Measuring the Dynamical State of the Internet: Large-Scale Network Tomography via the ETOMIC Infrastructure

Gábor Simon^{a, b} József Stéger^{a, b} Péter Hágá^a
István Csabai^{a, b} Gábor Vattay^{a, b}

^aDepartment of Physics of Complex Systems, Eötvös Loránd University and ^bCollegium Budapest Institute for Advanced Study, Budapest, Hungary

Key Words

Complex networks • Internet • Network tomography • Queuing delay distributions

Abstract

In this paper we show how to go beyond the study of the topological properties of the Internet, by measuring its dynamical state using special active probing techniques and the methods of network tomography. We demonstrate this approach by measuring the key state parameters of Internet paths, the characteristics of queuing delay, in a part of the European Internet. In the paper we describe in detail the ETOMIC measurement platform that was used to conduct

Simplexus

Since the 1970s, the Internet has evolved rapidly from a few interconnected local area networks to a complex global network of routers both wired and wireless. This network has rapidly become embedded in the fabric of society. People use it for everything from email and web browsing to file transfer and Internet telephony. And its importance is likely to increase.

Which is why, in recent years, the study of the Internet and the way it behaves has taken on an increasing importance. However, the decentralized way in which the Internet has grown means that it is hard to study.

Most of the work on characterizing the Internet has focused on properties such as the pattern in which routers are linked together and how this pattern affects the Internet's tolerance to attacks and failures. Less work has been done on the properties of the Internet that change over time, such as the way patterns of congestion change in time.

This paper describes one of these dynamic properties: the delays experienced by data packets as they move across the Internet. It describes how to measure these delays and the results of the measurements they have made. These measurements have allowed the authors to study the distribution of queuing delays over a large part of the Internet in Europe and to draw up a map of network congestion. The authors also analyze the spatial structure of the Internet and identify highly congested or faulty segments.

Information is sent across the Internet using a protocol known as TCP/IP. This divides data into packets, numbers each packet and marks it with a destination address which routers use for forwarding. At the destination the data is reassembled and checked for missing packets. No network is perfect so packets are often delayed and sometimes even lost. The rate at which information can be transmitted is determined by these factors so being able to

the experiments, and the applied method of queuing delay tomography. The main results of the paper are maps showing various spatial structure in the characteristics of queuing delay corresponding to the resolved part of the European Internet. These maps reveal that the average queuing delay of network segments spans more than two orders of magnitude, and that the distribution of this quantity is very well fitted by the log-normal distribution.

Copyright © 2006 S. Karger AG, Basel

1 Introduction

The Internet is a complex network of computers and routers connected by direct wired or wireless links, where information is transmitted in discrete sized packets. Since 1970 the Internet evolved rapidly and in a decentralized manner from a few interconnected local area networks into a highly heterogeneous global network spanning continents. The Internet provides various services and applications that became part of our everyday activity including web browsing, file transfer, multimedia, Internet telephony, and others. In the future it is expected that the reliance of the population on the firm functioning of the Internet and on the new emerging applications will increase dramatically, thus it is of utmost importance to study and to understand the properties of this huge network at various levels of abstraction. By exploring the characteristics of the present-day Internet, and the various phenomena associated with its traffic it may be possible to extrapolate that knowledge to predict the properties of the future Internet and to foresee its problems.

Traditionally the Internet was studied by various engineering approaches that concentrated mostly on practical aspects, however recently with the accumulation of large data sets, it was recognized that the Internet can also be studied from a complex systems perspective, which represents

a different level of abstraction. Complex systems [1] are large networks of interacting agents, where the emphasis of the studies are on the global properties, characterizing the system as a whole, by neglecting the fine details of the interactions. It is believed that microscopically vastly different complex systems have some universal features that are independent of the details. For example collective phenomena of congestion build-up are present both in the Internet and highway traffic, showing analogous dynamics on a global scale [2]. Also both of these systems are characterized by self-similarity [3, 4], and $1/f$ noise in the congested state [5–7].

Most of the existing work so far addressed questions concerning the topological properties of the Internet [8, 9] or related topics like tolerance to attacks or failures [10]. These studies established interesting characteristics like the power-law degree distributions of the Internet connectivity graph on the autonomous system and router level. Although as the network evolves this continues to be an open issue and a hot topic, nevertheless it is essential to extend the study of the Internet beyond the discovery of topological properties towards the characterization of its dynamical state. Steps have already been taken in this direction; for instance in Vazquez et al. [11] and de Menezes and Barabási [12] who revealed universal scaling in traffic intensity fluctuations.

From the point of view of the current, most widely used data transfer protocols like the variants of TCP, the most relevant state variables of the network are the loss rates and the delays encountered by data packets on a path, since these quantities control the transmission rate of the transfer protocol. There are two different techniques to measure loss rate and delay over the Internet; these are active probing and passive monitoring. Active probing involves injecting of probe packets into the network and analyzing the properties of the received probe stream.

measure them is an important tool for monitoring the Internet's health.

