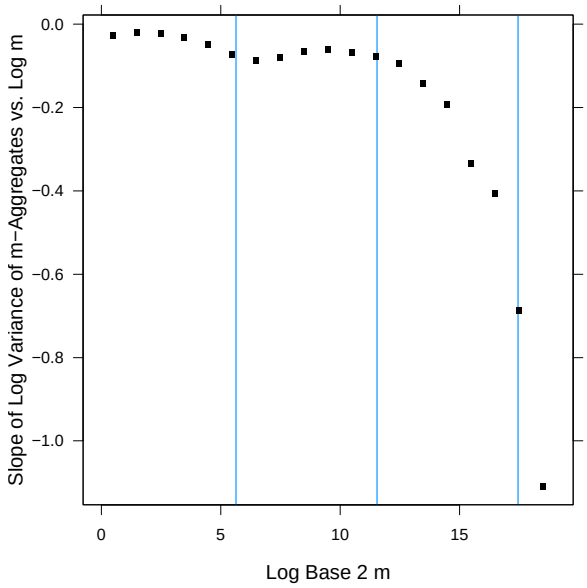

FIG 23. *The slopes of the line segments in the variance-time plot are graphed against the midpoints of the segment endpoints. The vertical lines are at values of log m corresponding to 1 hr, 1 min, and 1 sec.*

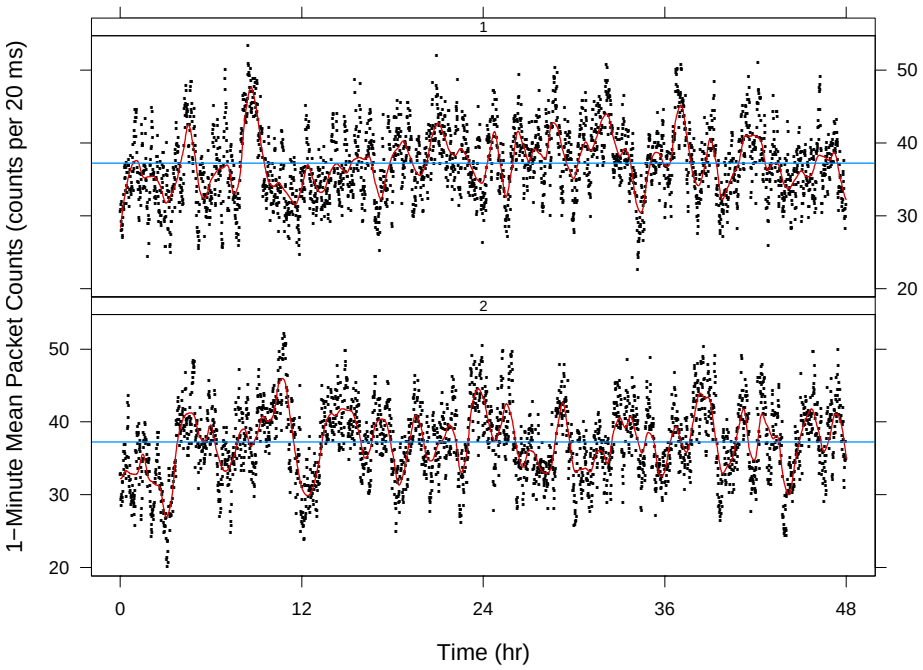

FIG 24. *1 minute means of the synthetic 20-ms counts are graphed against time, broken into two intervals, with a loess fit superimposed. The horizontal lines are at the overall mean of the 20-ms counts.*

The conclusion then from the duration distribution, the power spectrum, and the variance-time plot is that the VoIP traffic is not formally long-range dependent. This departs from best-effort Internet traffic where the connection durations are well approximated by a Pareto distribution, which results in long-range dependence (Willinger et al., 1997).

