Packet Loss Measurement Based on Sampled Flow

Haoliang Lan 1,*, Jie Xu 1, Qun Wang 1 and Wei Ding 2

1 Department of Computer Information and Network Security, Jiangsu Police Institute, Nanjing 210031, China; xujie@jspi.cn (J.X.); wangqun@jspi.cn (Q.W.)
2 School of Cyber Science and Engineering, Southeast University, Nanjing 211189, China; wding@njnet.edu.cn
* Correspondence: lanhaoliang@jspi.cn

Abstract: This paper is devoted to further strengthening, in the current asymmetric information environment, the informed level of operators about network performance. Specifically, in view of the burst and perishability of a packet loss event, to better meet the real-time requirements of current high-speed backbone performance monitoring, a model for Packet Loss Measurement at the access network boundary Based on Sampled Flow (PLMBSF) is presented in this paper under the premise of both cost and real-time. The model overcomes problems such as the inability of previous estimation to distinguish between packet losses before and after the monitoring point, deployment difficulties and cooperative operation consistency. Drawing support from the Mathis equation and regression analysis, the measurement for packet losses before and after the monitoring point can be realized when using only the sampled flows generated by the access network boundary equipment. The comparison results with the trace-based passive packet loss measurement show that although the proposed model is easily affected by factors such as flow length, loss rate, sampling rate, the overall accuracy is still within the acceptable range. In addition, the proposed model PLMBSF, compared with the trace-based loss measurement is only different in the input data granularity. Therefore, PLMBSF and its advantages are also applicable to aggregated traffic.

Keywords: information asymmetry; network performance measurement; refined network performance management; sampled flow; packet loss rate

1. Introduction

The current network is information asymmetric, and network users know much less than the network operator, Internet Service Provider (ISP) and content provider about what happens inside the network [1]. Around this information asymmetry, from the perspective of users, it needs a light and user-friendly end-system measurement to rebalance the information asymmetry and empower users [2], while from the perspective of network performance management, operators should also constantly improve their own refined network performance management level to keep informed and in control regarding the network operation status [3]. In our previous work [4], we proposed an application-level packet loss rate measurement model IM-L-Rex from the perspective of users to reverse the information asymmetry disfavoring users. In this paper, we construct a sampled-TCP-flow oriented packet loss measurement model PLMBSF from the perspective of network management, with a focus on further improving the refined network performance management level.

Access network boundary performance measurement not only supports the measurement of end-to-end and aggregate traffic but also distinguishes the performance differences between ISP management domain and external Internet, which provides operators and managers with the possibility to grasp network performance by fine-grained analysis. Moreover, as a connection oriented and reliable transmission protocol, TCP’s congestion control mechanism has a natural response to changes in network performance. Therefore,
it has become a research hotspot in current academic circles to conduct access network boundary performance measurement with TCP flow.

The basis of research on network performance measurement and evaluation [5] is the data source adopted. The finer the granularity of the information provided by the data source, the more accurate the measurement results. Correspondingly, a series of judgments and strategies built on the measurements will be more effective, which, meanwhile, also means that the price paid by leveraging such information is greater. In turn, the coarse-grained data information is a challenge for the measurement technology. Therefore, in the selection of data sources, it is not only necessary to consider that the information contained therein can meet the corresponding measurement requirements but also to take into account the overhead of data collection, storage and processing so as to strike a balance between practicality and cost. Among the common data source types, the management information base and system log are the data types required for traditional port-based network performance management, which have poor scalability and do not meet the actual requirements of refined network performance management; while packet trace [6] and sampled flow [7] are the types of network flow oriented data sources based on network paths, which are suitable for current refined network performance management. The trace-based access network boundary loss measurement sacrifices time and space in exchange for higher accuracy and thus is widely used in network performance diagnosis and analysis. In recent years, scholars have carried out a series of targeted research around trace-based access network boundary measurement [8–10]. However, from the perspective of refined network performance management, besides performance evaluation with accurate measurements [11], the real-time nature of performance monitoring must also be considered. Real-time network performance measurement requires comprehensive consideration of cost, real-time and accuracy. Considering the high load and large data amount from trace data, the storage resources and processing costs required for the real-time measurement on the high-speed backbone network will be very high. Although it can be improved to some extent through sampling, packet trace has not become a de facto industrial standard. On the contrary, sampled flow [12] represented by NetFlow, IPFIX, etc., obtains general support from current network equipment and is easier to collect and store. For this reason, the packet loss measurement based on sampled flow has attracted attention. For instance, Wu et al. believed that the internal performance of the monitored network would not usually become the bottleneck of the application and thereupon constructed measurement models based on TCP flow to estimate packet losses [13] and loss episodes [14] outside the monitored network. Gu et al. [15] achieved estimation for packet losses by comparing sampled NetFlow collected from different routers. Similarly, Ricciato et al. [16] realized packet loss measurement between two monitoring points by comparing IPFIX. Giannakou et al. [17] proposed a machine learning approach to predict packet losses in science flows. Liu et al. [18] designed a packet loss estimation method based on bidirectional ICMP NetFlow. By analyzing these researches, it can be found that: (1) The researches on packet loss measurement based on sampled flow is relatively few in the current literature of boundary performance measurement; (2) the existing research methods either require ISP to widely deploy flow recording and measurement equipment in their own Point of Presence (PoP) and share measurement results among various points, or do not distinguish the performance inside and outside the monitored network, or assume that the performance in the monitored network will not become an application bottleneck, while only measuring and estimating the performance outside the monitored network. Firstly, the extensive deployment of flow measurement equipment not only requires the cost of deployment, operation and maintenance but also has the problem of cooperative operation consistency. Secondly, combined with existing and our own research work [19], the performance of the monitored network is not always negligible relative to the entire end-to-end performance, which mainly depends on whether the congestion and bottleneck appear inside or outside the network [20].
From the discussion above, it can be seen that the current research does not fully fit the actual scenario of an access network boundary, so it is not applicable, or fully meet the actual demand for access network boundary packet loss measurement. To this end, this paper adds to the body of measurement techniques by constructing model \textit{PLMBSF}. Drawing support from the Mathis equation [21] and regression analysis [22], the model is able to estimate packet losses before and after the monitoring point when only the sampled flows generated by the access network boundary equipment are available. Compared with other research in this area, this work has the following contributions:

- The access network boundary performance measurement can be achieved with only a single point of data, and there are no problems of deployment difficulties and cooperative operation consistency.
- Without making pre-assumptions on the performance of the intra-domain network, the performance status of intra-domain and inter-domain can be given at the same time, which is more in line with the actual demand of the current refined network performance management.

The rest of the paper is organized as follows: Section 2 outlines the construction of \textit{PLMBSF}. Section 3 details the solving for model \textit{PLMBSF}. Section 4 discusses the processing strategy for flow sampling. Section 5 validates and analyzes \textit{PLMBSF} based on the practical measurement data. Finally, Section 6 concludes the paper and suggests future work.

\section{2. PLMBSF}

\textit{PLMBSF} stems from the famous Mathis equation that is able to infer the bandwidth of a TCP connection under steady state and with random loss rate. Through regression analysis to mine the hidden information in the data stream and ACK stream, \textit{PLMBSF} realizes the estimation for packet losses before and after the monitoring point.

