Model-based Identification of Dominant Congested Links *

Wei Wei, Bing Wang, Don Towsley, Jim Kurose
Department of Computer Science
University of Massachusetts, Amherst, MA 01003

ABSTRACT

In this paper, we propose a model-based approach that uses periodic end-end probes to identify whether a "dominant congested link" exists along an end-end path. Informally, a dominant congested link refers to a link that incurs the most losses and significant queuing delays along the path. We begin by providing a formal yet intuitive definition of dominant congested link and present two simple hypothesis tests to identify whether such a link exists. We then present and examine several novel model-based approaches for identifying a dominant congested link that are based on interpreting probe loss as an unobserved (virtual) delay. We develop parameter inference algorithms for Hidden Markov Model (HMM) and Markov model with a hidden dimension to infer this virtual delay. Our validation using ns simulation and live Internet experiments demonstrate that this approach can correctly identify a dominant congested link with only a small amount of probe data. We further estimate the maximum queuing delay of the dominant congested link, once we identify that a dominant congested link exists.

Categories and Subject Descriptors

C.2.3 [**Computer-Communication Networks**]: Network Operations-Network monitoring, Network management

General Terms

Performance, Measurement

*This research was supported in part by the National Science Foundation under NSF grants ANI-0085848, ANI-9980552, ANI-9973092, ANI-9977635, EIA-0080119, EIA-0087945 and under a subcontract with the University of Florida, grant UF-EIES-0205003-UMA. Any opinions, findings, and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the funding agencies.

IMC'03, October 27–29, 2003, Miami Beach, Florida, USA.

Keywords

Dominant Congested Link, Bottleneck Link, Path Characteristic, End-end Inference, Model-based Identification

1. INTRODUCTION

Measurement and inference of end-end path characteristics have attracted a tremendous amount of attention in recent years. Properties such as the delay and loss characteristics of the end-end path [1], the minimum capacity and available bandwidth of the path [2, 3, 4, 5, 6] and the stationarity of the network [7] have been investigated. These efforts have improved our understanding of the Internet. They have also proved valuable in helping to manage and diagnose heterogeneous and complex networks. Furthermore, they have been exploited by several applications, such as server selection, overlay networks and streaming applications, to improve performance [8, 9].

In this paper, we study a specific end-end path characteristic, namely, whether a *dominant congested link* exists along an end-end path. Informally, a dominant congested link is one that produces most of the losses and significant queuing delays on the end-end path. A formal definition is deferred to a later section in the paper. We avoid using the term "bottleneck link", since this term has been defined in many different ways in the literature (see e.g. [4, 10, 11, 12]) and there is no consensus on its meaning. Later in the paper, we relate our definition of dominant congested link to the notion of a bottleneck link.

Identifying the existence of a dominant congested link is useful for traffic engineering. For example, when there are multiple paths from one host to another and all are congested, improving the quality along a path with one dominant congested link may require fewer resources than along a path with multiple congested links. Identifying if a path has a dominant congested link also helps us understand the dynamics of the network since the behavior of a network with a dominant congested link differs dramatically from one with multiple congested links. When modeling a network, people usually assume that there is a single congested link (e.g. [13, 14]). One reason for this is that it is a good starting point, since this assumption usually simplifies the analysis significantly. Another more important reason might be that it is widely believed to be true. However, there exist no large scale measurements supporting this assumption, and to the best of our knowledge, there is no practical and efficient methodology for this purpose. Our work aims to provide such a methodology.

When a dominant congested link exists, identifying the

existence of such a link requires distinguishing its delay and loss characteristics from those of the other links. Identifying such a link via direct measurement is only possible to the organization in charge of that network. Commercial factors often prevent an organization from disclosing the performance of internal links. Furthermore, as the Internet grows in both size and diversity, one organization may only be responsible for a subset of links on an end-end path. Some measurement techniques obtain internal properties of a path by using ICMP messages to query internal routers. *Traceroute* and *ping* are two widely used tools in this category. Some more advanced techniques use ICMP messages to measure per-hop capacity or delay [15, 16, 17]. However, we are not aware of any work using ICMP messages to measure per-hop loss rate. This might be because ICMP queries can be dropped at a router for reasons other than buffer overflow. Contrary to direct measurements using responses from routers, a collection of techniques infer internal loss rate and delay characteristics using end-end measurement [18, 19, 20]. Most existing work in this category, however, are unable to identify behavior of individual links.

In this paper, we propose a novel model-based approach to identify whether a dominant congested link exists along an end-end path using end-end measurements. We periodically send probes from one host to another so as to obtain a sequence of delay and loss values. The key insight in our approach is to utilize the queuing delay properties of the lost probes. For example, if one link along the path is solely responsible for all losses, then all lost probes have the property that they "see" a full queue at this link. We interpret a loss as an unobserved delay and discretize the delay values. Afterwards, we model the discretized delay sequence of all probes including those with missing values to infer if a dominant congested link exists. This model utilizes the delay and loss observations *jointly* for inference instead of the common approach of treating them separately. Furthermore, the model makes use of the correlation in the *entire* observation sequence instead of the very limited temporal correlation present in back-to-back packets. As we will see, the identification procedure only requires a short probing duration, in terms of minutes.

The following are the main contributions of this paper:

- We present a formal yet intuitive definition of dominant congested link and provide two simple hypothesis tests to identify the existence of dominant congested link along a path.
- Our model-based approach fully utilizes the information from the probing packets and enables very fast identification. Validation using *ns* simulation [21] and Internet experiments demonstrate that this approach can correctly identify the existence of a dominant congested link within minutes.
- As a result of the identification procedure, we provide a statistical upper bound on the maximum queuing delay of the dominant congested link once we identify that a dominant congested link exists.

Related work [10, 11] study how to detect *shared* congested links over multiple paths. Their focus, however, is different from ours. The work closest in spirit to ours is the loss pair approach to discover network properties [12, 22]. A loss pair is formed when two packets are sent close in time and only one of the packets is lost. Assuming that the two packets experience similar behaviors along the path, the packets not lost in the loss pairs are used to provide insights on network conditions close to the time that loss occurs. Although our work also uses properties of lost packets, it has different objectives from [12, 22]. The authors of [12] do not address the issue of identifying whether a dominant congested link exists but start by assuming that a bottleneck link exists along the path and use loss pairs to determine the maximum queuing delay at the bottleneck link. The authors of [22] use a hidden Markov model to classify if a packet loss occurs at a wired or a wireless part of the network based on the measurements of loss pairs. We focus on whether a dominant congested link exists along a path. Furthermore, our model-based approach differs significantly from the loss pair approach: our approach infers the properties of lost packets by utilizing the delay and loss observations jointly and the correlation in the entire observation sequence, instead of direct measurements from the loss pairs.

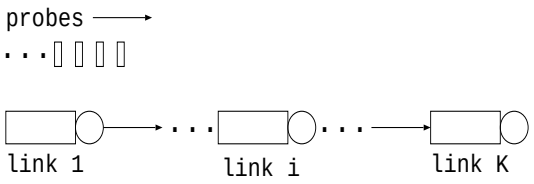


Figure 1: Periodic probes are sent along a path with K links to identify the existence of dominant congested link.

The rest of the paper is organized as follows. In Sections 2 and 3, we provide a formal definition of a dominant congested link and describe a methodology to identify whether a dominant congested link exists along a path. Section 4 presents the model-based approach. Section 5 validates the methodology using *ns* simulation and Internet experiments. Finally, Section 6 concludes the paper and describes future work.

2. DEFINITION OF DOMINANT CONGESTED LINK

In this section, we formally define a *dominant congested link* and relate it to the widely used term "bottleneck link". A bottleneck link is typically defined to be a link with either high loss rate, long queuing delay, high utilization or low link bandwidth; sometimes multiple perspectives are included in one definition [4, 10, 11, 12]. Although no consensus has been reached, the various definitions intuitively consider a bottleneck link (directly or indirectly) to be a link that produces most of the losses and/or significant queuing delays along a path.

We define dominant congested link based on the loss and end-end queuing delay experienced by the probes sent along the path of interest. Parameters are introduced to quantify the extent of loss and queuing delay that qualify a link as a dominant congested link. Furthermore, our definition from the end-end performance of the probes allows us to account for the dynamic nature of an end-end path.

Before providing a formal definition, we introduce some

Notation	Definition
K	Number of links/queues along the path
Q_k	The maximum queuing delay at queue k
D_t^k	Queuing delay for virtual probe t at link k
D_t	Aggregate queuing delay for virtual probe t over all the links along the path
L_k	Set of virtual probes marked as lost at link k
L	Set of virtual probes with loss marks
F_k	Set of virtual probes that experience the maximum queuing delay Q_k at link k
F	Set of virtual probes that experience the maximum queuing delay at some link along the path

Table 1: Key notations.

notations summarized in Table 1. Assume there are K links/routers along the path of interest, as shown in Figure 1. Each link/router is modeled by a droptail queue with a processing rate equal to the link bandwidth and the maximum queue size equal to the buffer size of the router. Let Q_k denote the maximum queuing delay at queue k, i.e., the time required to drain a full queue. Q_k is determined by the buffer size and the link bandwidth. Probes are sent periodically from the source to the destination. The existence of a dominant congested link in a time interval $[t_1, t_2)$, where t_2 can be ∞, is determined from measurements of the probes.

For ease of exposition, we introduce the concept of a *virtual probe*. A virtual probe goes through all the links along the path and records the delay (both propagation and queuing delay) at each link. If it "sees" a full queue when reaching router k, it records the maximum queuing delay Q_k and marks itself as lost. Otherwise, it calculates the queuing delay from the current queue length and the link bandwidth. The end-end delay for a virtual probe is the sum of its delays over all the links along the path. A virtual probe differs from a real probe in that it traverses all the links even if "lost" at some link. Furthermore, it does not occupy a position in the queue and hence does not affect packets that arrive afterwards. We refer to a virtual probe marked as lost at some link as a probe with a *loss mark*. Note that a virtual probe can be marked as lost at most once. This is consistent with the fact that a real probe can only be lost once.