Still, the VoIP traffic is highly persistent over time scales that can be expected to have a major impact on the queueing process. The spectrum suggests the persistence extends to about 75 min/cycle, the period at which the power spectrum becomes flat. 75 min is the 0.998-quantile of the durations of the *complete-call data*. It is at a lag of about 75 min that autocorrelations become close to 0. By 75 min, the variance-time plot is close to independence. In fact, we can see this period of dependence, although not with the same precision, in a plot of the data. Figure 24 displays 5760 1 minute means of the synthetic 20-ms counts against time, broken into two intervals, with a loess fit superimposed with locally quadratic fitting and a span of 2/96. The horizontal lines are at the overall mean of the 20-ms counts. There are sojourns of the means above or below the overall mean lasting up to about 75 min.

Analyses of the multiplexed VoIP traffic of Zhang et al. (2007) and of Ciullo et al. (2008) also lead to conclusions that the multiplexed packet traffic is not long-range dependent. Dang et al. (2004) ran a simulation using durations of their fitted model, and inserted sampled transmission and silence intervals into

each call from their packet trace collection on a corporate VoIP network. Their plot of the log periodogram against log frequency has a pattern much like that of Figure 21, with a decrease in slope to 0 as the frequency gets small, also suggesting a lack of long-range dependence. This, of course, is to be expected given their modeled duration distribution.

The following describes the estimation of the power spectra. For the IP-to-PSTN and PSTN-to-IP data, the first $n = 2^{23}$ counts were used, and for the synthetic data, the first $n = 2^{24}$ counts, conducive to the fast Fourier transform and variance-time plots to come below. For each, the sample mean was subtracted and the resultant divided by the sample standard deviation. The periodogram was computed at the Fourier frequencies, $k/n, k = 0, \ldots, n/2$. The first 30 frequencies of the IP-to-PSTN and PSTN-to-IP counts were deleted because testing with white noise showed the frequency response function of the high-pass loess adjustment method begins its consequential low-frequency attenuation at the largest of these frequencies. For the synthetic data, the periodogram at frequency 0 was deleted since its value is 0 due to subtracting the mean. Means of the resulting periodogram values were taken in non-overlapping blocks of size 6. The process starts at the lowest available frequency and ends when there are fewer than 6 frequencies available. Each periodogram mean has a frequency equal to the mean of the 6 frequencies of its averaged periodogram values. The power spectra were estimated by smoothing the log of the averaged periodogram values using loess with locally quadratic fitting and with a span of 0.01 (Cleveland and Devlin, 1988).

14. Specification of the models and discussion

The previous sections have provided the first comprehensive analysis and modeling of the statistical properties of multiplexed IP-inbound VoIP packet traffic on a large operational network carrying a wide range of VoIP applications such as voice calls and fax transmissions. The data analyzed were collected on a link of the Global Crossing international network. A VoIP call consists of two semi-calls, callee-to-caller and caller-to-callee. Semi-call signals arrive at a gateway where they are converted to packets and sent out over a single direction of link to a first IP router. The IP-inbound traffic consists of the multiplexing, or superposition, of the packets of the semi-calls on the link.

The analyses result in two validated models for the multiplexed IP-inbound packet traffic arrival process. A semi-empirical model uses empirical observations as part of its modeling. A mathematical model consists of parametric statistical specifications of mathematical components. Modeling is at the semi-call level: specification of the semi-call arrival times and packet inter-arrival times of arriving semi-calls.

The purpose of the modeling is generation of IP-inbound traffic as inputs for simulations of the queueing of VoIP traffic on a network of routers. Simulations are carried out to investigate the effect on delay and jitter of VoIP implementation factors — e.g., priority queueing, silence suppression, signal capture rate,

compression, and accumulation interval — and traffic engineering factors — e.g., link speeds and multiplexed traffic bit-rates.

In this section we provide full specifications of the models and discuss their use in practice. Specifications are given by describing how the models are used to generate the IP-inbound multiplexed packet arrivals, and then describing assumptions made about independence and identical distribution. Notation will be used in the descriptions. Let a_k be the semi-call arrival time of the k-th generated semi-call, let c_k be its duration, and let $\dot{t}_{kr}$ and $\dot{s}_{kr}$ for $r = 1, 2, \ldots$ be the lengths of the sequences of transmission (on) and silence (off) intervals, respectively, of the kth call.