There are two ways of measuring loss rate and delay – actively and passively. Passive monitoring assesses the traffic over a specific segment of the Internet by measuring the loss rates and delays directly. This is tricky because it requires direct access to the segment in question. And since many segments of the Internet are commercially owned fibres, it is not always practical to get access to every part of it.

Active monitoring offers a way round this. It involves transmitting a stream of data packets over the Internet and measuring what arrives at the other end. A number of systems have been developed to do this such as Surveyor, Felix and AMP [see 13]. While better than passive monitoring, these systems have an important limitation – they measure characteristics such as losses over the entire journey but they cannot resolve these characteristics for the different segments that make up the route. What's more, these systems rely on cooperation from routers along the path to process the data packets. It will not be possible to rely on this in future. The rapid, decentralized nature of the Internet's expansion means that there is no way to ensure that routers will have the necessary capability to cooperate.

But there is a way to study the Internet at higher resolution using an active monitoring technique known as unicast network tomography. This works by sending data packets from a single source to many receivers at various locations on the Internet. By monitoring the arrival time of the packets it is possible to determine the end-to-end characteristics of the journey but the authors explain how it is also possible to tease apart the contribution made by each segment along the way.

The authors first consider a simplified network consisting of a source, a single branching node and two receivers, labelled 2 and 3 in figure 7. The difference between the transmission and arrival times of any

This technique is flexible, has a wide range of applicability, however its extensive usage may impose an unnecessary load on the network. On the contrary, passive monitoring does not impose any load, but it relies on the fact that the user must have access to the network element under investigation, which is generally not the case. In the past years several measurement platforms have been developed for conducting active measurements over the Internet (e.g. Surveyor, Felix, AMP [13]), however these platforms can provide only end-to-end information between the participating nodes, where the measured characteristic cannot be resolved on the parts of the end-to-end path. Also most of the existing active probing tools rely on extra cooperation of the routers in the path to process their packets. As the Internet continues to evolve towards more decentralized and heterogeneous administration, in the future the cooperation of the network elements can be foreseen to be limited to the basic process of just storing and forwarding incoming probe packets.

The solution to these problems is provided by network tomography, which is a special class of active-probing measuring techniques, that is able to resolve the end-to-end delay statistics [14, 15] and packet loss rates [16, 17] to internal segments of the paths. In general a tomography measurement made from a single source to a set of receivers admits the determination of the delay statistics and loss rates on each segment of an underlying logical tree that is spanned by the source and receiver nodes, and the branching nodes (nodes where the path of probes destined for different receivers diverge). By increasing the number of sources and receivers involved in a tomography measurement, the portion of the network for which state information can be resolved grows dramatically. Initially network tomography techniques were developed for the use with multicast probes [18–21], which requires the extra cooperation of the routers to sup-

port multicast functionality; however, later these approaches were also extended to the case of using unicast probes [14, 16, 17], and performing the measurements from multiple sources [22, 23], which makes unicast network tomography the most general tool to measure spatially resolved characteristics of an uncooperative Internet.

The main idea of unicast network tomography is to use back-to-back packet pairs, where each packet of a pair is destined to different receivers. As the packets of a pair traverse their paths, they experience the same network conditions on the common segment from the source to the branching node, which brings correlation into the time series of the end-to-end characteristics.

This inherent correlation property of such probe streams is the key to resolve the internal characteristics from the end-to-end measurements.

The delay experienced by a packet over an Internet path sums up from two non-negative components, a constant propagation delay and a time-varying component due to queuing in the buffers of routers. In this paper we are concerned with large-scale inference of queuing delay distributions in the Internet by performing extensive unicast network tomography measurements. The large-scale study of queuing delay distributions is motivated by the fact that this observable carries a vast amount of information about traffic properties and the state of congestion on the measured path. By resolving the queuing delay distributions from end-to-end measurements we can draw a map of congestion of the network segments, analyze spatial structure, and identify highly congested or faulty segments.

The rest of the paper is organized as follows. In the following section we describe ETOMIC, the measurement platform, where the experiments were conducted. Next we proceed with the presentation of the results of our preliminary large-scale

packet is a straightforward measure of the overall delay along the route. This delay is the result of two factors – the propagation time for the signal which is a constant for a given distance and delays caused by queuing at routers along the way which varies with time.

The authors assume that if they measure the travel time of many packets, at least one will experience no queuing delays at all on its journey. This shortest travel time equates to the propagation delay. Subtracting this figure from the travel time of another packet gives its queuing delay, a figure that is the sum of all the queuing delays experienced by the packet along its journey.

The power of network tomography is that it allows the authors to obtain the contribution to the delay made by each segment. The key to this is that although packets sent from a single source have different destinations, they travel along the same segments for at least part of their journey.

So let's go back to the simplified network consisting of a source, a branching node and two receivers labelled 2 and 3. The source sends a packet to receiver 2 and then immediately fires off a second packet to receiver 3 – a 'back-to-back packet pair'. Both packets first travel along the same segment to the branching node which then routes them respectively to receivers 2 and 3. Each packet experiences a delay in the first segment and then an additional delay on the way to either receiver 2 or 3.