Consider an arbitrary virtual probe sent at time t from the source, where $t \in [t_1, t_2)$. We use the sending time t to index the virtual probe. That is, we use virtual probe t to refer to a virtual probe that is sent at time t from the source. Let D_t^k be the queuing delay for virtual probe t at link k, $1 \leq k \leq K$. Let D_t be the aggregate queuing delay for virtual probe t over all the links along the path. That is, $D_t = \sum_{k=1}^{K} D_t^k$. Let L_k denote the set of the virtual probes marked as lost at link k. Define $L = \bigcup_{k=1}^{K} L_k$, the set of all virtual probes with loss marks. For virtual probe t, $t \in L_k$ indicates that this probe is marked as lost at link k; $t \in L$ indicates that this probe has a loss mark. We further define F_k to be the set of the virtual probes experiencing the maximum queuing delay, Q_k, at link k. That is, $F_k = \{t \mid D_t^k = Q_k\}$. We assume that the queuing delay for a probe taking the last available position at router k is Q_k. In practice, these two values are very close, since a probe is very small relative to the full queue size. Therefore, F_k contains all the probes that are either marked as lost at link k or take the last free buffer position at link k. Since $t \in L_k \Rightarrow D_t^k = Q_k$, we have $L_k \subseteq F_k$. Define $F = \bigcup_{k=1}^{K} F_k$, the set of the virtual probes that experience the maximum queuing delay at some link along the path. We then have $L \subseteq F$.

DEFINITION 2.1. *Link k is a* **strongly dominant congested link** *in time interval $[t_1, t_2)$ if and only if for a virtual probe sent at any time $t \in [t_1, t_2)$, the following two conditions are satisfied:*

$$P(t \in L_k \mid t \in L) = 1, \tag{1}$$

$$P(D_t^k \geq \sum_{i \neq k} D_t^i \mid t \in F_k) = 1. \tag{2}$$

If link k is a strongly dominant congested link, then condition (1) states that all losses occur at link k; condition (2) states that if a virtual probe experiences a maximum queuing delay at link k, this delay is no less than the aggregate queuing delays over all the other links. From this definition, it is easy to see that a strongly dominant congested link is unique.

The above definition incorporates both loss and delay into consideration, reflecting our sense that a dominant congested link is one that causes most losses and leads to significant queuing delays. Note that the condition on queuing delay is defined only over the virtual probes experiencing the maximum queuing delay at link k and not over all probes. This definition accounts for the dynamic nature of the network, since even a congested link may sometimes have very low queue occupancy. We next relax the strict delay and loss requirements in definition 2.1 and define a weaker notion of a dominant congested link.

DEFINITION 2.2. *Link k is a* **weakly dominant congested link** *with parameters θ and ϕ, where $0 \leq \theta < 0.5$ and $0 \leq \phi < 1$, in time interval $[t_1, t_2)$ if and only if for a virtual probe sent at any time $t \in [t_1, t_2)$, the following two conditions are satisfied:*

$$P(t \in L_k \mid t \in L) \geq 1 - \theta, \tag{3}$$

$$P(D_t^k \geq \sum_{i \neq k} D_t^i \mid t \in F_k) \geq 1 - \phi. \tag{4}$$

If link k is a weakly dominant congested link, then condition (3) states that a virtual probe is lost at link k with a probability no less than $1 - \theta$; condition (4) states that if a virtual probe experiences a maximum queuing delay at link k, this queuing delay is no less than the aggregate queuing delays over all the other links with a probability no less than $1 - \phi$. Since $0 \leq \theta < 0.5$, that is, more than half of the losses occur at a weakly dominant congested link, a weakly dominant congested link is unique.

Note that the lower the values of θ and ϕ, the more stringent are the requirements on being a weakly dominant congested link. In particular, the definition of a weakly dominant congested link is the same as that of a strongly dominant congested link when $\theta = \phi = 0$. A link identified as a weakly dominant congested link with θ and ϕ is also a weakly dominant congested link with θ' and ϕ', where $\theta' \geq \theta$ and $\phi' \geq \phi$. In particular, a strongly dominant congested link is a weakly dominant congested link with any $\theta \geq 0$ and $\phi \geq 0$.

2.1 Some special forms of dominant congested link

Definition 2.1 and 2.2 are based on the loss and queuing delays experienced by virtual probes. Some special forms of

dominant congested link can be easily identified if loss and queuing delay properties of the path are known.

PROPOSITION 2.1. *If link k is the link responsible for all of the losses and $Q_k \geq \sum_{i \neq k} Q_i$, then link k is a strongly dominant congested link.*

PROPOSITION 2.2. *If link k is responsible for at least $(1-\theta) \times 100\%$ of the losses and $Q_k \geq \sum_{i \neq k} Q_i$, then link k is a weakly dominant congested link with parameters θ and $\phi = 0$.*

We give an example to illustrate the above propositions. Suppose a path has 11 links. One link has bandwidth of 10 Mbps, denoted as link k. All the other links have bandwidth of 100 Mbps, each with maximum queuing delays of Q ms. Suppose all of the routers along the path have the same amount of buffer space. Then we have $Q_k = 10Q \geq \sum_{i \neq k} Q_i = 10Q$. If we furthermore know that losses only occur at link k, we classify link k as a strongly dominant congested link according to Proposition 2.1. If at least $(1-\theta) \times 100\%$ of the losses occur at link k, we classify link k as a weakly dominant congested link with parameters θ and $\phi = 0$ using Proposition 2.2.

2.2 Dominant congested link versus bottleneck link

Dominant congested link differs from bottleneck link [4, 10, 11, 12] mainly in the following aspects:

- Whether or not a link is a dominant congested link is relative. A link with a low loss rate is a dominant congested link as long as it satisfies the corresponding delay and loss requirements, despite the low loss rate.
- By definition, dominant congested link is unique while there may exist multiple bottleneck links along a path.
- Neither strongly nor weakly dominant congested link can describe links that do not have losses. Therefore, a bottleneck link without losses is not a dominant congested link. For instance, a link with the lowest bandwidth or highest utilization is not a dominant congested link if no loss occurs at that link.

3. IDENTIFICATION OF DOMINANT CONGESTED LINKS

In this section, we describe (i) two hypothesis tests to identify if a dominant congested link exists along a path and (ii) if it exists, how to estimate the maximum queuing delay of the dominant congested link.

We identify the existence of a dominant congested link using the queuing delay distribution of the virtual probes with loss marks, that is, the probes in set L. Since virtual probes do not exist in practice, we describe how to apply the identification methodology using *real* probes in Section 4.

Before describing the identification methodology, we first provide some intuition why the virtual probes in L are helpful for dominant congested link identification. First, the virtual probes in L will exhibit some common characteristics. For instance, if link k is a strongly dominant congested link, then any virtual probe t in L has $D_t \geq Q_k$ since this probe experiences the maximum queuing delay at link k. Secondly, the fact that L is a subset of F can be utilized for hypothesis testing. Suppose the null hypothesis is that a dominant congested link exists. If we observe that probes in L violate the delay condition for the dominant congested link (i.e., condition (2) or (4)), then we can reject the hypothesis. This is because if the hypothesis were true, then all probes in F, in particular, all probes in L, must satisfy the delay condition, which contradicts the observation.

For the purpose of dominant congested link identification, the queuing delays are discretized as follows. Let D_0 denote the end-end propagation delay along the path. For all virtual probes sent in the time interval $[t_1, t_2)$ (including those with and without a loss mark), we denote the largest end-end delay as D_{max}. The maximum queuing delay is therefore $D_{max} - D_0$. We then divide the interval $[0, D_{max} - D_0]$ into M equal length bins with bin width $b = (D_{max} - D_0)/M$. A queuing delay takes value in $\{1, 2, \ldots, M\}$, where i corresponds to an actual delay value between $(i-1)b$ and ib. Let W be a random variable representing the end-end queuing delay of virtual probes with loss marks. Let $F_W(w)$ represent the cumulative distribution function (CDF) of W, where $w = 1, 2, \ldots, M$. That is $F_W(w) = P(D_t \leq w \mid t \in L)$ for any virtual probe sent at time $t \in [t_1, t_2)$. We next summarize the properties of $F_W(w)$; these properties will form the basis for the hypothesis tests used to identify the existence of dominant congested link.

