Available at
www.ElsevierComputerScience.com
POWERED BY SCIENCE DIRECT®
Computer Networks 44 (2004) 211–233
COMPUTER NETWORKS
www.elsevier.com/locate/comnet

Edge-to-edge measurement-based distributed network monitoring

Ahsan Habib *, Maleq Khan, Bharat Bhargava

Department of Computer Sciences, Center for Education and Research in Information, Assurance and Security (CERIAS), Purdue University, West Lafayette, IN 47907, USA

Received 1 January 2003; received in revised form 27 June 2003; accepted 15 August 2003

Responsible Editor: G. Pacifici

Abstract

Continuous monitoring of a network domain poses several challenges. First, routers of a network domain need to be polled periodically to collect statistics about delay, loss, and bandwidth. Second, this huge amount of data has to be mined to obtain useful monitoring information. This increases the overhead for high speed core routers, and restricts the monitoring process from scaling to a large number of flows. To achieve scalability, polling and measurements that involve core routers should be avoided. We design and evaluate a distributed monitoring scheme that uses only edge-to-edge measurements, and scales well to large network domains. In our scheme, all edge routers form an overlay network with their neighboring edge routers. The network is probed intelligently from nodes in the overlay to detect congestion in both directions of a link. The proposed scheme involves only edge routers, and requires significantly fewer number of probes than existing monitoring schemes. Through analytic study and a series of experiments, we show that the proposed scheme can effectively identify the congested links. The congested links are used to capture the misbehaving flows that are violating their service level agreements, or attacking the domain by injecting excessive traffic.

Keywords: Network monitoring; Network security; Quality of service; Denial of service

1. Introduction

Continuous monitoring of a *network domain* is necessary to ensure proper operation of the network by detecting possible service violations and attacks. Attackers can impersonate a legitimate customer by spoofing flow identities. Network filtering [18] at routers can detect such spoofing if the attacker and the impersonated customer are in different domains. Otherwise, the attacks remain undetected. The quality of service (QoS) enabled networks face QoS attacks. In this setting, the attacker is a regular user of the network trying to get more resources (a better service class) than what it has signed (paid) for. A QoS network provides different classes of service for different prices, which can entice attackers to steal bandwidth and other network resources. Such attacks involve injecting traffic into the network with the intent to

* Corresponding author.
E-mail addresses: habib@cs.purdue.edu (A. Habib), mmkhan@cs.purdue.edu (M. Khan), bb@cs.purdue.edu (B. Bhargava).

1389-1286/$ - see front matter © 2003 Published by Elsevier B.V.
doi:10.1016/j.comnet.2003.08.002

steal bandwidth or to cause QoS degradation, by causing other customers' flows to experience longer delays, higher loss rates, and lower throughput. Taken to an extreme, this may result in a denial of service (DoS) attack.

A large variety of network monitoring tools can be found in [22]. Many tools use SNMP [10], RMON [34], or NetFlow [12], which are built-in functionality for most routers. Using these mechanisms, a centralized or decentralized model can be built to monitor a network. The centralized approach to monitor network latency, jitter, loss, throughput, or other QoS parameters suffers from scalability. One way to achieve scalability is to use a hierarchical architecture [2,33]. Subramanyan et al. [33] design a SNMP-based distributed network monitoring system that organizes monitoring agents and managers in a hierarchical fashion. Both centralized or decentralized models obtain monitoring data by polling each router of a network domain, which limits the ability of a system to scale for large number of flows. The alternative way of polling is to use an event reporting mechanism that sends useful information typically in a summarized format only when the status of a monitored element changes. A more flexible way of network monitoring is by using mobile agents [25] or programmable architecture [4]. However, periodic polling or deploying agents in high speed core routers put non-trivial overhead on them. We propose a very low overhead monitoring scheme that does not involve core routers for any kind of measurements. Our assumption is that if a network domain is properly provisioned and no user is misbehaving, the flows traversing through the domain should not experience a high delay or a high loss. An excessive traffic due to attacks changes the internal characteristics of a network domain. This change of internal characteristics is a key point to monitor a network domain.

Edge-to-edge monitoring scheme is studied in [20], where we devise a network monitoring mechanism to detect attacks on QoS domains using network tomography [7,13,17]. This monitoring mechanism measures the service level agreement (SLA) parameters, and compares these measurements with the values negotiated between a service provider and a user. To infer SLA parameters, the tomography-based scheme constructs a tree from the network topology, and probes the leaves from the root. Probing all leaves from the root cannot infer SLA parameters in both directions of a link. We need to measure loss in both directions of a link because they can be very different. This path asymmetry phenomenon is shown in [30]. The stripe-based monitoring can achieve this with very high overhead. Our goal is to devise a low overhead monitoring scheme that can detect attacks in both directions of all links in a network domain.

The proposed monitoring scheme has two phases. In the first phase, we continuously measure edge-to-edge link delays to observe any unusual delay pattern. All ingress routers (entry points) sample the incoming traffic to probe latency of the paths followed by a user packet. This measures the delay experienced by a user inside the domain. If the delay is higher than a pre-defined threshold (SLA value), the edge routers conduct intelligent probing for loss measurements. For this probing, an overlay network is formed using all edge routers on top of the physical network. The probing does not calculate loss ratio for each individual link, instead, the congested links (having high losses) are identified using edge-to-edge loss measurements. Our solution consists of two methods: simple method and advanced method. In the simple method, all edge routers probe their neighbors in clockwise and counter-clockwise direction. This method requires only $O(n)$ probing, where n is the number of edge routers. Through extensive analysis, both analytical and experimental, we show that the simple method is very powerful to identify the congested links to a close approximation. If necessary, we use the advanced method to refine the solution of the simple method. The advanced method searches the topology tree intelligently for probes that can be used to identify the status of the undecided links from the simple method. When the network is less than 20% congested the advanced method requires $O(n)$ probes. If the congestion is high, it requires more probes, however, it does not exceed $O(n^2)$.

In the second phase of our monitoring process, we use the congested links as a basis to identify edge routers through which traffic are entering into and exiting from the domain. From exiting edge routers, we identify the flows that are violating any

SLA agreement. If the SLA is violated for delay and loss, the network is probed to detect whether any user is stealing bandwidth. The service violations can indicate a possible attack on the same domain, or on a downstream domain. In case of a DoS attack, numerous flows from different sources are destined to a victim. These flows aggregate on their way as they get closer to the victim. Monitoring an upstream network domain can detect these high bandwidth aggregates that could result in DoS attacks on downstream domains [20,26]. To control the attacks, filters are activated at edge routers through which flows are entering into a network domain. We restrained ourselves from discussing on other techniques to detect and prevent DoS attacks. The primary focus of this paper is monitoring. A detailed discussion and analysis among different techniques to detect and prevent DoS attacks can be found in our paper [21].

Using simulation, we conduct a series of experiments to evaluate the proposed monitoring scheme. We conclude that the distributed monitoring scheme shows a promise for efficient and scalable monitoring of a domain. This scheme can detect service violations, bandwidth theft attacks, and tell when many flows are aggregating towards a downstream domain for a possible DoS attack. The scheme requires low monitoring overhead, and detects service violations in both directions of any link in a network domain.

The rest of the paper is organized as follows. The related work is discussed in Section 2. Measuring all necessary network parameters for monitoring purposes is presented in Section 3. It discusses our proposed monitoring scheme, and analyzes its strength and limitations. Section 4 explains how to use the monitoring scheme to detect service violations and DoS attacks. Experimental results and discussions are provided in Section 5. We discuss the advantages of the distributed monitoring over stripe-based monitoring in Section 6. We conclude the paper in Section 7.

2. Related work

One common way of monitoring is to log packets at various points throughout the network and then extract information to discover the path of any packet [29]. This scheme is useful to trace an attack long after the attack has been accomplished. The effectiveness of logging is limited by the huge storage requirements especially for high speed networks. Stone [32] suggested to create a virtual overlay network connecting all edge routers of a provider to reroute *interesting* flows through tunnels to central tracking routers. After examination, suspicious packets are dropped. This approach also requires a great amount of logging capacity.

Many proposals for network monitoring [6,14] give designs to manage the network and ensure that the system is operating within desirable parameters. In efficient reactive monitoring [14], the authors discuss ways to monitor communication overhead in IP networks. Their main idea is to combine global polling with local event driven reporting to monitor a network. Breitbart et al. [6] identify effective techniques to monitor bandwidth and latency in IP networks. The authors present *probing-based* techniques where path latencies are measured by transmitting probes from a single point of control. They describe algorithms for computing an optimal set of probes to measure latency of paths in a network, whereas we focus on measuring parameters without the involvements of the core routers.

In [11], a histogram-based aggregation algorithm is used to detect SLA violations. The algorithm measures network characteristics on a hop-by-hop basis, uses them to compute end-to-end measurements, and validates end-to-end SLA requirements. In large networks, efficient collection of management data is a challenging task. The authors propose an *aggregation* and *refinement* based monitoring approach. The approach assumes that the routes used by SLA flows are known, citing VPN and MPLS [9] provisioning. Though routes are known for double-ended SLAs that specify both ingress and egress points in the network, they are unknown in cases where the scope of the service is not limited to a fixed egress point.

Duffield and Grossglauser [15] propose trajectory sampling to infer traffic flows through a domain. In this process, each link samples packets

based on a hash function computed over the content of the packets. Then, the trajectory of a packet is reconstructed using the same sample set of packets. This provides a neat way of monitoring a network domain, which does not depend on the network status information. However, all routers (edge and core) participate in sampling that might put large overhead on the high speed core routers. Our goal is to devise a scheme that does not involve the core routers for any measurement.

Kim and Hong [24] collect statistical data from every single router for each service class and then analyze the data to compute edge-to-edge QoS of aggregate IP flows. This approach has very high overhead, and not suitable for real time monitoring.

Duffield et al. [16] use packet "stripes" (back-to-back probe packets) to infer link loss by computing the correlation of packet loss within a stripe at the destinations. This work is an extension of loss inference for multicast traffic, described in [1,8]. To infer loss, a series of probe packets, called a stripe, are sent from one edge router to two other edge routers with no delay between the transmissions of successive (usually three) packets. For example, in a two-leaf binary tree spanned by nodes 0, k, R_1, R_2, stripes are sent from the root 0 to the leaves to estimate the characteristics of one link, say $k \rightarrow R_1$ (Fig. 1). The first two packets of a 3-packet stripe are sent to R_2 and the last one to R_1. If a packet reaches any receiver, we can infer that the packet must have reached the branch point k. If R_2 gets both packets of a stripe, it is likely that R_1 will receive the last packet of the stripe. Using number of packets reach to R_1 and R_2, we can calculate the successful transmission probability of the link $k \rightarrow R_1$. Similarly, a complementary stripe is sent to estimate the characteristics of link $k \rightarrow R_2$. By combining estimates of stripes down each such tree, the characteristics of the common path from $0 \rightarrow k$ is estimated. This inference technique extends to general trees by sending probes from root to each ordered pair of leaves [16]. Ji and Elwalid [23] show that measurement-based monitoring using tree is scalable when the probe packets reach the edge routers with high probability. If the internal loss is very high, the solution does not scale for a large network.