\subsection*{2.1. Model Description}

Mathis is one of the simplest renewal-theory-based TCP models, which describes the macroscopic behavior of the congestion avoidance algorithm. In addition to Mathis, there exist more complex and advanced models. For instance, PFTK [23] modeled timeout and window limits, and C-PFTK [24] extended PFTK to simulate the impact of a TCP delay and slow start strategy, while GRZ [25] further modeled the congestion avoidance scenarios after the retransmission timer fires. However, the reason why Mathis is still chosen in our research is that its parameters are relatively easy to obtain at the access network boundary. Mathis’s simplicity is built on a number of assumptions, and here list only the ones that are relevant to \textit{PLMBSF}:

- When Sack is enabled, multiple packet losses in one \textit{RTT} represent a single congestion event, and only one packet loss is recorded;
- The receiver window is large enough, and the sender always has data to send;
- The connection is long enough to reach a steady state;
- The \textit{RTT} is constant over the path.

Based on the assumptions above, Mathis is expressed as:

\begin{equation}
\text{goodput} = \frac{MSS \cdot C}{\text{RTT} \cdot \sqrt{\text{PLR}}} \quad (1)
\end{equation}

where:

- \text{goodput} represents the effective throughput, indicating the useful traffic [26] transferred from the source to the sink in a unit time;
- \text{MSS} represents maxim segment size, typically 1460 bytes;
- \text{C} is the smoothing parameter, with a value of 0.93;
- \text{RTT} is the round-trip time;
- \text{PLR} is the application-level packet loss rate.
Transform (1) to obtain PLR:

\[ \text{PLR} = \left( \frac{\text{MSS} \cdot \text{C}}{\text{RTT} \cdot \text{goodput}} \right)^2 \] (2)

2.2. Problem Setting and Challenges

The effective acquisition of several unknown parameters involved in PLMBSF is critical. Suppose that for a TCP connection, the packet loss rates before and after the monitoring point are \( \text{LB} \) and \( \text{LA} \), respectively. Meanwhile, Equation (2) only gives the end-to-end application-level packet loss rate. To further distinguish the packet loss rate before and after the monitoring point indicated by \( \text{LB} \) and \( \text{LA} \), the following problems need to be solved:

- Mathis is suitable for the sink application layer, how to realize the utilization for Mathis at the monitoring point, viz., how to determine goodput;
- After obtaining PLR, how to determine \( \text{LB} \) and \( \text{LA} \);
- Based on the sampled flow at the monitoring point, how to determine the RTT corresponding to a TCP connection;
- In practical application, how to deal with flow sampling to ensure the model accuracy.

In view of the problems above, this paper next establishes statistical models with regression analysis to dynamically determine the target parameters. On this basis, with the aid of heavy-tailed distribution [27] and Pareto distribution [28], the influence of sampling is further considered by increasing the number of flows and performing a weighted average for the estimation results.

3. Methodology

The proposed method leverages regression analysis and the ratio of the number of data packets to the number of ACK packets to determine the unknown parameters of PLMBSF and finally obtains the target parameters \( \text{LA} \) and \( \text{LB} \). Specifically, the steps to determine \( \text{LA} \) and \( \text{LB} \) are as follows:

Step 1: Establish a regression model to mine the relationship hidden between \( \text{AD} \) and goodput to determine goodput;
Step 2: Determine RTT by estimating the number of packets transmitted in each transmission round;
Step 3: Determine PLR by Equation (1);
Step 4: Establish a regression model between \( \text{AD} \) and \( \text{LA} \) to determine \( \text{LA} \);
Step 5: Determine \( \text{LB} \) by PLR and \( \text{LA} \).

Since the method of determining the target parameters that is described later is based on regression analysis, the regression analysis is first presented in this section. In addition, to determine the target parameters of PLMBSF, it needs the TCP transfers used for regression analysis to cover various characteristics related to packet losses. To this end, after presenting regression analysis, the platform generating the wanted TCP transfers are further described in this section. Finally, on this basis, the method for determining the target parameters is detailed in this section.

3.1. Regression Analysis

In statistics, regression analysis refers to a set of statistical processes used to estimate the relationship between variables. Regression analysis contains numerous techniques for modeling when inferring the relationship between dependent variables and independent variables. Correspondingly, the goal of regression analysis is to construct an optimum fitting model of the statistical data, i.e., to solve the weight \( \theta \) of each characteristic independent variable.
Assume that there are eigenvalues $x_1, x_2, \ldots, x_n$ ($n$ variables), and then the prediction model can be expressed as:

$$h_\theta x = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_n x_n$$  \hspace{1cm} (3)

Convert to matrix form:

$$h_\theta x = \sum_{i=0}^{n} \theta_i x_i = \theta^T X$$  \hspace{1cm} (4)

The error between the predicted value and the true value is represented by $\epsilon$, then:

$$y^{(i)} = \theta^T x^{(i)} + \epsilon^{(i)}$$  \hspace{1cm} (5)

Assume that residual error $\epsilon^{(i)}$ is independent and identically distributed, and it is generally believed that $\epsilon^{(i)}$ obeys a normal distribution with a mean value of 0 and a variance of $\sigma^2$. Thus:

$$P(\epsilon^{(i)}) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{\epsilon^{(i)}_2}{2\sigma^2}}$$  \hspace{1cm} (6)

Substitute $\epsilon^{(i)}$ into:

$$P(y^{(i)}|x^{(i)}; \theta) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y^{(i)}-\theta^T x^{(i)})^2}{2\sigma^2}}$$  \hspace{1cm} (7)

Since $\epsilon$ is independent identically distributed, the output $y$ of each sample is also independent and identically distributed. Accordingly, the maximum likelihood estimation is used to approximate $\theta$ so that the value of Equation (7) is most likely to be $y^{(i)}$ under the condition $x^{(i)}$.

$$L(\theta) = \prod_{i=1}^{m} P(y^{(i)}|x^{(i)}; \theta) = \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y^{(i)}-\theta^T x^{(i)})^2}{2\sigma^2}}$$  \hspace{1cm} (8)

The log-likelihood function is:

$$\ell(\theta) = \log L(\theta) = m \log \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{1}{\sigma^2} \frac{1}{2} \sum_{i=1}^{m} (y^{(i)}-\theta^T x^{(i)})^2$$  \hspace{1cm} (9)

Accordingly, the cost function is:

$$J(\theta) = \frac{1}{2} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2$$  \hspace{1cm} (10)

Next, the goal is to solve $\theta$ to minimize $J(\theta)$.

\begin{enumerate}
  \item[(a)] When the matrix is full rank
  
  In this case, the least squares method (LSM) is used. $J(\theta)$ is expressed as:
  
  $$J(\theta) = \frac{1}{2} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2 = \frac{1}{2} (X\theta - y)^T (X\theta - y)$$  \hspace{1cm} (11)