14.1. The semi-call arrival process

The a_k are a non-homogeneous Poisson process with rate λ in arrivals/sec; the arrivals provide the times of the first packets of the semi-calls. Both the semi-empirical model and the mathematical model use this model for the semi-call arrivals. The changing λ is not specified since this depends on the specific network to which the model applies. However, one often carries out traffic engineering to accommodate a sustained peak traffic load, so simulation with a fixed λ is quite relevant to practice.

λ determines the average VoIP multiplexed traffic bit-rate. Let ϕ be the average semi-call bit rate, and let d be the average semi-call duration. Then the multiplexed traffic bit-rate is $\lambda\phi d$. For example, the average bit-rate across all semi-calls of the *complete-call data*, which is the sum of the bits of all semi-calls divided by sum of their durations, is $\phi = 0.04465$ megabits/sec. The average duration of all semi-calls is $d = 165.24$ sec. To generate the multiplexed packet traffic with a bit-rate of 3 megabits/sec in Section 13, λ was taken to be 0.4066 arrivals per sec.

14.2. The semi-empirical model

The semi-empirical model generates traffic in the following way.

- Semi-call arrival times a_k are generated by the above non-homogeneous Poisson process.
- Empirical semi-call k for arrival a_k is randomly sampled from the 277540 semi-calls of the *complete-call data*. Its packet arrival times are translated so that the time for the first packet is 0.
- The packet arrival times for the kth generated semi-call are a_k plus the translated empirical times of empirical semi-call k.
- The generated multiplexed traffic packet arrival times are the superposition of the packet arrival times of the generated semi-calls.

There are a number of possibilities for the sampling frame for the semi-calls of the semi-empirical model. To follow the independence assumptions of the model, verified from our analyses, we need to have the vector of random variables c_k,

t_{kr}, and s_{kr} for $r = 1, 2, \ldots$ be independent from one semi-call k to the next. We can insure this by randomly choosing one of the two empirical semi-calls of each empirical call (caller-to-callee or callee-to-caller), and then sampling without replacement from these semi-calls. This sampling frame does not allow a semi-call to appear more than once, and does not allow both semi-calls of a single call to appear. The two semi-calls of a single call are associated because the durations are the same, and when one side is transmitting the other side tends not to. Of course, sampling a semi-call two or more times also results in dependence between different semi-calls.

However, random sampling with replacement, desirable because it can serve simulations of long time periods, is unlikely to create a dependency that has more than minor effects on the statistical properties of the multiplexed traffic process. The on-off intervals have no significant synchronized behavior and the on and off intervals very quickly become out of phase. So in generation, a sampled semi-call followed by itself or by the other semi-call of the same call, would not cause significant positive or negative correlations unless the second call appeared almost simultaneously with the first. This is highly unlikely. The semi-call arrival rate that would result in 100 megabits/sec of multiplexed traffic, a large load for a simulation study, has 13.55 arrivals per second, and there are 277540 semi-calls in the *complete-call data*.

14.3. The mathematical semi-call model

The mathematical model generates traffic in the following way.

1. Semi-call arrival times a_k are generated by the above non-homogeneous Poisson process.
2. For the kth semi-call, generate a semi-call duration c_k from the mixture of piece-wise Weibull distributions specified in Table 2.
3. Generate a log base 2 shape parameter, α_k, and a log base 2 scale parameter, β_k, from the bivariate normal distribution for the parameters of $\sqrt{t_{kr}}$ described in Table 3.
4. Form $\alpha_k^\star = 2^{\alpha_k}$ and $\beta_k^\star = 2^{\beta_k}$, the shape and scale parameters for the $\sqrt{t_{kr}}$.
5. Generate values of $\sqrt{t_{kr}}$ from a gamma distribution with parameters $\alpha_k^\star$ and $\beta_k^\star$, square to get t_{kr}, and add 60 ms to get $\dot{t}_{kr}$, the sequence of on lengths for the kth semi-call.
6. Carry out steps, similar to 3 to 5, to generate values of $\dot{s}_{kr}$, the sequence of silence durations, but with 60 ms replaced by 15 ms.
7. The values of $\dot{t}_{kr}$ and $\dot{s}_{kr}$ alternate; generation stops when their sum is approximately c_k.
8. Packets for a transmission interval begin at the start of the interval, occur every 20 ms, and end when there is less than 20 ms remaining.
9. Packets for a silence interval begin at the start of the interval, occur every 2 sec, and end when there is less than 2 sec remaining.