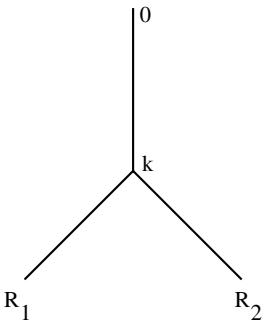

Fig. 1. Binary tree to infer loss from source 0 to receivers R_1 and R_2.

The stripe-based probing mechanism is adopted to monitor loss characteristics inside a QoS network domain without relying on the core routers in [20]. The stripe-based monitoring scheme requires less overhead than core-assisted monitoring. In this paper, we propose a scalable scheme that requires much less probes than stripe-based scheme.

3. Measurements with distributed probing

The service level agreement (SLA) parameters such as delay, packet loss, and throughput are measured to ensure that all users are getting their target share of resources. Delay is the edge-to-edge latency. Packet loss is the ratio of total number of packets dropped from a flow [1] to the total number of packets of the same flow entering into the domain. Throughput is the total bandwidth consumed by a flow inside a domain. Delay and loss are important parameters to monitor in a network domain. Bandwidth measurement is used to detect whether any flow is getting more than its share of resources, which causes other flows to suffer. Although jitter (a delay variation) is another important SLA parameter, it is flow-specific and, therefore, is not suitable to use in network monitoring. A large body of research has focused on

[1] A flow is a microflow with five tuples (addresses, ports, and protocol) or an aggregate flow that is combined of several microflows.

measuring delay, loss, and throughput in the Internet [27,28]. In this section, we describe techniques to measure each parameter. Delay and throughput measurements are discussed in details in [20]. This paper proposes an efficient way that detect links with high losses using edge-to-edge measurements.

The distributed monitoring scheme measures SLA components and compares the measurements to the pre-defined values to detect service violation. There is one monitoring agent that gets feedback from all other edge routers about delay, loss, and throughput. The monitoring agent can sit on any edge router in the network domain. We note that the monitoring algorithm can be implemented in a measurement box behind each edge router assuming there is not congestion in this connection (measurement box to edge router). This enables us to deploy our monitoring scheme without changing the existing infrastructure.

3.1. Delay measurements

To measure delay, the ingress routers copy the IP header of incoming packet into a new packet with a certain pre-configured probability. Copying the header from user traffic to measure delay has a couple of benefits. First, the probe packet follows the same path as the user traffic because the route inside a domain does not change too often. Hence, the probe delay is similar to the delay experienced by the user. Second, if some links do not have any traffic, the links will not be probed, which saves the probing overhead.

The routers encode the current timestamp t_{ingress} into the payload and mark the protocol field of the IP header with a new value. An egress router recognizes such packets and removes them from the network. Additionally, the egress routers compute the edge-to-edge link delay for a packet from the difference between its own time and t_{ingress}. We ignore minor drifts of the clocks since all routers are in one administrative domain and can be synchronized fairly accurately. The egress classifies the probe packets as belonging to flow i of user j, and update the average packet delay, avg_delay_j^i, for delay sample $\text{delay}_j^i(t)$ at time t using an exponential weighted moving average (EWMA):

$$\text{avg\_delay}_j^i(t) = \alpha \times \text{avg\_delay}_j^i(t-1) + (1-\alpha) \times \text{delay}_j^i(t), \quad (1)$$

where α is a small fraction $0 \leqslant \alpha < 1$ to emphasize recent history rather than the current sample alone.

The egress routers send the average delay to the monitor. If the average packet delay exceeds the delay guarantee in the SLA ($\text{SLA}_{\text{delay}}^i$) for flow i, i.e. $\text{avg\_delay}^i > \text{SLA}_{\text{delay}}^i$, it is an indication of an SLA violation. If the network is properly provisioned and flows do not misbehave, there should not be any delay greater than $\text{SLA}_{\text{delay}}^i$ for any flow i. A high delay can be caused by some flows that are violating their SLAs or bypassing the SLA checking, which is an attack.

If the delay exceeds a certain threshold, the monitor needs to probe the network for loss and throughput. We discuss the throughput measurement first, and then discuss the loss estimation to isolate congested links. Identifying the congested links is necessary to detect egress and ingress routers involved in high traffic paths, which helps to detect and control attacks.

The frequency of delay probing is important because it determines the overhead of the monitoring process. A detailed discussion on this issue can be found in [20]. The main idea here is that each path is probed with a probability, instead of a deterministic fashion. This probability changes with time, which puts uncertainty to the attackers (they do not know and cannot predict it beforehand). If there are N edge routers in a domain, $\mathscr{L}$ different paths for each router, and each router select a path for delay probing with a probability p_{probe}, the total number of paths need to be probed for delay is $N \times \mathscr{L} \times p_{\text{probe}}$. Each probe packet is 20 bytes in size and our experimental results show that 15–20 probes/s can estimate the delay accurately. Another suggestion is to use recent history of delay to search for any pattern that helps to predict an attack.

3.2. Throughput measurements

The objective of checking throughput violation is to ensure that nobody is consuming extra

bandwidth (beyond the SLA). The attackers can send excessive amount of best effort (BE) traffic to consume bandwidth, because BE traffic is not controlled at the ingress routers. Consumption of excess bandwidth by any flow can deteriorate the QoS for many others. This cannot be detected by a single ingress or egress router, if the user sends through multiple ingress routers at a rate lower than the SLA. For each ingress router, the user does not violate the SLA but as a whole he does. The service provider may allow a user to take extra bandwidth as long as everybody else is not harmed. This can depend on the policy of the service provider.

The monitor measures throughput by probing globally all egress routers when the monitor suspects any violation in delay and loss. Egress routers of a QoS domain maintain the aggregate flow rate for each user. This rate is a close approximation of the bandwidth consumption by each flow inside the domain [20]. The throughput for user j is $B^j = \sum_{i=1}^{N} B^{ij}$, where B^{ij} is bandwidth consumed by user j at edge router i and N is the total number of edges. If $\mathrm{SLA}^j_{\mathrm{bw}}$ is the bandwidth guarantee for user j, $B^j > \mathrm{SLA}^j_{\mathrm{bw}}$ indicates an SLA violation by user j. To detect bandwidth theft that does not change delay or loss pattern, the monitor can periodically poll egress routers.

Throughput measurement is flow-aware, which raises the concern about scalability. Our solution to this problem is to conduct this procedure only for a certain number of flows. If the polling for throughput measurement is done only for the flows that are consuming bandwidth higher than a given threshold. The question is how to set the threshold so that it is not vulnerable for a potential DoS attack. The threshold is computed based on the link capacity and number of active flows in such a way that crossing the threshold really means the flows are consuming considerable amount of bandwidth. If an attacker wants to send a large number of flows with low bandwidth, it will not impact on the throughput measurement to launch DoS attacks because those will not be considered for throughput polling.

3.3. Loss measurements

Packet loss guarantees made by a provider network to a customer are for the packet losses experienced by its conforming traffic inside the provider domain. Measuring loss by observing packet drops at all core routers is an easy task. It imposes, however, an excessive overhead on the core routers by forcing them to record each drop entry, and periodically sending it to the monitor. The *stripe-based* approach is an edge-to-edge mechanism, described in Section 2, to measure loss in a domain.

An interesting observation is that service violation can be detected without exact loss value of each internal link, instead, it requires to check whether a link has loss higher than the specified threshold. Like [5,19], we measure loss using average values in a recent time frame. The link with a high loss is referred to as a *congested link*. It is defined in Definition 1. The similar congestion measure is used in [3]. This congestion model is simple, and enables us to provide an in-depth analysis of the system. In future, we plan to use the model that considers loss correlation [35] among successive packets.

We propose a new approach to detect links with high losses by edge-to-edge measurements. The distributed probing detects all congested links using edge-to-edge loss measurements. These links are used to detect flows that pose threats to other flows by consuming extra resources.

To apply our distributed probing, we convert the network topology into tree structure. This converting process is discussed later in this section. The tree contains core routers as internal nodes and edge routers as leaf nodes. Monitoring agents are deployed in the leaves to collect statistics from other edge routers to check SLA violations. The probing agents sit only at the edge routers or at the measurement boxes and know their neighbors. The neighbors are determined by visiting the tree using depth first search algorithm starting from any edge router, and putting all edge routers in an ordered sequence. All probing agents form a virtual network on top of the physical network. The probes follow edge-to-edge path in the virtual network. We equivalently refer the tree topology

or the virtual network to an *overlay* network. A typical spanning tree of the topology, the corresponding overlay network, and direction of all internal links for each probe are shown in Fig. 2.

The following definitions and observations are used to describe the properties of the overlay network, and to identify the congested links.

Definition 1 (*Congested link*). A link is congested if all loss measurement samples in a given time frame exceed a specified loss threshold.

Definition 2 (*Terminal core router*). A core router, which is connected to only one other core router in an overlay network is called a terminal core router. In Fig. 2, the core routers $C4$ and $C5$ are terminal core routers.

Definition 3 (*Probe path*). A probe path $\mathscr{P}$ is a sequence of routers (either core or edge) $\langle E_1, C_1, C_2, \ldots, C_n, E_n \rangle$ where a router exists in the sequence only once. A probe packet originates at the edge router E_1, passes through the core routers $C_1, C_2, \ldots, C_{n-1}$, and C_n, in the given order, and terminates at the edge router E_n. We also represent the probe path $\mathscr{P}$ by the set of links, $\langle E_1 \rightarrow C_1, C_1 \rightarrow C_2, \ldots, C_n \rightarrow E_n \rangle$.

Definition 4 (*Link direction*). For link $u \rightarrow v$, we say link from node u to v, is in inward direction (IN) with respect to node v. Similarly, the same link is in outward (OUT) direction with respect to node u.

