FAD and SPA: End-to-end Link-level Loss Rate Inference without Infrastructure

Yao Zhao, Yan Chen
{*yzhao and ychen*}*@cs.northwestern.edu*

It is highly desirable and important for end users, with no special privileges, identify and pinpoint faults inside the network that degrade the performance of their applications. However, existing tools are inaccurate to infer the link-level loss rates and have large diagnosis granularity (in terms of the number of hops). To address these problems, we propose a suite of user-level diagnosis approaches in two categories: (1) only need to be deployed at the source and (2) deployed at both source and destination. For the former, we propose two fragmentation aided diagnosis approaches (FAD), Algebraic FAD and Opportunistic FAD, which uses IP fragmentation to enable accurate link-level loss rate inference. For the latter category, we propose Striped Probe Analysis (SPA) which significantly improves the diagnosis granularity over those of the source-only approaches. Internet experiments are applied to evaluate each individual schemes (including an improved version of the state-of-the-art tool, Tulip [1]) and various hybrid approaches. The results indicate that our approaches dramatically outperform existing work (especially for diagnosis granularity) and provide not only the best performance but also smooth tradeoff among deployment requirement, diagnosis accuracy and granularity.

I. Introduction

It is highly desirable and important for end users, with no special privileges, identify and pinpoint faults inside the network that degrade the performance of their applications. However, the modern Internet is heterogeneous and largely unregulated, which renders Internet fault diagnosis an increasingly challenging problem. The servers and routers in the network core are usually operated by businesses, and those businesses may be unwilling or unable to cooperate in collecting the network traffic measurements vital for Internet fault diagnosis.

Internet Tomography denotes a class of techniques that infer link level properties [2]–[5] based on end-to-end measurements. Generally, Internet tomography requires *a measurement infrastructure*, usually a set of end hosts, on which special measurement tools are deployed. However, normal users or companies usually only have access to a few end hosts and thus tomography is not available to them. More importantly, the diagnosis problem often demands *on-demand online* measurement for a particular path. Thus it is desirable to have a handy diagnosis tool that only needs to be deployed on one or both end hosts of the target path. Such design often has to leverage router response, such as Tulip [1] and cing [6]. However, as shown in Section II, the existing tools are inaccurate and cannot give very fine-level diagnosis. For example, the state-of-the-art tool Tulip accurately infers the loss rate of the forward path only when (1) the reverse path is not lossy and (2) there is no strong correlation on the loss of the forward path, because the control packets and data packets have be sent within very short period (*e.g.* 3.5ms suggested in [1]).

We improve Tulip by fixing the first problem, but the second one is inherent in its design. To address these challenges, in this paper, we propose a suite of schemes for link-level loss rate inference from end-to-end measurements without any infrastructure. We consider two categories: (1) measurement tools can only be deployed at the source (called *source only*) and (2) tools can be deployed at both source and destination (called *source+destination*).

For the first category, we propose *fragmentation aided diagnosis (FAD)* approaches, which use packet fragmentation to obtain extra measurement information to differentiate loss on the forward path vs. loss on the reverse path. We design two variants of FAD, called *Algebraic FAD (AFAD)* and *Opportunistic FAD (OFAD)* respectively. Note that such fragmentation happens at the network layer, which leaves us full flexibility to choose upper layer (*e.g.*, transport layer) protocols for link-level diagnosis. In addition, IP fragmentation is well defined in IPv4 and most routers and hosts support IP fragmentation. In addition, we discuss some practical issues, such as prefix subpath problem and packet loss correlations, as well as solutions to overcome these problems.

For the second category, with extra destination support, we propose Striped Probe Analysis (SPA) which improves the diagnosis granularity to the lower bound (close to each

physical link).

We implemented and deployed these tools on the PlanetLab testbed, and designed our experiments carefully through: 1) studying the correlation between probe packet size and loss rate to get representative probe packet size; and 2) calibrating the link loss rate inference results and excluding measurement outliers.

Then we evaluated the diagnosis granularity and accuracy of each individual schemes as well as those of various combination schemes. For source-only diagnosis, we found that FAD approaches have comparable diagnosis granularity to Tulip but they are much more accurate than even the improved version of Tulip. Furthermore, the combination of FAD and Tulip can significantly reduce the diagnosis granularity and also provide improved accuracy (especially for OFAD+Tulip).

When adding support from the destination, SPA achieves the best possible diagnosis granularity, but its accuracy is not as good as OFAD and AFAD. However, we found the combination of SPA and OFAD can effectively solve this problem and offers both good accuracy and granularity.

For the rest of the paper, we first introduce the related work, especially Tulip in Section II. Then we present the source-only diagnosis approaches in Section III and the source+destination scheme in Section IV. We discuss the evaluation methodology in Section V and show the results in Section VI. Finally, we conclude in Section VII.

II. Related Work

Packet loss rate is an important metric of the QoS of a network. For example, throughput of TCP streams is severely affected even by very small loss rate, because packet loss is used as the signal of the existence of congestion now. Ping, Zing [7] are two well known path-level loss rate measurement tools, and recently Badabing [8] is proposed to improve the accuracy of loss rate measurement. These tools only measure the end-to-end loss property, but do not attempt to locate where the lossy links are. Our focus in this paper is more challenging, *i.e.*, inferring the loss rates in link level. Link-level diagnosis can be put into two categories: *infrastructure based approaches* [2]–[5, 9] and *router response based approaches* [1, 6].

Traditional Internet tomography approaches fall into the first class. With an overlay network infrastructure, Internet tomography can infer the loss rate with the granularity up to each virtual link (*i.e.*, sequence of consecutive links without a branching point) with high probability. Multicast-based tomography can achieve unbiased inference on the loss rates of each virtual link [3]. However, IP multicast is not widely available in the Internet and thus unicast-based tomography [4] was proposed as an approximation. The unicast-based tomography tries to mimic multicast by exploiting the transmission correlation. But it only works well when the two back-to-back probes are always both lost or are both transmitted successfully, *i.e.*, perfect transmission correlation.

Router-based approaches rely on response packets sent by routers on the path to be diagnosed. Tulip is the latest representative of this category [1]. Basically, some routers use an increasing counter for the IP-ID field of the response packets generated by the router (*i.e.*, consecutive IP-IDs) and Tulip uses that feature for diagnosis as follows. The sender sends multiple probes to each router (with the appropriate TTL) on the path. In each probe, three packets, two short control packets (Packets 1 and 3) separated by one long data packet (Packet 2), are sent with certain intervals between the packets. The router then sends back corresponding responses to the sender.

By checking the IP-ID fields of the response packets, Tulip can infer the loss rate of the forward path. Let binary random variable $X_i = 1$ if packet i is received by the probed router and $X_i = 0$ otherwise. Similarly, we denote $Y_i = 1$ if the response to packet i is received by the sender and $Y_i = 0$ if such response is lost on the reverse path. Assume that Tulip sends out m probes (3 packets in each probe) to a router. Among the response, n of them only contain response triggered by the control packets (Packets 1 and 3) with the right IP-IDs so that we know the data packet is lost on the forward path. Then Tulip uses n/m as the estimated data packet loss rate on the forward path. That is, it uses the probability $P(X_1 = 1, X_2 = 0, X_3 = 1, Y_1 = 1, Y_3 = 1)$ to approximate $P(X_2 = 0)$. Obviously, Tulip tends to underestimate loss rate of the forward path, unless the short packets 1 and 3 are never dropped on the forward path and their responses are never lost on the reverse path. In reality, short packets can also be lost and Tulip may severely underestimate the loss rates as we show in Section VI-C.2.

In addition, Tulip has conflicting requirements for probe packet correlation which further compounds the inference accuracy problem. On one hand, Tulip desires the transmission independence among the the three packet in a probe. Otherwise, if they are 100% correlated, the three probe packets are either all lost or all go through. Then the event of $(X_1 = 1, X_2 = 0, X_3 = 1)$ will never happen no matter how large $P(X_2 = 0)$ is. That is, Tulip will always get zero loss rate estimation. On the other hand, to avoid the interference of the cross traffic on the continuous IP-ID in the response packets, the three packets in a probe of Tulip are sent within 3.5 msec, which means the transmission (or loss) of the three packets in a probe is probably correlated. This conflict is fundamental to the design of Tulip. Our FAD can also be affected by the loss correlation. However, as we will discuss in Section III-B.2, we can put enough interval between the fragmented packets so that their loss correlation is small.

III. Source-only Diagnosis

In this section, we first present our design to improve Tulip. Then we propose to apply fragmented packet based

measurements to do link-level diangosis. We propose two such FAD schemes. One is based on an extra algebraic loss rate equation introduced by packet fragmentation. We call it *Algebraic FAD*, or $AFAD$. The other explores the opportunity that most of the loss rates on the Internet paths are not very large. We term it *Opportunistic FAD*, or $OFAD$. We first introduce the basic ideas for each approach to distinguish the forward loss and reverse loss. Then we extend them to achieve link-level diagnosis, and discuss practical issues, such as the packet loss correlation as well as our solutions.