The goal is to determine the delay on each segment using the data from a large number of packet pairs. The authors develop a method of inference for determining these delays which depends on the assumption that both packets experience the same delay on the common leg of the journey. They then generalize this method for larger trees with a greater number of branching nodes and receivers.

That's the theory. In practice, measuring the time it takes for a packet to travel from a source to a receiver is a difficult business

tomography measurement, while the detailed description of the queuing delay tomography method is given in the last section.

2 The ETOMIC Measurement Platform

To perform unicast queuing delay tomography in a real network environment poses several challenges. First in order to be able to measure true end-to-end delay, source and receiver nodes need to be synchronized to a common clock reference, and must stay in the synchronized state during the measurements. Second, the measuring infrastructure has to be very precise in order to be able to resolve the microsecond-scale queuing delay components associated with high-bandwidth (multi-Gigabit) links. The precision of commercial workstations are insufficient for this task, thus to achieve submicrosecond precision, a hardware solution is inevitable.

In the subproject of the European Union-sponsored EVERGROW [24] Integrated Project, we are developing a state of the art high-precision, synchronized measurement platform, the Evergrow Traffic Observatory Measurement InfrastruCture (ETOMIC) [25]. This platform among oth-

ers provides the ability to perform large-scale delay and loss tomography based on unicast probing techniques, and will be generally available and open to the public. Currently ETOMIC consists of 15 measuring nodes deployed at different locations in various European countries (see table 1), while in the future we plan to extend this number to 50 participating nodes. The measurement nodes and the network experiments with them are managed through a central management system that is accessible to the researchers through a web-based graphical user interface [26].

The schematics of an ETOMIC measurement node is displayed in figure 1. It is based on standard PC hardware, but also includes an Endace DAG 3.6GE card as the network monitoring interface, which is specifically designed for precise active and passive measurements [27]. These DAG cards provide very accurate time stamping of the probe packets, with a time resolution of 60 ns, and also advanced capabilities for transmission. A burst composed of several packets can be transmitted with precise user-defined interpacket timings. The measuring nodes are synchronized by GPS (Garmin GPS 35 HVS), that provides a PPS (pulse per second) reference signal directly to the DAG card. The accuracy of one-

since the transmission and arrival time of each data packet must be noted with great accuracy.

To do this job, the authors have developed a device called the Evergrow Traffic Observatory Measurement InfrastruCture or ETOMIC which allows accurate synchronized packet timing. The system currently consists of 15 measuring nodes placed at various locations around Europe but the authors plan to increase this number to 50 in future. The devices are controlled remotely over the Internet using a web interface and this allows the experiments to be carried out from a central location.

The nodes, essentially modified PCs, are time synchronized using GPS signals and capable of time stamping each data packet to a resolution of 60 ns. In the lab, the authors have found that the accuracy of delay measurements between two ETOMIC nodes is about 0.5 μ s, a result that is mainly limited by the performance of the GPS receivers. In reference 33, the authors have measured the performance of their analytical method in extensive realistic simulations as well as in controlled local area network experiments with ETOMIC nodes. These experiments showed an impressive agreement between the estimated and real queuing delays.

In this paper, the authors used 9 ETOMIC nodes to carry out a large scale tomography measurement involving 38 branching nodes and 93 network segments of the Internet in Europe. Of these 9 ETOMIC nodes, 8 operated simultaneously as senders and receivers of packets while one worked purely as a receiver. Each of the sources sent back-to-back packet pairs to all possible pairs of receivers with a gap between each pair of 1 ms and repeated this process many times.

The results show a number of interesting features:

- The average queuing delay across the network ranges over three orders of magnitude. The shortest delay of only 0.5 μ s

Table 1. The list of current ETOMIC nodes

Abbreviation	IP address	Location
SICS	193.10.64.81	Stockholm, Sweden
TELI	217.209.228.122	Stockholm, Sweden
ERIC	192.71.20.150	Stockholm, Sweden
UNAV	130.206.163.165	Pamplona, Spain
ASTN	134.151.158.18	Birmingham, UK
HUJI	132.65.240.105	Jerusalem, Israel
OVGU	141.44.40.50	Magdeburg, Germany
ROME	141.108.20.7	Rome, Italy
UNIV	193.6.205.10	Budapest, Hungary
COLB	193.6.20.240	Budapest, Hungary
ELTE	157.181.172.74	Budapest, Hungary
UPAR	193.55.15.203	Paris, France
SALZ	212.183.10.184	Salzburg, Austria
UBRU	193.190.247.240	Brussels, Belgium
CRET	147.27.14.7	Chania, Greece

Fig. 1. The measuring node is a standard PC equipped with a DAG 3.6GE network interface card, connected to a GPS unit.

Fig. 2. One-way delay measurement between two GPS-synchronized ETOMIC nodes over an empty link. The FWHM of the histogram indicates that the attainable precision of end-to-end delay measurements within ETOMIC is around 0.5 μ s.

way delay measurements between two ETOMIC nodes is found to be $\approx 0.5 \mu$ s, which is mainly limited by the performance of the GPS receivers. This result was obtained in a lab experiment before the deployment of the nodes, where we connected two ETOMIC nodes by an empty link, and transferred a long stream of probe packets between them. Figure 2 shows the histogram of the measured delay, where the bin size reflects the time resolution of the DAG cards.