Lemma 1. *If a core router C is connected to two routers (core or edge) R_1 and R_2 only, the duplex path $R_1 \leftrightarrow C \leftrightarrow R_2$ can be replaced with the duplex link $R_1 \leftrightarrow R_2$, and both links are functionally equivalent in the distributed probing scheme.*

Proof. See Appendix A. □

Lemma 2. *In an overlay network, every core router is connected to at least three other routers.*

Proof. If a core router C is connected to two routers only, C together with its two connecting links can be replaced by a single link (Lemma 1). If C is connected to only one other router, C can

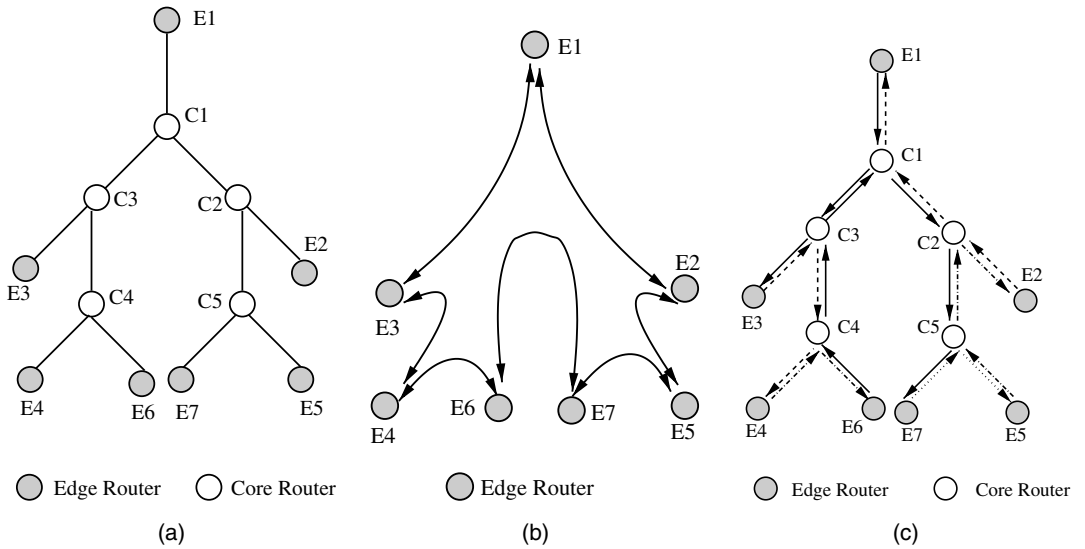

Fig. 2. (a) Tree topology transformed from a network domain. (b) All probing agents at the edge routers form a virtual network with both neighbors in an ordered sequence. (c) Direction of internal links for each probing.

never be included in an edge-to-edge probe path, and can simply be removed. Hence, all core routers are connected to at least three other routers. □

3.3.1. Simple method

In this solution, we conduct total two rounds of probing. One in the counter-clockwise direction, and one in the clock-wise direction starting from any edge router. The former one is referred to as *first round* of probing, and the latter one is referred to as *second round* of probing. In each round, probing is done in parallel.

We describe the loss monitoring scheme with a simple network topology. In this example, Fig. 3b, edge router 1 probes the path $1 \to 3$, router 3 probes the path $3 \to 4$, and 4 probes the path $4 \to 1$. Let $P_{i,j}$ be a boolean variable that represents the outcome of a probe between edge routers i to j. $P_{i,j}$ is 1 if the measured loss exceeds the threshold in any link within the probe path, and 0 otherwise. Notice that $P_{i,j} = 0$ for $i = j$. We express the outcome of a probe in terms of combination of all links status. Let $X_{i,j}$ be a boolean variable to represent the congestion status of an internal link $i \to j$. We refer X to a *congestion variable* throughout the rest of this paper. For Fig. 3c, we can write equations as follows:

$$\begin{aligned} X_{1,2} + X_{2,3} &= P_{1,3}, \\ X_{3,2} + X_{2,4} &= P_{3,4}, \\ X_{4,2} + X_{2,1} &= P_{4,1}, \end{aligned} \tag{2}$$

where (+) represents a boolean "OR" operation.

Similarly, for the second round of probing,

$$\begin{aligned} X_{1,2} + X_{2,4} &= P_{1,4}, \\ X_{4,2} + X_{2,3} &= P_{4,3}, \\ X_{3,2} + X_{2,1} &= P_{3,1}. \end{aligned} \tag{3}$$

For an arbitrary topology,

$$X_{i,k} + \sum_{n=k}^{n=l-1} X_{n,n+1} + X_{l,j} = P_{i,j}. \tag{4}$$

Note that loss in path $1 \to 3$ might not be same as loss in path $3 \to 1$. This path asymmetry phenomenon is shown in [30]. In general, $X_{i,j}$ is independent of $X_{j,i}, \forall_{ij}, i \neq j$.

The set of Eqs. (2) and (3) are used to detect congested links in the network. For example, if the outcome of the probing shows $P_{1,3} = 1$, $P_{1,4} = 1$, and rest are 0, we get the following:

$$X_{1,2} + X_{2,3} = 1, \quad X_{1,2} + X_{2,4} = 1. \tag{5}$$

All other probes do not see congestion on its path, i.e., $X_{3,2} = X_{2,4} = X_{4,2} = X_{2,1} = X_{2,3} = 0$. Thus, the equation set (5) reduces to $X_{1,2} = 1$. Similarly, if any of the single link is congested, we can isolate the congested link. Suppose, two of the links, $X_{1,2}$ and $X_{2,3}$, are congested. The outcome of probing will be $P_{1,3} = 1$, $P_{1,4} = 1$, and $P_{4,3} = 1$, which makes $X_{3,2} = X_{2,4} = X_{4,2} = X_{2,1} = 0$. This leaves the solution as shown in Eq. (6). Thus, the distributed scheme can isolate the congested links in this topology:

$$X_{1,2} + X_{2,3} = 1, \quad X_{1,2} = 1, \quad X_{2,3} = 1. \tag{6}$$

Analysis of simple method: The strength of simple method comes from the fact that congestion variables in one equation of any round of probing is distributed over several equations in the other

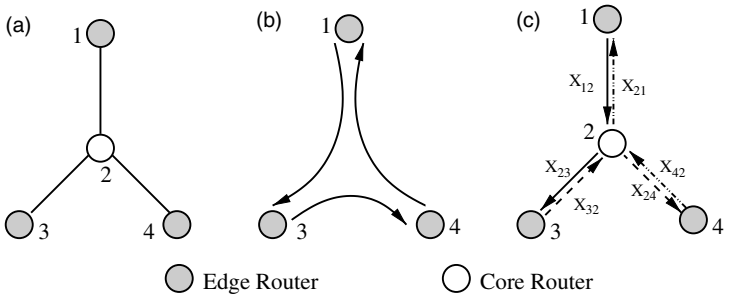

Fig. 3. (a) Spanning tree of a simple network topology. (b) Each edge router probes its neighbor edge router in counter-clockwise direction (c) Direction of internal links for each probing.

round of probing. If n variables appear in one equation in the first round of probing, no two (out of this n) variables appear in the same equation in the second round of probing (Lemma 3) or vice versa. This property helps to solve the equation sets efficiently. Theorem 1 shows that if any single probe path is congested with arbitrary number of links, the simple method can identify all the congested links. In Theorem 2, we show that the simple method determines the status of a link with very high probability when the network is less congested.

Lemma 3. *If $\mathscr{P}$ and $\mathscr{P}'$ are any probe paths in the first and the second round of probing respectively, $|\mathscr{P} \cap \mathscr{P}'| \leqslant 1$.*

Proof. See Appendix A. □

Lemma 4. *For any arbitrary overlay network, the average length of the probe paths in the Simple Method is $\leqslant 4$.*

Proof. In an overlay network, the number of links are $2(e + c - 1)$ considering both directions of a link. The edge routers are leaves of the topology tree whereas the core routers are the internal nodes of the tree. Number of leaf nodes is greater than the number of internal nodes. Thus, the number of links is $\leqslant 2(e + e - 1) = 4e$. Number of probe paths in first (or second) round of probing is e, and every link appears exactly once in each round. Hence, the average length of a path $\leqslant 4e/e = 4$. □

Theorem 1. *If only one probe path $\mathscr{P}$ is shown to be congested in the first round of probing, the simple method identifies each congestion link in $\mathscr{P}$.*

Proof. Let, the congested probe path be $\mathscr{P} = \langle l_1, l_2, \ldots, l_k \rangle$ and X_i is the congestion variable for link $l_i, 1 \leqslant i \leqslant k$. X_i appears once in the equations for each round of probing. Let, X_m is in equation $X_m + f(S) = 1$ in the second round of probing, where S is a set of congestion variables excluding X_m that appear in the equation. The expression $f(S)$ does not contain any of the variables X_i for $1 \leqslant i \leqslant k, i \neq m$ (Lemma 3). From first round of probing, we obtain $f(S) = 0$, because the outcome of all probe paths except $\mathscr{P}$ is zero in this round. Thus, we can determine X_m, which is 1, hence the status of the link l_m, for any $1 \leqslant m \leqslant k$. □

Theorem 2. *Let p be the probability of a link being congested in any arbitrary overlay network. The simple method determines the status of any link of the topology with probability* $2(1-p)^4 - (1-p)^7 + 2p(1-p)^{12} - p(1-p)^{24}$.

Proof. Let a particular link l appears in probe paths $\mathscr{P}_1$ and $\mathscr{P}_2$ in first and second round of probing. The status of a link can be either non-congested or congested. We consider both cases separately and then combine the results.

When l is non-congested. The status of l can be determined if the rest of the links in either $\mathscr{P}_1$ or $\mathscr{P}_2$ are non-congested. Let the length of probe paths $\mathscr{P}_1$ and $\mathscr{P}_2$ are i and k respectively. The probabilities that the other links in $\mathscr{P}_1$ and $\mathscr{P}_2$ are non-congested are $(1-p)^{i-1}$ and $(1-p)^{k-1}$ respectively. Since, only common link between paths $\mathscr{P}_1$ and $\mathscr{P}_2$ is l (Lemma 3), the following two events are independent: Event_1 = all other links in $\mathscr{P}_1$ are non-congested and Event_2 = all other links in $\mathscr{P}_2$ are non-congested. Thus, for a non-congested link,

$$\begin{aligned}&\Pr\{\text{status of } l \text{ be determined}\}\\ &= (1-p)^{i-1} + (1-p)^{k-1} - (1-p)^{i-1}(1-p)^{k-1}\\ &= (1-p)^{i-1} + (1-p)^{k-1} - (1-p)^{i+k-2}.\end{aligned}$$

Using the average length for the probe paths (Lemma 4), i.e., $i = k = 4$, $\Pr\{\text{Status of } l \text{ be determined}\} \approx 2(1-p)^3 - (1-p)^6$.

When l is congested. If l is a congested link, its status can be determined when all other links that appear on the probe path of l are non-congested and their status is determined. Let link l appears on a path in the first round of probing with l_1, l_2, and l_3 (considering the average path length is 4). The probability that l_1 (l_2 or l_3) is non-congested and determined is $(1-p)^4$. The probability to determine the status of these three links is $(1-p)^{12}$. This is true for the equations set in the

second round, where l appears with variables other than l_1, l_2, and l_3. Thus, Pr{Status of l be determined} $= 2(1-p)^{12} - (1-p)^{24}$.