  Take the derivative of $\theta$:
  
  $$\nabla_\theta J(\theta) = \nabla_\theta \left(\frac{1}{2} (X\theta - y)^T (X\theta - y)\right)$$
  $$= \nabla_\theta \left(\frac{1}{2} \theta^T X^T - y^T\right) (X\theta - y)$$
  $$= \nabla_\theta \left(\frac{1}{2} \theta^T X \theta - y^T X \theta + y^T y - \theta^T X y\right)$$
  $$= \frac{1}{2} (2X^T X \theta - X^T y - (y^T X)^T)$$
  $$= X^T X \theta - X^T y$$
\end{enumerate}
Let $X^T X \theta - X^T y = 0$, then:
\[
\theta = (X^T X)^{-1} X^T y \tag{13}
\]

(b) When the matrix is not full rank

In this case, LSM may not be able to directly find the extreme value. Accordingly, the batch gradient descent is used to solve the local optimum, and the specific steps are as follows:

1. Assign a random value to $\theta$, which can be a vector of all zeros;
2. Change $\theta$ to make $J(\theta)$ decrease in the direction of gradient descent, i.e., for each component of $\theta$, update as the following manner until convergence:
\[
\theta_i := \theta_i - \alpha \frac{\partial J(\theta)}{\partial \theta_i} ; \quad i = 0, 1, \ldots, n \tag{14}
\]

where:
\[
\frac{\partial}{\partial \theta_i} J(\theta) = \frac{\partial}{\partial \theta_i} \frac{1}{2} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2 = \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}) \frac{\partial}{\partial \theta_i} (h_\theta(x^{(i)}) - y^{(i)}) \tag{15}
\]

Since:
\[
h_\theta(x^{(i)}) = \theta_0 + \theta_1 x_1^{(i)} + \theta_2 x_2^{(i)} + \cdots + \theta_n x_n^{(i)} \tag{16}
\]

Then:
\[
\frac{\partial}{\partial \theta_i} J(\theta) = \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}) \cdot x_i^{(i)} \tag{17}
\]

Finally, take the appropriate $\alpha$ as the step size to obtain:
\[
\theta_i := \theta_i - \alpha \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}) \cdot x_i^{(i)} \tag{18}
\]

3.2. Platform

In order to obtain the data required for regression analysis and PLMBSF evaluation, it is necessary to generate the target TCP flow and make it flow through the preset capture point so as to create a data capture scenario shown in Figure 1. In detail, the platform consists of four servers and ten clients located inside and outside JSERNET [26], respectively. Among them, the server configuration information is shown in Table 1, and number 1 to number 3 are communication servers, and number 4 is the storage server. The reason for setting 3 communication servers is to simplify the client operation to easily form TCP transfers with different acknowledgment mechanisms in the network [29]. Specifically, No. 1 server implements SCA by turning off SACK and DSACK, No. 2 server implements SACK by turning off DSACK and No. 3 server implements DSACK by default configuration. Moreover, to effectively carry out regression analysis and fully evaluate PLMBSF, in addition to the pre-placed files of different sizes on the communication server, we also irregularly arrange file uploads and downloads to obtain TCP transmissions with different sizes and loss rates.

![Figure 1. Data capture scenario.](image)
Thus far, we have arranged more than 10,000 TCP transmissions. Particularly, we selected the semantically complete TCP transmissions to generate the dataset RED_SET, focusing on the effectiveness evaluation of PLMBSF itself. Here, the semantically complete TCP transmission refers to the TCP transmission containing connection establishment (three-way handshake) and connection release (four-way handshake).

3.3. Target Parameter Determination

Regression analysis is mainly used to reveal the quantitative relationship between related variables. Therefore, before constructing the regression model, appropriate dependent variables should be set, and the related independent variables need to be selected in conjunction with the dependent variables. Combined with the research objective, in order to obtain goodput, the independent variables are set as SPR and LA. Then, goodput can be expressed as:

$$goodput = \frac{(1 - SPR - LA) \cdot TN \cdot AVG - TN \cdot (IH + TH)}{T}$$

(19)

where SPR represents the proportion of the number of spurious retransmissions at the monitoring point to the total number of data packets, TN represents the total number of data packets at the monitoring point, T indicates the flow duration, AVG represents the average number of bytes of the data packets, IH and TH represent the overhead of IP header and TCP header of the data packet respectively, and the default values are both 20 bytes.

For SPR and LA, the corresponding independent variable selected is AD, i.e., the ratio of the number of data packets to the number of ACK packets. For SPR, the reason why AD is selected as an independent variable is that the spurious retransmission will trigger redundant ACK and change the status of AD. Specifically, the number of spurious retransmissions will affect AD, and in turn, AD can reflect SPR to a certain extent. Furthermore, due to the existing relationship between TCP packet losses and AD [13], AD is also selected as the independent variable of LA. Accordingly, the specific steps to determine target parameters are: establish the regression model between AD and SPR to determine SPR → establish the regression model between AD and LA to determine LA → determine goodput, LB and RTT.

According to empirical analysis, whether SPR and AD, or LA and AD, there exists a multivariate nonlinear relationship between them. Therefore, we intend to linearize the multivariate nonlinear model first and then solve the target parameters with linear regression. Moreover, the polynomial model plays a very important role in nonlinear regression analysis. According to the theory of the series expansion method, any problem of the curve, surface and hypersurface can be approximated arbitrarily with the polynomials in a certain range. Therefore, when the correlation between the regression variables is unknown, it can be approximated with the polynomials of appropriate power. Since only one independent variable is involved here, namely AD, the corresponding polynomial function is the univariate polynomial, and its general form is:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_n x^n + \epsilon$$

(20)

Linearization, let:

$$x_i' = x_i, \ i = 1, 2, \ldots, n$$

(21)
Obtain the multiple linear regression equation:

$$y = \beta_0 + \beta_1 x_1' + \beta_2 x_2' + \cdots + \beta_n x_n'$$  \hspace{1cm} (22)

In order to establish the regression model between SPR and AD, we randomly selected twenty TCP flows from RED_SET to make statistics of SPR and AD, and the statistical result is shown in Table 2. As can be seen, there is a positive correlation between AD and SPR as a whole. Moreover, Table 2 also gives the corresponding values after linearizing AD.

**Table 2.** SPR under different AD.

| ID | X_1 | X_1' | X_2 | X_2' | X_y | Y   |
|----|-----|------|-----|------|-----|-----|
| 01 | 1.52| 1.52 | 2.3104 | 3.5118 | 0.1520750 |
| 02 | 1.80| 1.80 | 3.2400 | 5.8320 | 0.0110000 |
| 03 | 1.92| 1.92 | 3.6864 | 7.0779 | 0.0006875 |
| 04 | 1.74| 1.74 | 3.0276 | 5.2680 | 0.0242000 |
| 05 | 1.65| 1.65 | 2.7225 | 4.4921 | 0.0589875 |
| 06 | 1.97| 1.97 | 3.8809 | 7.6454 | 0.0000000 |
| 07 | 1.79| 1.79 | 3.2041 | 5.7353 | 0.0127900 |
| 08 | 1.85| 1.85 | 3.4225 | 6.3316 | 0.0046700 |
| 09 | 1.70| 1.70 | 2.8900 | 4.9130 | 0.0371200 |
| 10 | 1.60| 1.60 | 2.5600 | 4.0960 | 0.0880000 |
| 11 | 1.82| 1.82 | 3.3124 | 6.0286 | 0.0079700 |
| 12 | 1.90| 1.90 | 3.6100 | 6.8590 | 0.0013800 |
| 13 | 1.63| 1.63 | 2.6569 | 4.3307 | 0.0697125 |
| 14 | 1.55| 1.55 | 2.4336 | 3.7964 | 0.2821500 |
| 15 | 1.77| 1.77 | 3.1329 | 5.5452 | 0.0167700 |
| 16 | 1.72| 1.72 | 2.9929 | 5.1777 | 0.1920900 |
| 17 | 1.88| 1.88 | 3.5344 | 6.6447 | 0.0023400 |
| 18 | 1.58| 1.58 | 2.4964 | 9.9443 | 0.1018900 |
| 19 | 1.67| 1.67 | 2.7889 | 4.6575 | 0.0493600 |
| 20 | 1.94| 1.94 | 3.7636 | 7.3014 | 0.0002800 |