Figure 3 depicts the connection topology between 9 measuring nodes, obtained by extensive traceroute measurements, where we extracted from the detailed traces the segments for which characteristics can be resolved by network tomography measurements. This arrangement involves 104 network segments and 42 branching nodes (among these 19 are situated in the GÉANT Multi-Gigabit European Academic Network [28]) with link speeds ranging from 2 to 15 Gb/s.

3 Large-Scale Queuing Delay Tomography

Here we report a large-scale tomography measurement that involved 9 ETOMIC measurement nodes, and a connectivity graph consisting of 38 branching nodes and 93 network segments. Among these 9 ETOMIC nodes 8 were simultaneously sources of outgoing back-to-back probe pairs, and receivers of the incoming probe packets, while one of the nodes was only used as a receiver. Each of the source nodes sent probe pairs consisting of small-sized UDP packets to all the possible pairs of receivers in a round-robin fashion with an inter-pair time of 1 ms, and repeated this process many times. This procedure finally resulted in data sets, each containing two time series of end-to-end delays with an approximate length of 10,000 elements.

These data sets comprised the input to the tomography method described in section 4 that yielded as an output the queuing delay distributions resolved for each segment contained in the connectivity graph.

represents the limit of the measurements, while the longest of 1 ms occurred over the segment that connects the Multi-Gigabit European Academic Network (GÉANT) to the Hebrew University of Jerusalem in Israel.

- The southern part of the European Internet tends to have the longest queuing delay with the longest delays occurring on segments ending at nodes at the Hebrew University of Jerusalem and the University of Navarra in Spain.
- The incoming data along a segment to any given node is subject to greater delays than the outgoing data. This is usually because the amount of incoming data requested by users at that node is greater than the outgoing load requested by users across the Internet.
- The shortest delays occur between branching nodes. This is not surprising since these internal segments are part of the gigabit backbone of the network.
- In plotting the length of queuing delays against their frequency (actually their complementary cumulative distribution) the authors find that average queuing delay follows a log-normal distribution.

The data also allowed the authors to create a map of a large part of the European Internet showing the average and standard deviations of queuing delays (see fig. 4 and 5). This, they say, is their main result.

Justin Mullins
enquiries@justinmullins.com

Fig. 3. The connection topology between 9 ETOMIC measurement nodes. The ellipse-shaped nodes on the edge are ETOMIC measurement nodes with abbreviations given in table 1, while box-shaped nodes in the interior of the graph are branching nodes. The arrows indicate the direction of probe packet flow on a given network segment. The boxes with an IP address of 62.40.X.X are nodes within GÉANT.

Since a given segment can be a part of different end-to-end paths, this fact made it possible to test the consistency of the results, as well as the averaging of the distributions obtained from different data sets, but attributed for the same segment. For

better visualization of the results we extracted the mean and the standard deviation of the queuing delay distributions. These dynamical state parameters are shown on top of the connectivity graph in figures 4 and 5.

3.1 Analysis of the Results

The results of figures 4 and 5 reveal some interesting structure. First of all the state variables of the segments span three orders of magnitude, ranging from the error limit of the delay measurements

STDV

Fig. 5. The connectivity graph coloured and labelled by the mean queuing delay, given in units of microseconds, for each network segment.

downloaded from the servers of the organization by clients situated elsewhere in the Internet. Looking at the spatial arrangement of the state variables, one can see that the internal segments that are connections between branching nodes consti-

tute a core which is characterized by the smallest values of the state variables. This is not surprising, since these are network segments in the gigabit backbone.

To analyze further the results in figure 6a we plot the complementary cumulative

distribution function of the average queuing delays on the different segments, that is closely fitted by

$$g(x) = \frac{1}{2} \left[1 - \operatorname{erf} \left(\frac{\ln x - m}{\sigma \sqrt{2}} \right) \right], \quad (1)$$

Fig. 6. **a** The complementary cumulative distribution of the average queuing delays is given by filled boxes with error bars, while the continuous line is a fit given by equation 1. **b** The standard deviations (STDV) as a function of the average queuing delays are plotted with error bars for all the resolved network segments of figure 4. Here the continuous line marks the diagonal.

Fig. 7. The labelling of the different nodes and network segments of the measurement tree.

where x is the average queuing delay of a network segment in microseconds, while σ and m are the fitting parameters. The implication of the very good fit with $\sigma \approx 1.42$ and $m \approx \ln(37.8 \mu\text{s})$ in figure 6a is that the average queuing delay of the different segments follows a log-normal distribution.

$$P(x) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{(\ln x - m)^2}{2\sigma^2}}$$

It is interesting that despite the features and spatial structures identified earlier, the data follows a smooth continuous function without segregation of the average queuing delays into clearly visible groups. In figure 6b we plot the standard deviations as a function of the average values of the queuing delay.

The figure indicates that the majority of the segments, with a data point near the diagonal, are characterized by a queuing delay distribution that is close to being exponential, while the segments with a large deviation above the diagonal may be characterized by a self-similar traffic flow and heavy-tailed distributions.

