Scalable Deterministic End-to-End Probing and Analytical Method for Overlay Network Monitoring

Yanjie Ren, Yan Qiao, Xue-song Qiu, Shun-an Wu
State Key laboratory of Networking and Switching Technology
Beijing University of Posts and Telecommunications
Beijing, P.R.China, 100876
Email: {yjren, qiaoyan, sawu}@metarnet.com xsqiu@bupt.edu.cn

***Abstract*—Overlay network monitoring is one of the most important issues in the design and operation of overlay networks. It faced with increasing demand for better throughput and response time performance. Given an overlay network with n end hosts, existing systems either require $O(n^2)$ measurements, and thus lack scalability, or can only estimate the latency but not failures or congestion. We propose a scalable deterministic network monitoring scheme that selectively monitors k linearly independent paths which can fully describe all the $O(n^2)$ paths. We use the loss rates and latency of the collected k path to deduce the loss rate and latency of the rest paths. Our method only assumes knowledge of the underlying IP topology, with links dynamically varying between normal and lossy.**

In this paper, we propose a novel Path Loss Inference Algorithm (in short, PLIA) which improves implements and extensively estimates such a monitoring method. We carry out our experiments in two different scenarios and the results from the experiments demonstrate that our approach increases accuracy in monitoring congested paths compared to other representative schemas'. Especially in a router-level topology, our approach can deliver almost a 12.5% increase.

Keywords-Overlay; Network measurement and monitoring; Numerical linear algebra

I. INTRODUCTION

The rigidity of the Internet architecture and its access into many aspects of daily life has led to a great dependency on its services. Multimedia and peer-to-peer file sharing distribution applications require Quality of Service (QoS) guarantees in terms of delay, loss, and jitter to maintain a certain level of performance. The most practical way to realize such applications is through the using of overlay and peer-to-peer systems. These systems flexibly choose their targets and communication paths, and thus can benefit from end-to-end network distance estimation (e.g., loss rate and latency).

Accurate loss rate monitoring system can detect periods of degraded performance and path outages within seconds. It can both help build adaptive overlay applications and facilitate distributed system management.

It is desirable to build a scalable deterministic overlay loss rate monitoring system which could detect anomaly, accurately and timely. However, existing network monitoring systems are insufficient for this goal. Most of them can be classified into two categories based on the targeted metrics: general metrics [1] and latency only [2, 3, 4]. As a new schema, network tomography has been well studied ([5] provides a good survey) by researchers. Most tomography systems assume limited measurements are available (often in a multicast tree-like structure), and then try to estimate link characteristics [6, 7] or shared congestion [8] in the middle of the network. However, the issue is under-constrained: there exist unidentifiable links [9] for their loss rates cannot be uniquely computed.

Chen et al. [10] dispatch probes from one host to the other hosts on the overlay network and collect the routing matrix. Then they select the paths which are linear independent with a variant of the QR decomposition to monitor. Using the loss rates of selected paths as input, they calculate the outcomes of equations related with links' loss rate and paths' loss rate as loss rates of all other paths.

Zhao et al.[9] note that the above inference results are not deterministic or unique. They propose virtual links, which are linearly independent as well as the columns in routing matrix.

In this paper, we described the idea of a tomography-based overlay monitoring system in which we selectively monitor a basis set of k paths which are linear independent. Hence, by monitoring loss rates of the paths in the basis set, we infer loss rates for all end-to-end paths. It can also be extended to other additive metrics, such as latency. Moreover, with our method, the end-to-end path loss rates can be computed even when they contain unidentifiable links (i.e. the properties that cannot be uniquely determined), which enables our method to outperform the former ones.

Our simulation experiments results demonstrate that our method can achieve high accuracy on the computing the loss rates of all paths. The average absolute error of loss rate estimation is only 0.2%; and the average error factor is 100%.