For any link l (congested or non-congested)

$$\begin{aligned}&\Pr\{\text{Status of } l \text{ be determined}\}\\&= (1-p)[2(1-p)^3 - (1-p)^6]\\&\quad + p[2(1-p)^{12} - (1-p)^{24}]\\&= 2(1-p)^4 - (1-p)^7 + 2p(1-p)^{12}\\&\quad - p(1-p)^{24}. \qquad \square\end{aligned}$$

Fig. 4 shows the probability to determine status of a link when certain fraction of the links are actually congested. This figure shows that simple method determines status of a link with probability close to 0.90 when 10% links of a network are congested. For 20% and 30% congestion, the probabilities are 0.64 and 0.40 respectively. This result is validated with the simulation result for two different topologies. The simple method does not help much when 50% or more links are congested. In that case, we use the advanced method to find probes that can decide the status of undecided links in the simple method.

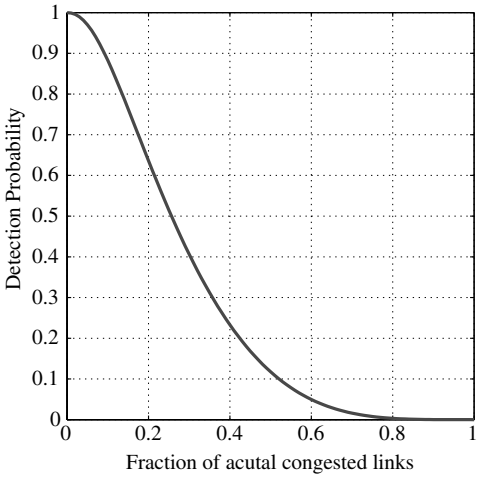

Fig. 4. Probability that the simple method determines the status of a link of any arbitrary topology. X-axis is the fraction of total links that are actually congested. The simple method performs extremely well when less than 20% links of a network are congested. If a network is more than 50% congested, the simple method cannot contribute much.

Having congestion on links that affect multiple probe paths might eventually lead to some boolean equations that do not have unique solutions. Thus, the solution of simple method usually have some links undecided. If we report these undecided links as congested, they will be referred to as false positive because some non-congested links will be reported as congested. The false positive is calculated as a ratio of undecided links labeled as congested by simple method to the total number links in the network. Fig. 5 shows false positive for two topologies; Topology 1 shown in Fig. 2(b) and Topology 2 shown in Fig. 9(b). The false positive is a small percentage of all links of a domain. The number of links that are marked as false positive is very close to the number of actually congested links. The reason we get false positive is that some good (non-congested) links sit on the same probes of congested links, and the simple method does not have enough probes to isolate them. Notice that the solution does not have any false negative.

We further analyze the simple method when a network has congestion that spreads from one edge router to any other edge routers. In real network, numerous flows come from different edge routers, and make a series of links to be congested. In this case, the simple method performs very well.

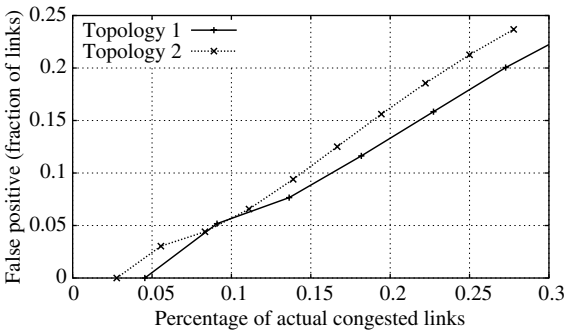

Fig. 5. The solution of the simple method cannot decide about some links. If those links are considered as congested links, the solution of the simple method provides false positive by declaring some links as congested. The graph is shown for two topologies; Topology 1 shown in Fig. 2(b) and Topology 2 shown in Fig. 9(b). This figure does not compare the two topologies, instead, it shows the false positive as a percentage of total links with respect to the percentage of links that are really congested. The solution does not have any false negative.

We observe that for edge-to-edge congested paths, the simple method does not add any link as false positive. We plot this behavior for Topologies 1 and 2 with all possible edge-to-edge paths in Fig. 6. On the average, the simple method can isolate more than 50% of the congested links for edge-to-edge congestion scenario. Rest of the cases, the solutions have some equations with more than one variable. The percentage of identified links is little high for the path length = 6 in case of Topology 1, Fig. 6. Because this path has more shared links comparing to other paths.

3.3.2. Advanced method

The advanced method is used to identify the status of the undecided variables in the simple method. Therefore, the output of the simple method is used as the input of the advanced method. We traverse the topology tree to find probes that can help to decide about the values for each undecided variable.

The algorithm of the advanced method is shown in Fig. 7. First, we conduct the simple method. Let the set of equations with undecided variables be $\mathbb{E}$. For each variable in equation set $\mathbb{E}$, we need to find two nodes that can be used to probe the network. Each probe needs one start node and one end node.

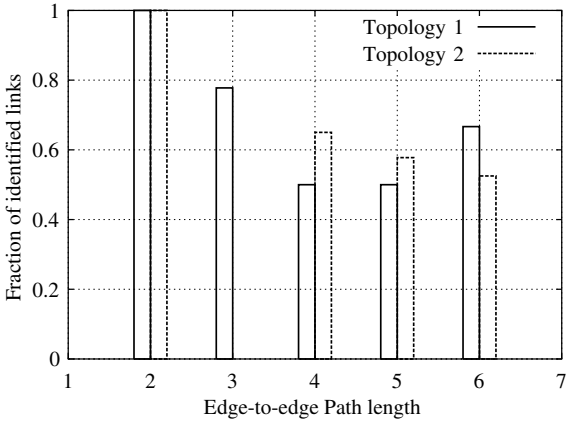

Fig. 6. Fraction of identified links by the simple method for all edge-to-edge congested paths. The X-axis shows all paths with a specific length. All solutions for edge-to-edge congestion path do not have any false positive. Topology 2 does not have any path of length 3.

The algorithm shows functions to find start and end probe nodes. Link direction (Definition 4) plays an important role to find these probes. For example, in Fig. 2, if link $C1 \rightarrow C3$ is congested, the start probe node can be $E2$, $E5$, or $E7$. On the other hand, if link $C3 \rightarrow C1$ is congested, the start probing node can be $E3$, $E4$, or $E5$.

For an undecided link $v_i \rightarrow v_j$, the function *FindNode* looks for leaves descended from node v_i and v_j. First, the algorithm searches for a node in IN direction on a subtree descended from v_i and then in OUT direction on a subtree descended from v_j. For any node v, the *DecidePath* explores all siblings of v to choose a path in the specified direction. The function avoids previously visited path and known congested path. It marks already visited paths so that the same paths will not be considered in exploration of an alternate path.

If the network is congested in a way that no solution is possible, the *AdvancedMethod* cannot add anything to the simple method. If there is a solution, the *AdvancedMethod* can obtain probes to decide about links because this probe finding is an exhaustive search on the topology tree to find leaf-to-leafs path that are not already congested.

Analysis of advanced method: The number of probes required in the advanced method depends on the number of congested links existing in a network. The advanced method starts with the undecided links in the simple method. When the network is sparely congested or densely congested, the algorithm exits within few runs, and the number of trial for each congestion variable is low. To obtain how many trials we need to identify the status of each link, we need the average length of a probe path d and on how many paths b a link lies on. For an arbitrary overlay network, we calculate the approximated value of d and b in Lemmas 6 and 5 respectively. Using these two values we show that the advanced method identifies the status of a link in $O(n)$ probing with a very high probability when the network is 20% congested or less.

Lemma 5. *For an arbitrary overlay network with e edge routers, on the average, a link lies on $\frac{e(3e-2)}{8 \ln e}$ edge-to-edge paths.*

```
AdvancedMethod()
begin
  Conduct the simple method. Outcome is an unsolved equation set 𝔼.
  for Each undecided variable X_ij of 𝔼 do
    node1 = FindNode(Tree T, v_i, IN) /*See Definition 4 for description of IN and OUT direction.*/
    node2 = FindNode(Tree T, v_j, OUT)
    if node1 ≠ NULL AND Node2 ≠ NULL then
      Probe(Node1, Node2). Update equation set 𝔼.
    end if
    Stop if no probe exists
  end for
end

FindNode(Tree T, Node v_i, dir)
begin
  if v_i is leaf then
    return v_i
  end if
  v_k = DecidePath(v_i)
  if v_k = NULL then
    return NULL
  else
    node = FindNode(T, v_k, dir)
  end if
end

DecidePath(Node v_i, integer dir)
begin
  𝕍 ← siblings(v_i)
  for Each v of 𝕍 do
    if (dir=IN AND good(v → v_i)) OR (dir=OUT AND good(v_i → v)) then
      return v /*good(L) ⇔ L is neither congested nor visited.*/
    end if
  end for
  return NULL
end
```

Fig. 7. Advanced method to obtain probes to decide about the status of a congestion variable.

Proof. See Appendix B. □

Lemma 6. *For an arbitrary overlay network with e edge routers, the average length of all edge-to-edge paths is* $\frac{3e}{2\ln e}$.

Proof. See Appendix B. □

Theorem 3. *Let p be the probability of a link being congested. The advanced method can detect the status of a link with probability* $1-(1-(1-p)^d)^b$, *where* $d=\frac{3e}{2\ln e}$ *is the average path length and* $b=\frac{e(3e-2)}{8\ln e}$ *is the average number of paths a link lies on.*

Proof. The probability that a path of length d is non-congested $(1-p)^d$. The probability of having all b paths congested is $(1-(1-p)^d)^b$. Thus, the probability that at least one non-congested path exists is $1-(1-(1-p)^d)^b$. □

The detection probability in the advanced method (Theorem 3) is plot in Fig. 8 for Topology 1. This figure shows the probability that a good (non-congested) path exists for any link. The congestion status of the network is varied on the X-axis. Two graphs are shown: one shows the probability that a good path exists. It provides the upper bound because the solution cannot be better than this limit. If no path exists, the advanced method cannot do

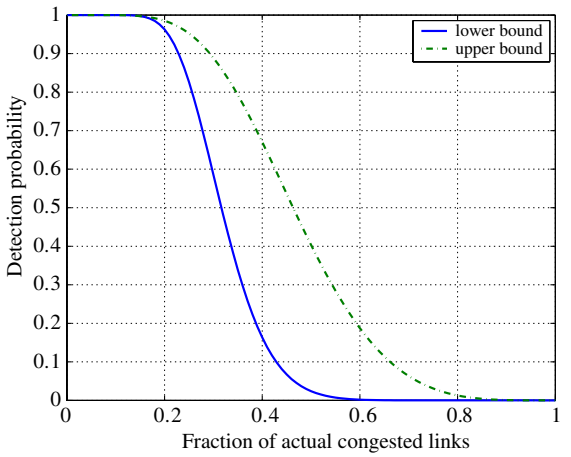

Fig. 8. Probability that the advanced method determines the status of a link of topology shown in Fig. 2a. The X-axis is the fraction of links that are actually congested. The Y-axis is the probability to identify the status of a link. The dotted graph is plot with existing good paths. The solid graph is plot with good and decided path from the first round. These two provide the bound of the solution.

anything. The other graph shows the probability that a good as well as decided path exists. This provides the lower bound because it uses the decided links from the simple method and the solution cannot be worst than this. The advanced method needs only *one probe* on the average to identify the status of the link when the network is less than 20% congested. In this case, the total required probes is $O(n)$. Some links might need more than one, which is not high because a good and decided path exists. If the network is 20–50% congested, the advanced method might need multiple probes to decide the status of one unknown variable in $\mathbb{E}$. If the network is more than 50% congested, the advanced method cannot find a good path easily because the path does not exist, and the advanced method terminates quickly. When the network is highly congested, we need to check almost all the flows. We can go to the detection phase instead of wasting time to rule out very few good links.