4 Description of the Method for Queuing Delay Tomography

In this section we provide the detailed description of the queuing delay tomography method that was applied to the measured end-to-end data in obtaining the results of the previous section.

To achieve a unique labelling of the different network segments of a measurement tree we use the arrangement illustrated in the example of figure 7. The source node is situated on the top of the hierarchy, labelled with '0', while branching nodes are put on a given level according to their 'distance' from the source. This distance is measured by the number of branching nodes a probe packet must pass to reach the target. Different nodes on the same hierarchical level are enumerated in the incremental order from the left to right. Using this scheme one can identify each node by a unique index, while the segments, in-

dicated by arrows in figure 7, can be identified by the indices of their target nodes. We denote the one-way delay experienced by the n -th probe on the i -th segment by $\tilde{X}_i(n)$ and the corresponding end-to-end delay from the source to the i -th receiver node by $\tilde{Y}_i(n)$. The queuing delays associated with these quantities are denoted by $X_i(n)$ and $Y_i(n)$. By dropping the n index we indicate the time series of the respective quantities.

4.1 Inference in the Two-Leaf Tree

Here we describe the method of queuing delay distribution inference in the basic case of the two-leaf tree, which is comprised of a source, two receivers, and a single branching node (see the part of fig. 7 until level 1). The generalization to larger trees will be treated in the next subsection. The known quantities are the time series of end-to-end queuing delays Y_2 and Y_3 obtained by minimum filtering the $\tilde{Y}_2$ and $\tilde{Y}_3$ time series:

$$\begin{aligned} Y_2(n) &= \tilde{Y}_2(n) - \min(\tilde{Y}_2), \\ Y_3(n) &= \tilde{Y}_3(n) - \min(\tilde{Y}_3). \end{aligned} \quad (3)$$

These are measured by a big number of packet pairs, where the packets in a pair are sent back-to-back from the source, and are destined to the two distinct receivers. We

count only those pairs where each of the packets has reached its destination, and assign the number of successful pairs by N . Above we made the assumption that in every end-to-end delay measurement at least one probe will experience no queuing, thus the minimum can be identified with the propagation delay on the given path, which can be subtracted to yield the delay component due to queuing. Measurement studies of references [29, 30] as well as our own experience show that in the current Internet link utilizations are low enough, so that the above assumption is justified. Using the labelling introduced in the previous subsection we have the trivial relations

$$\begin{aligned} Y_2(n) &= X_1(n) + X_2(n), \\ Y_3(n) &= X'_1(n) + X_3(n), \end{aligned} \quad (4)$$

where $X'_1(n)$ stands for the queuing delay on segment 1, experienced by that particular probe in the n -th pair, which was destined towards receiver node (equation 3). From this point we will assume perfect correlation of the delays experienced by the two packets in a pair on the common segment $X_1(n) = X'_1(n)$. In practice this ideal situation is approached if for all n , $X'_1(n) - X_1(n) \ll X'_1(n)$. The goal is to estimate the distribution of the unknown X_1 , X_2 , and X_3 time series, based on the knowledge of Y_2 and Y_3 .

This can be achieved by introducing the quantized versions of the queuing delays. By using the quantization rule $Y_i^d(n) = jq$, if $(j - 1/2)q < Y_i(n) \leq (j + 1/2)q$ one can map the end-to-end time series Y_i to their quantized versions Y_i^d that take values from the set $\{0, q, 2q, \dots, Bq\}$. With the similar quantization we can also introduce the quantized versions of the unknown time series of X_i^d . We choose the quantization parameters q and B so that they satisfy the relation $\max(Y_2, Y_3) < (B + 1)q$, which ensures that all the possible values of Y_i and X_i will be included in one of the bins defined by the quantization rule, and thus probabilities can be assigned to the

bins. We denote the probability of a queuing delay falling in the j -th bin on the i -th segment by $P_{i,j}$, and for each $i \in \{1, 2, 3\}$ we have

$$\sum_{j=0}^B P_{i,j} = 1.$$

The reason behind the introduction of the discrete quantities above was that using the $P_{i,j}$ quantities one can express the probabilities of finding the quantized end-to-end queuing delays in a given bin $P(l, m) = P(Y_2^d = lq, Y_3^d = mq)$, which can be estimated from the measured data. Assuming temporal and spatial independence of the probe delays these probabilities can be expressed by the convolution

$$P(l, m) = \sum_{k \in H} P_{1,k} P_{2,(l-k)} P_{3,(m-k)}, \quad (5)$$

where the set H is given by $\{B \geq k \geq 0\} \cap \{B \geq (l-k) \geq 0\} \cap \{B \geq (m-k) \geq 0\}$. Labelling the observed variables by $\mathbf{Y} \equiv \{Y_2^d(n), Y_3^d(n)\} (n \in \{1 \dots N\})$ and the set of unknown probabilities by $\boldsymbol{\theta} \equiv \{P_{i,j}\}$ finally the unknown probabilities $P_{i,j}$ can be determined from the property that their true value maximizes the log likelihood function

$$\log L(\mathbf{Y} | \boldsymbol{\theta}) = \sum_{n=1}^N \log P(Y_2^d(n), Y_3^d(n)), \quad (6)$$

where the probabilities $P(Y_2^d(n), Y_3^d(n))$ are given by equation 5 evaluated at the known quantized end-to-end queuing delay pairs. Although the log-likelihood function equation 6 cannot be maximized analytically, there are several numerical procedures to accomplish this goal. Coates and Nowak [14] suggest to apply the expectation-maximization (EM) algorithm [31].