According to Table 2, X and y are:

$$X = \begin{bmatrix}
1 & 1.52 & 2.3104 & 3.5118 \\
1 & 1.80 & 3.2400 & 5.8320 \\
1 & 1.92 & 3.6864 & 7.0779 \\
1 & 1.74 & 3.0276 & 5.2680 \\
1 & 1.65 & 2.7225 & 4.4921 \\
1 & 1.97 & 3.8809 & 7.6454 \\
1 & 1.79 & 3.2041 & 5.7353 \\
1 & 1.85 & 3.4225 & 6.3316 \\
1 & 1.70 & 2.8900 & 4.9130 \\
1 & 1.60 & 2.5600 & 4.0960 \\
1 & 1.82 & 3.3124 & 6.0286 \\
1 & 1.90 & 3.6100 & 6.8590 \\
1 & 1.63 & 2.6569 & 4.3307 \\
1 & 1.55 & 2.4336 & 3.7964 \\
1 & 1.77 & 3.1329 & 5.5452 \\
1 & 1.72 & 2.9929 & 5.1777 \\
1 & 1.88 & 3.5344 & 6.6447 \\
1 & 1.58 & 2.4964 & 9.9443 \\
1 & 1.67 & 2.7889 & 4.6575 \\
1 & 1.94 & 3.7636 & 7.3014 \\
\end{bmatrix}, \quad y = \begin{bmatrix}
0.1520750 \\
0.0110000 \\
0.0006875 \\
0.0242000 \\
0.0589875 \\
0.0000000 \\
0.0127900 \\
0.0046700 \\
0.0371200 \\
0.0880000 \\
0.0079700 \\
0.0013800 \\
0.0697125 \\
0.2821500 \\
0.0167700 \\
0.1920900 \\
0.0023400 \\
0.1018900 \\
0.0493600 \\
0.0002800 \\
\end{bmatrix}$$
According to LSM, $\theta$ is:

$$\theta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} = (X^T X)^{-1} X^T y$$

$$= \begin{bmatrix} 1816.325349 & -2802.949119 & 1405.326402 & -227.896575 \\ -2802.949118 & 4866.286495 & -2783.836972 & 1788.08126 \\ 1405.326403 & -2783.836971 & 4866.286495 & -374.91797 \\ -227.896575 & 1788.08126 & -374.91797 & 84.9521 \end{bmatrix} \times \begin{bmatrix} 9.502045874 \\ -13.69684417 \\ 6.531289683 \\ -1.028488695 \end{bmatrix}$$

Finally, the regression model between $SPR$ and $AD$ is:

$$SPR = -1.02849 AD^3 + 6.53129 - 13.69684 AD + 9.50205$$

(24)

After obtaining the regression model indicated by Equation (24), a statistical test is needed to further confirm the correlation between $SPR$ and $AD$. Therefore, we next carry out a goodness-of-fit test and significance test.

The indicators used for goodness-of-fit are $R^2$ and adjusted $R^2$. Before the goodness-of-fit test, introduce the following indicators:

Total Sum of Squares (TSS):

$$TSS = \sum (Y_i - \bar{Y})^2$$

(25)

Explained Sum of Squares (ESS):

$$ESS = \sum (\hat{Y}_i - Y)^2$$

(26)

Residual Sum of Squares (RSS):

$$RSS = \sum (Y_i - \hat{Y}_i)^2$$

(27)

Then, $R^2$ is calculated as:

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum e_i^2}{\sum (Y_i - \bar{Y})^2} = 1 - \frac{0.0013636}{0.109766} = 0.988$$

(28)

Divide RSS and TSS by their respective degrees of freedom to eliminate the effect of the number of variables on the goodness-of-fit:

$$R^2 = 1 - \frac{\sum e_i^2 / (n - k)}{\sum (Y_i - \bar{Y})^2 / (n - 1)} = 1 - \frac{n - 1 RSS}{n - k TSS} = 1 - \frac{19 \times 0.0013636}{16 \times 0.109766} = 0.985$$

(29)

where $n - k$ and $n - 1$ are the degrees of freedom of RSS and TSS, respectively.

As can be seen from the above goodness-of-fit, $R^2$ and $R^2$ are both close to 1 and greater than 0.96. Therefore, the constructed regression model has a good fitting effect and meets the actual application requirements.

Next, to evaluate the joint significance, it is necessary to perform an F-test on the basis of variance analysis. To this end, introduce a null hypothesis and alternative hypothesis:

Null hypothesis: $H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$.

Alternative hypothesis: $H_1 : \beta_j (j = 1, 2, \ldots, k)$ is not all 0.
Construct statistics $F$:

$$F = \frac{\sum (\hat{Y}_i - \bar{Y})^2 / (k - 1)}{\sum (Y_i - \hat{Y}_i)^2 / (n - k)} = \frac{(n - k) \text{ESS}}{(k - 1) \text{RSS}} = \frac{16}{3} \times \frac{1.1078024}{0.0013636} = 4332.853 \tag{30}$$

Given the significance level $\alpha = 0.05$, check the critical value with degrees of freedom of 3 and 16 in the $F$ distribution table [30]:

$$F_{0.05}(3, 16) = 3.34 \tag{31}$$

According to Equation (31), $H_0$ is rejected, i.e., the regression relationship between the explanatory variable and the explained variable is significant.

In unary regression, the $F$-test and $t$-test are equivalent. However, in multiple regression, the $F$-test is significant, which does not mean that every explanatory variable has a significant influence on the explained variable. Therefore, for the constructed regression model, the $t$-test is also needed. To this end, introduce the null hypothesis and alternative hypothesis:

- Null hypothesis: $H_0 : \beta_j = 0$, $j = 1, 2, \ldots, k$.
- Alternative hypothesis: $H_1 : \beta_j \neq 0$, $j = 1, 2, \ldots, k$.

Construct statistics $t^*$:

$$t^* = \frac{\hat{\beta}_j - \beta_j}{\hat{\sigma} \sqrt{c_{jj}}} \tag{32}$$

where $\hat{\sigma}$ is the standard deviation of the random disturbance term, and its estimated value is:

$$\hat{\sigma} = \frac{\sum e_j^2}{n - k} \tag{33}$$

$c_{jj}$ is the element of row $j$ and column $j$ in matrix $(X^TX)^{-1}$.