The performance of the advanced method is not significant when the network is heavily congested. It raises the question whether it is worth to use the advanced method when the network is highly congested. Instead, we can apply only the simple method, and go to the second phase of monitoring directly after that. In this case, we might need to check flows at most of the routers any way. Thus, we should go to the advanced method if the congestion is below a certain level. The question is how do we know about the congestion level. Fortunately, the simple method can do it. Even though, the simple method cannot identify all the congested links, it can give a good idea about the congestion using Fig. 4. For example, Fig. 4 shows that the detection probability is 12% when the network is 50% congested. Therefore, if the simple method can detect the status of 12% links, we know that the network is 50% congested, and skip the advanced method. Thus, the algorithm to auto select simple and advanced method as follows: First, we conduct the simple method. Then, we determine the level of congestion from Fig. 4. If congestion level is less than a specified threshold (50%), only then we go to the advanced method. We proceed to the second phase (Section 4 of monitoring with this outcome.

3.4. General network topology

The simple and the advanced methods are applicable to a network topology with a tree structure only. If there is any loop in the topology or multiple paths from one edge router to another edge router, we need to preprocess the topology before applying the algorithm. A related work for multicasting can be found in [7], which can be plugged in to our work. In this section, we describe a simple approach to solve this problem.

First, we split the topology into a spanning tree, and a set of subtrees that may be connected or not. The algorithm is applied to all separate trees to identify the congested links. We might have multiple probe paths from one edge router to another. In this case, we apply source routing for probe packets to follow the specified route. Some of the subtrees may not be connected to edge routers, i.e., some parts of subtrees may consist of only core routers. To probe those links, we need to connect them to edge routers. We should be careful to connect these internal links with non-congested links. When all subtrees are probed, we need to combine them. As probing any path does not affect other paths, applying our scheme on any tree will

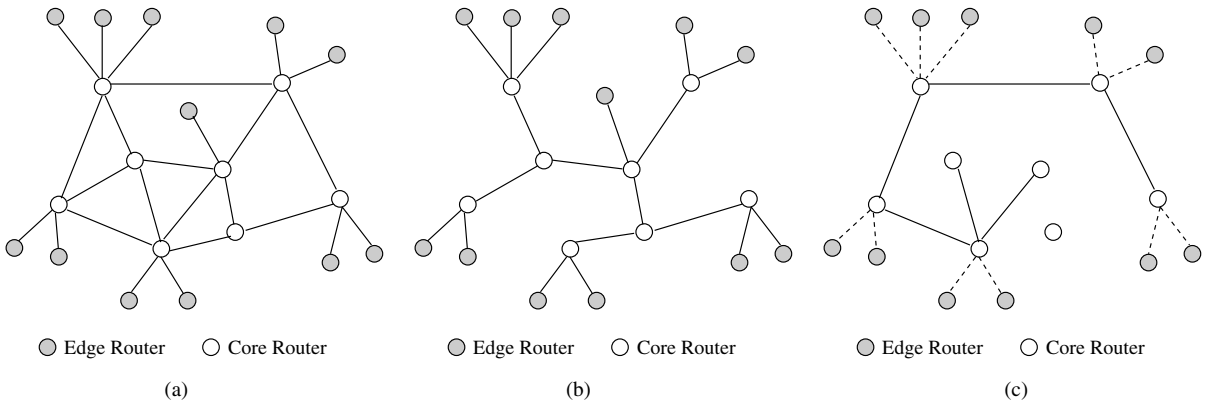

Fig. 9. Preprocessing of a general tree topology to apply distributed probing. The original topology is split into tree topologies. Then, the results are aggregated to get overall picture of a network. (a) Original topology, (b) converted tree topology and (c) rest of the topology.

not affect the others. We obtain the union of all congested links from each topology as a final set of congested links for the whole topology.

In Fig. 9, the general topology (Fig. 9a) is split into two trees. The first one (Fig. 9b) is a spanning tree for the general topology. The other one (Fig. 9c) is a tree where two core routers are not connected to any edge router. We need to add links to these core routers so that we can access this link from edge routers. When probing on Fig. 9b is done we select some good links to connect these core routers with the edge routers. At the end, all results can be combined together to reflect the overall status of the topology. The topology preprocessing is done infrequently only when a network is setup, and when any link or router is added.

We note that Fig. 9a follows a similar pattern of the Sprint topology reported by Spring et al. [31]. For simplicity, we use this one instead of the actual backbone topology of Sprint. However, we can convert any arbitrary topology into tree structure to apply our monitoring algorithm.

3.5. Limitations of distributed monitoring

There are some limitations for the distributed monitoring approach. For example, in Fig. 3a, if both $X_{2,3}$ and $X_{2,4}$ are congested, we cannot decide about $X_{1,2}$. Because we need at least one non-congested outgoing link from core router 2 to decide about the link $X_{1,2}$. The argument is the same for $X_{2,1}$ when both $X_{3,2}$ and $X_{4,2}$ are congested. If all links have the same bandwidth, we can report all three links as congested. Even if $X_{1,2}$ ($X_{2,1}$) has the combined capacity of the two outgoing (incoming) links, the argument is still valid. As long as any non-congested core $\rightarrow$ edge link exists, our method can provide partial solution. If not, the algorithm will report one non-congested link as congested, which is a close approximation of actual result.

The worst case is when all links from the edge routers to the core routers are congested in a network domain. In this case, the outcome of all probes will be congested. The final solution of the simple and advanced method is *all links are congested*. This solution is useful because it is very likely that the whole network is congested when all edge $\rightarrow$ core links are congested. Thus, we can also go to the detection phase considering the whole network is congested. This is also true when all core $\rightarrow$ edge links are congested. If some combinations of $E \rightarrow C$ or $C \rightarrow E$ are not congested, we can use them to provide a partial solution for the network.

4. Detecting violations and attacks

4.1. Violation detection

Violation detection is the second phase of our monitoring process. When delay, loss, and band-

width consumption exceed the pre-defined thresholds, the monitor decides whether the network experiences a possible SLA violation. The monitor knows the existing traffic classes and the acceptable SLA parameters per class. For each service class, we obtain bounds on each SLA parameter that is used as a threshold. A high delay is an indication of abnormal behavior inside a network domain. If there is any loss for the guaranteed traffic class, and if the loss ratios for other traffic classes exceed certain levels, an SLA violation is flagged. This loss can be caused by some flows consuming bandwidths above their SLA_{bw}. Bandwidth theft is checked by comparing the total bandwidth obtained by a user against the user's SLA_{bw}. The misbehaving flows are controlled at the ingress routers.

4.2. Detecting DoS attacks

To detect DoS attacks, set of links L with high loss are identified. For each congested link, $l(v_i, v_j) \in L$, the tree is divided into two subtrees: one is formed by leaves descendant from v_i and the other is formed by the leaves descendant from v_j. The former subtree has egress routers as leaves through which high aggregate bandwidth flows are leaving. If many exiting flows have the same destination IP prefix, either this is a DoS attack or they are going to a popular site [26]. Decision is taken by consulting the destination entity. In case of an attack, we control it by triggering filters at the ingress routers, which are leaves of the subtree descendant from v_i and feeding flows to the congested link. For each violation, the monitor takes action such as throttling a particular user's traffic using a flow control mechanism.

A scenario of detecting and controlling DoS attack is now illustrated using Fig. 10a. Suppose, the victim's domain $\mathscr{D}$ is connected to the edge router $E6$. The monitor observes that links $C3 \rightarrow C4$ and link $C4 \rightarrow E6$ are congested for a specified time duration Δt s. From both congested links, we obtain the egress router $E6$ through which most of these flows are leaving. The destination IP prefix matching at $E6$ reveals that an excess amount of traffic is heading towards $\mathscr{D}$ connected to $E6$. To control the attack, the monitor

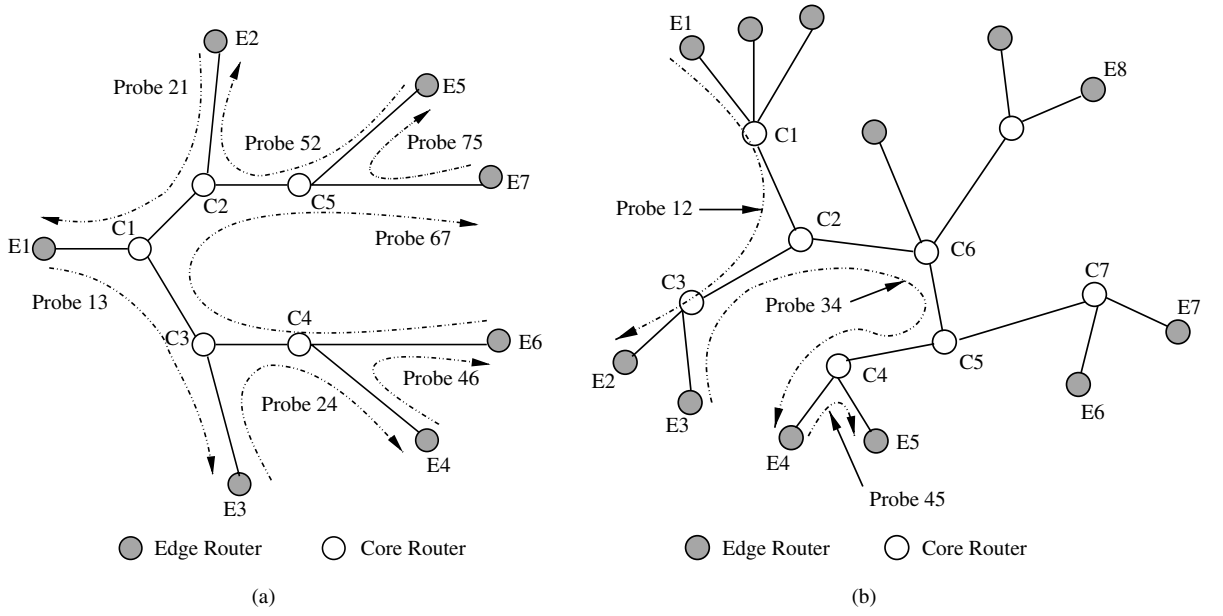

Fig. 10. Topology used to detect service violations using distributed probing. All edge routers are connected to one or multiple domains. All core to core router links are 20 Mbps with 30 ms delay and core to edge router links are 10 Mbps with 20 ms delay. The probes are named with the subscripts of the edge routers. (a) Topology 1 and (b) Topology 2.

needs to identify the ingress routers through which the suspected flows are entering into the domain. The algorithm to identify these ingress routers is discussed in next subsection.