The EM algorithm is an iterative method to find the maximum likelihood estimate of the unknown probabilities in the presence of some suitably defined hidden or unobserved variables. In our case the hidden variables are the quantized queuing delays on the internal segments of the two-leaf tree that we denote by $\mathbf{X} \equiv \{X_1^d(n), X_2^d(n), X_3^d(n)\} (n \in \{1 \dots N\})$. Marginalizing

over the hidden variables the condition to maximize equation 6 can be expressed as, for $\forall(i, j)$

$$\begin{aligned} & \frac{\partial \log L(\mathbf{Y} | \boldsymbol{\theta})}{\partial P_{i,j}} \\ &= \frac{\partial}{\partial P_{i,j}} \left[\sum_{n=1}^N \log \sum_{\mathbf{x} \in \mathbf{X}(n)} P(\mathbf{Y}(n), \mathbf{X}(n) | \boldsymbol{\theta}) + \sum_{k=1}^3 \lambda_k \left(1 - \sum_{j=0}^B P_{k,j} \right) \right] = 0. \end{aligned} \quad (7)$$

Above the λ_b quantities are Lagrange multipliers to account for the normalization constraint of the $P_{i,j}$ probabilities. Performing the derivations we arrive at

$$\begin{aligned} & \sum_{n=1}^N \frac{1}{\sum_{\mathbf{x} \in \mathbf{X}(n)} P(\mathbf{Y}(n), \mathbf{X}(n) | \boldsymbol{\theta})} \frac{\partial}{\partial P_{i,j}} \\ & \left(\sum_{\mathbf{x} \in \mathbf{X}(n)} P(\mathbf{Y}(n), \mathbf{X}(n) | \boldsymbol{\theta}) \right) - \lambda_i = \\ & \sum_{n=1}^N \left[\sum_{\mathbf{x}' \in \mathbf{X}(n)} \left(\frac{P(\mathbf{Y}(n), \mathbf{X}(n) | \boldsymbol{\theta})}{\sum_{\mathbf{x} \in \mathbf{X}(n)} P(\mathbf{Y}(n), \mathbf{X}(n) | \boldsymbol{\theta})} \frac{\partial}{\partial P_{i,j}} \right. \right. \\ & \left. \left. \log P(\mathbf{Y}(n), \mathbf{X}(n) | \boldsymbol{\theta}) \right) - \lambda_i = \right. \\ & \left. \sum_{n=1}^N \left[\sum_{\mathbf{x}' \in \mathbf{X}(n)} P(\mathbf{X}(n) | \mathbf{Y}(n), \boldsymbol{\theta}) \frac{\partial}{\partial P_{i,j}} \right. \right. \\ & \left. \left. \log P(\mathbf{Y}(n), \mathbf{X}(n) | \boldsymbol{\theta}) \right) - \lambda_i = 0 \right] \quad (8) \end{aligned}$$

In the above equation the joint probabilities $P(\mathbf{Y}(n), \mathbf{X}(n) | \boldsymbol{\theta})$ can be factorized as $P(\mathbf{Y}(n) | \mathbf{X}(n)) P(\mathbf{X}(n) | \boldsymbol{\theta})$. Since $P(\mathbf{Y}(n) | \mathbf{X}(n))$ does not depend on $\boldsymbol{\theta}$ thus its derivative with respect to $P_{i,j}$ is zero. Using the assumption of spatial and temporal independence we have

$$P(\mathbf{X}(n) | \boldsymbol{\theta}) = \prod_{i=1}^3 P(X_i^d(n)), \quad (9)$$

where $P(X_i^d(n))$ is the probability that the n -th probe experienced a quantized queuing delay of X_i^d on the i -th segment. Substituting equation 9 into equation 8 we may write

$$\begin{aligned} & \sum_{n=1}^N \left[\sum_{\mathbf{x}' \in \mathbf{X}(n)} P(\mathbf{X}(n) | \mathbf{Y}(n), \boldsymbol{\theta}) \frac{\delta(X_i^d(n) = jq)}{P_{i,j}} \right] - \lambda_i = \\ & \frac{1}{P_{i,j}} \sum_{n=1}^N P(X_i^d(n) = jq | \mathbf{Y}(n), \boldsymbol{\theta}) - \lambda_i = 0, \end{aligned} \quad (10)$$