A. Improvement to Tulip

We design the following scheme to improve Tulip. We use the same notations in Section II. In addition, let n' be the expected number of cases that we received both response packets triggered by the two short packets (Packets 1 and 3). Thus n'/m gives $P(X_1 = 1, X_3 = 1) \times P(Y_1 = 1, Y_3 = 1)$. Then we can estimate the forward loss rate as:

$$\begin{aligned} \hat{p_f} &= \frac{n/m}{n'/m} = \frac{P(X_1=1,X_2=0,X_3=1)\times P(Y_1=1,Y_3=1)}{P(X_1=1,X_3=1)\times P(Y_1=1,Y_3=1)} \\ &= \frac{P(X_1=1,X_2=0,X_3=1)}{P(X_1=1,X_3=1)} \\ &= P(X_2 = 0|X_1 = 1, X_3 = 1) \end{aligned} \quad (1)$$

We use the assumption that the transmission (or loss) on the forward path and the reverse path is independent in the deduction [10, 11]. This estimate is no longer affected by the reverse path loss rate, and thus has better inference accuracy, especially for these paths which are very lossy as shown in Section VI-C.2. On the other hand, this improved Tulip scheme still suffers from the packet loss correlation problem which is inherent to the use of IP-ID to differentiate loss on the forward path vs. on the reverse path. So next, we will introduce our FAD schemes to overcome this problem.

B. AFAD for Forward Path Diagnosis

1) Basic Algebraic Idea: Assume p, p_f and p_r are the loss rate of the round-trip path, the forward path and the reverse path respectively. Given only support from the source, we can easily get the route-trip loss rate (p), *e.g.*, using ping. Thus we can obtain the following equation with two variables:

$$(1 - p_f) \times (1 - p_r) = 1 - p \quad (2)$$

However, if we can somehow change such "one request to one response" pattern, we can obtain more equations. For example, imagine there is an scenario that each probe from the source has i packets, and only when all of the packets are received, the destination will send back j replies ($i \neq j$). *When the loss of the packets are random and independent*, we get a new equation (p' is the new round-trip loss rate) as follows.

$$1 - p' = (1 - p_f)^i \times (1 - p_r)^j \quad (3)$$

Given Equation 2 and 3, we can easily solve the two variables p_f and p_r. However, we leverage on an assumption that the loss of packets are not correlated. This assumption may not always be true, and we will justify it in the next subsection.

There are many methods to implement the above idea, *i.e.,* to introduce Equation 3. For example, if the destination provides http service, we may be able to send one http request for a large web page (or a figure) and then get many replied packets. Unfortunately, Internet routers usually do not open any TCP service to unauthorized users. To achieve link level diagnosis, we need to seek some prevalent responses from routers as well as end hosts. IP fragmentation is the best candidate that we find so far.

Internet Protocol allows IP fragmentation so that datagrams can be fragmented into pieces small enough to pass over a link with a smaller MTU (Maximum Transmission Unit) than the original datagram size [12]. A router does not reassemble IP fragments while forwarding. But routers and end hosts reassemble fragments if they are the destination. Take ping for example, if an ICMP Echo Request datagram is split into two fragments at the prober and sent to a host (a router or an end host), the host will reply with an ICMP Echo Reply only when it can reassemble the ICMP Echo Request, *i.e.,* when it gets both the fragments. By doing so, we actually obtain a case of Equation 3, where $i = 2$ and $j = 1$.

2) The Impact of Packet Loss Correlation and its Solution: To achieve Equation 3, we make an assumption that the loss of packets are uncorrelated. In this section, we study such assumption and propose countermeasures when it is violated.

In our specific scheme of using IP fragmentation, we split a probing packet into two fragments F_1 and F_2. Let X_i be the random variable of whether F_i is received by the probed destination. We set $X_i = 0$ when F_i is lost, and $X_i = 1$ otherwise ($i = 1, 2$). Therefore,

$$P\{X_1 = 1\} = P\{X_2 = 1\} = 1 - p_f.$$

When we consider the correlation of the loss of the two fragments, Equation 3 is changed to be:

$$\begin{aligned} 1 - p' &= P\{X_1 = 1\} \times P\{X_2 = 1|X_1 = 1\} \times (1 - p_r) \\ &= (1 - p_f) \times P\{X_2 = 1|X_1 = 1\} \times (1 - p_r) \end{aligned} \quad (4)$$

By manipulating Equations 2 and 4, we can get the value of $P\{X_2 = 1|X_1 = 1\}$. We actually use $P\{X_2 = 1|X_1 = 1\}$ to estimate p_f, *i.e.*,

$$\begin{aligned} \hat{p_f} &= 1 - P\{X_2 = 1|X_1 = 1\} \\ &= P\{X_2 = 0|X_1 = 1\} \end{aligned} \quad (5)$$

From the above equation, it is clear that we use the conditional loss to estimate the loss rate. When there is no correlation between the loss of the two fragments, *i.e.,* X_1 and X_2 are independent, $P\{X_2 = 0|X_1 = 1\} = p_f$

and our estimation is unbiased. However, if X_1 and X_2 are dependent, our estimation is biased and the inaccuracy depends on the degree of correlation.

Many previous studies show that Internet packet loss has short-term correlation [10, 11, 13]. Bolot found that the loss of probe packets are essentially random when the probe traffic uses a small fraction of the available bandwidth [13]. Most recent work of Zhang *et al.* show that 27% of measured paths have uncorrelated loss, while the remaining paths show significant loss correlations under timescale of 500-1000ms. In addition, both [10] and [11] found that loss rates in a path's two directions are weakly correlated or completely independent.

One straight-forward way to break the loss correlation between two fragments is to have their sending interval sufficiently large. In [12], the lower bound on the reassembly waiting time is recommended as 15 seconds, which is much larger than the correlation timescale of packet loss. Thus if we choose the interval between two fragments as $1000ms$, we can have an unbiased and accurate estimate of the path loss rate.

However, in practice, the interval between two fragments of an IP datagram affects the probing frequency we can take. The buffer allocated by routers to reassemble fragments is limited, as the main task of routers is forwarding instead of receiving as an end host. This means a router can buffer only a few first-half fragments at the same time. For example, if a router can buffer 100 fragments of different IP datagrams, and the interval between the two fragments of a datagram is one second, then the first-half fragment of the 101st datagram should be sent later than the time that the first datagram is reassembled. This means in one second there are at most 100 probes sent out.

In fact, one second interval is a very conservative upper bound of the loss correlation. As we show in Section VI, even the interval of packets are much less than one second, the estimation error introduced by correlation is very small, and thus we can achieve accurate diagnosis while having enough probing frequency.

C. OFAD for Forward Path Diagnosis

Inspired by Tulip, we designed OFAD as follows. In one probe, we create two datagrams (datagram 1 and 2) with the same IP-ID, which share the same first k bytes of IP payload. Therefore, when we split each datagram into two fragments, we can have a common first-half fragment [1]. As shown in Figure 1, we send these three packets with certain interval: the first one (packet 1) is the common fragment, which is usually small; the second one (packet 2) is the second-half fragment of datagram 1; and the third one (packet 2') is the second-half fragment of datagram 2. If receiving packet 1, the destination will assemble a datagram (either datagram 1 or datagram 2) when receiving either packet 2 or packet 2'. The reassembled datagram will trigger a response to the source with enough information to tell which datagram has been reassembled. Once datagram 2 is assembled (case (b) in Figure 1), we are sure that both packet 1 and packet 3 are transmitted successfully while packet 2 is lost on the forward path.

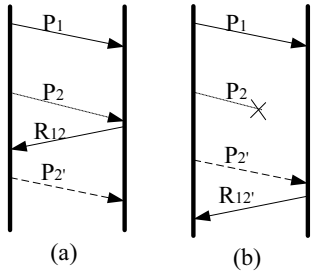

Fig. 1. Inferring loss in OFAD

Let's denote $X_i = 0$ when forward packet i is lost and $X_i = 1$ otherwise. $Y = 0$ means the response packet is lost and $Y = 1$ otherwise. Then the probability of receiving the response to datagram 1 is $P(X_1 = 1, X_2 = 1) \times P(Y = 1)$. The probability of receiving datagram 2 is $P(X_1 = 1, X_2 = 0, X_{2'} = 1) \times P(Y = 1)$. Let n be the expected number of received response of datagram 1 , and n' that of datagram 2. Then we have

$$\begin{aligned} & \frac{n}{n+n'} \\ = & \frac{P(X_1=1,X_2=1)\times P(Y=1)}{P(X_1=1,X_2=1)\times P(Y=1)+P(X_1=1,X_2=0,X_{2'}=1)\times P(Y=1)} \\ = & \frac{P(X_1=1,X_2=1)}{P(X_1=1)(1-P(X_2=0,X_{2'}=0|X_1=1))} \\ \approx & \frac{P(X_1=1,X_2=1)}{P(X_1=1)} = P(X_2=1|X_1=1) \end{aligned} \tag{6}$$

In the above deduction, we assume $P(X_2 = 0, X_{2'} = 0|X_1 = 1)$ to be close to 0, which is true when the loss is not very large. The loss rate of forward path p_f is $P(X_2 = 0)$. When there are loss independence between packets 1 and 2, we have $\hat{p_f} = 1 - \frac{n}{n+n'} = \frac{n'}{n+n'}$. The resulting conditional probability is similar to that of Tulip. However, unlike Tulip, we can choose large interval between these probe packets to achieve small loss dependence for OFAD. Thus OFAD is more accurate than the improved Tulip as verified by the evaluation in Section VI-C.2.