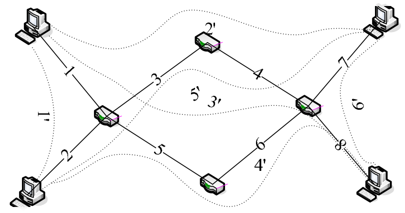
II. PROBLEM FORMULATION

A. Network Model

We model the overlay network as a undirected graph $G=(V,E)$ where the V is the set of nodes (including routers, hosts, etc) and the $E=\{e_1...e_m\}$ stands for the set of links connecting nodes in V. On the overlay network, the number of

This work was partly supported by t 973 Project of China (2007310703) , National S&T Major Project (2011ZX03003-002-01) and NSFC (60821001, 60802035, 60973108).

end hosts and edges is denoted by n_h and n_e . Let $P=\{p_1,...,p_n\}$ be the set of paths which are detected, thus we know $n_p=n_h(n_h-1)/2$ .The overlay network in Fig. 1 which has 4 end hosts and 8 links, has 6 end-to-end paths where $n_p=6$.



$$M=\begin{bmatrix}1&1&0&0&0&0&0&0\\1&0&1&1&0&0&1&0\\1&0&0&0&1&1&0&1\\0&1&1&1&0&0&1&0\\0&1&0&0&1&1&0&1\\0&0&0&0&0&0&1&1\end{bmatrix}$$

Figure 1. An overlay network sample.

Given an overlay network $G=(V,E)$ and a set of paths P, we compute the routing matrix M of dimension $n_p \times n_e$ as follows. We defined a path by a column vector $v \in \{0,1\}^{n_e}$, where the j th entry $v_j=1$ if link j is part of path, and $v_j=0$ otherwise. A row of M therefore corresponds to a path, and a column corresponds to a link. If one column contains only zeros, meaning none path monitored contain this link and we can't measure it. Hence, we drop these columns from the matrix M to obtain a matrix of dimensions $n_p \times n_s$ where $n_s (\leq n_e)$ is the real number of passed links at least contained by one path in P . We use E_s to denote the set of covered links, $n_s = |E_s|$.

B. Algebraic Model

Suppose in the overlay network topology above, let φ denote the loss rate of a path represented by v , then the loss rate l_j of link e_j contained in v is given by

$$1-\varphi = \prod_{j=1}^{n_s} (1-l_j)^{v_j} \qquad (1)$$

Equation (1) assumes that pack loss rates are independent among links. Chen et al. [10] show that the link loss dependence has little effect on the accuracy of (1) .

We take logarithms on both sides of (1) . Then we define a column vector x with elements $x_j=\log(1-l_j)$, and write v^T for the transpose of the column vector v , we can rewrite (1) as:

$$\log(1-\varphi) = \sum_{j=1}^{n_s} \log(1-l_j) v_j = \sum_{j=1}^{n_s} x_j v_j = xv^T \qquad (2)$$

There are $n_r = O(n^2)$ paths in the overlay network, By putting them together, we form a rectangular matrix $M \in \{0,1\}^{n_p \times n_s}$. Every row of M represents a path in the overlay network: $M_{ij}=1$ when path i contains link j , and $M_{ij}=0$ otherwise（e.g. Fig 1 is that kind of matrix）. Let φ_i be the end-to-end loss rate of the i th path, and let b be a column vector with $b_i=\log(1-\varphi_i)$. Then we rewrite the n_p equations in form (2) as

$$Mx=b \qquad (3)$$

In matrix M, some rows can be represented as the linear combinations of several other rows. That is to say, some of paths do not need to be measured. Therefore, we continue to select k linear independent paths from M . Then we obtain a reduced system from (3):

$$\bar{M}x=\bar{b} \qquad (4)$$

Where $\bar{M} \in \Re^{k \times n_s}$ and $\bar{b} \in \Re^k$, consisting of k rows of M and b , respectively.

Our goal is: once given the measurements of probes in $\bar{b}$, we could infer the rest probes in b .

To compute the loss rates of all paths, we must find a solution to the underdetermined linear system $\bar{M}x=\bar{b}$. The vector $\bar{b}$ comes from measurements of the linear independent paths. Mostly, the number of probes in $\bar{M}$ is smaller than the number of links (i.e. $k \leq n_s$). Thus, $\bar{M}$ is rank deficient.

By solving equations (4), we can obtain the general solution of link loss rate variables x , where x contains innumerable solutions to (4). Inserting x into (3), we could also obtain the general solution to the global link loss rate variable x , which contains innumerable solutions to b as well..

However, we would like to obtain the particular solution of b which is most close to the ground true. In the next section, we will propose an algebraic approach to obtain the particular solution of b .

III. ACCURATE PAHT LOSS INFERENCE ALGORITHM

Links in network tomography can be grouped by identifiable and unidentifiable. If M is rank deficient, we will be able to uniquely determine the loss rates of link sequences from(3). These link sequences are called identifiable link sequences. When the loss rate of some links can not be determined, we called unidentifiable link sequences. Only use the loss rates of identifiable links to infer the path loss rate is much more accurate than using the loss rates of both identifiable and unidentifiable links.