LEMMA 1. *If link k is a strongly dominant congested link, then $F_W(2Q_k) = 1$.*

PROOF. If k is a strongly dominant congested link, then virtual probe t satisfying $t \in L$ is lost at link k. This implies that it experiences the maximum queuing delay at link k. That is, $D_t^k = Q_k$. Therefore the end-end queuing delay of this probe $D_t = Q_k + \sum_{i \neq k} D_t^i \in [Q_k, 2Q_k]$. Since the probe is arbitrary, we have $W \in [Q_k, 2Q_k]$. Hence $F_W(2Q_k) = P(W \leq 2Q_k) = 1$. □

THEOREM 1. *Let $D = \min\{w \mid F_W(w) > 0\}$, where $w = 1, 2, \ldots, M$. If link k is a strongly dominant congested link, then $D \geq Q_k$ and $F_W(2D) = 1$.*

PROOF. If k is a strongly dominant congested link, then virtual probe t satisfying $t \in L$ experiences a queuing delay of Q_k at router k. Therefore, we have $W \geq Q_k$. Since D is the minimum delay value such that $F_W(D) > 0$, we have $D \geq Q_k$. Lemma 1 indicates that $F_W(2Q_k) = 1$. Since CDF $F_W(w)$ is a non-decreasing function, we have $F_W(2D) = 1$. □

LEMMA 2. *If link k is a weakly dominant congested link with parameter θ and ϕ, then $F_W(2Q_k) \geq (1-\theta)(1-\phi)$.*

PROOF. For arbitrary virtual probe packet t, we have

$$\begin{aligned}
& P(D_t \le 2Q_k \mid t \in L) \\
= \; & P(D_t \le 2Q_k, D_t^k = Q_k \mid t \in L) \\
& + P(D_t \le 2Q_k, D_t^k \ne Q_k \mid t \in L) \\
\ge \; & P(D_t \le 2Q_k, D_t^k = Q_k \mid t \in L) \\
= \; & P(D_t \le 2Q_k, t \in F_k \mid t \in L) \\
= \; & P(t \in F_k \mid t \in L) P(D_t \le 2Q_k \mid t \in F_k, t \in L) \\
\ge \; & P(t \in L_k \mid t \in L) P(D_t \le 2Q_k \mid t \in F_k, t \in L) \\
\ge \; & (1-\theta) P(D_t \le 2Q_k \mid t \in F_k, t \in L) \\
= \; & (1-\theta) P(Q_k + \sum_{i \ne k} D_t^i \le 2Q_k \mid t \in F_k, t \in L) \\
= \; & (1-\theta) P(\sum_{i \ne k} D_t^i \le Q_k \mid t \in F_k, t \in L) \\
= \; & (1-\theta) P(Q_k \ge \sum_{i \ne k} D_t^i \mid t \in F_k, t \in L) \\
= \; & (1-\theta) P(D_t^k \ge \sum_{i \ne k} D_t^i \mid t \in F_k, t \in L) \\
\ge \; & (1-\theta)(1-\phi)
\end{aligned}$$

The last inequality follows from the condition on delays for weakly dominant congested link. The above implies $F_W(2Q_k) \ge (1-\theta)(1-\phi)$. □

THEOREM 2. *Let $D = \min\{w \mid F_W(w) > \theta\}$, where $w = 1, 2, \ldots, M$. If link k is a weakly dominant congested link with parameters θ and ϕ, then $D \ge Q_k$ and $F_W(2D) \ge (1-\theta)(1-\phi)$.*

PROOF. We first prove $D \ge Q_k$ by contradiction. Suppose $D < Q_k$. Then for an arbitrary virtual probe with loss mark that is sent at time t, i.e., $t \in L$,

$$\begin{aligned}
& P(D_t \le D \mid t \in L) \\
= \; & P(D_t \le D, t \in L_k \mid t \in L) + P(D_t \le D, t \notin L_k \mid t \in L) \\
\le \; & P(D_t < Q_k, t \in L_k \mid t \in L) + P(D_t \le D, t \notin L_k \mid t \in L) \\
= \; & P(D_t \le D, t \notin L_k \mid t \in L) \\
= \; & P(D_t \le D \mid t \notin L_k, t \in L) P(t \notin L_k \mid t \in L) \\
\le \; & P(t \notin L_k \mid t \in L) \\
= \; & 1 - P(t \in L_k \mid t \in L) \\
\le \; & 1 - (1-\theta) \\
= \; & \theta
\end{aligned}$$

Since the probe is chosen arbitrarily from set L, we have $F_W(D) = P(W \le D) \le \theta$. However by the definition of D, we have $F_W(D) > \theta$, a contradiction. Therefore $D \ge Q_k$.

By Lemma 2, $F_W(2Q_k) \ge (1-\theta)(1-\phi)$. Since $F_W(w)$ is a non-decreasing function, we have $F_W(2D) \ge F_W(2Q_k) \ge (1-\theta)(1-\phi)$. □

We next give an example where $F_W(2D) = (1-\theta)(1-\phi)$. It indicates that the bound on $F_W(2D)$ in Theorem 2 can not be reduced. Suppose link k is a weakly dominant congested link with parameter θ and ϕ along a path. Figure 2 plots the probability mass function (PMF) of W for this path. The values of Q_k and D are marked on the x-axis. Any virtual probe marked as lost at link k has a queuing delay of at least Q_k. Since at least a fraction $1-\theta$ of the losses occur at link k, at least a fraction $1-\theta$ of the mass is to the right of Q_k. On the other hand, by the definition of D, less than a fraction $1-\theta$ of the mass is to the right of D. This provides an intuitive explanation why $Q_k \le D$ in Theorem 2. In Figure 2, we observe that the support of the PMF of W consists of three non-overlapping intervals. Suppose the left two intervals correspond to queuing delays of the virtual probes "lost" at link k containing $(1-\theta)(1-\phi)$ and $(1-\theta)\phi$ of the mass respectively. The total mass contained in these two intervals is $1-\theta$. By Lemma 2, the value of $2Q_k$ is to the right of the first interval. In Figure 2, the values of $2Q_k$ and $2D$ are marked on the x-axis. Both of them are between the first and the second interval. Therefore we have $F_W(2D) = (1-\theta)(1-\phi)$ in this example.

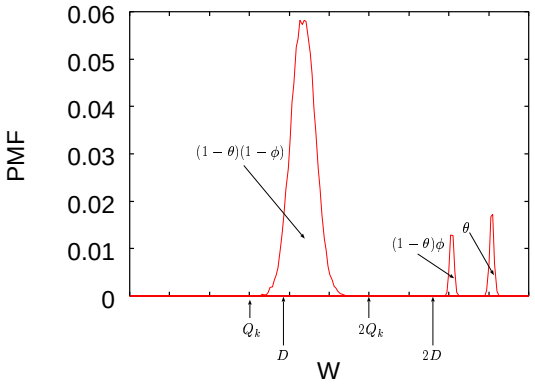


Figure 2: An example of the PMF of W for a path where link k is a weakly dominant congested link with parameter θ and ϕ.

The property stated in Theorem 1 forms the basis for the hypothesis test described in Figure 3 for a strongly dominant congested link, referred to as **Test 1**. The null hypothesis H_0 is that a strongly dominant congested link exists along a path. When the property in Theorem 1 is violated, we reject H_0. Otherwise, we accept it. Similarly, we have a hypothesis test for a weakly dominant congested link based on Theorem 2, as described in Figure 4 and is referred to as **Test 2**.

H_0: A strongly dominant congested link exists along the path.
Step 1: From $F_W(w)$, find $D = \min\{w \mid F_W(w) > 0\}$, where $w = 1, 2, \ldots, M$.
Step 2: If $F_W(2D) < 1$, we reject H_0. Otherwise, we accept H_0.

Figure 3: Test 1: Hypothesis test for a strongly dominant congested link.

Note that, as in all hypothesis tests, accepting the null hypothesis H_0 means that there is not sufficient evidence to reject H_0. It does not mean that H_0 is definitely true. However, for a non-trivial network ($K > 2$), when the probing duration is sufficiently long, accepting H_0 indicates that there is strong evidence that H_0 is true.

H_0: A weakly dominant congested link with parameters θ and ϕ exists along the path.

Step 1: From $F_W(w)$, find $D = \min\{w \mid F_W(w) > \theta\}$, where $w = 1, 2, \ldots, M$.
Step 2: If $F_W(2D) < (1-\theta)(1-\phi)$, we reject H_0. Otherwise, we accept H_0.

Figure 4: Test 2: Hypothesis test for a weakly dominant congested link.

3.1 Generalization of dominant congested link

We can generalize the definitions of the strongly and weakly dominant congested link by introducing a parameter $\rho > 0$ in their delay conditions. For strongly dominant congested links, the delay requirement in the generalized definition is

$$P(D_t^k \geq \frac{1}{\rho} \sum_{i \neq k} D_t^i \mid t \in F_k) = 1. \quad (5)$$

For weakly dominant congested links, the delay requirement in the generalized definition is

$$P(D_t^k \geq \frac{1}{\rho} \sum_{i \neq k} D_t^i \mid t \in F_k) \geq 1 - \phi. \quad (6)$$

Definition 2.1 and Definition 2.2 are special forms of the generalized definition using $\rho = 1$. From (5) and (6), it is clear that the smaller the value of ρ, the more stringent is the requirement of being a strongly or weakly dominant congested link. Parameter ρ makes the definition of dominant congested link more descriptive and flexible. For instance, when using $\rho = 1$, a link with a very low maximum queuing delay while being responsible for all losses along a path may not be classified as a strongly dominant congested link. However, when using a large value of ρ, this link can be classified as a strongly dominant congested link.

All the results described before can be extended to a more generalized form for the generalized definition of dominant congested link: Proposition 2.1 and 2.2 are extended by replacing $\sum_{i \neq k} Q_i$ with $\frac{1}{\rho} \sum_{i \neq k} Q_i$; Lemma 1 and 2 are extended by replacing $2Q_k$ with $(1+\rho)Q_k$; Theorem 1, 2 and the hypothesis tests are extended by replacing $2D$ with $(1+\rho)D$. These results of the generalized form can be proved in a similar manner. We only use $\rho = 1$ in this paper.

3.2 An example

We give an example to illustrate our hypothesis test for a strongly dominant congested link. Suppose a path contains $K (K > 2)$ links and at least two of which are lossy. Let I denote this set of lossy links. We assume all the links are independent and $P(D_t^i > 0) \in (0, 1)$, where $1 \leq i \leq K$. Therefore as time goes to infinity, for all the virtual probes with loss marks, the smallest queuing delay is $\min_{i \in I} Q_i$. That is, $D = \min_{i \in I} Q_i$ by the definition of D in Theorem 1. We will also observe a queuing delay of $\sum_{i \in I} Q_i + \epsilon$, where $\epsilon > 0$ is the queuing delay from all the links not in I. It is clear that $\sum_{i \in I} Q_i + \epsilon > 2D$ since there are at least two lossy links and the links are independent. Therefore, $F_W(2D) < 1$, which indicates that there is no strongly dominant congested link. Note that the above argument is true even if the multiple lossy links are identical in the sense that virtual probes experience the same queuing delays and loss at the links. This example also shows that the hypothesis test for strongly dominant congested link asymptotically provides a correct identification. In practice, we do not require time to go to infinity, but require that the duration be "sufficiently long", an issue we will return to in Section 5. The probing duration is important since observations made in short time intervals are correlated. Therefore, to obtain sufficiently many independent observations, we need a sufficiently long probing duration.