Packet reordering may introduce some potential false loss rate detection in OFAD. For example, if no packets are lost but packet 2' arrives before packet 2, OFAD considers this as a case of packet loss. In practice, the reordering problem does not affect OFAD too much. As the intervals between packet 2 and packet 2' in OFAD can be relatively large (*e.g.* 100ms) to reduce the packet loss correlation, the reordering problem seldom happens. Actually, if the response to datagram 2 is received by the source before packet 2' is sent out, packet 2' is saved. For example, the program can first estimate the round-trip latency with a few probes and set the interval between packet 2 and packet 2' to be larger than the round-trip latency, if the latency is relatively small (*e.g.* $< 500ms$).

D. Link-level Diagnosis of FAD (both AFAD and OFAD)

Given the capability of identifying the loss rate of forward paths, it is straightforward to achieve link-level diagnosis. For example in Figure 2, suppose we can identify the loss

[1]Checksum of ICMP, UDP or TCP payload of the two datagrams should be the same, which can be achieved by carefully padding the payload. Alternatively, we can create two datagrams with the same second fragments.

rate of forward paths p_1 and l_1. We can infer the loss rate of link l_2 by solving the equation $1-p_1 = (1-l_1)\times(1-l_2)$. [2] Similarly, we can identify the loss rate of each link along the end-to-end path from the source to the destination, no matter how many hops the path has. However, there are three practical issues we need to address: selection of FAD probes to be supported by routers, security problem of using IP fragmentation and the prefix subpath problem. Next, we discuss them as well as their solutions.

1) FAD Probes Widely Supported: If a router does not respond to the probes or respond in an unexpected manner, its related link cannot be diagnosed. Then we can only diagnose some link sequence and the diagnosis granularity will be affected. Fortunately, IP fragmentation is executed at the network layer and it does not have any limit on the higher layer protocols. This gives us the flexibility to explore any kind of probes that a router or an end host reacts. Some possible probes are ICMP Echo Request, ICMP Timestamp Request, UDP probe and TCP probe.

Our IP fragmentation based approaches only have two requirements: 1) routers support IP fragmentation, 2) routers respond to any of the four probes listed above. In Section VI-A, we show that more than 80% of routers satisfy both requirements.

It is worth mentioning that IPv6 does not support fragmentation any more, which means FAD cannot be applied to IPv6 network. Actually, this big evolution will invalidate many measurement tools, such as Tulip. IPv6 does not allow fragmentation and there is no IP-ID field in IPv6 packet header. However, it is predicted that there will be a slow adoption of IPv6 (especially in North America) and even adopted, both will co-exist for a long time [14].

2) Normal Amount of Normal Fragmented Packets Acceptable: Another concern on using IP fragmentation is the security issue. There were some security problems related to IP fragmentation, such as Ping of Death Fragmentation Attack and the Teardrop Attack [15]. Also, tiny fragments or overlapping fragments were used to bypass firewalls to gain access to victim hosts. However, normal IP fragments as those we use in measurements, will not cause any security problems. Considering the fact that routers and firewalls may spend more time processing fragmented packets than normal packets, large amount of probing should be avoided. Since we do not want our measurements to cause any congestion, we always send small amount of fragmented probes such as five per second, which has been proved to work well in most routers as shown in Section VI-A.3.

3) Prefix Subpath Problem and Its Solutions: Internet routing often does not take the shortest path and there is big difference between intra- vs. inter-AS routing. Therefore the routing path from the source to an in-between router on an end-to-end path may not be the prefix subpath of the end-to-end path from source to destination (we call it the *prefix subpath problem*). Figure 3 shows a case that the target end-to-end path is $S \to A \to B \to C \to D$, while the routing path from S to B is $S \to E \to B$ and the path from S to C is $S \to E \to B \to C$. In this case, the loss rate of link $C \to D$ can not be inferred by the loss rate of path S to D and path S to C. This prefix subpath problem is pretty common in Internet routing, because Internet routing is a combination of inter-domain routing (BGP) and intra-domain routing. In this example, we can only diagnosis the link $S \to A$ and the link sequence $A \to B \to C \to D$ on the end-to-end path. In FAD, traceroute is executed toward each router in the end-to-end path to identify the prefix subpath problem.

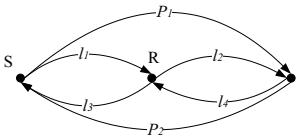

Fig. 2. Example of link level diagnosis.

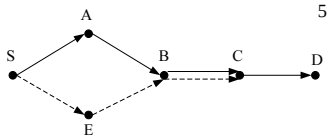

Fig. 3. Example of the prefix subpath problem.

To solve this prefix subpath problem, we propose the following approach. We can infer the loss rate of a prefix subpath of the end-to-end path indirectly. Considering the above example, we show how to infer the loss rate of path $S \to A \to B$. Assume the routing path from B back to S is $B \longrightarrow S$. As mentioned in Section III-B, we can get the loss rate of the reverse path ($B \longrightarrow S$), say p_α, as well as that of the forward path ($S \to E \to B$) by sending probes to router B. At the same time, we also send non-fragmented probes to the end host D while limiting TTL to be 2 (the same as traceroute). This new probes will traverse the path $S \to A \to B$ to B and be replied via path $B \longrightarrow S$. This probe will tell us the total loss rate of the round-trip $S \to A \to B \longrightarrow S$ (say p_β).

Some routers have severe rate limit on the generation of ICMP TTL Exceeded packets. To solve this problem, we also send a small control packet with the same TTL after each probe to check if rate-limiting happens. This is similar to the approach used in Tulip [1] to measure round-trip loss rate in face of rate-limiting. Note that we assume in all the probes the reverse path $B \longrightarrow S$ is always the same. This is generally true, if Internet route is stable. Thus given the loss rate of the round-trip $S \to A \to B \longrightarrow S$, p_β, and the loss rate of $B \longrightarrow S$, p_α, we can compute the loss rate of path $S \to A \to B$ as $1-(1-p_\beta)/(1-p_\alpha)$. Clearly, This approach solves the prefix routing problem and thus the diagnosis granularity is not affected. The tradeoff is that we need to measure more paths and hence with larger measurement overhead.

IV. Source+Destination Diagnosis

As shown in Section VI, FAD is highly accurate to infer the link-level loss rate. However, not all routers support fragmentation, and hence FAD cannot diagnose each individual links. In this section, we present the Striped Probe Analysis (SPA) to achieve the diagnosis granularity up to

[2] We use the same notation for the path (or link) and its loss rate.

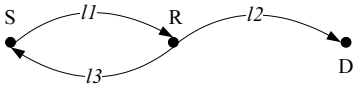

Fig. 4. Striped probe analysis.

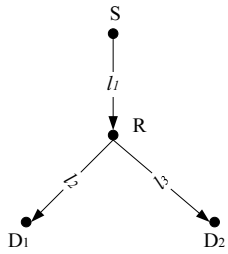

Fig. 5. Two leaf tree

each physical link. SPA requires support from both source and destination, but not any internal routers.

As shown in Figure 4, in SPA, we send two packets in a stripe: packet 1 is sent to the destination D with large enough TTL, and packet 2 is sent to D with a pre-configured TTL so that router R will send back "ICMP TTL-Exceeded error" message. By checking the logs on host D and the received ICMP messages, SPA can estimate the forward transmission (or loss) rate from S to R as follows.

Let X, Y and Z be corresponding random variables of the path segments $S \rightarrow R$, $R \rightarrow D$ and $R \rightarrow S$. The random variable is one if the packet goes through the corresponding path segment successfully and is zero otherwise. For example, $X_1 = 0$ means the packet 1 in the stripe is lost on path segment $S \rightarrow R$.

Based on received packets and responses on end hosts S and D, we know the transmission success rates on path $S \rightarrow R \rightarrow D$ ($P(X_1 = 1, Y_1 = 1)$) and on path $S \rightarrow R \rightarrow S$ ($P(X_2 = 1, Z_2 = 1)$), and the probability of the whole stripe is transmitted successfully is $P(X_1 = 1, Y_1 = 1, X_2 = 1, Z_2 = 1)$.

If we assume that the transmission on different path segments are independent, the transmission rate of path segment $S \rightarrow R$ can be estimated as:

$$\begin{aligned} \hat{q_1} &= \frac{P(X_1=1,Y_1=1) \times P(X_2=1,Z_2=1)}{P(X_1=1,Y_1=1,X_2=1,Z_2=1)} \\ &= \frac{P(X_1=1)}{P(X_1=1|X_2=1)} \end{aligned} \quad (7)$$

If the transmission success correlation of the packets in a stripe on path segment $S \rightarrow R$ is 1 (*i.e.*, $P(X_1 = 1|X_2 = 1) = 1$), this estimation is unbiased. To achieve such strong correlation, the two packets in a stripe should be sent back-to-back.

Since R is selected by the TTL value, SPA can actually infer the loss rate from S to every intermediate router on the path $S \rightarrow D$ as long as the router can generate "ICMP TTL-Exceeded" messages. Hence, every router that responds to traceroute supports SPA and the diagnosis granularity of SPA reaches the lower bound of any diagnosis scheme that relies on traceroute.