4.3. Flow aggregation and filtering

An important question is how to identify ingress routers through which the flows are entering into the domain. To identify the flow aggregation, we use delay probes. An ID is assigned to each router. An ingress router puts its ID on the delay probe packet. The egress router knows from which ingress routers the packets are coming. For example, in Fig. 10a, say egress router $E6$ is receiving flows from $E1$, $E2$, $E3$, and $E5$. These flows aggregate during their trip to $E6$, and makes the link $C4 \rightarrow E6$ congested. We traverse the path backwards from the egress router to the ingress routers to obtain the entry points of the flows that are causing attacks. In this example, all edge routers can feed the congested links, and they all will be candidates for activating filters. Knowing the ingress routers and congested links, we figure out the entering routers for the flows that are causing the attacks.

5. Simulation results

The performance of our monitoring mechanism is evaluated using simulation. Attacks are simulated by injecting excessive amount of traffic through multiple edge routers. First, we provide experiments to measure SLA parameters that shows the algorithms described in Section 3 work properly. Then, we conduct experiments on detecting service violations and attacks.

5.1. Measuring parameters and monitoring

We use a network topology shown in Fig. 10a, which is similar to the one used in [16,20] to evaluate stripe-based loss ratio approximations. We compare our distributed monitoring with the stripe-based monitoring scheme [20]. Fig. 10b is a more complex topology, which is used to show what happen when multiple attacks happen simultaneously, and one changes the behavior of the others. Multiple domains (not shown in the Fig. 10) are connected to the edge routers for both topologies to create flows along all links in the domain. In Topology 1, flows coming through $E1$, $E2$, $E3$ are destined to edge router $E6$ to make the link $C4 \rightarrow E6$ congested. Many other flows are created to ensure that all links carry a significant number of flows.

Interested readers are referred to [20] for detail analysis of our delay and throughput measurement experiments. In this paper, we show how delay pattern changes with excessive traffic in a domain. We measure delay when the network is properly provisioned or over-provisioned (and thus experiences little loss). When idle, the edge-to-edge delay of $E1 \rightarrow E6$ link is 100 ms. When there is an attack, the average delay of the $E1 \rightarrow E6$ link is increased to as high as 180 ms. Fig. 11 shows how the delay increases in presence of attacks. When there is no attack, the edge-to-edge delay is close to the link transmission delay. If the network path $E1 \rightarrow E6$ is lightly loaded, for example with a 30% load, the delay does not go significantly higher than the link transmission delay. Even when the path is 60% loaded (medium load in Fig. 11), the edge-to-edge delay of the link $E1 \rightarrow E6$ increases by 30%. Some instantaneous values of delay go as high as 50% of the link transmission delay, however, the EWMA does not fluctuate a lot. Excess

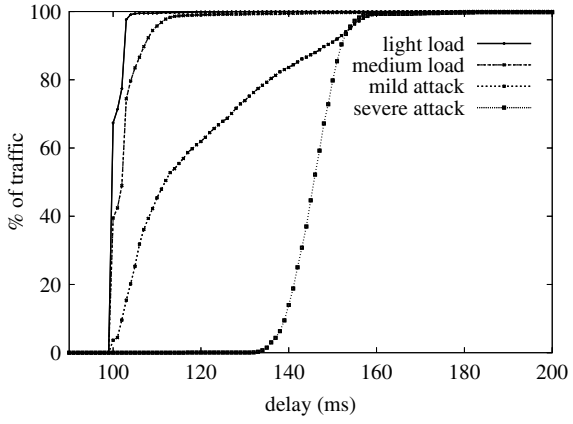

Fig. 11. Cumulative distribution function (CDF) of edge-to-edge link delay for link $E1 \rightarrow E6$. The delay changes with network traffic load.

traffic introduced by attackers increases the edge-to-edge delay inside a network domain. Most of the packets of attack traffic experience a delay 40–70% higher (Fig. 11) than the link delay. Delay measurement is thus a good indication of the presence of excess traffic inside a network domain.

Now, we demonstrate how the distributed probing detects congested links in a network domain. Some of the hosts that are connected to domains attached with the edge routers violate SLAs. They inject more traffic through multiple ingress routers to conduct an attack on the link $C4 \rightarrow E6$. The intensity of the attack is increased during the interval from $t = 15$ to $t = 45$ s. The attack causes around 35% of packet drops except an initial jump at 15 s.

To identify the congested links, the edge routers probe their neighbors. Fig. 12 shows that Probe 46 in counterclockwise direction and Probe 76 in clockwise direction experience high losses. Other probes do not face high losses, that is, most of the internal links are not congested. It is important to note that Probe 46 experiences high loss, however, Probe 64—which is in the opposite direction to Probe 46—faces very small amount of loss. It verifies the properties shown by Savage [30] that the link loss in both directions of a link can be very different, based on the traffic load on each direction. Using algorithm specified in Section 3, we detect that link $C4 \rightarrow E6$ is the only congested link in the domain. We conduct the same experiment for stripe-based monitoring to infer the loss of all individual links. The experiment shows that only link $C4 \rightarrow E6$ has high losses (30%), which means only link $C4 \rightarrow E6$ is congested.

All points in Fig. 12 are calculated by taking averages of samples over one second time period. If we take the average over a longer time period, we can avoid this high fluctuations of loss. Fig. 13 shows that taking averages over a longer time period reduces the chance of considering a non-congested link as congested. It helps more in reducing the fluctuations than in increasing the number of probes per second. The actual loss for this congested link is high (Fig. 14), which verifies the results of the distributed probing.

5.2. Local vs. global congestion

We address the question what happens if the congestion status is changed during the probing. To show an example, we use the Topology 2 (Fig. 10b). This topology is more complex, and we simulate congestion in such a way that congestion in one area might affect the congestion of another area. Two attacks are simulated in this case. The first attack (Attack 1) is due to excessive flows coming from different edge routers to make the

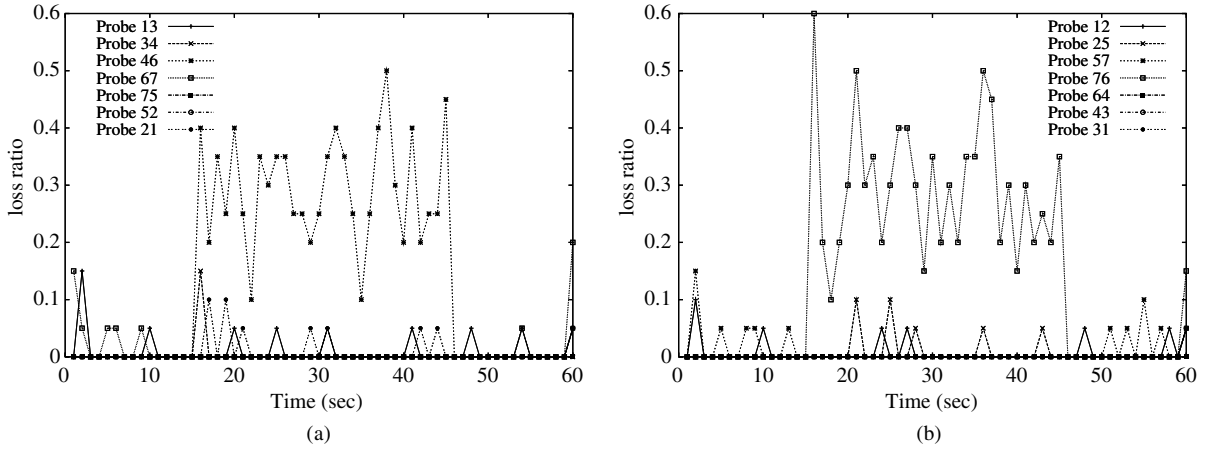

Fig. 12. Probe outcome both for counterclockwise and clockwise direction. Probe 46 in (a) and Probe 76 in (b) have high losses, which means that link $C4 \rightarrow E6$ is congested. (a) Counterclockwise probing. (b) Clockwise probing.

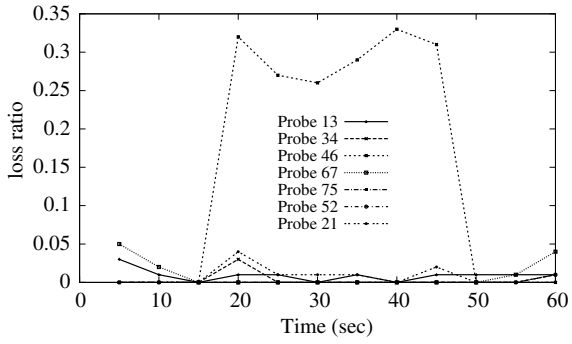

Fig. 13. Probe outcome using 5-s averages for the same experiments shown in Fig. 12a.

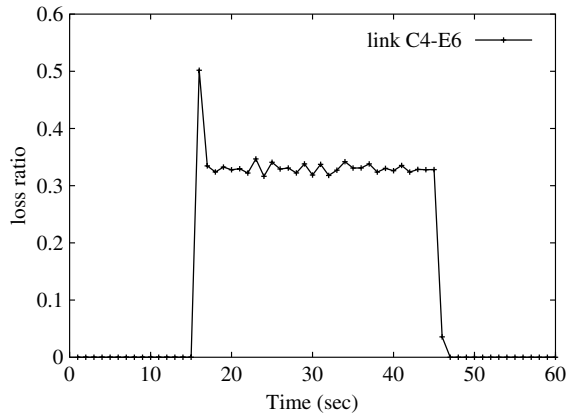

Fig. 14. Actual loss in link $C4 \rightarrow E6$. Other links have low losses. This verifies that our monitoring scheme detects the congestion properly.

link $C4 \rightarrow E5$ congested. All of the probes in the first round are good except "Probe 45". This attack continues up to time $T = 50$ s (Fig. 15). At time 50 s, we have another attack (Attack 2), which is more severe than Attack 1. This attack causes several links on "Probe 34" path congested. It is interesting to note that Attack 2 actually causes Attack 1 to be disappeared. Because most of the traffic that causes Attack 1 on the link $C4 \rightarrow E5$ are now dropped earlier in their path due to Attack 2 (Fig. 15).