3.3 Estimating the maximum queuing delay at a dominant congested link

It is easy to estimate the maximum queuing delay at a strongly dominant congested link from $F_W(w)$, once we identify that such a link exists. Suppose we identify link k as a strongly dominant congested link. Then all losses occur at link k. From $F_W(w)$, we find the smallest value of delay D such that $F_W(D) > 0$. Since all losses occur at link k, by the definition of $F_W(w)$, $D \geq Q_k$, where Q_k is the maximum queuing delay at link k. Therefore, D is an upper bound of Q_k. In general, the above estimation method applies to any link (not necessarily a strongly dominant congested link) if we know that all losses occur at that link.

Suppose link k is weakly dominant congested with parameters θ and ϕ. Then D defined in Theorem 2 can be used as an upper bound of the maximum queuing delay at link k, Q_k, since $D \geq Q_k$ by the theorem. However, for link k with a very small value of θ, we can usually apply the following heuristic to obtain a tighter bound on Q_k from the PMF of W. When plotting the PMF of W, we choose the number of bins such that in the PMF we observe a connected component with most of the mass and the other components are as separated as possible from this component. Then this connected component corresponds to the distribution from the weakly dominant congested link by the nature of the weakly dominant congested link. We use the smallest delay value that has probability significantly larger than 0 in this component as the estimate of the maximum queuing delay of the weakly dominant congested link. If no separated components are distinguished from the PMF, D is used as the estimate. We illustrate this method using an example in Section 5.1.2 .

4. MODEL-BASED IDENTIFICATION OF DOMINANT CONGESTED LINK

In Section 3, the identification of a dominant congested link relies on $F_W(w)$, the queuing delay distribution of virtual probes with loss marks. In this section, we describe how this distribution can be obtained using *real* probes. A virtual probe differs from a real probe in that it has an end-end delay even if it is marked as lost in the middle while a real probe does not have a delay if it is lost (indeed, it is not received at the end point). Analogous to the queuing delay of a virtual probe, we associate a *virtual queuing delay* with a lost real probe as follows. When a probe is lost at link k, we assign it the maximum queuing delay at link k and a queuing delay based on the queue occupancy at each of remaining links.

Let W represent the virtual queuing delay of a lost probe and $F_W(w)$ present the CDF of W. As in Section 3, W is represented using discretized values. Denote the smallest and the largest end-end delays of all the probes not lost as D_{min} and D_{max} respectively. If the end-end propaga-

tion delay along the path D_0 is known, we divide the range $[0, D_{max} - D_0]$ into M equal length bins. Otherwise, we use D_{min} to approximate D_0. Our simulation and experiments in Section 5 shows that the inaccuracy caused by this approximation is negligible when the probing duration is longer than several minutes.

Two methods can be used to obtain $F_W(w)$: an *empirical method* and a *model-based method*. An example of the empirical method is by loss pairs [12] where the queuing delay of the packet not lost is used as the virtual queuing delay of the lost packet. This is not always accurate since, as shown in [12], the measurement noise (e.g., caused by cross traffics) renders the distribution obtained from loss pairs inaccurate. If a strongly dominant congested link k exists, the minimum delay D such that $F_W(D) > 0$ is required to satisfy $D \geq Q_k$ by Theorem 1. However, measurements from loss pairs cannot guarantee this property, since the successful packet in a loss pair does not necessarily experience the maximum queuing delay at link k because of cross traffic. A similar argument applies to weakly dominant congested links.

In this paper, we focus on the model-based approach, where we infer the virtual queuing delay distribution $F_W(w)$ using both measurements and a model. We experiment with three models. The first model uses simple linear interpolation. Suppose a probe sent at time t is lost. Assume the successful probes immediately before and after it are sent at time t_1 and t_2 respectively. Then the virtual queuing delay for probe t is $D_t = \frac{D_{t_2} - D_{t_1}}{t_2 - t_1}(t - t_1) + D_{t_1}$. This model calculates the virtual queuing delay for each lost probe, which is then used to obtain the distribution. We find that the results are inaccurate for some cases. An example is shown in Section 5.1.1. The second model is based on a hidden Markov model (HMM) [23]; the third one is based on the model introduced in our previous work [24], which in this paper we refer to as a Markov model with a hidden dimension. In both models, we interpret a loss as a delay with a missing value and develop expectation and maximization (EM) algorithms to obtain $F_W(w)$. We find that the results from the second model (i.e., HMM) are not accurate (see one example in Section 5.1.3). This is mainly because the state space of this model does not contain delay observations. Hence the correlation in the delay observations is not well captured. The third model captures this correlation more accurately. Due to space limitations, we only describe the third model in detail.

4.1 An EM algorithm to infer the virtual queuing delay distribution $F_W(w)$

We now describe the inference procedure for $F_W(w)$, using the third model, a Markov model with a hidden dimension. We refer to a discretized queuing delay observation as a delay symbol. Suppose there are M delay symbols and N hidden states in this model. Each state of the model Z_t contains two components: the hidden state $X_t \in \{1, 2, \ldots, N\}$ and the delay symbol $Y_t \in \{1, 2, \ldots, M\}$. That is, $Z_t = (X_t, Y_t)$. This state representation differs from what is used in a traditional HMM [23], where each state contains only the hidden component but not the observation.

Let π denote the initial distribution of the states. Let P denote the probability transition matrix. An element in the transition matrix P is denoted as $p_{(i,j)(k,l)}$, which represents the transition probability from state (i, j) to state (k, l). Note that the model reduces to a Markov model when $N = 1$. This is because when $N = 1$, every state in the model contains the same hidden state and is only differentiated by the delay symbol. Let y_t be the observation value for Y_t. If at time t, the observation is a loss, we regard it as a delay with a missing value and use $y_t = *$ to denote it. A loss observation has a certain probability of having a delay symbol of j, $1 \leq j \leq M$. Let $s(j)$ be the conditional probability that an observation is a loss given that its delay symbol is j. That is, $s(j) = P(y_t = * \mid y_t = j)$. Let $\lambda = (P, \pi, s)$ denote the complete parameter set of the model. An EM algorithm is an iterative procedure to infer λ from a sequence of T observations. It ends when a certain convergence threshold is reached. The detailed description of the EM algorithm is in Appendix A.

After obtaining the model parameters, we obtain $F_W(w)$ from $f_W(w)$, which denotes the PMF of W. That is, $f_W(w) = P(y_t = w \mid y_t = *)$ and $f_W(w)$ is computed by

$$f_W(w) = \frac{s(w) \sum_{t=1}^{T} \mathbf{1}(y_t = w)}{\sum_{t=1}^{T} \mathbf{1}(y_t = *)} \tag{7}$$

where $\mathbf{1}(\cdot)$ is the indicator function. This equation follows from Bayes formula: the numerator corresponds to the probability that a loss has delay symbol of w and the denominator corresponds to the probability of loss in the sequence of T observations. Note that $s(w)$ is obtained from the the EM algorithm, where the entire observation sequence is utilized as shown in the derivation. Therefore $F_W(w)$ is obtained by using the information in the entire observation sequence, not only the loss observations.

4.2 An alternative interpretation of the virtual queuing delay distribution

We now provide an alternative interpretation of the virtual queuing delay distribution. Suppose the observation sequence $\{y_t\}_{t=1}^{T}$ is generated by a Markov model with a hidden dimension. The model has the parameter space of $\lambda = (P, \pi, s)$. At time t, the model generates a delay symbol of j. With probability of $s(j)$, this delay symbol generates a loss observation; with probability of $1 - s(j)$, this delay symbol generates a delay observation of j. The virtual delay of a loss observation can then be alternatively regarded as a delay symbol from which the loss is generated.

5. VALIDATIONS

In this section, we validate the model-based identification method using both *ns* simulations and live Internet measurements. We further explore the effects of the parameters in the model (e.g., M, N, the convergence threshold in the EM algorithm, etc.) and the probing duration required for correct identification.

5.1 Validation using *ns* simulations

We use a topology containing four routers r_0, r_1, r_2 and r_3 in the *ns* simulation, as shown in Fig. 5. Link (r_i, r_{i+1}) denotes the link from router r_i to r_{i+1}, where $0 \leq i \leq 2$. The bandwidth and the maximum queue length of link (r_i, r_{i+1}) are varied to create different scenarios. For all other links (from a source or a sink to a router), the bandwidth is 10 Mbps and the maximum queue size is set so that no loss occurs. We add TCP flows from router r_i to r_j, where $0 \leq i < j \leq 3$. The number of TCP flows from router r_i to r_j ranges from 1 to 10. In addition, we create HTTP traffic

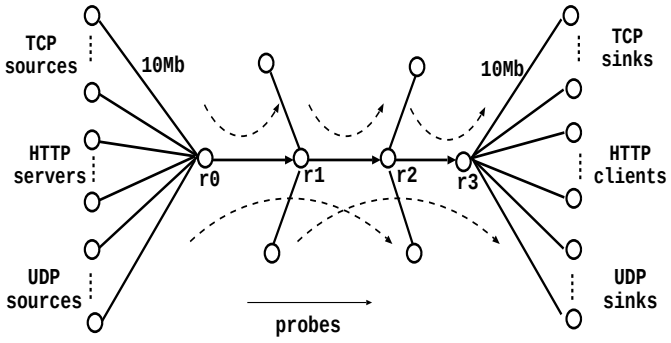


Figure 5: Topology used in *ns*.