Such correlation based statistical inference is also used for network tomography [4] which inspired our design of SPA. As the classical two leaf tree topology shown in Figure 5, S, D_1 and D_2 are the end hosts while R is a router. To infer the loss rates from $S \rightarrow R$, S sends a few back-to-back packets (called a stripe) to D_1 and D_2 respectively. This is analogous to the D and S in our problem. However, their approach usually requires an infrastructure (*i.e.*, multiple destinations) to cooperate for diagnosis. The diagnosis granularity is the unit of virtual link (*i.e.*, the sequence of links without branching point), which depends on the size of the infrastructure and is usually much larger than the close-to-1 granularity (each physical link) achieved by SPA.

V. Measurement Evaluation Methodology

In this section, we describe some of our measurement methodologies. We first discuss the choice of probing packet length, then how to calibrate the measurement results given the statistic nature of measurements. Finally, we list the evaluation metrics of link-level diagnosis, which will be further used in the next evaluation section.

A. Packet Probe Size Selection

In this section, we study whether the packet length affects the loss rates and what length we should use for representative loss rate measurements. Obviously, to save the active measurement overhead, we prefer to send out short probing packets. For example, we save about 97% measurement costs if we use 40-byte probes to take the place of 1500-byte probes. If the loss rates of different packet lengths on the same paths are different, we may have to use the probes of the same length as used by the targeting application.

We deployed UDP senders and sinkers on 120 randomly chosen PlanetLab hosts and measured the 14,280 paths between them. For each measured path, the sender sends out UDP packets of five different lengths, *i.e.*, 40, 200, 576, 1000 and 1500 bytes. The measurement of a path takes about 200 seconds, and we send 1000 probes for each packet size. As shown in [16], the majority of the packets seen are one of three sizes: 40 byte packets (the minimum packet size for TCP) which carry TCP acknowledgements but no payload, 1500 byte packets (the maximum Ethernet payload size) from TCP implementations that use path MTU discovery, and 576 byte packets from TCP implementations that don't use path MTU discovery. We also consider the packet lengths of 200 bytes and 1000 bytes, which are the middle lengths between these three outstanding packet lengths.

We consider both the direct loss rate difference and relative loss rate difference. For example, let the loss rate of 40-byte packets be l_{40} and that of 1500-byte packets be l_{1500}. The loss rate difference is $l_{40} - l_{1500}$ (we always subtract the loss rate of long packets from that of short packets) and the relative loss difference is $(l_{40} - l_{1500})/\max(l_{40}, l_{1500})$. Figures 6 and 7 show the histograms of the loss rate difference and relative loss rate difference between packet sizes of 40 bytes, 576 byes and 1500 bytes. We ignored the results involving packet sizes of 200 bytes and 1000 bytes, because they follow the clear relationship between packet size and loss rate, which can be observed in our

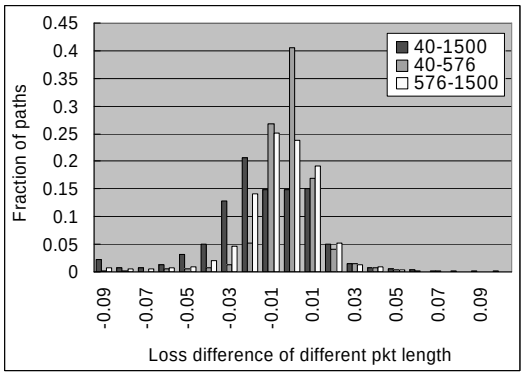

Fig. 6. Histogram of loss rate difference between size 40, 576 and 1000.

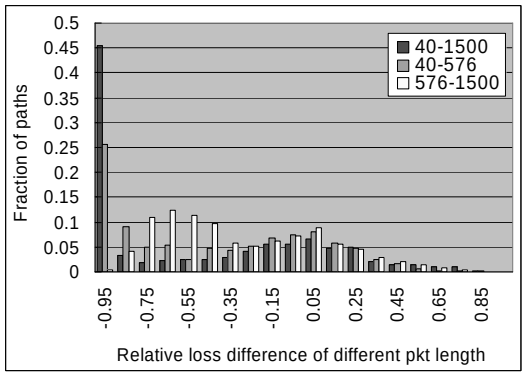

Fig. 7. Histogram of relative loss rate difference between size 40, 576 and 1000.

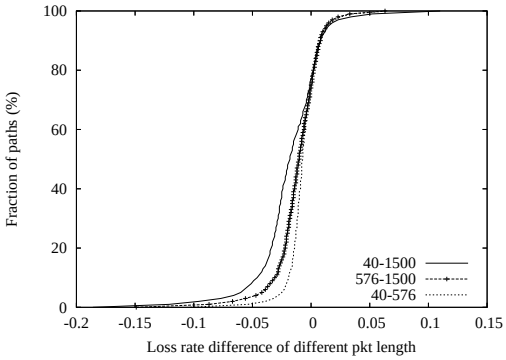

Fig. 8. CDF of loss rate difference between size 40, 576 and 1000.

presented results. From Figures 6 and 7, it is clear that in some paths the loss rate of the long packets is larger than that of the short packets. In the extreme cases, the loss rate of 40-byte packets is 0 while that of 576-byte or 1500-byte packets is significantly larger than 0. For example, in about 45% of cases, the relative loss difference between 40-byte packets and 1500-byte packets is about -0.95%. There are some interpretations on the phenomena. In [1], the authors mentioned that routers are more likely to drop long packets, perhaps due to the lack of buffer space. Another possible reason is the artifact of the bandwidth limiting policy of PlanetLab hosts. PlanetLab hosts are more likely to drop long packets when (shared) bandwidth is limited or the nodes are overloaded. To filter out the potential bias introduced by PlanetLab, we also used PING to measure the round-trip loss rate from the source to the first hop router for each sender. We remove all the paths that loss are observed on the round-trip paths from the senders to the first hop routers. We find that the relationship between packet size and loss rate does not remarkably change, no matter whether we do this filtering. Also shown in the figure, there are some paths on which short packets have similar or even larger loss rates than long packets. From the cumulative distribute function of (relative) loss rate difference (See Figures 8 and 9), on about 20-30% of paths the (relative) loss difference is non-negative. And in a few cases, the relative loss difference is close to 0.8. Currently, we do not have an explanation other than measurement errors for these extreme cases, as we are not aware of any scheme that prefers to drop short packets over long packets. In summary, in PlanetLab network, the loss rate of a path is likely to be related to the packet size. While using short probes may reduce measurement cost, you may miss some loss events that can be observed by large probing packets. In our evaluation part (Section VI), we infer the loss rate of 1500 bytes unless otherwise mentioned. And the default probe rate is 1000 probes sent in 200 seconds.

B. Calibration of Loss Rate Inference

Under ideal circumstances, as we extend the path segments, the loss rates only increase (if the newly included link or links are lossy) or stay the same (if the new link or links have no loss), but they should never decrease. Therefore, the loss rates of the path segments on a end-to-end path make a step function of the length of the path segment (See Figure 10). In reality, loss rate measurements may have some measurement errors, and hence the inferred loss rates of forward path segments may also have some errors. If we assume that the loss rate measurements are independent binomial experiments (as in [17]), then the measured loss rate has a binomial distribution. If the loss rate of the path is p and n probes are sent with Poisson intervals, the variance is $p(1-p)/n$, and therefore the standard deviation is $\sqrt{p(1-p)/n}$. For example, if $p = 0.05$ and $n = 1000$, then the standard deviation is about 0.7%. With a probability of 95%, measured loss rates are within $(0.0365, 0.0635)$.

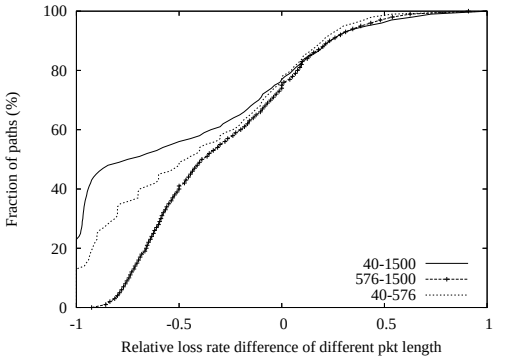

Fig. 9. CDF of relative loss rate difference between size 40, 576 and 1000.

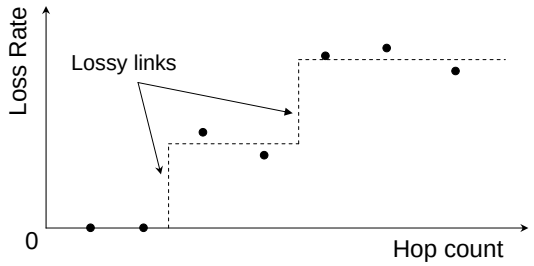

Fig. 10. Matching the loss rates to a step function

The variance of the inference schemes is related to the variance of the simple loss rate measurement process, but it could be much more complicated. Meanwhile, some routers may have unexpected behaviors, which may result in some remarkable inaccuracy in loss rate inference. For example, if a router has strict rate-limiting on ICMP packets and we can only receive 10 responses out of 1000 probes, we cannot trust any inference result from this measurement. Figure 10 shows an example of the inferred loss rates of the forward path segments. As shown in Section VI-C.3, these loss rate diagnosis schemes are consistent in most cases, which means a path segment usually has no less loss rate than its sub-path segments. But there are exceptions, which are probably caused by measurement variance or some unexpected events. Our objective of calibration is to fit the data into a step function, which helps remove measurement outliers, and reveals the location as well as the loss rate of lossy links (as in Figure 10).