Under hypothesis $H_0$, the $t$ statistics are:

$$t^*(9.502045874) = \frac{9.502045874}{0.0013636 \sqrt{1816.325349}} = 2616.094$$

$$t^*(AD) = \frac{-13.69684417}{0.0013636 \sqrt{4866.286495}} = -2303.852$$

$$t^*(AD^2) = \frac{6.531289683}{0.0013636 \sqrt{1788.08126}} = 1812.334$$

$$t^*(AD^3) = \frac{-1.028488695}{0.0013636 \sqrt{84.9521}} = -1309.319$$

Given significance level $\alpha = 0.05$, check the critical value with degrees of freedom of 16 in the $t$ distribution table [31]:

$$t_{0.05/2}(20 - 4) = t_{0.025}(16) = 2.1199$$

According to $|t^*(9.502045874)|, |t^*(AD)|, |t^*(AD^2)|$ and $|t^*(AD^3)|$, $H_0$ is rejected, i.e., each explanatory variable in the constructed regression has a significant impact on the explained variable.

Based on the discussion above, the regression model indicated by Equation (24) is optimum.

Likewise, referring to the statistical results in Table 3, the optimum regression between $AD$ and $LA$ can be built:

$$LA = 0.73652AD^2 - 2.85400AD + 2.77176 \tag{34}$$
Table 3. LA under different AD.

| ID | X₁  | X₁' | X₂  | Y  | ID | X₁  | X₁' | X₂  | Y  |
|----|-----|-----|-----|----|----|-----|-----|-----|----|
| 01 | 1.52| 1.52| 2.3104 | 0.132131 | 11 | 1.82| 1.82| 3.3124 | 0.017170 |
| 02 | 1.80| 1.80| 3.2400 | 0.021200 | 12 | 1.90| 1.90| 3.6100 | 0.005300 |
| 03 | 1.92| 1.92| 3.6864 | 0.003392 | 13 | 1.63| 1.63| 2.6569 | 0.072557 |
| 04 | 1.74| 1.74| 3.0276 | 0.035828 | 14 | 1.55| 1.55| 2.4336 | 0.132500 |
| 05 | 1.65| 1.65| 2.7225 | 0.064925 | 15 | 1.77| 1.77| 3.1329 | 0.028040 |
| 06 | 1.97| 1.97| 3.8809 | 0.023370 | 16 | 1.88| 1.88| 3.5344 | 0.007630 |
| 07 | 1.79| 1.79| 3.2041 | 0.047700 | 17 | 1.58| 1.58| 2.4964 | 0.093490 |
| 08 | 1.85| 1.85| 3.4225 | 0.011930 | 18 | 1.67| 1.67| 2.7889 | 0.057720 |
| 09 | 1.70| 1.70| 2.8900 | 0.047700 | 19 | 1.60| 1.60| 2.5600 | 0.01910 |

Combined with the application scenarios of Mathis, LA' is calculated as:

\[
LA' = \frac{LA \cdot TN}{TN \cdot (1 - SPR - LA)} = \frac{LA}{(1 - SPR - LA)}
\]

Correspondingly, LB' is calculated as:

\[
LB' = PLR - LA'
\]

Then, LB is calculated as:

\[
LB = \frac{TN \cdot (1 - SPR - LA) \cdot (PLR - LA')}{TN + TN \cdot (1 - SPR - LA) \cdot (PLR - LA')} = \frac{(1 - SPR - LA) \cdot (PLR - LA')}{1 + (1 - SPR - LA) \cdot (PLR - LA')}
\]

Finally, for RTT, we can estimate it with the existing models [32,33].

4. Sampling Analysis and Processing

Considering the system burden, the flow generated by the router is generally packet-sampled. Accordingly, the sampling needs to be processed to reduce the estimation error. The unknown variables involved in PLMBSF include: TDN, TAN, T, AVG and T. For this reason, how to determine these variables through sampled flow becomes the key to sampling analysis and processing.

In a sampling environment, if the number of sampled packets is known to be NS, then TDN can be estimated as follows [32,34].

Let A denote the number of packets and B denote the number of sampled packets, then:

\[
P_a = P(A = a \mid B = N_S) = \frac{P(B = N_S \mid A = a) \cdot P(A = a)}{P(B = N_S)}
\]

\[
P_a = \frac{P(B = N_S \mid A = a) \cdot P(A = a)}{\sum_{a' = N_S}^{\infty} P(B = N_S \mid A = a') \cdot P(A = a')}
\]

Since the network flow obeys a heavy-tailed distribution, TDN can be estimated by the Pareto distribution with a parameter of 1 [35], i.e., \(\beta = 1\) and \(A_{min} = N_S\), then:

\[
P(A = a) = \begin{cases} 
0, & a < N_S \\
\frac{\beta \cdot N_S^\beta}{a^{\beta+1}}, & a \geq N_S
\end{cases}
\]

Correspondingly, TDN is calculated as:

\[
TDN = \sum_{a = N_S}^{\infty} a \cdot P_a = \sum_{a = N_S}^{\infty} a \cdot \frac{P(B = N_S \mid A = a) \cdot P(A = a)}{\sum_{a' = N_S}^{\infty} P(B = N_S \mid A = a') \cdot P(A = a')}
\]
\[
= \sum_{\alpha = N_S}^{\infty} \alpha \cdot \left[ \frac{\alpha}{N_S} \right] S^{N_S} \cdot (1 - S)^{\alpha - N_S} \cdot \frac{N_S}{\alpha^2} \cdot \left[ \frac{1}{N_S} \right] \cdot (1 - S)^{\alpha}
\]

Although the numerator and denominator of Equation (40) contain infinite items, since the value of the expansion term is degressive, only the first few terms need to be calculated according to the actual accuracy requirements. Similarly, \(TAN\) can be obtained. As for \(T\) and \(AVG\), the general processing method is:

\[
T \approx T_S \quad (41)
\]

where \(T_S\) is the sampling flow duration.

\[
AVG = \frac{T_B}{T_N} \quad (42)
\]

where \(T_B\) and \(T_N\) represent the total number of bytes and packets in the sampled flow, respectively.

5. Experimental Verification

In this section, PLMBSF is compared with the classical trace-based algorithm of Benko-Veres [36] from four aspects: flow length, packet loss rate, acknowledgment mechanism and sampling rate. Since Benko-Veres does not support sampling, the corresponding unsampled packet trace is taken as the input. In addition, the flow length interval division standard takes the original flow length instead of the sampled one.