We form matrix $\bar{M}'$ by extracting the minimal set of identifiable link sequences corresponding columns in $\bar{M}$, we obtain the equations (5). Obviously, Equation (5) has a unique solution $\bar{x}$.

$$\bar{M}\bar{x}=\bar{b} \qquad (5)$$

Where, $\bar{x}$ is the loss rate of the identifiable link sequences.

We also form matrix M' by extracting the minimal set of identifiable link sequences corresponding columns in M . Once we obtain $\bar{x}$ by solving (5), we can obtain the unique solution of $\hat{b}$ as well, through solving equation (6).

$$M\bar{x}=\hat{b} \qquad (6)$$

Next, we will provide how we get the minimal set of identifiable link sequences. "minimal" means there cannot exit subset of the minimal set. If and only if $\|v\|=\|Q^T v\|$, v will lie in the path space, where v represents identifiable link sequence and Q is an orthonormal basis of $\Re(M^T)$ as mentioned in [14].

Here, we apply an algebraic approach to separate the identifiable and unidentifiable links. We define *virtual links*（also called identifiable links）as link sequences whose properties can be uniquely identified from end-to-end measurements. All identifiable virtual links belong to the column space of M, i.e., a subspace containing all the vectors can be written as linear combination of the column vectors of M. For example, link 1 is a virtual link, of which the loss rate can be inferred through the linear combination of path success rates as $(b_1+b_3-b_2)$. We design an efficient algorithm (Minimal Virtual Links Set Selection Algorithm, see in section V) to identify the virtual links and monitor their loss rates.

IV. PATH LOSS INFERENCE ALGORITHM

In this section, we present our Path Loss Inference Algorithm, or PLIA. The PLIA includes three steps. First, we select a basis set of k paths to monitor. And then, we select the basis of links to compute the loss rates of all the other paths. Last, we calculate and update the loss rates of all other paths based on continuous monitoring of the selected paths.

A. Monitoring Paths Selection

To select k linear independent paths from M to monitor, we use standard rank-revealing decomposition techniques [12]. Our algorithm is a variant of QR decomposition as bellow.

```
Procedure PathSelection( $M$ )
for every row(path) $v$ in $M$ do
    $\hat{R}_{12} = R^{-T}\bar{M}v^T$
    $\hat{R}_{22} = \|v\|^2 - \|\hat{R}_{12}\|^2$
    if $\hat{R}_{22} \neq 0$ then
        Select $v$ as a measurement path
        Update $R = \begin{bmatrix} R & \hat{R}_{12} \\ 0 & \hat{R}_{22} \end{bmatrix}$ and $\bar{M} = \begin{bmatrix} \bar{M} \\ v \end{bmatrix}$
    end
end
```

Figure 2. PathSelection algorithm

B Minimal Virtual Links Set Selection

Our selection of Minimal Virtual Links Set (MVLS in short) is on a given path in creasing order of size. We can track whether every link is the starting link in some already-discovered MVLS. To test whether a link sequence is identifiable, we just need to find out that the corresponding path vector v lies in the path space. Since Q which is pre-computed above is an orthonormal basis for the path space, v will lie in the path space if and only if $\|v\| = \|Q^T v\|$. If there are i links contained in the link sequence, and then v will contain only i nonzeros.

```
Procedure MVLS_Selection( $\bar{M}$ )
for every row(path) $r$ in $\bar{M}$ do
    boolean start_mvls [ length (r) ]
    Clear start_mvsl to all false
    for i=1 to length(r) do
        for every segment $S = r_k \ldots r_l$ of length i do
            if start_mvls[k] then
                continue
            else
                Let $v$ be the corresponding vector of $S$
                if $\|v\| = \|Q^T v\|$ then
                    start_mvls[k] = true
                    $S$ is an identifiable link sequence
                else
                    $S$ is not an identifiable link sequence
                end
            end
        end
    end
end
```