We first preprocess the data and filter out some outliers that may be caused by unexpected behaviors of routers (such as rate-limiting on replying ICMP packets). Here are some heuristic schemes as follows.

- For a path segment, we infer the forward loss rate while at the same time measuring the round-trip loss rate. If the round-trip loss rate of a path segment is much larger than those of other path segments (especially longer path segments), then it is quite possible that this measurement suffers rate-limiting or there is outstanding loss rate on the reverse path. In either case, the loss rate inference on the forward path segment is prone to be inaccurate, and we discard the data point before fitting to the step function.
- For each diagnosis scheme, there are assumptions. If an assumption is violated in a measurement, this measurement will be discarded. For example, if obvious reordering of packets is found in a Tulip measurement, this measurement will be filtered. For AFAD, if the inferred forward loss rate is much larger than the round-trip loss rate, this may be caused by the special rate-limiting only on fragmented packets, and hence the corresponding data pointer is removed.

After preprocessing the data, the next step is to fit the loss rates of the path segments to a step function. In [18], a dynamic programming approach is introduced to fit gap sequences into a step function. The situation in [18] is similar to ours, and we adopt the approach proposed in [18].

C. Metrics

We consider the following two metrics:

- *Diagnosis granularity*: We define a router on an end-to-end path to be *diagnosable* if we can infer the loss rate of the forward subpath (of the end-to-end path) from the source to the router. A path segment between two diagnosable routers is a *diagnosable segment*. Diagnosis granularity of a single path is defined as the weighted average of the lengths of its diagnosable segments, as used in [1]. For example, if an 8-hop path has two diagnosable segments of length 3 and 5, the granularity of the path is $(3^2 + 5^2)/8$=4.25. This metric represents the expected length of diagnosable lossy segments if a lossy link distributes in the path randomly.
- *Accuracy*: To compare the inferred loss rate $\hat{p_f}$ with the real loss rate p_f of the forward path, we use both the *estimation error* ($\hat{p_f} - p_f$) and the *relative error* $F_\varepsilon(p_f, \hat{p_f})$ defined as follows:

$$F_\varepsilon(p_f, \hat{p_f}) = \frac{|\hat{p_f} - p_f|}{\max(\varepsilon, p_f)}$$

We choose ε as 1%, which is used to avoid the division by zero problem.

VI. Measurement Evaluation Results

We implemented AFAD, OFAD, Tulip and SPA, and deployed them in the Internet for evaluation. In this section, we first present the results on the prevalence of router support, and then analyze the diagnosis granularity and accuracy results. We consider both individual diagnosis schemes and various combinations of them for evaluation.

A. Prevalence of Router Support

In this subsection, we study how widely FAD is supported by Internet routers. In the next subsection, we compare the diagnosis granularity of FAD and other schemes.

1) Router Collections: To obtain a large number of router IP addresses, we randomly generate destination IP addresses, run traceroute to these IPs and collect routers on the paths from a computer in our institute. We filter the paths with length less than 8 hops because in most cases, short paths are due to failures of route to the random IP (*e.g.*, the IP is in an unassigned IP block). We measured altogether 72,874 paths in March 2006, which involved 64,320 router IP addresses. The number of routers is smaller than the number of paths, because some paths may find same routers (Note that end hosts are not counted). In most cases (93.3% of paths), traceroute cannot give all routers on the path. The last several hops are usually "* * *". The main reason is that these randomly generated IPs can be unused IPs, and thus traceroute cannot find the destinations. The average length of all the traceable sub-paths is about 15.1 hops, which is close to the typical path length in the Internet [19].

	Echo	Timestamp	UDP	TCP	Any
1 source	85.3%	69.2%	64.5%	71.7%	88.2%
11 sources	87.3%	72.3%	70.7%	73.3%	90.1%

TABLE I
ROUTER RESPONSE TO DIFFERENT PROBES

2) Support of Different Probes: We sent 5 packets for each of the four types of probes listed in Section III-D.1 to these 64,320 IPs from multiple sources. We use 10 PlanetLab nodes and one PC in out institute as the sources in our experiments. Table I shows the fraction of responsive routers for different types of probes. For example, if only the source in one major university is used, we find that 85.3% of routers respond to ICMP Echo and 69.2% of routers support ICMP Timestamp requests. About 88.2% of routers reply to at least one type of probe. If all these sources are used, the number of responsive routers will increase about 2% to 5%. For UDP, responsive routers increase most when the number of sources increases. This is partially because UDP probes are severely rate limited by routers and thus are likely to be affected by cross traffic.

3) Support of IP Fragmentation: Although IP fragmentation is required to be supported by both routers and end hosts in IPv4 networks, we find that in practice about 90.3% of routers that respond to at least one type of probes support IP fragmentation. Thus altogether, about 80% of routers support FAD. This means FAD is widely supported by the current IPv4 Internet. By examining the routers that do not support IP fragmentation, we find many of them are from `sprintlink.com`, `verizon-gni.net`, `cox.net`, `wcg.net`, `telia.net` and `atlas.cogentco.com`. However, these routers usually do not filter IP fragments, which means they forward IP fragments as common IP datagrams. Also, we find that the buffer size of routers to reassemble IP fragments are usually larger than or equal to 10 packets.

4) Degree of Rate Limiting on Responses: It is well known that ICMP packets are prone to being rate limited. ICMP rate limiting may significantly affect the router response based approaches because the dropped ICMP packets will be counted in the packet loss of the paths. We select 8,000 routers that support all four kinds of probes from the router pool and send probes to them with a frequency of 100Hz. We find that for ICMP Echo, ICMP Timestamp and TCP probes, more than 99% of routers allow a rate of 100 probes/sec, as we receive the responses with negligible losses. However, UDP probes to more than 60% of routers suffer severe rate limiting. Therefore it seldom suffers rate-limiting to use ICMP Echo, ICMP Timestamp and TCP probes for diagnosis, especially when the probe frequency is low (*e.g.*10 probes/s). In addition, when one type of probe suffers rate limiting, we may still be able to switch to another kind of probe which does not have this problem.

B. Diagnosis Granularity Results

1) Results of Individual Schemes: Based on the definition, in the ideal case, if all the routers that are discovered by traceroute are diagnosable, we achieve the finest diagnosis granularity. This is the lower bound of any diagnosis approach that leverages on traceroute, at least to find the router. This is also the lower bound of almost all Internet tomography approaches. SPA can reach this lower bound because SPA only requires that the routers support the ICMP TTL-Exceeded packets.

By testing whether a router supports fragmented probes, we get the diagnosis granularity of FAD approaches. We exclude all the routers that only support UDP probes because "ICMP Port Unreachable" packets are usually rate limited (See Section VI-A.4). Note that a Planetlab host does not allow users to send out fragmented packets, but it does the fragmentation itself when the packet length is larger than the MTU (usually 1500 bytes). Similarly, we obtain the diagnosis granularity of Tulip by checking which routers support IP-ID.

To compare the diagnosis granularity of Tulip, FAD and SPA, we check the granularity of these schemes on the same set of paths. Note that if all the tested paths are from the same source, these paths usually share long prefix paths. Thus the diagnosis granularity comparison is biased and heavily depends on the routers close to the source. Therefore, we randomly select 60 PlanetLab hosts, and each of them runs traceroute to 500 random IP addresses. Then altogether we have 60 sources and 30,000 destinations.

Figure 11 shows the cumulative distribution function (CDF) of the diagnosis granularity of different schemes. The average diagnosis granularity of SPA is about 1.09 hops, which means nearly every physical link is diagnosable. For FAD, the average diagnosis granularity is 2.74 hops because not all the routers support IP fragments. The median of diagnosis granularity is 1.88 hops. With Tulip, the average diagnosis granularity is 2.71 hops, which is slightly better than that of FAD, but the median is 2.24 hops, which is larger than that of FAD. That is, although FAD has better diagnosis granularity for most of the paths, for about 10%-20% of paths FAD has very large granularity. This is because, as shown in Section VI-A.3, some ISPs tend to disable IP fragmentation on most of their routers. So when a path goes through such ISPs, it includes a long diagnosable path segment composed of routers in such ISPs. This leads to large path diagnosis granularity because the metric of granularity gives more weight for long diagnosable path segment than shorter ones, as in its definition.

2) Results of Combined Schemes: Different diagnosis schemes rely on different probe response support of routers. Thus the combination of them is very likely to have better diagnosis granularity. In Section VI-C.2, we also show that such combinations can improve the loss rate inference accuracy.