5.1. Effect of Flow Length

In order to cover as many flow lengths as possible, the distribution of the file size (original flow length) on the server side during data capture is shown in Figure 2. Among them, each flow length interval corresponds to 10 files with different sizes. Correspondingly, for each flow length interval, 10 TCP flows were selected for experiment. Since the presentation of the estimation results is grouped in flow length intervals, and each group of results is the packet loss estimation between different hosts, the traditional statistical-based error analysis is not applicable here. To this end, cosine similarity (CS) is used to measure the difference in estimation results between PLMBSF and Benko-Veres.

\[
\text{Cosine Similarity} = CS = \frac{A \cdot B}{||A|| \cdot ||B||} = \frac{\sum_{n} a_i \cdot b_i}{\sqrt{\sum_{n} a_i^2} \cdot \sqrt{\sum_{n} b_i^2}} \quad (43)
\]

where \(A = [a_i]_n\) and \(B = [b_i]_n\) represent two sets of estimated results, respectively. \(CS\) ranges from 0 to 1. The closer \(CS\) is to 1, the more similar \(A\) and \(B\) are, and the better the estimation result.

In order to focus on the effect of flow length on PLMBSF, the selected public parameters are shown in Table 4. Figure 3 shows the CS of PLMBSF and Benko-Veres under different flow length intervals. It can be seen that the flow length has no significant effect on PLMBSF after the monitoring point. Instead, with the increase in flow length, there is a significant difference in the estimation before the monitoring point, i.e., the estimation of long flow is better than that of short flow. Through analysis, the difference before the monitoring point mainly stems from the limitations of Mathis itself. Specifically, the premise of Mathis’s reliability is that the flow should be long enough to have enough interaction time to reach
steady state. When this assumption is not met, underestimation may occur. Actually, it is not the flow duration time that matters, but, rather, the transmission round (a transmission round equals the RTT in slow-start stage, or equals, in congestion avoidance, the time needed to increase the window by one segment). For the case of small loss rates, a short connection will not be able to interact enough times to average the effect of overestimated bandwidth caused by a slow start. Again, through limited rounds, short connections cannot exclude the effect of few rounds with fewer packet losses.

![Flow length distribution](image)

**Figure 2.** Flow length distribution.

| ID | Parameter                              | Description                      |
|----|----------------------------------------|----------------------------------|
| 1  | Original flow length                   | (0.1 MB, 100 MB)                 |
| 2  | Loss rate before the monitoring point  | (0.4%, 0.5%)                     |
| 3  | Loss rate after the monitoring point   | (1.2%, 1.3%)                     |
| 4  | Sampling rate                          | 0                                |
| 5  | Acknowledgment mechanism               | SCA                              |

**Table 4.** Public parameters.

![Cosine Similarity](image)

**Figure 3.** Effect of flow length interval on PLMBSF.

### 5.2. Effect of Loss Rate

The public parameters selected for investigating the effect of loss rate on PLMBSF are shown in Table 5. The selected 100 flows are divided into 10 groups on average according to loss rates, and the ranges of loss rate before and after the monitoring point corresponding to group \( i \) are \( (0.001 \cdot (i - 1), 0.001i] \) and \( (0.005 \cdot (i - 1), 0.005i] \), respectively. From the results
In Figure 4, it can be seen that the estimation after the monitoring point tends to be stable on the entire parameter space. In contrast, as the loss rate increases, the estimate accuracy before the monitoring point rises first and then falls. Specifically, when the loss rate is less than or equal to 1.5%, CS fluctuates between 0.7 and 0.8, when the packet loss rate rises to between 1.5% and 3%, CS remains above 0.8, and when the packet loss rate is greater than or equal to 3%, CS falls back to around 0.7.

Table 5. Public parameters.

| ID | Parameter                  | Description          |
|----|----------------------------|----------------------|
| 1  | Original flow length       | (5 MB, 7 MB)         |
| 2  | Loss rate before the monitoring point | (0, 1%)             |
| 3  | Loss rate after the monitoring point | (0, 5%)             |
| 4  | Sampling rate              | 0                    |
| 5  | Acknowledgment mechanism   | SCA                  |

![Figure 4. Effect of loss rate on PLMBSF.](image)

Actually, the estimation error mainly stems from Mathis itself. First, Mathis assumes that the receiver window is large enough and does not limit the data transmission rate. However, when the loss rate is too low to meet this assumption, what PLMBSF estimates is not the real loss rate but the “virtual” one imposed by the receiver buffer. Secondly, Mathis assumes that the loss event is random and is repaired by fast retransmission. However, the type of packet loss in a real network usually depends on the actual network conditions. In particular, when the tail drop strategy is implemented at the bottleneck, a series of packets may be lost. In this case, the sender will wait until the timeout retransmission is triggered. Correspondingly, such situation will increase with the increase in loss rate and eventually result in the inaccuracy of PLMBSF.

5.3. Effect of Acknowledgment Mechanism

The public parameters selected for investigating the effect of acknowledgment mechanism on PLMBSF are shown in Table 6. The selected 300 flows are divided into three groups on average according to the acknowledgment mechanism. For the TCP flows of each acknowledgment mechanism: First, calculate their CS separately; secondly, calculate their maximum CS, minimum CS, and average CS before and after the monitoring point. The experiment results in Figure 5 show that the acknowledgment mechanism has no significant effect on PLMBSF estimation after the monitoring point. Specifically, the overall CS is maintained between 0.7 and 0.8, while the average CS fluctuates around 0.74. Actually, for any acknowledgment mechanism, the spurious retransmissions will not update the
receiver buffer while increasing the number of data packets, thereby causing their differences in AD. Unlike the situation after the monitoring point, before the monitoring point, the performance of PLMBSF on SCA is significantly better than that on SACK and DSACK. Regardless of SACK or DSACK, when SACK is enabled, Mathis will treat multiple packet losses within one RTT as a single congestion, which will cause PLMBSF’s underestimation before the monitoring point.

Table 6. Public parameters.

| ID | Parameter                             | Description              |
|----|---------------------------------------|--------------------------|
| 1  | Original flow length                  | (5 MB, 8 MB)             |
| 2  | Loss rate before the monitoring point | (0.3%, 0.5%)             |
| 3  | Loss rate after the monitoring point  | (1.2%, 1.5%)             |
| 4  | Sampling rate                         | 0                        |
| 5  | Acknowledgment mechanism              | SCA, SACK, DSACK         |

Figure 5. Effect of acknowledgment mechanism on PLMBSF.

5.4. Effect of Flow Sampling

The selected 100 flows are divided into 10 groups on average according to flow length, and the range of flow length corresponding to group \(i\) is \((5 \cdot i \text{ MB}, 5 \cdot (i+1) \text{ MB})\). With reference to the effect of loss rate and flow length on PLMBSF, the selection and grouping of flows according to the parameters in Table 7 will have a relatively small interference on the investigation of the effect of the sampling rate. Correspondingly, the sampling rates are set to 1/4 and 1/16. From the results in Figure 6, it can be seen that the sampling results in certain fluctuations, and the estimation becomes more and more inaccurate as the sampling rate increases. In fact, this is mainly related to inaccurate statistical values of some parameters (e.g., flow length, flow duration time and so on) caused by sampling. In view of this, we try to reduce the adverse effect by increasing the number of flows. To this end, set the sampling rate to 1/32 and select 10, 50 and 100 flows within the same flow length interval for experiments, respectively. The experiment results in Figure 7 show that the accuracy and stability of PLMBSF, after increasing the number of sampled flows, are improved. This indicates that although sampling will bias the statistics of the relevant parameters, the effect on the expectation value of estimations is relatively small. Therefore, when evaluating overall performance, the weighted average of the estimates of enough sampled flows within a specific time granularity will reduce the effect of sampling.
Table 7. Public parameters.

| ID | Parameter                                | Description       |
|----|------------------------------------------|-------------------|
| 1  | Original flow length                     | (5 MB, 7 MB)      |
| 2  | Loss rate before the monitoring point    | (0.7%, 1.0%)      |
| 3  | Loss rate after the monitoring point     | (1.0%, 3.0%)      |
| 4  | Sampling rate                            | 1/4, 1/6, 1/32    |
| 5  | Acknowledgment mechanism                 | SACK              |

Figure 6. Effect of flow sampling on PLMBSF.