After rearranging and using the normalization constraints

$$\sum_{j=0}^P P_{i,j} = 1.$$

finally we arrive at the following system of coupled equations. For $\forall(i,j)$

$$P_{i,j} = \frac{1}{N} \sum_{n=1}^N P(\mathbf{X}_i^d(n) = jq \mid \mathbf{Y}(n), \boldsymbol{\theta}), \quad (11)$$

where

$$P(X_i^d(n) = jq \mid \mathbf{Y}(n), \boldsymbol{\theta})$$

stands for the conditional probability of

$$\mathbf{X}_i^d(n) = jq,$$

given the observed values

$$\mathbf{Y}_2^d(n), \mathbf{Y}_3^d(n)$$

of the n -th packet pair. According to Bayes law these conditional probabilities can be expressed as

$$\begin{aligned} P(\mathbf{X}_1^d(n) = jq \mid \mathbf{Y}_2^d(n) = lq, \mathbf{Y}_3^d(n) = mq, \boldsymbol{\theta}) &= \frac{P_{1,j} P_{2,(l-j)} P_{3,(m-j)}}{P(l, m)}, \\ P(\mathbf{X}_2^d(n) = jq \mid \mathbf{Y}_2^d(n) = lq, \mathbf{Y}_3^d(n) = mq, \boldsymbol{\theta}) &= \frac{P_{1,(l-j)} P_{2,j} P_{3,(m-l+j)}}{P(l, m)}, \\ P(\mathbf{X}_3^d(n) = jq \mid \mathbf{Y}_2^d(n) = lq, \mathbf{Y}_3^d(n) = mq, \boldsymbol{\theta}) &= \frac{P_{1,(m-j)} P_{2,(l-m+j)} P_{3,j}}{P(l, m)}, \end{aligned} \quad (12)$$

where $P(l, m)$ is given by equation 5. The system of coupled equations 5, 11 and 12 can be solved iteratively. One starts by assigning initial values for all the $P_{i,j}$ probabilities (e.g. the evenly distributed probabilities

$$P_{i,j}^0 = 1/(B+1),$$

then equation 11 together with equations 5 and 12 provides the improved estimates

$$P'_{i,j}.$$

This process is iterated until a suitable criterion for convergence is met. We terminate the iterations if for $\forall(i,j)$

$$|P'_{i,j} - P_{i,j}| \leq \varepsilon \quad (13)$$

where ε is a chosen small number, e.g. 0.0001.

4.2 Inference in Arbitrary Trees

For an arbitrary large tree it can be shown that the queuing delay distribution on any segment can be either estimated directly by the inference algorithm developed for the two-leaf tree, or can be generated by numerical deconvolution of two directly inferred distributions. To resolve all the network segments by a minimal approach one must perform packet pair measurements in a way that all the branching nodes and all the receivers are visited at least in one of the measurements. We demonstrate this in the following example.

Consider the tree shown in figure 8. Performing the back-to-back packet pair measurements to the receiver pairs of (2,3) and (5,6) in a round-robin fashion, and applying the inference algorithm developed for the two-leaf tree estimates directly the probabilities $P_{1,j}$, $P_{2,j}$, $P_{3,j}$ and of $P_{(0 \rightarrow 4),j}$, $P_{5,j}$, $P_{6,j}$ ($j \in \{0,1, \dots, B\}$). Here we denoted by $(0 \rightarrow 4)$ the common segment originating in 0 and ending in branching node 4. The probabilities $P_{(0 \rightarrow 4),j}$ can be expressed by the convolution

$$P_{(0 \rightarrow 4),j} = \sum_{k \in G} P_{1,(j-k)} P_{4,k}, \quad (14)$$

where the set G is given by $\{B \geq k \geq 0 \mid B \geq (j-k) \geq 0\}$. This convolution can also be viewed as a matrix operation

$$P_{(0 \rightarrow 4),j} = \sum_{k=0}^B A_{j,k} P_{4,k}, \quad (15)$$

where the matrix element is $A_{j,k} = P_{1,(j-k)}$, if $k \in G$, and $A_{j,k} = 0$, if $k \notin G$. This way the problem of numerical deconvolution to obtain the remaining probabilities on segment 4 can be mapped to a matrix inversion problem. For this task the best-suited numerical method is the non-negative least squares (NNLS) algorithm [32], which guarantees that the resulting probabilities will be strictly non-negative.

In a previous study [33] we investigated the performance of the inference method described so far in extensive realistic simulation studies, as well as in controlled

Fig. 8. An example of a measurement tree with 4 receiver nodes.

LAN experiments with ETOMIC measurement nodes. These investigations revealed an impressive agreement of the estimated and the real queuing delay distributions, and an error in the estimation of the first moments being less than the value of the chosen bin size.

5 Conclusion

This paper presented dynamical state measurements of a part of the European Internet, conducted via the ETOMIC measurement infrastructure, using special active probing techniques and the methods of network tomography. This very precise and fully synchronized infrastructure meets the requirements needed to perform large-scale unicast tomography measurements, and can be viewed as the prototype of network testbeds that will be able to operate in the uncooperative Internet of the future. In this paper we have investigated a particularly important state variable of the Internet, the characteristics of queuing delay distributions on different network segments, which provides information about traffic properties and the state of congestion in the network. As the main results we presented maps of the averages and standard deviations of queuing delay, and identified in them various structures. We find that the average queuing delay of network segments spans three orders of magnitude, and its distribution function closely follows a log-normal distribution.