SPA has the finest possible diagnosis granularity and cannot be improved further. For the source-only schemes,

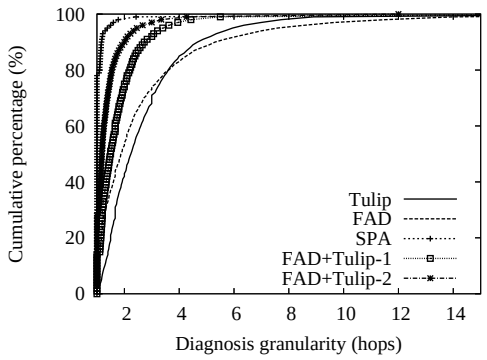

Fig. 11. Diagnosis granularity comparison.

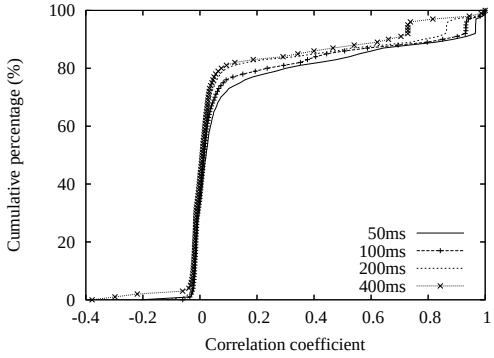

Fig. 12. The CDF of packet transmission correlation coeffi cient.

Tulip and FAD, there are multiple ways of combining them. One simple combination of FAD (OFAD or AFAD) and Tulip is to select the scheme with the finer diagnosis granularity for the target path. For example, given a path, if the diagnosis granularity of FAD is three and that of Tulip is two, we will use Tulip for that path and hence the diagnosis granularity is two. The resulting diagnosis granularity of this approach is shown in Figure 11 as "FAD+Tulip-1". The average diagnosis granularity is improved to 1.76 hops, and the median is 1.5 hops. The more aggressive combination is to consider a router diagnosable if the router supports the probe response packet either for FAD or for Tulip. This hybrid scheme gives the best diagnosis granularity that the combination of FAD and Tulip can offer, but as discussed in Section VI-C.2, it can be a complicated problem to determine the loss rate of each diagnosable path segment. We show its diagnosis granularity as "FAD+Tulip-2" in Figure 11 with mean being 1.38 hops and median 1.19 hops.

C. Diagnosis Accuracy Results

In this section, we first study the packet transmission correlation on the Internet and how it is affected by packet transmission intervals. This is important to understand the limitation of Tulip and to choose the interval parameter for FAD. Then we evaluate the accuracy of different schemes through path-level experiments due to lack of the ground truth on link-level loss rates. Finally, we run the link-level loss rate inference and show the consistency test results.

1) Packet Transmission Correlation: Packet transmission (or loss) correlation is a very important factor that affects the accuracy of SPA, Tulip, OFAD and AFAD. In this section, we study how it varies with different time intervals between packets.

We randomly choose 100 PlanetLab hosts and randomly measure 5000 paths between them with a probing frequency of 20Hz. We calculate the packet transmission correlation coefficient ($\hat{\rho_f}$) of certain time intervals (50ms, 100ms, 200ms and 400ms) of each experiment. As most Internet paths are no loss, we exclude all the paths with loss rate less than ε (1% as defined in Section V-C) and only consider the lossy paths. Figure 12 shows that when the interval is as small as 50ms, in about 80% of paths the transmission correlation coefficient is close to 0, which means the transmission correlation is negligible. A larger interval such as 400ms has a small correlation, however, the improvement is marginal. About 20% of paths show strong transmission correlation. After manually checking them, we found that these are caused by long loss episode in these paths. One possible reason is that PlanetLab hosts have strict traffic rate limiting, and thus long loss episode happens when network traffic is above the quota.

2) Path-Level Accuracy Evaluation: Basically, there are two major sources of inaccuracy: transmission correlation and measurement errors. SPA desires strong transmission correlation, while AFAD, OFAD and Tulip prefer transmission correlation to be as weak as possible. Since it is hard to get the real loss rates of Internet links as the ground truth for validation, we apply a path-level validation for these schemes. Once the schemes have accurate loss rate inference on the forward path, it is very likely that they can achieve accurate link-level diagnosis as well.

To this end, we implemented the application level FAD, Tulip and SPA with UDP packet probes. For example, one fragmented packet is implemented with two UDP packets, with each containing a certain payload to indicate its role. Similarly, the destination will simulate the reassembly process and respond with another UDP packet when two "UDP fragments" of one datagram are received successfully. Tulip control packets are also UDP in the application-level evaluation. The consecutive IP-ID is simulated by the payload of UDP packets. By introducing this application-level simulation, we bypass the problem that PlanetLab hosts do not allow IP fragmentation. In addition, we do not have the rate-limiting problem of routers.

We randomly selected 120 PlanetLab hosts and measured the 14,280 paths between these PlanetLab nodes. We did the experiments twice in Jul 2006, so 28,560 paths were measured altogether. We send 1000 probes [3] for each of the four schemes in 200 seconds. Tulip and the improved Tulip (marked as "I-Tulip") use the same measurements and they only differ in the loss rate inference. The probe frequency is low because we measure these four schemes at the same time for each path and higher probe frequency can easily trigger the rate-limiting of the PlanetLab hosts. The interval

[3]A probe may have different number of packets for different schemes.

between the two fragments of a probe in AFAD is 100ms, while the interval between the three fragments of a probe in OFAD is 50ms. Among all the measured paths, a majority of them have no loss or very low loss rate, and a few paths are obviously rate-limited by the PlanetLab hosts. By removing them, we have 5136 paths with round-trip loss rate of no less than 0.5% which are used for accuracy evaluation.

Results of Individual Schemes: Figures 13 and 14 show the inference errors of the individual schemes. I-Tulip is slightly better than Tulip in our experiments because most paths have very small loss rate, and hence the improvement of I-Tulip is not obvious. Clearly Tulip, I-Tulip, OFAD and SPA are prone to underestimate the loss rate. This is because of the intrinsic bias of these three schemes. AFAD does not show an obvious trend to underestimate or overestimate loss rate. First, this again confirms that the transmission correlation is small with reasonably large packet intervals. Second, the measurement error is the main source of inaccuracy for AFAD. By manually checking the data, we find that AFAD tends to be inaccurate in the case that the reverse path is quite lossy while the forward path is good. When there exists a certain measurement error on the reverse path, this measurement error is brought into the loss inference on the forward path in AFAD. For example, if the reverse path has a loss rate of 10% and there is no loss on the forward path, it is possible that the round-trip loss rate of a fragmented probe is 11% while that of a non-fragmented probe is 9%. Therefore, the forward path loss rate is inferred as about 2%. Comparing these four schemes, OFAD is the most accurate one, while SPA is the least accurate one. Overall, Tulip and AFAD are similar, although there are different reasons for their inaccuracies.

Figures 15 and 16 show the inference accuracy of these schemes on very lossy paths. We consider the paths with a round-trip loss rate larger than 10%, of which there are 322 paths. These figures clearly show that the major estimation errors of AFAD are caused by measurement errors. In some cases, the absolute errors of AFAD are not large while the relative errors are very large because the forward path has nearly no loss. I-Tulip is shown to be slightly better than the original Tulip, as the transmission correlation is still the major source of inaccuracy for Tulip. SPA has an accuracy similar to that of Tulip on these paths, and OFAD is still the most accurate scheme.

Results of Combined Schemes: In Section VI-B.2, we show that the combination of source-only diagnosis schemes can significantly improve their diagnosis granularity. In this section, we investigate whether a similar combination (including both source-only and source+destination schemes) can improve their accuracy as well.

The four individual schemes (SPA, AFAD, OFAD and Tulip) have different bias related to the transmission correlation. Since the improved Tulip has better accuracy than the original Tulip, we will use the improved scheme as the representative for the rest of the paper and just call it Tulip. SPA requires strong correlation to be accurate, while OFAD, AFAD and Tulip prefer the loss to be independent. Therefore, the question is how to combine these schemes to improve accuracy.

Since SPA, OFAD, and Tulip tend to underestimate the loss rates, we design a simple hybrid approach as follows. When both individual schemes (*e.g.*, SPA and OFAD) give loss inference for certain segments, we choose the larger one as the real loss rate estimation. However, when each scheme give different diagnosable path segments, it can be a complicated problem to combine the results from each scheme because they may have different characteristics of inference errors. Due to lack of the ground truth on link-level loss rates, in this paper, we evaluate their path-level performance for the combined scheme. It is part of our future work to study the link-level performance.

Figures 17 and 18 show the estimation errors and relative errors of combined schemes: OFAD+SPA, OFAD+Tulip, and Tulip+SPA. Clearly, these hybrid schemes outperform the single schemes, especially for the methods involving SPA. In other words, SPA is a good complement to OFAD or Tulip because when there exists long loss episode, the probe packets will become correlated in loss, which will affect the accuracy of OFAD and Tulip as shown in our statistical analysis in Section III. The OFAD+Tulip scheme has similar performance to that of OFAD because their loss rate estimation approaches are similar, but OFAD is more accurate because it can afford to choose a large interval between fragmented packets to reduce the correlation. Similarly, we do not show the performance of the SPA+OFAD+Tulip combination because it is similar to that of SPA+OFAD.