Figure 3. MVLS_Selection algorithm

C Path Loss Rate Calculations

In this paper, we provide an algebraic scheme as shown in Fig. 4 to infer the loss rates of the global paths with a small sub set of probes. Given measured loss rates for $\bar{b}$, we compute the solution $\bar{x}$ by $\bar{x} = R\backslash(Q\backslash\bar{b})$, If we get $\bar{x}$, we can solve $\hat{b} = M\bar{x}$, and from $\hat{b}$, we can infer the loss rates of the remain paths.

```
Procedure Unique_LossRates
Input: routing matrix $M$.
Output: all the path loss rate $b$
begin
    $\bar{M}$ = PathSelection( $M$ )
    $\bar{M}'$ =MVLS_Selection( $\bar{M}$ )
    $M'$ =MVLS_Selection( $M$ )
    [Q,R]=QR decomposition( $\bar{M}'$ )
    $\bar{x} = R\backslash(Q\backslash\bar{b})$
    $\hat{b} = M\bar{x}$
    return $\hat{b}$
end
```

Figure 4. Unique loss rates algorithm

V. EXPERIMENTS

In this section, we implement our new algorithm PLIA and a representative algorithm proposed in [10], then we compare the accuracies of the two algorithms in BA scenarios.

A. Simulator Experiments Settings

We consider the following dimensions for simulation:

- Topology type: in our experiments, network topologies are generated from BRITE. Since all hierarchical models have the similar results, we use Barabasi-Albert at the AS level.

- Topology size: the numbered of nodes is 500 and 1000 in our experiments respectively. The nodes include both internal nodes (i.e., routers) and end hosts.
- Link loss rate distribution: 90% of the links are classified as "good" and the rest as "bad". We use model ($LLRD1$)as in [12], the loss rate for good links is chosen uniformly at random in the 0-1% range and that for bad links is chosen in the 5-10% range.
- Loss model: After assigning each link a loss rate, we use a Gilbert model to simulate the loss processes at each link. According to Paxon's observed measurement of Internet [13], the probability of maintaining bad state is set to be 35% as in [15].

We then measure the loss rate between each pair of host points. Each host sends 1000 UDP packets of size 40 bytes to each other host. Time between probes follows an exponential distribution with a mean value of 0.2 seconds.

B. Results and analyze

To evaluate accuracy, we use both absolute error and error factor as proposed in [14]. The absolute error is $|p-\hat{p}|$ and the error factor $F_{\varepsilon}(p,\hat{p})$ where $\hat{p}$ is inferred loss rate and p is real loss rate is defined as follows:

$$F_{\varepsilon}(p,\hat{p})=\max\left\{\frac{p(\varepsilon)}{\hat{p}(\varepsilon)},\frac{\hat{p}(\varepsilon)}{p(\varepsilon)}\right\} \tag{6}$$

Where $p(\varepsilon)=\max(\varepsilon,p)$ and $\hat{p}=\max(\varepsilon,\hat{p})$. Thus p and $\hat{p}$ are treated as no less than ε , and then the error factor is the maximum ratio, upwards or downwards, by which they differ. In our experiments, we use the default value $\varepsilon=0.001$. If the estimation is perfect, the error factor is one.

We plot the cumulative distribution functions (CDFs) of absolute errors and error factors with the Gilbert model in Fig. 5 and Fig 6. We can find that our outcomes have the higher accuracy compared to existing path monitoring schemes.

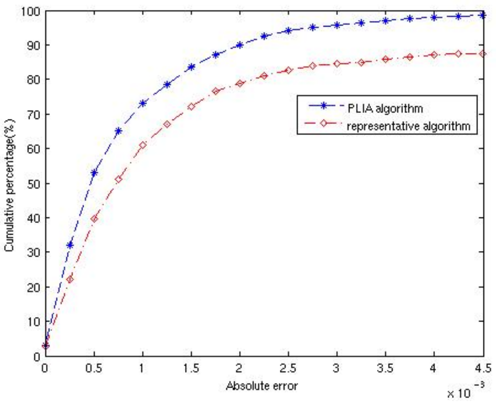


Figure 5. Cumulative distribution of absolute errors under Gilbert loss model for Barabasi-Albert at the AS level topologies.

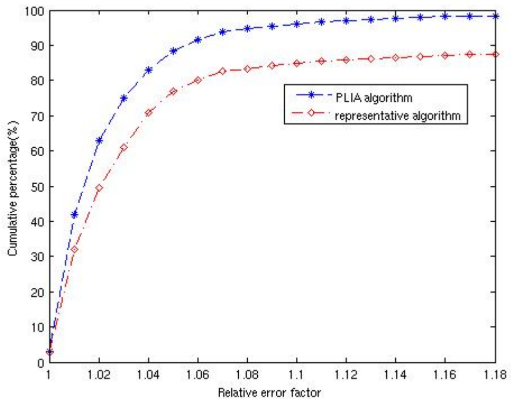


Figure 6. Cumulative distribution of error factors under Gibert loss model for Barabasi-Albert at the AS level topologies.