Figure 7. Effect of the number of sampling flows on PLMBSF.

6. Application Analysis

For PLMBSF, in order to focus on its estimation technique itself, we assume that the data source is a semantically complete TCP bidirectional flow. However, referring to the generation process of the sampled flow (e.g., NetFlow), the sampled flow, when taken as the data source of PLMBSF, must be preprocessed.

When the IP flow stored in the flow buffer meets one of the four conditions listed in Table 8, it is considered to be terminated. Subsequently, the corresponding one-way NetFlow is output. Combined with the generation process of NetFlow, the following two issues should be considered when taking NetFlow as input:

(a) Merging of truncated flows

As can be seen from Table 8, a NetFlow may correspond to a truncated IP flow in addition to a normally ended IP flow. If it is the latter, you need to merge the truncated NetFlow into a complete one.

(b) Bidirectional flow matching

The generated NetFlow is unidirectional. Therefore, when PLMBSF is applied in a real network, it is necessary to perform bidirectional flow matching based on 5-tuple consisting of protocol number, source IP address, destination IP address, source port and destination port so as to produce the required bidirectional flow.
Table 8. NetFlow output conditions.

| Condition | Description |
|-----------|-------------|
| 1         | No packet arrives within a period of time (default 15 s, configurable). |
| 2         | Flow duration is too long (default 30 min, configurable). |
| 3         | Flag indicating the end of the flow appears, such as FIN, RST, etc. |
| 4         | Flow buffer area is full. |

7. Conclusions

The model PLMBSF constructed and evaluated in this paper is able to realize the estimation for packet losses before and after the monitoring point when only sampled flow is available. Compared with traditional algorithms and models in this field, the advantages of our method are mainly reflected in two points: (1) Under the premise of considering cost, real-time and realistic-usability, the access network boundary performance measurement can be achieved with only a single point of data, and there are no problems of deployment difficulties and cooperative operation consistency. (2) Without making pre-assumptions on the performance of the intra-domain network, the performance status of intra-domain and inter-domain can be given at the same time, which is more in line with the actual demand of currently refined network performance management. Moreover, PLMBSF is suitable for continuous observation, at the large-scale network boundary, for packet loss behavior so as to realize fine-grained network performance monitoring. Moreover, PLMBSF is a good supplement to trace-based performance measurement, i.e., network operators and managers can make more targeted trace-based measurements on the basis of this macro measurement to fully understand and grasp network performance.

For PLMBSF, the estimation before the monitoring point tends to be stable, but the estimation after the monitoring point is easily affected by the flow length, loss rate and acknowledgment mechanism. To this end, in future work, it is considered to complete the model construction based on more complex and advanced PFTK, C-PFTK, GRZ, etc. Meanwhile, regardless of whether before or after the monitoring point, the expected value of the estimation is relatively less affected by sampling. Therefore, PLMBSF is suitable for overall performance measurement on the high-speed backbone network.

Furthermore, the impact of IP fragmentation [37] will spread to the flow record level. Accordingly, processing IP fragmentation is beneficial to improve the robustness of PLMBSF in practical applications. If one-way traffic [38] is supported, the number of available flows will be increased to further improve the availability of PLMBSF. Moreover, since wireless access networks [39] are now very popular and have noticeable effects on network performance [40], it will be very interesting to carry out relevant research in wireless networks.

Author Contributions: Conceptualization, H.L. and J.X.; investigation, H.L.; methodology, H.L.; resources, H.L.; software, H.L.; validation, H.L.; visualization, H.L.; writing—original draft, H.L.; writing—review and editing, W.D.; project administration, Q.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the JSPIGKZ Project (No. 2911121110), Innovative and Entrepreneurial Doctor of Jiangsu Province (No. JSSCBS20210598), Jiangsu Provincial University Nature Science Foundation Project (No. 2020KX007Z), Jiangsu Provincial Science and Technology Research Project (No. 20KJB413002), and National Key Research and Development Program Project (No. 2018YFB1800200).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.
References

1. Basso, S.; Meo, M.; Servetti, A.; De Martin, J.C. Estimating packet loss rate in the access through application-level measurements. In Proceedings of the ACM SIGCOMM Workshop on Measurements up the Stack, Helsinki, Finland, 17 August 2012; pp. 7–12.

2. Basso, S.; Meo, M.; De Martin, J.C. Strengthening measurements from the edges: Application-level packet loss rate estimation. ACM SIGCOMM Comput. Commun. Rev. 2013, 43, 45–51. [CrossRef]

3. Wu, H.; Liu, Y.; Cheng, G.; Hu, X. Real-time packet loss detection for TCP and UDP based on feature-sketch. In Proceedings of the IEEE INFOCOM Workshops, Vancouver, BC, Canada, 10–13 May 2021; pp. 1–6.

4. Lan, H.; Ding, W.; Deng, L. Application-level packet loss rate measurement based on improved L-Rex model. Symmetry 2019, 11, 44. [CrossRef]

5. Samba, A.; Yeremou, T.; Hermine, S.; Leandre, N.N. Networked iterative learning fault diagnosis algorithm for systems with sensor random packet losses, time-varying delays, Limited Communication and Actuator Failure: Application to the Hydroturbine Governor System. WSEAS Trans. Syst. Control 2021, 16, 244–252.

6. Demeter, R.; Kovari, A.; Katona, J.; Heldal, I.; Costescu, C.; Rosan, A.; Hathazi, A.; Thill, S. A quantitative study of using cisco packet tracer simulation software to improve IT students’ creativity and outcomes. In Proceedings of the IEEE International Conference onCogInfoCom, Naples, Italy, 23–25 October 2019; pp. 353–358.

7. Muoz, J.; Suárez-Varela, J.; Barlet-Ros, P. Detecting cryptocurrency miners with NetFlow/IPFIX network measurements. In Proceedings of the IEEE International Symposium on M&M, Catania, Italy, 8–10 July 2019; pp. 1–6.

8. Sierra, E.; Muelas, D.; Ramos, J.; de Vergara, J.E.L.; Morató, D.; Aracil, J. Online detection of pathological TCP flows with retransmissions in high-speed networks. Comput. Commun. 2018, 127, 95–104. [CrossRef]

9. Hark, R.; Richerzhagen, N.; Richerzhagen, B.; Rizk, A.; Steinmetz, R. Towards an adaptive selection of loss estimation techniques in software-defined networks. In Proceedings of the IFIP Networking Conference and Workshops, Stockholm, Sweden, 12–16 June 2017; pp. 1–9.