3) Link-level Accuracy Evaluation: In this section, we evaluate the link-level accuracy. Since we cannot get the ground truth of link-level loss rates, we use two indirect methods for evaluation. First, we take a similar consistency check method as in [1]. That is, if an approach is measuring the loss rate correctly, the inferred loss rates should not decrease as we move further along the path. We use this approach to evaluate each individual scheme. Second, since both SPA and Tulip tend to underestimate the loss rates of the links (which is shown both in theory and in the experiments in Section VI-C.2), we simply consider the one that infers larger loss rates to be more accurate. We apply this to compare SPA, Tulip and their combination.

Internal Consistency Check of FAD: Since PlanetLab hosts disable fragments generated by slice users, we cannot use PlanetLab hosts to do the experiments. Therefore, we use three common Linux hosts (one in an US university, one in CERNET of China and one in the Comcast residential network) as the sources, randomly generate 1500 IP addresses as the destination and use the 4500 paths for evaluation. 275 paths among the measured path have a loss rate larger than 0.5%, and we only consider these lossy paths in the consistency check.

Figure 19 shows the CDF for all the forward loss rate deltas of OFAD. It shows that the forward loss rate inference is consistent. Over 80% of loss rate deltas are non-negative,

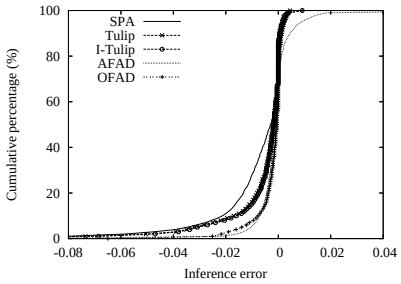

Fig. 13. Estimation error CDFs of five individual schemes for 5136 paths (with loss rates $> 0.5\%$).

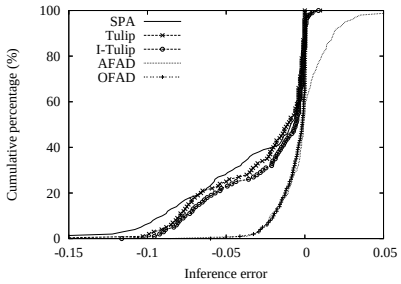

Fig. 15. Estimation error CDFs of combined schemes for 322 lossy paths (with loss rates $>$ 10%).

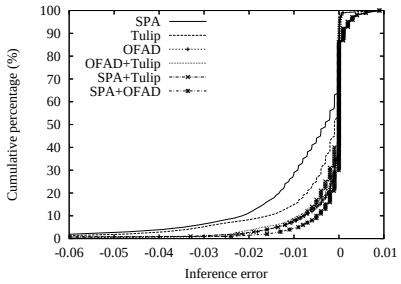

Fig. 17. Estimation error CDFs of the hybrid schemes for 5136 paths (with loss rates $>$ 0.5%).

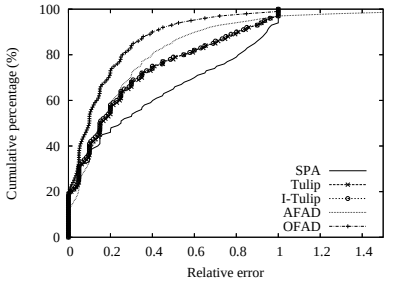

Fig. 14. Relative error CDFs of five individual schemes for 5136 paths (with loss rates $>$ 0.5%).

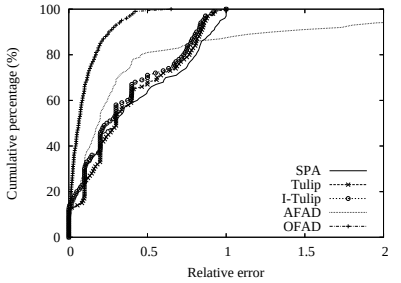

Fig. 16. Relative error CDFs of combined schemes for 322 lossy paths (with loss rates $>$ 10%).

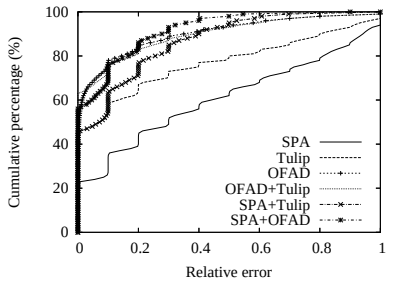

Fig. 18. Relative error CDFs of the hybrid schemes for 5136 paths (with loss rates $>$ 0.5%).

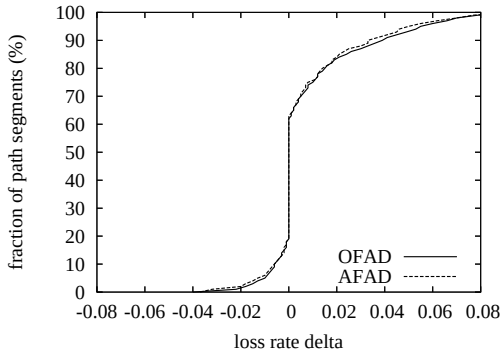

Fig. 19. Consistency check for AFAD and OFAD.

and most negative deltas are likely to be caused by the statistical nature of the measurement. Similar to [1], we use the Chi-squared test [20] to check whether the negative loss delta is statistically significant or not. The null hypothesis is that the two path segments have the same loss property and the loss difference can be considered the result of a statistical variation. With a 95% confidence interval, 92.9% of negative loss deltas are not statistically significant; 97.8% of them are not statistically significant with a 99% confidence interval. We manually checked these statistically significant negative loss rate deltas, and found that the measured path segments usually have a much larger reverse loss rate than forward loss rate. For example, if the forward loss rate is 2% while the reverse loss rate is 15%, the variance may be dominated by the reverse loss rate, which is very large compared to the forward loss rate.

Figure 19 shows the CDF for all the forward loss rate deltas of AFAD. It shows that the AFAD measurement is also consistent: about 80% of loss deltas are non-negative and the Chi-squared test shows that 91.2% of negative loss deltas are not statistically significant with a 95% confidence interval. Manually checking shows that large loss rate on the reverse path is an important cause of inconsistency. Furthermore, we observe a small number of cases in which the forward loss rates inferred by AFAD are much larger than those of the round-trip loss rates. Further probes show that some routers have a rate-limit on processing fragmented packets but not on common packets.

Internal Consistency Check of SPA: SPA does not rely on IP fragmentation, so we can use PlanetLab host to easily measure a large number of paths for a consistency check. These paths are more representive because of the divergence of the sources and destinations. In this consistency check, we employed 115 PlanetLab hosts randomly and used SPA to measure the paths between them. Finally measurement data of 10,493 paths were collected. Among them, 2,325 (about 22%) paths have loss rate larger than 0.5%.

Figure 20 shows the loss rate deltas measured by SPA. Again, the figure shows that the loss rate measured by SPA is internally consistent. Over 75% of loss rate deltas are non-negative and most negative loss rate deltas are not statistically significant. The Chi-squared test shows that with a 95% confidence interval, 94.3% of negative loss rates are not statistically significant; the number increases to 98.1% with a 99% confidence interval.

Internal Consistency Check of the Combination of SPA and Tulip: While we were checking the consistency of SPA (See the previous section), we also ran Tulip (as well as I-Tulip) at the same time. Therefore we can check the consistency of SPA, Tulip and I-Tulip respectively and also

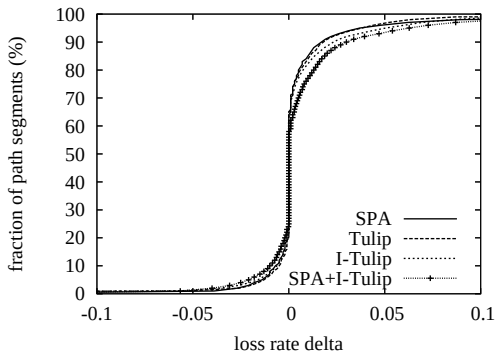

Fig. 20. Consistency check for SPA, Tulip and their combination.

the consistency of combined schemes, such as SPA+I-Tulip.

Figure 20 shows the CDF of the loss rate delta of the four schemes. First, I-Tulip has slightly larger absolute values of loss rate delta than Tulip, which is not surprising. From the figure, we cannot see an obvious difference in the degree of consistency between Tulip and I-Tulip. Also, the lines for SPA and Tulip are overlapping, which suggests that they are similar in terms of consistency. As for the combination of SPA and I-Tulip (See the line "SPA+I-Tulip" in Figure 20), we see that larger positive loss deltas are observed, as we expected. Since both SPA and I-Tulip tends to underestimate the loss rate, larger loss rate deltas suggest that the inferred loss by the combination scheme is closer to the true value. Meanwhile, the combination scheme also has a slightly smaller negative loss rate deltas. This is reasonable, because we may get the worst case of I-Tulip and SPA for a single path.

Accuracy Comparison of SPA and Tulip: In this section, we compare the link-level accuracy of SPA, Tulip and their combination through comparing their inferred loss rate values as introduced earlier. Note that one important reason that we do not compare AFAD/OFAD with Tulip or SPA is that both SPA and Tulip are easy to deploy on PlanetLab hosts, which prohibit users from sending out fragmented packets. Leveraging on the large scale PlanetLab testbed, we can easily employ hundreds of hosts simultaneously and let them measure the paths between them.