Acknowledgements

The authors thank the National Science Foundation (OTKA T37903), the National Office for Research and Technology (NKF 02/032/2004 and NAP 2005/KCKHA005) and the European Union IST FET COMPLEXITY EVERGROW Integrated Project for their partial support.

References

- Albert R, Barabási A-L: Statistical mechanics of complex networks. *Rev Mod Phys* 2002; 74: 47–97.
- Gábor S, Csabai I: The analogies of highway and computer network traffic. *Physica A* 2002; 307: 516–526.
- Leland WE, Taqqu MS, Willinger W, Willson DV: On the self-similar nature of ethernet traffic (extended version). *IEEE/ACM Trans Networking* 1994; 2: 1–15.
- Crovella ME, Bestavros A: Self-similarity in World Wide Web traffic: evidence and possible causes. *IEEE/ACM Trans Networking* 1997; 5: 835–846.
- Musha T, Higuchi H: The 1/f fluctuation of a traffic current on an expressway. *Jpn J Appl Phys* 1976; 15: 1271–1275.
- Csabai I: 1/f noise in computer network traffic. *J Phys A Math Gen* 1994; 27: 417–421.
- Takayasu M, Takayasu H, Sato T: Critical behaviors and 1/f noise in information traffic. *Physica A* 1996; 233: 824–834.
- Faloutsos M, Faloutsos P, Faloutsos C: On power-law relationships of the Internet topology. *Comput Commun Rev* 1999; 29: 251–262.
- Govindan R, Tangmunarunkit H: Heuristics for Internet Map Discovery. *Proceedings of the IEEE Infocom*, Tel Aviv, 2000.
- Albert R, Jeong H, Barabási A-L: Error and attack tolerance of complex networks. *Nature* 2000; 406: 378–382.
- Vazquez A, Pastor-Satorras R, Vespignani A: Large-scale topological and dynamical properties of the Internet. *Phys Rev E* 2002; 65: 066130.
- de Menezes MA, Barabási A-L: Fluctuations in network dynamics. *Phys Rev Lett* 2004; 92: 028701.
- Active Measurement Project, <http://amp.nlanr.net/>.
- Coates M, Nowak R: Network tomography for internal delay estimation. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Salt Lake City, 2001.
- Duffield N, Horowitz J, Presti FL, Towsley D: Network Delay Tomography from End-to-End Unicast Measurements. *Lecture Notes in Computer Science*. Berlin, Springer, 2001.
- Coates M, Nowak R: Network loss inference using unicast end-to-end measurement. *Proc ITC Conf IP Traffic, Modeling and Management*, Monterey, 2000.
- Duffield N, Presti FL, Paxson V, Towsley D: Inferring link loss using striped unicast probes. *Proceedings of the IEEE Infocom*, Anchorage, 2001.
- Cáceres R, Duffield NG, Moon SB, Towsley D: Inference of internal loss rates in the mbone. *Proceedings of the IEEE/ISOC Global Internet*, December 1999.
- Cáceres R, Duffield NG, Horowitz J, Towsley D, Bu T: Multicast-based inference of network-internal characteristics: accuracy of packet loss estimation. *Proceedings of the IEEE Infocom*, New York, 1999.
- Duffield N, Presti FL: Multicast inference of packet delay variance at interior network links. *Proceedings of IEEE Infocom*, Tel Aviv, 2000.
- Presti FL, Duffield N, Horowitz J, Towsley D: Multicast-based inference of network-internal delay distributions. *IEEE/ACM Transactions on Networking*, December 2002.
- Bu T, Duffield N, Presti FL, Towsley D: Network tomography on general topologies. *Proceedings of ACM Sigmetrics*, Marina del Rey, 2002.
- Rabbat M, Nowak R, Coates MJ: Multiple source, multiple destination network tomography. *Proceedings of the IEEE Infocom*, Hong Kong, 2004.
- EVERGROW homepage, <http://www.evergrow.org/>.
- ETOMIC homepage, <http://www.etomic.org/>.
- Morató D, Magaña E, Izal M, Aracil J, Naranjo F, Astiz F, Alonso U, Csabai I, Hágá P, Simon G, Stéger J, Vattay G: The European Traffic Observatory Measurement Infrastructure (ETOMIC): a testbed for universal active and passive measurements. *Proceedings of Tridentcom*, Trento, 2005.
- ENDACE homepage, <http://www.endace.com/>.
- GEANT homepage, <http://www.geant.net/>.
- Moon S, Skelly P, Towsley D: Estimation and removal of clock skew from network delay measurements. *Proceedings of Infocom*, New York, 1999.
- Paxson V: On calibrating measurements of packet transit times. *Proceedings of SIGMETRICS*, Madison, 1998.
- McLachlan GJ, Krishnan T: *The EM Algorithm and Extensions*. New York, Wiley, 1997.
- Lawson CL, Hanson BJ: *Solving Least Squares Problems*. Englewood Cliffs, Prentice Hall, 1974.
- Simon G, Hágá P, Vattay G, Csabai I: A flexible tomography approach for queueing delay distribution inference in communication networks. *Proceedings of IPSMoMe*, Warsaw, 2005.