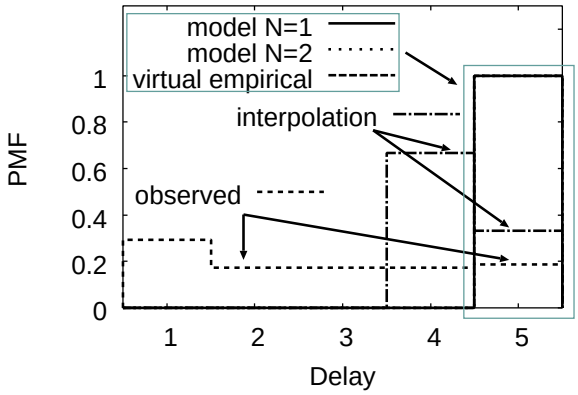


Figure 6: PMFs of the observed and virtual queuing delays for a setting in which link (r_0, r_1) is a strongly dominant congested link.

Bw(Mbps)	Loss rate	Max. queuing delay (ms)		
		Value	model	loss pair
0.1	3.3%	200	200	205
0.2	2.5%	100	101	101
0.4	0.04%	50	51	53
1.0	0.02%	20	22	21

Table 2: A strongly dominant congested link: Various bandwidths for link (r_0, r_1).

from router r_i to r_j, where $0 \leq i < j \leq 3$. The HTTP traffic is generated using the empirical data provided by *ns*. We further add UDP on-off traffic on link (r_i, r_{i+1}). The propagation delay of link (r_i, r_{i+1}) is 5 ms. The propagation delay from a source or a sink (TCP, HTTP or UDP) to its corresponding router is uniformly distributed in $[10, 20]$ ms. Probe packets are sent periodically along the path from r_0 to r_3 at an interval of 20 ms. The packet size of a probe packet is 10 bytes. Therefore, the traffic generated by the probing process is 4 Kbps, which is much smaller than the link bandwidths used in the simulation.

Each simulation runs for $2,000$ seconds. We first use the second $1,000$ seconds of the trace to identify whether there exists a dominant congested link. We then investigate the length of the probing duration required for accurate identification. We obtain the minimum and maximum end-end delay D_{min} and D_{max} from the probes in a selected interval. We assume the propagation delay along the path D_0 is unknown and use D_{min} to approximate D_0. We then discretize the range of queuing delay $[0, D_{max} - D_0]$ into 5 symbols, i.e., $M = 5$, with queuing delays taking values in $\{1, \ldots, 5\}$. We choose the number of hidden states N to be in the range of 1 to 4. The convergence threshold in the EM algorithm is 10^{-4} or 10^{-5}. Unless otherwise stated, we use the Markov model with a hidden dimension.

In *ns* simulation, we know the queue occupancy at each router at any time. Therefore, we are able to obtain the virtual queuing distribution empirically. We compare the virtual queuing distributions obtained empirically and from the models for validation. For a setting in which a dominant congested link exists, we estimate the maximum queuing delay at the dominant congested link using our model-based approach (see Section 3.3) and the loss pair approach [12]. To obtain an accurate estimation, we discretize delays more finely and use $M = 20$ or 40. When using the loss pair approach, packet pairs are sent along the path from r_0 to r_3 at an interval of 40 ms so that the estimations from our model-based approach and the loss pair approach are based on the same number of probes.

5.1.1 A strongly dominant congested link

We first investigate settings in which a strongly dominant congested link exists. In particular, we set the various parameters so that losses only occur at link (r_0, r_1). The buffer sizes at router r_0, r_1 and r_2 are 20 Kb, 80 Kb and 80 Kb respectively. The bandwidths of links (r_1, r_2) and (r_2, r_3) are both 10 Mbps. The bandwidth of link (r_0, r_1) is varied from 0.05 Mbps to 1 Mbps. Table 2 lists the loss rates at link (r_0, r_1) for various bandwidths settings. No loss occurs at other links. For all settings, the identification procedure accepts the null hypothesis that a strongly dominant congested link exists. These identification results are confirmed by Proposition 2.1. We further estimate an upper bound on the maximum queuing delay of the strongly dominant link, which is link (r_0, r_1). Table 2 lists the actual maximum queuing delay of link (r_0, r_1), and the estimates from our approach and the loss pair approach. Estimates from both approaches are close to the actual value in all the settings: the maximum error from our approach and the loss pair approach is 2 ms and 5 ms respectively.

We only describe one setting in Table 2 in detail. In this setting, the bandwidth of link (r_0, r_1) is 1 Mbps. Fig. 6 plots PMFs of the virtual queuing delay distributions from the models ($N = 1, 2$) and the empirical virtual queuing delay distribution. The various virtual queuing delay distributions are very close[1], demonstrating the accuracy of our model-based approach. From Fig. 6, $D = 5$ is the minimum delay such that $F_W(D) > 0$. Since $2D = 10 > M = 5$, we have $F_W(2D) = 1$. By Test 1, we accept the null hypothesis that a strongly dominant congested link exists.

Fig. 6 also plots the virtual queuing delay distribution obtained from the linear interpolation model. It deviates from the empirical virtual queuing delay distribution and leads to an underestimate of the maximum queuing delay. This indicates the inaccuracy of the linear interpolation model. In Fig. 6, the observed queuing delay distribution differs dramatically from the virtual queuing delay distribution. This is because the virtual queuing delay consists of the maxi-

[1]The three distributions all lie on $D = 5$.

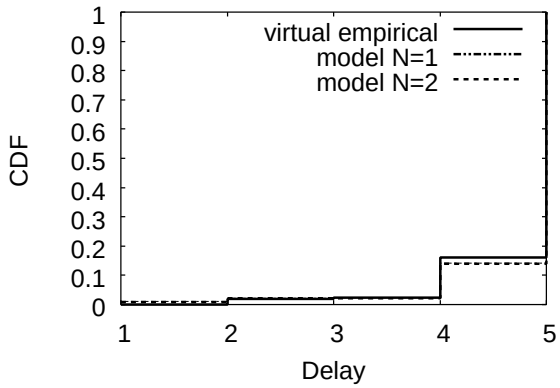


Figure 7: Distributions of the virtual queuing delays for a setting in which link (r_2, r_3) is a weakly dominant congested link.

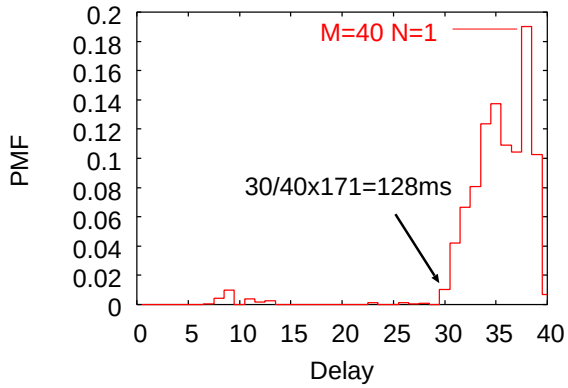


Figure 8: The virtual queuing delay distribution to estimate an upper bound on the maximum queuing delay for the weakly dominant congested link (r_2, r_3).

mum queuing delay of the link where it is marked as lost while the observed queuing delay does not have this property. Our identification method relies on the virtual queuing delay distribution, which cannot be replaced by the observed queuing delay distribution as shown by this example.

5.1.2 A weakly dominant congested link

We next investigate settings in which a weakly dominant congested link exists. In particular, we set the parameters such that losses occur only at link (r_0, r_1) and (r_2, r_3) and the loss rate at one link is significantly lower than that at the other link. The buffer sizes at router r_0, r_1 and r_2 are 25.6 Kb, 76.8 Kb and 25.6 Kb respectively. The link bandwidth of (r_1, r_2) is 1 Mbps. The link bandwidths of (r_0, r_1) and (r_2, r_3) with their corresponding loss rates are listed in Table 3.

The null hypothesis is that there exists a weakly dominant congested link with $\theta = 0.06$ and $\phi = 0$. That is, the requirements on the weakly dominant congested link are: at least 94% of the losses occur at this link; furthermore, when a probe experiences the maximum queuing delay at this link, 100% of the time this queuing delay is no less than the aggregate queuing delay that this probe experiences on the other links. The identification procedure accepts the null hypothesis for all the settings. These identification results are confirmed by Proposition 2.2.

We estimate an upper bound on the maximum queuing delay at the weakly dominant congested link, which is link (r_2, r_3). Table 3 lists the actual maximum queuing delay of link (r_2, r_3), and the estimates from our approach and the loss pair approach. The estimates from our approach are much closer to the actual value (with a maximum error of 9 ms) than those from the loss pair approach (with a maximum error of 100 ms). The reason for the deviation from the loss pair approach might be that this approach is sensitive to the queuing delays at links other than the weakly dominant congested link and/or requires longer probing durations.

We next describe one setting in Table 3 in detail. In this setting, the bandwidths of link (r_0, r_1) and (r_2, r_3) are 0.7 Mbps and 0.2 Mbps respectively. The average loss rates on link (r_2, r_3) is 3.8%, which accounts for 95% of the losses. Fig. 7 plots the distributions of the virtual queuing delays obtained empirically and from the various models. The various virtual queuing delay distributions are very similar. From Fig. 7, $D = 1$ is the minimum delay such that $F_W(D) > 0$, which is around 0.01 and not quite observable from the figure. Since $F_W(2D) = F_W(2) = 0.02 < 1$, by Test 1, no strongly dominant congested link exists along the path. For $\theta = 0.06$ and $\phi = 0$, $D = 4$ is the minimum delay such that $F_W(D) > \theta$. Since $2D = 8 > M = 5$, we have $F_W(2D) = 1$. By Test 2, we accept the hypothesis that there exists a weakly dominant congested link with $\theta = 0.06$ and $\phi = 0$. When using $\theta = 0.02$ and $\phi = 0$ as the parameters for the null hypothesis, the identification procedure rejects the hypothesis, which is correct since no link in this setting is responsible for more than 98% of the loss.