10. The generated multiplexed traffic packet arrival times are the superposition of the packet arrival times of the generated semi-calls.

14.4. Strengths and weaknesses

The strength of the mathematical model is that it reproduces statistical effects in the data reasonably well and is far more parsimonious than the semi-empirical model. The weakness is more assumptions and less verisimilitude; for example, the low-frequency trends in transmission and silence intervals are not modeled, although this is likely not critical since the magnitudes of the trend variations are small. The semi-empirical model is more robust, not as dependent on model assumptions; for example, there is no modeling of durations and the on-off process beyond establishing assumptions of independence and identical distribution. The weakness is that it is not parsimonious, requiring the values of the *complete-call data*. One way to check the accuracy of the mathematical model for QoS studies, beyond the scope of this work, would be to use the semi-empirical and the mathematical models for queueing simulations in test runs, and compare the resulting delay and jitter.

References

Agrawal, S., Narayan, P. P. S., Ramamirtham, J., Rastogi, R., Smith, M., Swanson, K., and Thottan, M. Voip service quality monitoring using active and passive probes. In *First International Conference on Communication System Software and Middleware (Comsware 2006)*, pages 1–10, 2006.

Agrawal, S., Kanthi, C. N., Naidu, K. V. M., Ramamirtham, J., Rastogi, R., Satkin, S., and Srinivasan, A. Monitoring infrastructure for converged networks and services. *Bell Labs Technical Journal*, 12:63–77, 2007.

Andrews, D. F., Bickel, P. J., Hampel, F. R., Huber, P. J., Rogers, W. H., and Tukey, J. W. *Robust Estimates of Location: Survey and Advances*. Princeton University Press, 1972. MR0331595

Anscombe, F. J. Graphs in Statistical Analysis. *American Statistician*, 27:17–21, 1973.

Arlos, P. and Fiedler, M. A comparison of measurement accuracy for DAG, tcpdump and windump. www.its.bth.se/staff/pca/, 2003. verified August 2009.

Avaya ExpertNet. Avaya ExpertNet voip assessment tool. www.avaya.com.

Babu, T. V. J. G. and Hayes, J. F. *Modeling and Analysis of Telecommunications Networks*. John Wiley & Sons, 2004. ISBN 9780471348450.

Barford, P. and Crovella, M. Generating representative web workloads for network and server performance evaluation. In *Proceedings of the 1998 ACM SIGMETRICS Joint International Conference on Measurement and Modeling of Computer Systems*, pages 151–160, 1998.

Baset, S. A. and Schulzrinne, H. G. An analysis of the Skype peer-to-peer Internet telephony protocol. In *Proceedings of the 25th IEEE International Conference on Computer Communications (INFOCOM 2006)*, pages 1–11, 2006.

Belottia, P., Caponeb, A., Carellob, G., and Malucelli, F. Multilayer mpls network design: The impact of statistical multiplexing. *Computer Networks*, 52:1291–1307, 2008.

Beran, J. *Statistics for Long-Memory Processes*. Chapman & Hall, 1994. ISBN 0412049015. MR1304490

Biernacki, A. VoIP source model based on the hyperexponential distribution. *Proceedings of World Academy of Science, Engineering and Technology*, 11:202–206, 2006.

Birke, R., Mellia, M., Petracca, M., and Rossi, D. Understanding VoIP from backbone measurements. In *INFOCOM 2007: The 26th IEEE International Conference on Computer Communications.*, pages 2027 – 2035, 2007.

Bolotin, V. A. Modeling call holding time distributions for CCS network design and performance analysis. *IEEE Journal on Selected Areas in Communications*, 12:433–438, 1994.

Box, G. E. P. Science and statistics. *Journal of the American Statistical Association*, 71:791–799, 1976. MR0431440

Brady, P. T. A statistical analysis of on-off patterns in 16 conversations. *Bell System Technical Journal*, 47:73–91, January 1968.