In this experiment, we use 115 PlanetLab hosts to measure the 13,110 paths between them. For each path, Tulip and SPA are run at the same period to diagnosis the path. Since SPA has finer diagnosis granularity and the diagnosable routers of Tulip is a subset of that of SPA, we downgrade the diagnosis granularity of SPA to that of Tulip in this experiment. Finally we collected measurement results for 10,585 paths out of 13,110 paths. Some paths are missing, simply because some PlanetLab hosts were down when we collected the data. For each path, we control the packet length of the probe packets and infer the loss rate of packet size of 40 bytes and 1500 bytes. We only consider the paths with loss rate larger than 1%. After calibration, SPA and Tulip infers the loss rates of some path segments in each path. We check the difference of the loss rate inferred by SPA and Tulip (defined as $l_{SPA} - l_{Tulip}$) as well as the relative difference of loss rate inference, defined as $(l_{SPA} - l_{Tulip})/\max(l_{SPA}, l_{Tulip})$.

From Figures 21 and 22, we see that for long packets, Tulip and SPA have similar accuracy overall. However, it is clear in Figure V-A that for a single path, SPA and Tulip may have a large divergence on the loss rate inference. This confirms our theoretical analysis, as Tulip and SPA have contradictory requirement on the loss correlation. As for short packets, SPA seems to be more accurate than Tulip, and the (relative) differences are more likely to be positive. This is mainly because the loss correlation between both short packets is larger than the loss correlation between a long and a short packet, especially when long packets usually have larger loss rates. Still, the combination of SPA and Tulip should underestimate the loss less and hence be more accurate.

We also studied the loss rate inferred by the combination of SPA and Tulip (called SPA+Tulip), simply choosing the larger loss rate inferred by Tulip and SPA. As we argued, since both of SPA and Tulip tends to underestimate loss rates, the combination scheme are supposed to be more accurate. Figure 23 shows the relative difference of the loss rate between SPA+Tulip and Tulip, and between SPA+Tulip and SPA with 1500-byte probes. Obviously, using either SPA or Tulip will miss a large portion of lossy links, and SPA+Tulip can improve each single scheme remarkably because of the divergent loss rate inference results of SPA and Tulip. For example, in about 30% paths, the relative loss difference between SPA+Tulip and Tulip is larger than 75%, which means Tulip estimates less than one-fourth loss rates than SPA+Tulip.

D. Lossy Link Distribution

Since SPA has close to one diagnosis granularity and it is easy to be deployed on PlanetLab hosts, we conducted a large scale experiment on PlanetLab to measure the location of the lossy links. We randomly selected 117 PlanetLab hosts and about 60% of them are within US. Since there are many long international paths, SPA used 50-byte probes instead of 1500-byte probes so as to probe all the diagnosable internal routers simultaneously. The probe rate is 5 probes per second and 1000 probes are sent for each path segment. Finally, data of 11,721 out of all the 13,572 paths were collected. Among these paths, there are 1,047 lossy paths with loss rate larger than 1%. After calibration, we got 110 unique lossy links as well as the IP addresses of both ends of the lossy links.

We derive the IP-to-AS mapping from the BGP tables published in Route Views [21]. Then we mapped the IP addresses of the ends of the lossy links to their AS numbers. By checking the AS numbers of the both ends of the lossy links, we found that about 25.5% lossy links have different AS numbers on their ends, and these links are likely to be inter-AS links. The remaining 74.5% lossy links are probably intra-AS links, as each of they connects two routers with the same AS number. The result is a bit different from previous research which shows half bottleneck links are

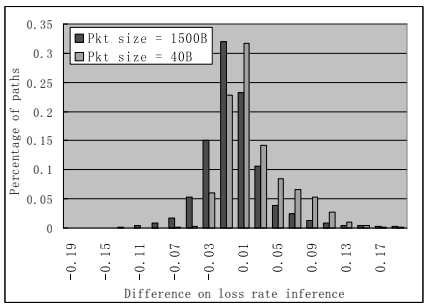

Fig. 21. Loss rate inferred by SPA minus loss rate inferred by Tulip.

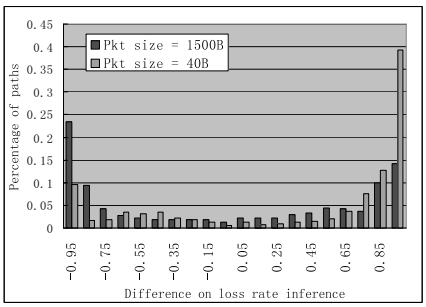

Fig. 22. Relative difference of loss rate inferred by SPA and Tulip.

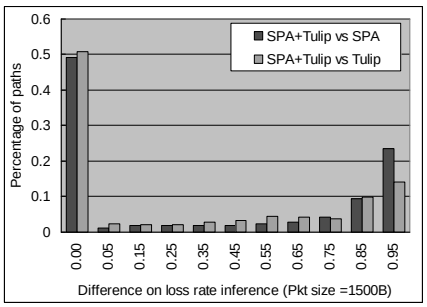

Fig. 23. Relative difference of loss rate inferred by SPA+Tulip, SPA and Tulip.

probably inter-AS links [22] and bottleneck link location is correlated with lossy link location [18]. Further investigation revealed that we measured much more international paths than [22] and the networks outside of US are more lossy. For example, we found 20 loss links in CERNET of China (AS 4538) alone. Thus we suspect that networks in different areas may have quite different characteristics due to different deployment status.

E. Summary and Recommendations to Users

In summary, based on the evaluation results above, we make the following recommendation for users. If the user can only deploy measurement tools at the source, OFAD+Tulip is recommended. If the user has control for both source and destination, we recommend SPA+OFAD which has both the finest granularity and best accuracy.

VII. Conclusions

In this paper, we propose a suite of user-level on-demand link-level loss rate inference schemes, and compare their accuracy and diagnosis granularity with existing tools. Internet experiments show that our approaches dramatically outperform existing work especially in terms of diagnosis granularity. Furthermore, the suite of schemes provide a smooth tradeoff among deployment requirement, diagnosis accuracy and granularity.

References

[1] R. Mahajan, N. Spring, D. Wetherall, and T. Anderson, "User-level internet path diagnosis," in *ACM SOSP*, 2003.

[2] M. Coates, A. Hero, R. Nowak, and B. Yu, "Internet Tomography," *IEEE Signal Processing Magazine*, vol. 19, no. 3, pp. 47–65, 2002.

[3] R. Caceres, N. Duffield, J. Horowitz, and D. Towsley, "Multicast-based inference of network-internal loss characteristics," *IEEE Transactions in Information Theory*, vol. 45, 1999.

[4] N.G. Duffield, F.L. Presti, V. Paxson, and D. Towsley, "Inferring link loss using striped unicast probes," in *IEEE INFOCOM*, 2001.

[5] V. Padmanabhan, L. Qiu, and H. Wang, "Server-based inference of Internet link lossiness," in *IEEE INFOCOM*, 2003.

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

[7] M. Mathis A. Adams, J. Mahdavi and V. Paxson, "Creating a scalable architecture for internet measurement," in *IEEE Network*, 1998.

[8] N. Duffield J. Sommers, P. Barford and A. Ron, "Improving accuracy in end-to-end packet loss measurement," in *ACM SIGCOMM*, 2005.

[9] Y. Zhao, Y. Chen, and D. Bindel, "Towards unbiased end-to-end network diagnosis," in *ACM SIGCOMM*, 2006.

[10] V. Paxon, "End-to-end Internet packet dynamics," in *ACM SIGCOMM*, 1997.

[11] Y. Zhang et al., "On the constancy of Internet path properties," in *Proc. of SIGCOMM IMW*, 2001.

[12] DARPA INTERNET PROGRAM, "Internet protocol," RFC 791, 1981.

[13] J. Bolot, "Characterizing end-to-end packet delay and loss in the Internet," in *ACM SIGCOMM*, 1993.

[14] J. Lyman, "China Starts Up World's Biggest Next-Gen Internet Network," `http://www.technewsworld.com/story/39233.html`.

[15] J. Anderson, "An analysis of fragmentation attacks," `http://www.ouah.org/fragma.html`.

[16] D. Moore C. Shannon and k claffy, "Beyond folklore: Observations on fragmented traffic," *IEEE/ACM Transactions on Networking*, vol. 10, no. 16, 2002.

[17] P. Barford and J. Sommers, "Comparing probe- and router-based methods for measuring packet loss," in *IEEE Internet Computing - Special issue on Measuring the Internet*, Sept/Oct, 2004.

[18] N. Hu and et. al., "Locating internet bottlenecks: Algorithms, measurements, and implications," in *ACM SIGCOMM*, 2004.

[19] R. Govindan and H. Tangmunarunkit, "Heuristics for internet map discovery," in *IEEE INFOCOM*, 2000.

[20] J. McClave and F. Dietrich, *Statistics*, Macmillan Publishing Company, 6th edition, 1994.

[21] University of Oregon Route Views Archive Project, "`http://www.routeviews.org/`," .

[22] S. Seshan A. Akella and A. Shaikh, "An empirical evaluaton of wide-area Internet bottlenecks," in *ACM SIGCOMM Internet Measurement Conference (IMC)*, 2003.