We use $M = 40$ and $N = 1$ to estimate an upper bound of the maximum queuing delay at the weakly dominant congested link, which is link (r_2, r_3). The PMF of the virtual queuing delay is shown in Fig. 8. According to the heuristic described in Section 3.3, the connected component with most of the mass (i.e., the one in the right of the figure) corresponds to the virtual queuing delay from the weakly dominant congested link. From Fig. 8, $D = 30$ is the minimum delay that is significantly different from 0. The queuing delay ranges from 0 to 171 ms. Therefore, an upper bound on the maximum queuing delay at link (r_2, r_3) is $30/40 * 171 = 128$ ms, which is the same as the actual maximum queuing delay.

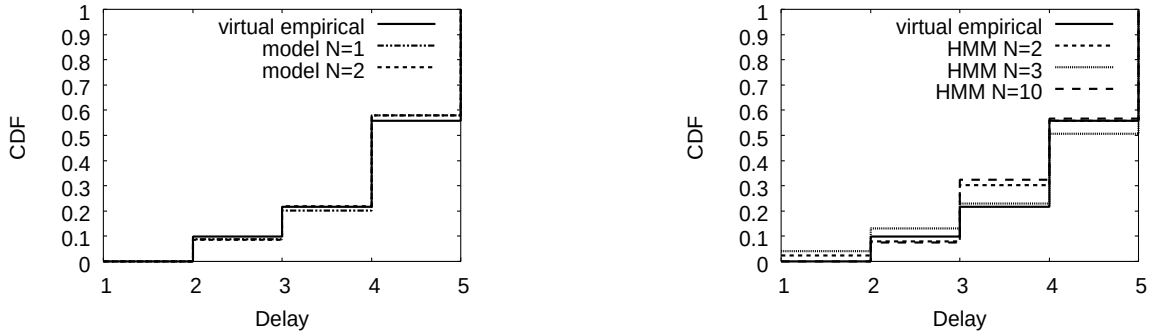
5.1.3 No dominant congested link

We next investigate settings in which no dominant congested link exists. In particular, we vary the parameters such that losses occur at links (r_0, r_1) and (r_2, r_3) and the loss rates at the two links are comparable. The buffer size at routers r_0, r_1 and r_2 are 25.6 Kb, 128 Kb and 25.6 Kb respectively. The link bandwidth of (r_1, r_2) is 1 Mbps. The link bandwidths of (r_0, r_1) and (r_2, r_3) are varied, as listed in Table 4. The loss rates at links (r_0, r_1) and (r_2, r_3) in each setting are comparable, as shown in Table 4. The null hypothesis is that there exists a weakly dominant congested link with $\theta = 0.03$ and $\phi = 0$. For all settings, the identification procedure rejects the hypothesis.

We describe one setting in detail. In this setting, the bandwidths of link (r_0, r_1) and $r_2, r_3)$ are 0.1 Mbps and 0.2 Mbps respectively. The average loss rates on link (r_0, r_1)

(r_0, r_1)		(r_2, r_3)		Max. queuing delay (ms)		
Bw (Mbps)	Loss rate	Bw (Mbps)	Loss rate	Value	Est. (model)	Est. (loss pair)
0.7	0.2%	0.2	3.8%	128	128	164
0.5	0.2%	0.2	3.9%	128	128	181
0.25	0.2%	0.1	7.1%	256	265	356
1.0	0.1%	0.1	3.8%	256	258	281

Table 3: A weakly dominant congested link: various bandwidths for links (r_0, r_1) and (r_2, r_3).



(a) Markov model with a hidden dimension

(b) HMM

Figure 9: Distributions of the virtual queuing delays for a setting with no dominant congested link.

(r_0, r_1)		(r_2, r_3)	
Bw (Mbps)	Loss rate	Bw (Mbps)	Loss rate
0.4	0.26%	0.8	0.04%
0.2	0.3%	0.5	0.2%
0.1	2.2%	0.3	1.0%
0.1	2.3%	0.2	2.0%

Table 4: No dominant congested link: various bandwidths for links (r_0, r_1) and (r_2, r_3).

and (r_2, r_3) are similar (2.3% and 2.0% respectively). We therefore have two lossy links and no dominant congested link. Fig. 9(a) also shows that the virtual queuing delay distributions obtained empirically and from the various models are very close. For $\theta = 0.03$ and $\phi = 0$, $D = 2$ is the minimum delay such that $F_W(D) > \theta$. However $F_W(2D) = F_W(4) = 0.58 < (1-\theta)(1-\phi) = 0.97$. We therefore conclude that there is no weakly dominant congested link with $\theta = 0.03$ and $\phi = 0$. Of course, there is no weakly dominant congested link with lower values of θ and ϕ either.

Fig. 9(b) plots the virtual queuing delay distributions obtained empirically and from HMMs in this setting. The virtual queuing delay distributions from the various HMMs are different and deviate from the empirical distribution, indicating that the distributions from HMMs are sensitive to the parameters and do not match the empirical result. Note that even for a large value of N ($N = 10$), the virtual queuing delay distributions from the HMM still differs from the empirical distribution.

5.1.4 The requirement on the probing duration for accurate identification

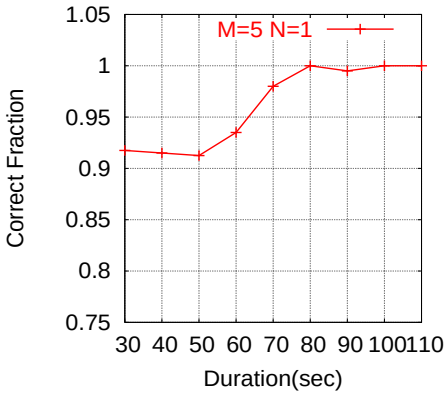
In the above, we used a trace 1,000 seconds in length containing 50,000 observations. We next investigate the probing duration required to obtain an accurate identification. We randomly choose a segment from the 1000-second trace as the probing sequence to identify if a dominant congested link exists. Following that, we check if the identification is correct or not. In the model, the number of delay symbols M is chosen to be 5 and the number of hidden states N is 1. We repeat the process 400 times and obtain the fraction of correct identifications.

In cases where a strongly dominant congested link exists, the probing duration of several tens of seconds suffices to achieve correct identification. Fig. 10(a) plots the correct fraction versus the probing duration for the setting with a weakly dominant congested link described in detail in Section 5.1.2. We observe that a probing duration of 80 seconds suffices to correctly identify the existence of a dominant congested link. Fig. 10(b) depicts the correct fraction versus the probing duration for the setting with no dominant congested link described in detail in Section 5.1.3. The correct fraction is very close to 1 with a probing duration of 250 seconds.

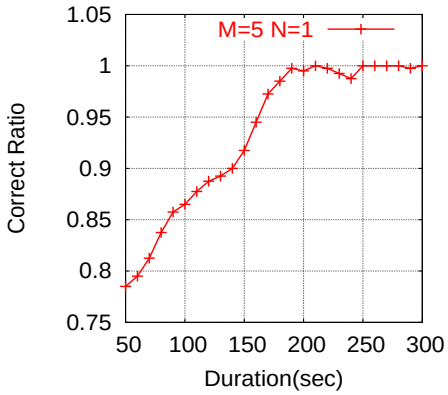
5.2 Internet experiments

We next describe validation results using Internet experiments over three paths. In the first path, the source is at University of Massachusetts, Amherst (UMass) and the destination is at Universidade Federal do Rio de Janeiro (UFRJ), Brazil. In the second and third paths, the source is at a resident house in Amherst, Massachusetts and the destinations are at UMass and University of Southern California (USC) respectively. The resident house uses a cable modem for Internet connection. It is physically very close to UMass and far from USC. We use Linux-based machines for all the experiments.

In each experiment, we run *tcpdump* [25] to capture the timestamps of the probes at the source and the destination to obtain one-way delay. We then use the method proposed in [26] to remove clock offset and skew. Each experiment lasts for one hour. For each experiment, we select a stationary probing sequence of 20 minutes for model-based identification. The minimum end-end delay in the entire hour is used as D_0. The queuing delays are discretized into 5 delay symbols, that is, $M = 5$. We vary the number of hidden



(a) A setting with a weakly dominant congested link.



(b) A setting with no dominant congested link.

Figure 10: Correct fraction versus the probing duration for two settings in *ns*.

states N from 1 to 4. At the end, we investigate the effect of probing duration, the choice of M and D_0 on the identification results. For all the experiments, the null hypothesis is that there exists a weakly dominant congested link with $\theta = 0.05$ and $\phi = 0.05$.

It is very difficult to validate the results from the Internet experiments since we do not have access to the internal routers to measure per-hop delay and loss for each probe. We therefore use some existing measurement tools. We use *pathchar* [15] to estimate link bandwidth along a path. During the experiments, we ran *ping* to each internal router to obtain a crude estimate of the loss rate from the source to the internal routers.

5.2.1 UMass to Brazil

For the several experiments from UMass to UFRJ, the identification procedure accepts the hypothesis that a dominant congested link exists along the path. This is consistent with the observations from the network operators in UFRJ that the UFRJ gateway was lossy at the time we ran the experiments. Results from *pathchar* indicate that the link from UFRJ gateway has the lowest bandwidth along the path, which is consistent with our identification. It is worth mentioning that the link bandwidth estimates from *pathchar* alone can not be used as a reliable basis for dominant congested link identification. Furthermore, it takes several hours when running *pathchar* on this path while only minutes of probes are sufficient to identify whether a dominant congested link exists, a point we will return to in Section 5.2.3. We next describe one experiment in detail.