This experiment shows that a local congestion might disappear due to a global and severe congestion. The main objective of our work is to pin

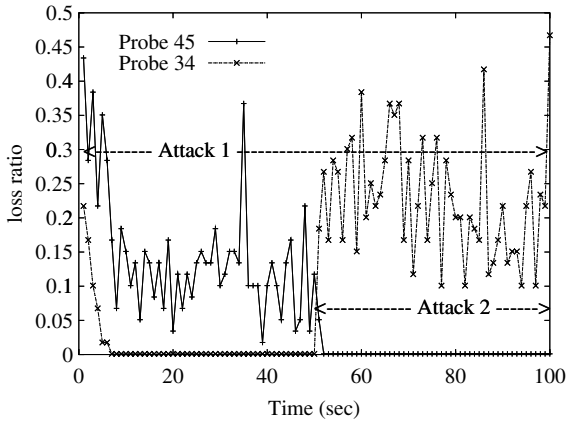

Fig. 15. Attack 1 causes link $C4 \rightarrow E5$ congested. However, Attack 2 comes from all different edge routers to E4, which causes the traffic of Attack 1 to drop early. As a result Probe 45 is not congested after 50 s.

point a congestion. However, if the congestion is changed while an experiment is being conducted, it catches the latest congestion. The simple method can complete two rounds of probing within 10–20 s. If both rounds of probing are done in parallel, it takes only 10 sec. If a congestion does not last for 20 s, we believe that no action is necessary to alleviate it.

5.3. Detecting attacks

A major advantage of using the SLA monitor is that it is able to detect denial of service (DoS) and Distributed DoS (DDoS) attacks in a network domain. When the monitor detects an anomaly (a high delay or a high loss), it polls the edge devices to obtain the throughput of existing flows. The egress routers measure the outgoing rate of each flow. Using these rates, the monitor computes the total bandwidth consumption by any particular user. The bandwidth obtained by an user is compared to SLA_{bw} of that user. If any flow gets very high bandwidth than it should, a DDoS attack is flagged. A DoS attack in a downstream domain can be detected by identifying the congested links, and the egress routers connected to the congested links. Using destination IP address prefix matching [26], we check whether many flows are aggregating

towards a specific network or host. Consulting with the destination object, we control these flows at the ingress routers, if necessary.

We demonstrate the detection of *no attack* and *severe attack*. "No attack" means no significant traffic in excess of the capacity. This scenario has little loss inside the network domain. This is the normal case of proper network provisioning and enforcing traffic conditioning at the edge routers. A severe attack injects excessive traffic into the network domain from different ingress points. At each ingress point, the flows do not violate the profiles but overall they do. The intensity of the attack is increased during $t = 15$ s to $t = 45$ s. The severe attack causes packet drops of more than 35%. Fig. 16 shows that the edge-to-edge delay is increased more than 100% in presence of severe attack. The outcome of one round of loss probing is shown in Fig. 17. The distributed schemes detects high losses in links $E2 \rightarrow C2$, $C1 \rightarrow C3$, $C3 \rightarrow C4$, and $C4 \rightarrow E6$. The link $C4 \rightarrow E6$ has a high loss for a short period of time. Since, some TCP flows adjusted their rates, and it causes the link to be non-congested again. The egress router for the exiting flows is $E6$, and ingress routers through which flows enter into the domain are $E1$, $E2$, $E3$, $E4$, and $E5$, where the filters are activated to control DoS attacks. No traffic came from $E7$.

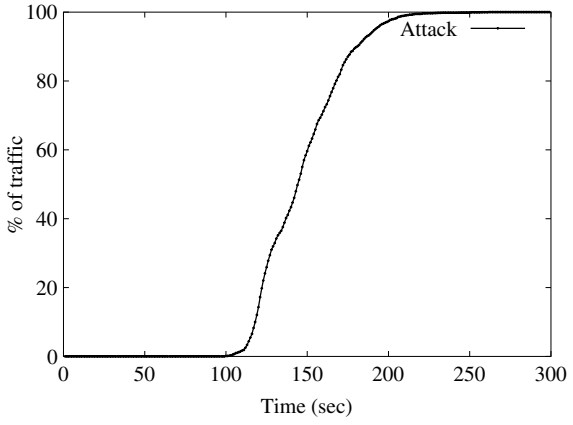

Fig. 16. Cumulative distribution function of edge-to-edge delay for link $E1 \rightarrow E6$. High delay indicates presence of severe attack in the domain.

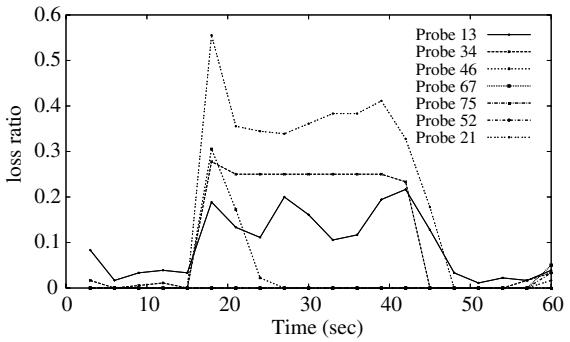

Fig. 17. Congestion on multiple probe paths due to severe attack. It indicates multiple links are having high losses.

6. Advantages of distributed monitoring

A detailed comparison among core-assisted monitoring, stripe-based monitoring, and overlay network-based distributed network monitoring is provided in [21]. In this paper, we provide several important advantages of the distributed monitoring over the stripe-based monitoring. These are as follows:

1. The simple method of distributed probing requires $O(n)$ probes to identify congested links whereas the stripe-based scheme requires $O(n^2)$ [20], where n is number of edge routers in the domain. The advanced method requires $O(n)$ probes when the network is less than 20% congested, however, it does not exceed $O(n^2)$ in worst case.
2. The distributed scheme is able to detect violations in both directions for any link in the domain, whereas the stripe-based method can detect any violation only if the flow direction of the misbehaving traffic is the same as the probing direction from the root. To achieve the same result as the distributed monitoring, the stripe-based method needs to probe the whole tree from several different points requiring $O(n^3)$ probes.
3. The distributed scheme can use TCP-based loss measurements (e.g. Savage [30]) to detect losses in both directions in one probe cycle.
4. In the stripe based scheme, two leaves/receivers are probed at a time. It takes a long time to

complete probing the whole tree. If all leaves are probed simultaneously, in our example, $E1 \rightarrow C1$ link will face huge amount of traffic at the same time. On the other hand, the distributed scheme can do parallel probing quite naturally.

7. Conclusions

We have developed a distributed network monitoring scheme to keep a domain safe from service violations and bandwidth theft attacks. We do not measure actual loss of all internal links, instead, we identify all congested links with high losses using network tomography and overlay networks. Our analytic analysis (verified by simulation) shows that even if 20% links of a network are congested, the status of each link can be identified with probability $\geqslant 0.98$. If the network is 40% congested, this probability is still high (0.65). However, if the network is more than 60% congested, this method cannot achieve anything significant since almost every edge-to-edge path has one or more congested links. This new tomography scheme requires only $O(n)$ probes when less than 20% links are congested, where n is the number of edge routers. For an OC3 link, the probe traffic to identify the congested links is 0.002% of link capacity. The distributed monitoring requires $O(n^2)$ in worst case in contrast to $O(n^3)$ probes required by the stripe-based monitoring to detect attacks in both directions of all links. The distributed monitoring conducts probing in parallel enabling the system to perform real time monitoring. The simulation results indicate that the proposed scheme detects service violations, bandwidth theft attacks, and DoS attacks caused by flow aggregation towards a victim network domain.

Acknowledgements

The authors thank Mohamed Hefeeda and Leszek Lillen for their valuable comments. This research is sponsored in part by the NSF grants ANI 0219110, CCR-001712, and CCR-001788, CERIAS and IBM SUR grant.

Appendix A

Proof of Lemma 1. Let the core router C is connected to only two other routers R_1 and R_2 (Fig. 18). No probe path can be constructed that either includes the link $R_1 \rightarrow C$ and does not include $C \rightarrow R_2$ or vice versa; or includes $R_2 \rightarrow C$ and does not include $C \rightarrow R_1$ or vice versa. The traffic that passes through the link $R_1 \rightarrow C$ also passes through $C \rightarrow R_2$. The traffic that passes through the link $R_2 \rightarrow C$ also passes through $C \rightarrow R_1$. Therefore, for the purpose of probing, a logically equivalent overlay network can be constructed by replacing $R_1 \leftrightarrow C \leftrightarrow R_2$ with $R_1 \leftrightarrow R_2$. We say that the link $R_1 \rightarrow R_2$ is congested if and only if at least one of the links $R_1 \rightarrow C$ and $C \rightarrow R_2$ is congested, i.e. the bandwidth of $R_1 \rightarrow R_2$ is the minimum of the bandwidths of $R_1 \rightarrow C$ and $C \rightarrow R_2$. Similarly, the bandwidth of $R_2 \rightarrow R_1$ is the minimum of the bandwidths of $R_2 \rightarrow C$ and $C \rightarrow R_1$. □

Proof of Lemma 3. Let the link $R_1 \rightarrow R_2$ (Fig. 19) appears in path $\mathscr{P}$ in the first round of probing and path $\mathscr{P}'$ in the second round of probing. If R_2 is a core router, it is connected to at least two other routers, say R_3 and R_4 (Lemma 2). $\mathscr{P}$ passes through the link $R_2 \rightarrow R_4$ and $\mathscr{P}'$ passes through the link $R_2 \rightarrow R_3$. Since the tree cannot have any cycle, $\mathscr{P}$ and $\mathscr{P}'$ never meet again. If R_2 is an edge router, both $\mathscr{P}$ and $\mathscr{P}'$ terminates at R_2. Therefore, $\mathscr{P}$ and $\mathscr{P}'$ cannot have any common link in their

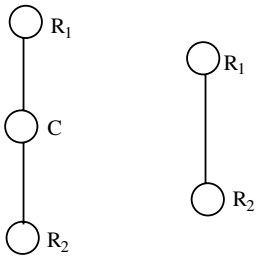

Fig. 18. Merging links that do not contribute in distributed probing.