Fig 5 and Fig 6 illustrate the results of our method comparing with the previous typical method [10]. We apply the two method under Gibert loss model in BA at the AS level topology. The top line represents the result of our method, and the bottom line represents the result of previous typical method. Since our algorithm is unique but their algorithm is not, our path loss rates are more accurate than their loss rates.

The loss path inference results are shown in Table I And in Barabasi-Albert topology, the most accuracy increase percentage compared to previous schema is 12%.

TABLE I. SIMULATION RESULTS FOR BA TOPOLOGY

nodes	end hosts		paths	links	increase percentage(%)
	total	OL(n)			
500	344	20	190	499	1.78
		50	1225		4.28
1000	665	50	1225	999	4.23
		100	4950		12.43

The above table shows the simulation results for BA topology. OL is the number of end hosts on overlay. The "increase percentage" is the increased accurate of our method than the previous typical method [10]. We can also see that in larger topology, the increase percentage is higher. So, our method is more suitable for large scale overlay network.

VI. CONCLUSIONS

In this paper, we improve, implement and extensively estimate a method for monitoring overlay network path QoS anomalies. For an overlay network with n end hosts, we selectively monitor a basis set of paths and links which can fully describe all the paths. Then the measurements of the basis set are used to infer the loss rates of all the remaining paths. Our approach works in accuracy, real time and handles topology measurement. It is shown by experiments that our approach can increase almost a 12.5% accuracy in monitoring congested paths compared to the existing typical path monitoring schemes.

REFERENCES

[1] D. G. Andersen et al., “Resilient overlay networks,” in Proc. of ACM SOSP, 2001.

[2] T. S. E. Ng and H. Zhang, “Predicting Internet network distance with coordinates-based approaches,” in Proc.of IEEE INFOCOM, 2002.

[3] S. Ratnasamy et al., “Topologically-aware overlay construction and server selection,” in Proc. of IEEE INFOCOM, 2002.

[4] Hitesh Ballani and Paul Francis, \Fault Management Using the CONMan Abstraction," in Proc. of IEEE INFOCOM, April 2009.

[5] Mark Coates, Alfred Hero, Robert Nowak, and Bin Yu, “Internet Tomography,” IEEE Signal Processing Magazine, vol. 19, no. 3, pp. 47–65, 2002.

[6] H.H.Zhao, M.Chen, “Topology inference based on network tomography,” Ruan Jian Xue Bao (Journal of Software). Vol. 21, no. 1, pp. 133-146. Jan. 2010.

[7] V. Padmanabhan, L. Qiu, and H. Wang, “Server-based inference of Internet link lossiness,” in IEEE INFOCOM, 2003.

[8] D. Rubenstein, J. F. Kurose, and D. F. Towsley, “Detecting shared congestion of flows via end-to-end measurement,” ACM Transactions on Networking, vol. 10, no. 3, 2002.

[9] Y. Zhao, Y. Chen, and D. Bindel, “Scalable deterministic overlay network diagnosis,” in Proceedings of ACM SIGCOMM, Pisa, Italy,2006.

[10] Y. Chen, D. Bindel, H. Song, and R. H. Katz, “An algebraic approach to practical and scalable overlay network monitoring,” in ACM SIGCOMM, 2004.

[11] H. X. Nguyen and P. Thiran, “Active measurement for failure diagnosis in IP networks,” in Proc. of PAM 2004, Juan-les-Pins, France, 2004.

[12] G.H. Golub and C.F. Van Loan, Matrix Computations, The Johns Hopkins University Press, 1989.

[13] V. Paxon, “End-to-end Internet packet dynamics,” in ACM SIGCOMM, 1997.

[14] Y. Zhao, Y. Chen, and D. Bindel, “Towards Unbiased End-to-End Network Diagnosis,” IEEE/ACM Transactions on Networking, v 17, n 6, p 1724-1737, December 2009.

[15] V. Padmanabhan, L. Qiu, and H. Wang, “Server-based inference of Internet link lossiness,” in IEEE INFOCOM,2003.