This experiment started at 3:21pm on Feb. 24, 2003 with a probing interval of 20 ms. The loss rate of this probing sequence is 0.2%. Fig. 11 shows the PMF of the virtual queuing delays for models with $N = 1, 2, 3$. The distributions from different models are very similar. For $\theta = 0.01$ and $\phi = 0.01$, $D = 1$ is the minimum delay such that $F_W(D) > \theta$. Since $F_W(2D) = F_W(2) > (1-\theta)(1-\phi) \approx 0.98$, by Test 2, we accept the hypothesis that there is a weakly dominant congested link with $\theta = 0.01$ and $\phi = 0.01$. Of course, we accept the hypothesis that there is a weakly dominant congested link with higher values of θ and ϕ.

5.2.2 Resident house as the source

We next describe validation results for the second and third paths, both using the resident house as the source.

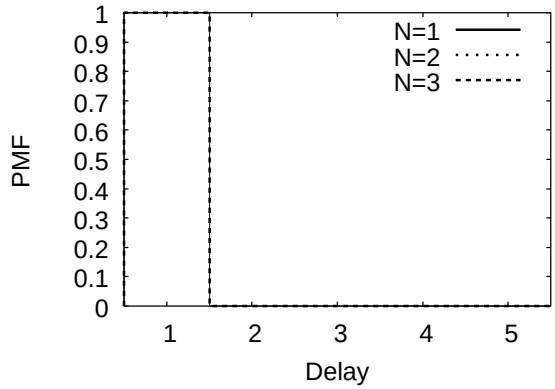


Figure 11: The virtual queuing delay distribution for an experiment from UMass to UFRJ.

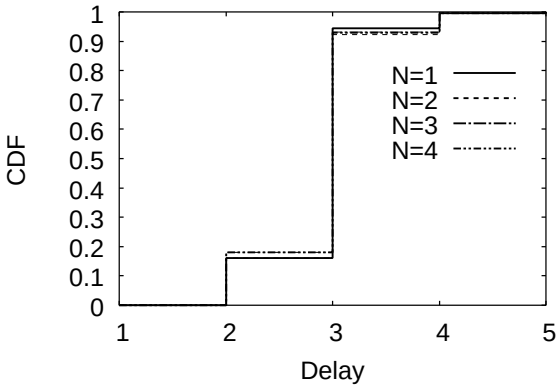


Figure 12: The virtual queuing delay distribution for an experiment from the resident house to UMass.

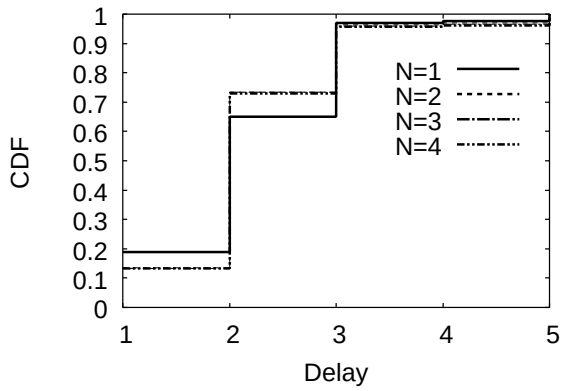


Figure 13: The virtual queuing delay distribution for an experiment from the resident house to USC.

The outgoing bandwidth of the cable modem in the resident house is around 250 Kbps. The path from the resident house to UMass contains 12 hops, over one commercial network. The path from the resident house to USC contains 28 hops, over commercial networks administrated by three ISPs. We next describe one experiment for each path in detail.

The experiment from the resident house to UMass started at 11:40pm on Feb. 20, 2003 with a probing interval of 20 ms. The average loss rate is 5.5%. Fig. 12 shows the virtual queuing delay distributions from different models. The distribution using $N = 1$ differs slightly from those using $N \geq 2$. For $\theta = 0.05$ and $\phi = 0.05$, $D = 2$ is the minimum delay such that $F_W(D) > \theta$. Since $F_W(2D) = F_W(4) > (1-\theta)(1-\phi) \approx 0.90$, by Test 2, the distribution does not contradict the hypothesis that there exists a weakly dominant congested link with $\theta = 0.05$ and $\phi = 0.05$. Results from *pathchar* indicate that the link from the UMass gateway has much lower bandwidth than other links (except the one from the resident house), which is consistent with our identification.

The experiment from the resident house to USC started at 11:16am on Feb. 22, 2003 with a probing interval of 20 ms. The average loss rate is 4%. Fig. 13 shows the virtual queuing delay distribution for different models. We observe that the distribution when $N = 1$ differs from others, which indicates that $N = 1$ is not sufficient for this setting. When using $\theta = 0.1$ and $\phi = 0.1$, $D = 1$ is the minimum delay such that $F_W(D) > \theta$. However, $F_W(2D) = F_W(2) < (1-\theta)(1-\phi) = 0.81$. By Test 2, we conclude that there is no weakly dominant congested link with $\theta = 0.1$ and $\phi = 0.1$. Of course, there is no weakly dominant congested link for lower values of θ and ϕ. Results from *pathchar* show two low-bandwidth links along the path (except the one from the resident house), which is consistent with our identification.

5.2.3 The effect of the probing duration and the choice of M and D_0

We next investigate the effect of the probing duration and the choice of M and D_0 on the identification results. We use the two experiments from the resident house since the loss rate from UMass to UFRJ is too low. For each experiment, we randomly choose a segment from the 20-minute trace as a probing sequence to identify if there exists a weakly dominant congested link with $\theta = \phi = 0.05$ using the model-based approach. Afterwards, we check if the identification from the segment is consistent with that from the 20-minute trace. In the model, the number of delay symbols M is 5, 7 or 9 and the number of hidden states N is 1. We repeat the process 400 times and obtain the fraction of consistent identifications. For each probing sequence, we examine two cases: using the minimum end-end delay in the probing sequence and and that in the entire hour as D_0. We believe the minimum end-end delay in one hour (over 10^5 probes) is very close to the real D_0. We therefore refer to the first case as D_0 unknown and the second case as D_0 known.

Fig. 14(a) and (b) depict the consistency fraction versus the probing duration for the two experiments. For $M = 5$, both the cases D_0 known and unknown are shown in the figure. For both experiments, the results for D_0 known and unknown are very close, especially for relatively long probing durations. This demonstrates that the approximation by using the minimum end-end delay in a probing sequence as D_0 does not affect the identification results for these two settings. We also observe in Fig. 14 that a probing duration on the order of minutes is sufficient to achieve a high consistency fraction in both experiments for $M = 5$, 7, or 9.

5.3 Summary results

We summarize the key results from the *ns* simulation and Internet experiments as follows:

- The virtual queuing delay distributions inferred with different number of hidden states, N, are similar. $N = 2$ is sufficient for all the settings we examined. The distribution when $N = 1$ is very close to those when $N \geq 2$ in most cases. The number of delay symbols M between 5 and 9 is sufficient.

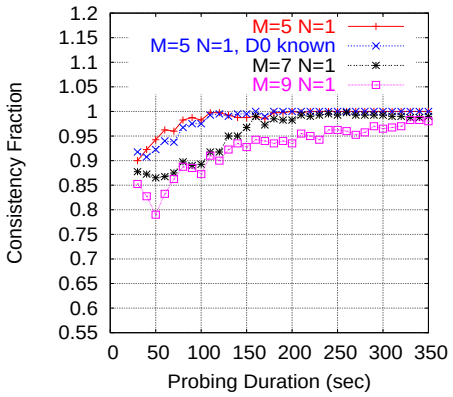
- The probing duration required for a correct identification needs to be on the order of minutes for the various settings we studied. The inaccuracy caused by using the minimum end-end delay to approximate D_0 is negligible when the probing duration is longer than several minutes.

- The computational requirements of the inference procedure are small (seconds) since the procedure only requires small M, N and short probing duration.

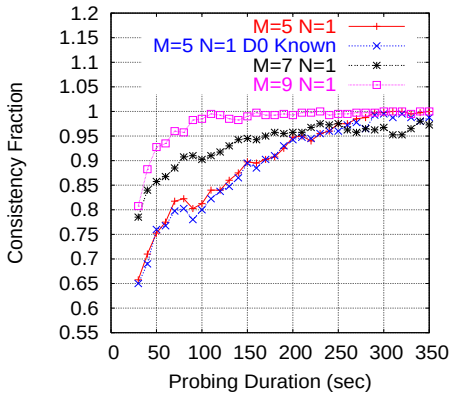
- The estimates of the maximum queuing delays at the weakly dominant congested link from our model-based approach are more accurate than those from the loss pair approach.

6. CONCLUSIONS AND FUTURE WORK

In this paper, we study a specific end-end path characteristic, namely, whether a dominant congested link exists along a path. We propose two simple hypothesis tests for identifying this characteristic and develop a model-based approach for identification from one-way end-end measurements. Our validation in *ns* simulation and Internet experiments shows that the model-based approach requires only minutes of probing for accurate identification. As future work, we are pursuing in several directions: (i) conduct controlled test-bed experiments and more/richer Internet experiments for validation; (ii) study how to pinpoint the dominant congested link after identifying a dominant congested link exists.



(a) With a dominant congested link.



(b) With no dominant congested link.

Figure 14: Consistency fraction versus probing duration for two experiments using the resident house as the source.

Acknowledgments

We would like to thank Subhabrata Sen for helpful discussions and the anonymous reviewers for their insightful comments. We would also like to thank L. Golubchik, C. Papadopoulos and E. A. de Souza e Silva for providing us accounts for the Internet experiments.

7. REFERENCES

[1] V. Paxson, "End-to-end Internet packet dynamics," in *IEEE/ACM Transactions on Networking*, p. 7(3), June 1999.

[2] K. Lai and M. Baker, "Measuring link bandwidths using a deterministic model of packet delay," in *SIGCOMM*, pp. 283–294, 2000.