10. Braun, L.; Didebulidze, A.; Kammenhuber, N.; Carle, G. Comparing and improving current packet capturing solutions based on commodity hardware. In Proceedings of the ACM SIGCOMM Internet Measurement, Melbourne, VIC, Australia, 1–3 November 2010; pp. 206–217.

11. Dong, Z.; Liang, T.; Luo, H. Stability analysis of networked control systems with multi-packet dropout based on switched system approach. Int. J. Circuits Syst. Signal Process. 2020, 14, 13–20.

12. Perdices, D.; Muelas, D.; Prieto, I.; de Pedro, L.; Vergara, J.E.L. On the modeling of multi-point RTT passive measurements for network delay monitoring. IEEE Trans. Netw. Serv. 2019, 16, 1157–1169. [CrossRef]

13. Wu, H.; Gong, J. Packet loss estimation of TCP flows based on the delayed ACK mechanism. In Proceedings of the APNOMS, Berlin/Heidelberg, Germany, 23–25 September 2009; pp. 540–543.

14. Wu, H.; Gong, J.; Ma, Z. Estimating network path loss episode frequency by passive measurement. In Proceedings of the International Conference on Biomedical Engineering and Informatics, Chongqing, China, 16–18 October 2012; pp. 1457–1461.

15. Gu, Y.; Breslau, L.; Duffield, N.; Sen, S. On passive one-way loss measurements using sampled flow statistics. In Proceedings of the INFOCOM, Rio de Janeiro, Brazil, 19–25 April 2009; pp. 2946–2950.

16. Ricciato, F.; Strohmeier, F.; Dorfinger, P.; Coluccia, A. One-way loss measurements from IPFIX records. In Proceedings of the IEEE International Workshop on Measurements and Networking Proceedings, Catania, Italy, 10–11 October 2011; pp. 158–163.

17. Giannakou, A.; Dwivedi, D.; Peisert, S. A machine learning approach for packet loss prediction in science flows. Future Gener. Comput. Syst. 2020, 102, 190–197. [CrossRef]

18. Liu, R.; Yang, S.; Zhang, Q.; Li, X. ICMP NetFlow Records Based Packet Loss Rate Estimation. In Proceedings of the Eighth International Conference on IMCCC, Harbin, China, 19–21 July 2018; pp. 1238–1241.

19. Cheng, G.; Tang, Y.; Gyires, T. A lightweight approach to manifesting responsible parties for TCP packet loss. In Proceedings of the International Conference on Networks, Barcelona, Spain, 23–25 January 2015; pp. 211–217.

20. Fatima, S.; Kasmir, Y. Network monitoring of the MHT company using the DUDE. WSEAS Trans. Commun. 2020, 19, 89–97.

21. Mathis, M.; Semke, J.; Mahdavi, J.; Ott, T.J. The macroscopic behavior of the TCP congestion avoidance algorithm. ACM SIGCOMM Comput. Commun. Rev. 1997, 27, 67–82. [CrossRef]

22. Fernández-Delgado, M.; Sirsat, M.S.; Cerndas, E.; Alawadia, S.; Barroa, S.; Febrero-Bande, M. An extensive experimental survey of regression methods. Neural Netw. 2019, 111, 11–34. [CrossRef]

23. Padhye, J.; Firoiu, V.; Towsley, D.; Kurose, J. Modeling TCP throughput: A simple model and its empirical validation. ACM SIGCOMM Comput. Commun. 1998, 28, 303–314. [CrossRef]

24. Cardwell, N.; Savage, S.; Anderson, T. Modeling TCP latency. In Proceedings of the IEEE INFOCOM, Tel Aviv, Israel, 26–30 March 2000; pp. 1742–1751.

25. Guillemin, F.; Robert, P.; Zwart, B. Performance of TCP in the presence of correlated packet loss. In Proceedings of the IFIP Specialist Seminar on Internet Traffic Engineering and Traffic Management, Wurzburg, Germany, 22–24 July 2002; pp. 1–10.

26. Lan, H.; Ding, W.; Gong, J. Useful Traffic Loss Rate Estimation Based on Network Layer Measurement. IEEE Access 2019, 7, 33289–33303. [CrossRef]

27. Liebeherr, J.; Burchard, A.; Ciucu, F. Delay bounds in communication networks with heavy-tailed and self-similar traffic. IEEE Trans. Inf. Theory 2012, 58, 1010–1024. [CrossRef]
28. Ciaccia, F.; Romero, I.; Arcas-Abella, O.; Montero, D.; Serral-Gracià, R.; Nemirovsky, M. Sabes: Statistical available bandwidth estimation from passive tcp measurements. In Proceedings of the IFIP Networking Conference, Paris, France, 22–26 June 2020; pp. 743–748.
29. Lan, H.; Ding, W.; Zhang, Y. Passive overall packet loss estimation at the border of an ISP. KSII Trans. Internet Inf. Syst. 2018, 12, 3150–3171.
30. Yoo, S.; Cotton, S.; Sofotasios, P.; Matthaiou, M.; Valkama, M.; Karagiannidis, G.K. The fisher–snedecor $F$ distribution: A simple and accurate composite fading model. IEEE Commun. Lett. 2017, 21, 1661–1664. [CrossRef]
31. Thompson, G.; Maitra, R.; Meeker, W.; Bastawros, A. Classification with the matrix-variate-t distribution. J. Comput. Graph. Stat. 2020, 29, 668–674. [CrossRef]
32. Zhang, X.; Gong, J.; Wu, H. A method of estimating average round-trip latency based on specific flow records in NetFlow. Comput. Appl. Softw. 2010, 27, 64–67.
33. Su, Q.; Gong, J.; Su, Y. RTT estimation based on sampled flow data. Chin. J. Softw. 2014, 25, 2346–2361.
34. Duffield, N.; Lund, C.; Thorup, M. Estimating flow distributions from sampled flow statistics. IEEE/ACM Trans. Netw. 2005, 13, 933–946. [CrossRef]
35. Hosking, J.; Wallis, J. Parameter and quantile estimation for the generalized Pareto distribution. Technometrics 1987, 29, 339–349. [CrossRef]
36. Benko, P.; Veres, A. A passive method for estimating end-to-end TCP packet loss. In Proceedings of the Global Telecommunications Conference, Taipei, Taiwan, 17–21 November 2002; pp. 2609–2613.
37. Gilad, Y.; Herzberg, A. Fragmentation considered vulnerable. ACM Trans. Inf. Syst. Secur. 2013, 15, 1–31. [CrossRef]
38. Glatz, E.; Dimitropoulos, X. Classifying internet one-way traffic. In Proceedings of the Internet Measurement Conference, Boston, MA, USA, 14–16 November 2012; pp. 37–50.
39. Belghachi, M.; Debab, N. The adaptation of vehicle assisted data delivery protocol in IoV networks. Int. J. Appl. Math. Comput. Sci. Eng. 2020, 2, 25–30.
40. Fadallah, C.; Walid, F.; Jamal, H.; Khoukhi, L.; Khatoun, R. Source fabrication detection model based on key-value variables in reactive protocols of VANET. Int. J. Circuits Syst. Signal Process. 2020, 14, 959–965.