Brady, P. T. A model for generating on-off speech patterns in two-way conversation. *Bell System Technical Journal*, 48:2445–2472, September 1969.

Brown, L., Gans, N., Mandelbaum, A., Sakov, A., Shen, H., Zeltyn, S., and Zhao, L. Statistical analysis of a telephone call center: A queueing-science perspective. *Journal of the American Statistical Association*, 100:36–50, 2005. MR2166068

Cao, J., Cleveland, W. S., Lin, D., and Sun, D. X. On the nonstationarity of Internet traffic. In *Proceedings of the 2001 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems*, pages 102–112, 2001.

Cao, J., Cleveland, W. S., Lin, D., and Sun, D. X. Internet traffic tends toward Poisson and independent as the load increases. In *Nonlinear Estimation and Classification*, pages 83–109. Springer, 2002. MR2005785

Casilari, E., Montes, H., and Sandoval, F. Modelling of voice traffic over IP networks. In *Third International Symposium on Communications Systems Networks and Digital Signal Processing (CSNDSP 2002)*, pages 411–414, Staffordshire, UK, 2002.

Choe, J. and Shroff, N. B. Queueing analysis of high-speed multiplexers including long-range dependent arrival processes. In *Proceedings of the 18th IEEE International Conference on Computer Communications (INFOCOM)*, pages 617–624, 1999.

Ciullo, D., Mellia, M., and Meo, M. Traditional IP measurements: What changes in a today multimedia IP network. In *Telecommunication Networking*

Workshop on QoS in Multiservice IP Networks, pages 262–267. IT-NEWS 2008, 2008.

Cleveland, W. S. and Devlin, S. J. Locally-weighted fitting: An approach to fitting analysis by local fitting. *Journal of the American Statistical Association*, 83:596–610, 1988.

Cox, D. R. Long-range dependence: A review. In *Statistics: An Appraisal*, pages 55–74. The Iowa State University Press, 1984.

Cox, D. R. and Isham, V. *Point Processes*. Chapman & Hall, 1992. ISBN 0412219107. MR0598033

Crovella, M. E. and Bestavros, A. Self-similarity in World Wide Web traffic: Evidence and possible causes. *IEEE/ACM Transactions on Networking*, 5:835–846, 1997.

Dang, T. D., Sonkoly, B., and Molnar, S. Fractal analysis and modeling of VoIP traffic. In *11th International Telecommunications Network Strategy and Planning Symposium*, pages 217–222, Vienna, Austria, 2004.

Daniel, C. and Wood, F. *Fitting Equations to Data*. Wiley, New York, 1971.

de Pereira, F. M., da Fonseca, N. L. S., and Arantes, D. S. On the performance of generalized processor sharing servers under long-range dependent traffic. *Computer Networks*, 40:413–431, 2002.

Erramilli, A., Narayan, O., and Willinger, W. Experimental queueing analysis with long-range dependent packet traffic. *IEEE/ACM Transactions on Networking*, 4:209–223, 1996.

Fraleigh, C., Moon, S., Lyles, B., Cotton, C., Khan, M., Moll, D., Rockell, R., Seely, T., and Diot, C. Packet-level traffic measurements from the Sprint IP backbone. *IEEE Network*, 17:6–16, 2003a.

Chuck Fraleigh, Fouad Tobagi, and Christophe Diot. Provisioning ip backbone networks to support latency sensitive traffic. In *Proceedings of the 22nd IEEE International Conference on Computer Communications (INFOCOM)*, 2003b.

Gruber, J. A comparison of measured and calculated speech temporal parameters relevant to speech activity detection. *IEEE Transactions on Communications*, COM-30:728–738, 1982.

Guha, S., Hafen, R. P., Kidwell, P., and Cleveland, W. S. Visualization databases for the analysis of large complex datasets. *Journal of Machine Learning Research*, 5:193–200, 2009.

Guha, S. RHIPE: The R and Hadoop integrated programming environment. http://ml.stat.purdue.edu/rhipe.

Hafen, R. P. and Cleveland, W. S. The ed method for nonparametric density estimation and diagnostic checking. Technical report, Department of Statistics, Purdue University, 2009.