[3] V. Ribeiro, M. Coates, R. Riedi, S. Sarvotham, and R. G. Baraniuk, "Multifractal cross-traffic estimation," in *Proc. ITC Specialist Seminar on IP Traffic Measurement. Modeling and Management*, September 2000.

[4] B. Melander, M. Bjorkman, and P. Gunningberg, "A new end-to-end probing and analysis method for estimating bandwidth bottlenecks," in *Proc. IEEE GLOBECOM*, November 2000.

[5] C. Dovrolis, P.Ramanathan, and D. Moore, "What do packet dispersion techniques measure?," in *Proc. IEEE INFOCOM*, 2001.

[6] M. Jain and C. Dovrolis, "End-to-end available bandwidth: measurement methodology, dynamics, and relation with TCP throughput," in *Proc. ACM SIGCOMM*, 2002.

[7] Y. Zhang, N. Duffield, V. Paxson, and S. Shenker, "On the constancy of Internet path properties," in *Proceedings of ACM SIGCOMM Internet Measurement Workshop*, November 2001.

[8] Z. Fei, S. Bhattacharjee, E. W. Zegura, and M. H. Ammar, "A novel server selection technique for improving the response time of a replicated service," in *INFOCOM (2)*, pp. 783–791, 1998.

[9] Y. Guo, K. Suh, J. Kurose, and D. Towsley, "P2cast: Peer-to-peer patching scheme for VoD service," in *Proceedings of the 12th World Wide Web Conference (WWW-03)*, May 2003.

[10] D. Katabi, I. Bazzi, and X. Yang, "A passive approach for detecting shared bottlenecks," in *Proc. International Conference on Computer Communications and Networks*, 2001.

[11] D. Rubenstein, J. F. Kurose, and D. F. Towsley, "Detecting shared congestion of flows via end-to-end measurement," in *Measurement and Modeling of Computer Systems*, pp. 145–155, 2000.

[12] J. Liu and M. Crovella, "Using loss pairs to discover network properties," in *ACM SIGCOMM Internet Measurement Workshop 2001*, November 2001.

[13] M. Mathis, J. Semke, and J. Mahdavi, "The macroscopic behavior of the TCP congestion avoidance algorithm," *Computer Communications Review*, vol. 27, no. 3, 1997.

[14] J. Padhye, V. Firoiu, D. Towsley, and J. Krusoe, "Modeling TCP throughput: A simple model and its empirical validation," in *Proc. ACM SIGCOMM*, (Vancouver, CA), pp. 303–314, 1998.

[15] V. Jacobson, "pathchar - a tool to infer characteristics of internet paths." ftp://ftp.ee.lbl.gov/pathchar, April 1997.

[16] A. B. Downey, "Using pathchar to estimate Internet link characteristics," in *Measurement and Modeling of Computer Systems*, pp. 222–223, 1999.

[17] K. G. Anagnostakis, M. B. Greenwald, and R. S. Ryger, "cing: Measuring network-internal delays using only existing infrastructure," in *Proc. IEEE INFOCOM*, April 2003.

[18] R. Cáceres, N. Duffield, J. Horowitz, and D. Towsley, "Multicast-based inference of network-internal loss characteristics," *IEEE Transactions on Information Theory*, November 1999.

[19] T. Bu, N. Duffield, F. L. Presti, and D. Towsley, "Network tomography on general topology," in *Proc. ACM SIGMETRICS*, 2002.

[20] M. Coates and R. Nowak, "Network tomography for internal delay estimation," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, May 2001.

[21] S. McCanne and S. Floyd, "ns-LBNL network simulator," http://www-nrg.ee.lbl.gov/ns/.

[22] J. Liu, I. Matta, and M. Crovella, "End-to-end inference of loss nature in a hybrid wired/wireless environment," in *Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks*, March 2003.

[23] L. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," in *Proceedings of the IEEE*, vol. 77(2), pp. 257–285, February 1989.

[24] W. Wei, B. Wang, and D. Towsley, "Continuous-time hidden Markov models for network performance evaluation," *Performance Evaluation*, vol. 49, 2002.

[25] "tcpdump." http://www.tcpdump.org/.

[26] S. Moon, P. Skelly, and D. Towsley, "Estimation and removal of clock skew from network delay measurements," in *Proc. IEEE INFOCOM*, March 1999.

APPENDIX

A. THE EM ALGORITHM TO INFER λ

We next describe the EM algorithm to infer λ from a

sequence of T observations for the Markov model with a hidden dimension, introduced in Section 4.1. We first define some notations conforming to those used in [23]. Define $\alpha_t(i,j)$ to be the probability of the observation sequence up to time t and the state being in (i,j) at time t, given λ. That is

$$\alpha_t(i,j) = P(Y_1 = y_1, Y_2 = y_2, \cdots Y_t = y_t, Z_t = (i,j) \mid \lambda)$$

Define $\beta_t(i,j)$ to be the probability of the observation sequence from time $t+1$ to T, given state being in (i,j) at time t, given λ. That is

$$\beta_t(i,j) = P(Y_{t+1} = y_{t+1}, \cdots, Y_T = y_T \mid Z_t = (i,j), \lambda)$$

Define $\xi_t(i,j,k,l)$ to be the probability of state being in (i,j) at time t and in (k,l) at time $t+1$, given the observation sequence and λ. That is

$$\xi_t(i,j,k,l) = P(Z_t = (i,j), Z_{t+1} = (k,l) \mid Y_1 = y_1, \cdots, Y_t = y_t, \lambda)$$

Define $\gamma_t(i,j)$ to be the probability of being in state (i,j) at time t, given the observation sequence and λ.

$$\gamma_t(i,j) = P(Z_t = (i,j) | Y_1 = y_1, \cdots, Y_T = y_T, \lambda)$$

We derive $\xi_t(i,j,k,l)$ from $\alpha_t(i,j)$ and $\beta_{t+1}(k,l)$ as follows

$$\xi_t(i,j,k,l) = \frac{\alpha_t(i,j) p_{(i,j)(k,l)} \beta_{t+1}(k,l)}{\sum_{i=1}^{N}\sum_{j=1}^{M}\sum_{k=1}^{N}\sum_{l=1}^{M} \alpha_t(i,j) p_{(i,j)(k,l)} \beta_{t+1}(k,l)}$$

Observe that γ_t can be calculated from ξ_t as

$$\gamma_t(i,j) = \sum_{k=1}^{N}\sum_{l=1}^{M} \xi_t(i,j,k,l).$$

The EM algorithm is an iterative algorithm in which each iteration consists of two steps: the expectation step and the maximization step. During the expectation step, we compute the expected number of transitions from state (i,j) and the expected number of transitions from state (i,j) to state (k,l) using the model parameters obtained during the previous iteration. We also compute the expected number of times that a loss observation has delay symbol of j and the expected number of symbol j. During the maximization step, we calculate the new model parameters from the expected values from the expectation step. The iteration ends when the difference between parameters of the new model and the previous model is less than a certain convergence threshold.

A.1 The expectation step

Without loss of generality, we assume y_1 and y_T are not losses. In the expectation step, we first calculate α and β using the procedures referred to as forward and backward steps respectively [23]. The procedure to calculate $\alpha_t(i,j)$, where $1 \le t \le T, 1 \le i \le N, 1 \le j \le M$, consists of the following steps:

1. Initialization

$$\alpha_1(i,j) = \begin{cases} \pi(i, y_1), & j = y_1. \\ 0, & j \ne y_1. \end{cases}$$

2. Induction

$$\alpha_{t+1}(i,j) = \begin{cases} \sum_{k=1}^{N}\sum_{l=1}^{M} \alpha_t(k,l) p_{(k,l)(i,j)} s(j), & y_{t+1} = * \\ \sum_{k=1}^{N}\sum_{l=1}^{M} \alpha_t(k,l) p_{(k,l)(i,j)}, & y_{t+1} = j \\ 0, & \text{o.w.} \end{cases}$$

where $t = 1, 2, 3, \cdots, T-1$.

The procedure to calculate $\beta_t(i,j)$, where $1 \le t \le T, 1 \le i \le N, 1 \le j \le M$, contains the following steps:

1. Initialization

$$\beta_T(i,j) = \begin{cases} 1, & j = y_T. \\ 0, & j \ne y_T. \end{cases}$$

2. Induction

$$\beta_t(i,j) = \begin{cases} 0, & y_t \ne *, j \ne y_t \\ \sum_{k=1}^{N}\sum_{l=1}^{M} p_{(i,j)(k,l)} \beta_{t+1}(k,l), & \text{o.w.} \end{cases}$$

where $t = T-1, T-2, \cdots, 1$.

Once α and β are obtained, we calculate ξ and γ as shown before. Afterwards, we calculate the various expectations using ξ and γ, which is omitted here and can be found in the computation in the maximization step.

A.2 The maximization step

The new model parameter estimates are obtained in the maximization step as follows

$$\begin{aligned} & \hat{\pi}_{(i,j)} = \gamma_1(i,j) \\ & \hat{p}_{(i,j)(k,l)} \\ = \; & \frac{\text{expected no. of transitions from } (i,j) \text{ to } (k,l)}{\text{expected no. of transitions from } (i,j)} \\ = \; & \frac{\sum_{t=1}^{T-1} \xi_t(i,j,k,l)}{\sum_{t=1}^{T-1} \gamma_t(i,j)} \\ & \hat{s}(j) \\ = \; & \frac{\text{expected no. of times that a loss has delay of } j}{\text{expected no. of delay } j} \\ = \; & \frac{\sum_{t=1}^{T} \mathbf{1}(y_t = *) \sum_{i=1}^{N} \gamma_t(i,j)}{\sum_{t=1}^{T}\sum_{i=1}^{N} \gamma_t(i,j)} \end{aligned}$$