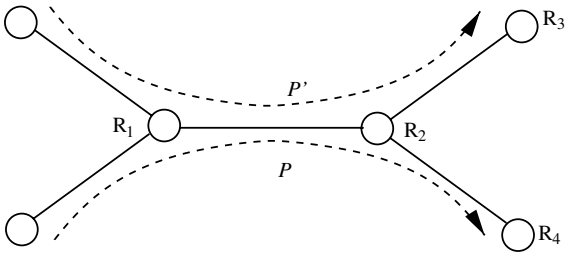

Fig. 19. Intersection of probe paths $\mathscr{P}$ and $\mathscr{P}'$. If R_2 is an edge router, $R_2 \to R_3$ and $R_2 \to R_4$ do not exist.

paths after node R_2. Similarly, it can be shown that $\mathscr{P}$ and $\mathscr{P}'$ cannot have common links before they meet at node R_1. That is $|\mathscr{P} \cap \mathscr{P}'| \leqslant 1$. □

Appendix B

Proof of Lemma 5. Consider an arbitrary link l in a tree T. If l is removed from T, it forms two subtrees, say T_1 and T_2. The link l lies on an edge-to-edge path whose one end belongs to T_1 and another end belongs to T_2. Let the number of edge routers in T_1 and T_2 be i and $e-i$ respectively. The total possible paths through l is $i(e-i)$. We observe that the probability that T_1 contains i edge routers is, $q_i \propto 1/i$, (approximately, if the tree is not heavily skewed), i.e. $q_i = k/i$. The average number of paths the link l lies on, $b = \sum_{i=1}^{e/2} q_i.i.(e-i)$.

Now, $\sum_{i=1}^{e/2} q_i = \sum_{i=1}^{e/2} k/i = 1$, i.e. $k = 1/\ln e/2$. Therefore,

$$b = \sum_{i=1}^{e/2} k(e-i) = \frac{e(3e-2)}{8\ln e/2}$$
$$= \frac{e(3e-2)}{8\ln e - 8\ln 2} \approx \frac{e(3e-2)}{8\ln e}. \qquad \square$$

Proof of Lemma 6. There are $e(e-1)$ edge-to-edge paths exist for the advanced method. The number of links in a topology is $\approx 4e$ (see the proof of Theorem 2). The average length of a path $d = b \times \frac{4e}{e(e-1)}$, where $b = \frac{e(3e-2)}{8\ln e}$ is the average number of paths a link lies on (Lemma 5).

Now,

$$d = \frac{e(3e-2)}{8\ln e} \times \frac{4e}{e(e-1)} = \frac{3e-2}{2\ln e} \times \frac{e}{e-1} \approx \frac{3e}{2\ln e}$$

(for large e). □

References

[1] A. Adams, T. Bu, R. Caceres, N. Duffield, T. Friedman, J. Horowitz, F. Lo Presti, S.B. Moon, V. Paxson, D. Towsley, The use of end-to-end multicast measurements for characterizing internal network behavior, IEEE Communications Magazine 38 (5) (2000) 152–159.

[2] E. Al-Shaer, H. Abdel-Wahab, K. Maly, HiFi: a new monitoring architecture for distributed systems management, in: Proceedings of the IEEE 19th International Conference on Distributed Computing Systems (ICDCS '99), Austin, Taxas, May 1999, pp. 171–178.

[3] K.G. Anagnostakis, M.B. Greenwald, R.S. Ryger, On the sensitivity of network simulation to topology, in: Proceedings of the 10th IEEE/ACM Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunications Systems (MASCOTS 2002), October 2002.

[4] K.G. Anagnostakis, S. Ioannidis, S. Miltchev, J. Ioannidis, M. Greenwald, J.M. Smith, Efficient packet monitoring for network management, in: Proceedings of the IEEE Network Operations and Management Symposium (NOMS), Florence, Italy, April 2002.

[5] D. Anderson, H. Balakrishnan, F. Kaashoek, R. Morris, Resilient overlay network, in: Proceedings of the ACM Symp on Operating Systems Principles (SOSP), Banff, Canada, October 2001.

[6] Y. Breitbart, C.Y. Chan, M. Garofalakis, R. Rastogi, A. Silberschatz, Efficiently monitoring bandwidth and latency in IP networks, in: Proceedings of the IEEE INFOCOM, Anchorage, Alaska, April 2001.

[7] T. Bu, N.G. Duffield, F. Lo Presti, D. Towsley, Network tomography on general topologies, in: Proceedings of the ACM SIGMETRICS, Marina del Rey, California, June 2002.

[8] R. Cáceres, N.G. Duffield, J. Horowitz, D. Towsley, Multicast-based inference of network-internal loss characteristics, IEEE Transactions on Information Theory 45 (1999) 2462–2480.

[9] R. Callon, P. Doolan, N. Feldman, A. Fredette, G. Swallow, A. Viswanathan, A framework for multiprotocol label switching, Internet draft, November 1997.

[10] J. Case, M. Fedor, M. Schoffstall, J. Davin, A Simple Network Management Protocol (SNMP), IETF RFC 1157, May 1990.

[11] M.C. Chan, Y.-J. Lin, X. Wang, A scalable monitoring approach for service level agreements validation, in:

Proceedings of the International Conference on Network Protocols (ICNP), Osaka, Japan, November 2000, pp. 37–48.
[12] Cisco, Netflow services and applications, Available from <http://www.cisco.com/>, 2002 May 2000.
[13] M. Coates, R. Nowak, Network tomography for internal delay estimation, in: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Salt Lake City, Utah, May 2001.
[14] M. Dilman, D. Raz, Efficient reactive monitoring, in: Proceedings of the IEEE INFOCOM, Anchorage, Alaska, April 2001.
[15] N.G. Duffield, M. Grossglauser, Trajectory sampling for direct traffic observation, IEEE/ACM Transactions on Networking 9 (3) (2001) 280–292.
[16] N.G. Duffield, F. Lo Presti, V. Paxson, D. Towsley, Inferring link loss using striped unicast probes, in: Proceedings of the IEEE INFOCOM, Anchorage, Alaska, April 2001.
[17] N.G. Duffield, J. Horowitz, F. Lo Presti, D. Towsley, Network delay tomography from end-to-end unicast measurements, in: Proceedings of the 2001 International Workshop on Digital Communications 2001 Evolutionary Trends of the Internet, September 2001.
[18] P. Ferguson, D. Senie, Network Ingress Filtering: Defeating Denial of Service Attacks which Employ IP Source Address Spoofing Agreements Performance Monitoring, IETF RFC 2827, May 2000.
[19] S. Floyd, K. Fall, Promoting the use of end-to-end congestion control in the Internet, IEEE/ACM Transactions on Networking 7 (4) (1999) 458–472.
[20] A. Habib, S. Fahmy, S.R. Avasarala, V. Prabhakar, B. Bhargava, On detecting service violations and bandwidth theft in QoS network domains, Computer Communicatons 26 (8) (2003) 861–871.
[21] A. Habib, M. Hefeeda, B. Bhargava, Detecting service violations and DoS attacks, in: Proceedings of the Network and Distributed System Security Symposium (NDSS '03), San Diego, California, February 2003, pp. 177–189.
[22] IEPM, Internet End-to-end Performance Monitoring, Available from <http://www-iepm.slac.stanford.edu/> 2002.
[23] C. Ji, A. Elwalid, Measurement-based network monitoring and inference: scalability and missing information, IEEE Journal on Selected Areas in Communications 20 (4) (2002) 714–725.
[24] J. Kim, J.W. Hong, Distributed QoS monitoring and edge-to-edge QoS aggregation to manage end-to-end traffic flows in differentiated services networks, Journal of Communications and Networks 3 (4) (2001) 324–333.
[25] A. Liotta, G. Pavlou, G. Knight, Exploiting agent mobility for large-scale network monitoring, IEEE Network 16 (3) (2002) 7–15.
[26] M. Mahajan, S.M. Bellovin, S. Floyd, J. Ioannidis, V. Paxson, S. Shenker, Controlling high bandwidth aggregates in the network, ACM Computer Communication Review 32 (3) (2002) 62–73.
[27] V. Paxson, Measurement and analysis of end-to-end Internet dynamics, Ph.D. thesis, University of California, Berkeley, Computer Science Division, 1997.
[28] V. Paxson, G. Almes, J. Mahdavi, M. Mathis, Framework for IP Performance Metrics, IETF RFC 2330, May 1998.
[29] G. Sager, Security fun with OCxmon and cflowd, Internet2 Working Group Meeting, November 98.
[30] S. Savage, Sting: a TCP-based network measurement tool, in: Proceedings of the USENIX Symposium on Internet Technologies and Systems (USITS '99), Boulder, Colorado, October 1999.
[31] N. Spring, R. Mahajan, D. Wetherall, Measuring ISP topologies with rocketfuel, in: Proceedings of the ACM SIGCOMM, Pittsburgh, Philadelphia, August 2002.
[32] R. Stone, Centertrack: an IP overlay network for tracking DoS floods, in: Proceedings of the USENIX Security Symposium, Denver, Colorado, August 2000.
[33] R. Subramanyan, J. Miguel-Alonso, J.A.B. Fortes, A scalable SNMP-based distributed monitoring system for heterogeneous network computing, in: Proceedings of the High Performance Networking and Computing Conference (SC 2000), Dallas, Texas, 2000.
[34] S. Waldbusser, Remote Network Monitoring Management Information Base, IETF RFC 2819, May 2000.
[35] Y. Zhang, N.G. Duffield, V. Paxson, S. Shenker, On the constancy of Internet path properties, in: Proceedings of the ACM SIGCOMM Internet Measurement Workshop, November 2001.

Ahsan Habib received B.S. in Computer Science and Engineering from Bangladesh University of Engineering and Technology, Bangladesh. He received M.S. in Computer Science from Virginia Tech, Blacksburg and Ph.D. in Computer Science from Purdue University, West Lafayette in 1999 and 2003 respectively. His research interests include network security, network economics, peer-to-peer networks, and distributed systems. Currently, he is a postdoctoral researcher in the School of Information and Management Systems, University of California at Berkeley.

Maleq Khan received B.S. in Computer Science and Engineering from Bangladesh University of Engineering and Technology, Bangladesh and M.S. in Computer Science from North Dakota State University, ND. He is currently working toward the Ph.D. degree in Computer Science at Purdue University, IN. His research interests include wireless sensor networks, communication networks, and data mining. His main research concern is developing energy-efficient routing scheme for self-configuring sensor networks.

Bharat Bhargava received his B.E. degree from Indiana Institute of Science and M.S. and Ph.D. degrees in EE from Purdue University. He is a professor of computer sciences at Purdue University. His research involves both theoretical and experimental studies in distributed systems. Currently, he is working in secure mobile systems, multimedia security and Quality of Service (QoS) as a security parameter. He has proposed schemes to identify vulnerabilities in systems and networks, and assess threats to large organizations. He has developed techniques to avoid threats that can lead to operational failures. These ideas and scientific principles are being applied to the building of peer-to-peer systems, cellular assisted mobile ad hoc networks, and to the monitoring of QoS-enabled network domains. He is a Fellow of the Institute of Electrical and Electronics Engineers and of the Institute of Electronics and Telecommunication Engineers. He has been awarded the charter Gold Core Member distinction by the IEEE Computer Society for his distinguished service. In 1999 he received IEEE Technical Achievement award for a major impact of his decade long contributions to foundations of adaptability in communication and distributed systems.