Hosking, J. R. M. Fractional differencing. *Biometrika*, 68:165–176, 1981. MR0614953

Jiang, W. and Schulzrinne, H. Analysis of on-off patterns in VoIP and their effect on voice traffic aggregation. In *Proceedings of the Ninth IEEE*

International Conference on Computer Communication and Network, pages 82–87, Las Vegas, Nevada, October 2000.

Karapantazis, S. and Pavlidou, F.-N. VoIP: A comprehensive survey on a promising technology. *Computer Networks*, 53:2050–2090, 2009.

Kesidis, G. *An Introduction to Communication Network Analysis*. Wiley-IEEE Press, 2007.

Lee, H. H. and Un, C. K. A study of on-off characteristics of conversational speech. *IEEE Transactions on Communications*, COM-34:630–637, 1986.

Leland, W. E., Taqqu, M. S., Willinger, W., and Wilson, D. V. On the self-similar nature of Ethernet traffic (extended version). *IEEE/ACM Transactions on Networking*, 2:1–15, 1994.

Markopoulou, A., Tobagi, F. A., and Karam, M. J. Assessing the quality of voice communications over Internet backbones. *IEEE/ACM Transactions on Networking*, 11:747–760, 2003.

Massoulie, L. and Simonian, A. Large buffer asymptotics for the queue with fbm input. *Journal of Applied Probability*, 36:894–906, 1999. MR1737061

Norwine, A. C. and Murphy, O. J. Characteristic time intervals in telephone conversation. *Bell System Technical Journal*, 17:281–291, 1938.

Paxson, V. and Floyd, S. Wide-area traffic: The failure of Poisson modeling. *IEEE/ACM Transactions on Networking*, 3:226–244, 1995.

Peterson, L. L. and Davie, B. S. *Computer Networks: A Systems Approach*. Morgan Kaufmann, 1999.

Riedi, R. H., Crouse, M. S., Ribeiro, V. J., and Baraniuk, R. G. A multifractal wavelet model with application to network traffic. *IEEE Transactions on Information Theory*, 45:992–1018, 1999. MR1682524

Rolls, D. A., Michailidis, G., and Hernández-Campos, F. Queueing analysis of network traffic: Methodology and visualization tools. *Computer Networks*, 48(3):447–473, 2005.

Rosenberg, J., Schulzrinne, H., Camarillo, G., Johnston, A., Peterson, J., Sparks, R., Handley, M., and Schooler, E. SIP: Session initiation protocol. IETF RFC 3261, June 2002.

Schulzrinne, H., Casner, S., Frederick, R., and Jacobson, V. RTP: A transport protocol for real-time applications. IETF RFC 3550, July 2003.

Shin, S. and Schulzrinne, H. Measurement and analysis of the voip capacity in ieee 802.11 wlan. *IEEE Transactions on Mobile Computing*, 8:1265–1279, 2009.

Suh, K., Figueiredo, D. R., Kurose, J. F., and Towsley, D. F. Characterizing and detecting Skype-relayed traffic. In *Proceedings of the 25th IEEE International Conference on Computer Communications (INFOCOM 2006)*, pages 1–12, 2006.

Tobagi, F. A., Markopoulou, A. P., and Karam, M. J. Is the internet ready for voip? In *Proceedings of Distributed Computing, Mobile and Wireless Computing 4th International Workshop (IWDC)*, pages 49–57. Microsoft Press, 2002.

Toral-Cruz, H. and Torres-Roman, D. Traffic analysis for ip telephony. In *Proceedings of the 2nd International Conference on Electrical and Electronics Engineering*, 2005.

Willinger, W., Taqqu, M. S., Sherman, R., and Wilson, D. V. Self-similarity through high-variability: Statistical analysis of Ethernet LAN traffic at the source level. *IEEE/ACM Transactions on Networking*, 5:71–86, 1997.

Zhang, G., Xie, G., Yang, J., Zhang, D., and Zhang, D. Self-similar characteristic of traffic in current metro area network. In *Proceedings of the 15th IEEE Workshop on Local and Metropolitan Area Networks*, pages 176–181, Princeton, New Jersey, USA, 2007.