Rigorous Statistical Analysis of Internet Loss Measurements

Hung X. Nguyen
The University of Adelaide, Australia
hung.nguyen@adelaide.edu.au

Matthew Roughan
The University of Adelaide, Australia
matthew.roughan@adelaide.edu.au

ABSTRACT

In this paper we present a rigorous technique for estimating confidence intervals of packet loss measurements. Our approach is motivated by simple observations that the loss process can be modelled as an alternating renewal process. We use this structure to build a Hidden Semi-Markov Model (HSMM) for the measurement process, and from this estimate both loss rates, and their confidence intervals. We use both simulations and a set of more than 18000 hours of real Internet measurements (between dedicated measurement hosts, PlanetLab hosts, web and DNS servers) to cross-validate our estimates, and show that they are significantly more accurate than any current alternative.

Categories and Subject Descriptors: C.2.3 [Computer-Communications Networks]: Network Operations-network monitoring

General Terms: Performance, Measurement

Keywords: Performance Measurement, Accuracy, Loss Rate

1. INTRODUCTION

It is now common for network operators to perform ongoing performance measurements. The most common method to obtain these is to inject probe packets into the network to sample the end-to-end performance, and there are now standards for such measurements [2, 4], commercial devices available to perform these tests, and measurements are even supported by some routers. The missing element is a rigorous statistical methodology for the analysis of the results.

In this paper we present a rigorous technique for analyzing performance measurements, in particular, for estimating confidence intervals of packet loss rates. Various models for packet loss processes have been presented (for instances see [1,5]), usually with the aim of using these models for simulation or for modelling application protocols. These studies have unanimously found correlations between packet probes, but have suggested different approaches for modelling these losses. Our model draws on these past experiences of modelling packet loss. We model the loss process as an alternating renewal process. This model is more general than the commonly used simple exponential models and is needed to capture non-Poissonian loss behaviours that occur due to route reconvergence, etc. We can use this structure to build a Hidden Semi-Markov Model (HSMM) of the measurement process, and from this estimate both loss rates and their confidence intervals. While our approach does use simplifying assumptions about the underlying losses, the proof is in the pudding. We use a set of real Internet performance measurements to cross-validate our estimates, and show that they are more reasonable than any current alternative.

SIGMETRICS'10, June 14–18, 2010, New York, New York, USA.
ACM 978-1-4503-0038-4/10/06.

2. SAIL: STATISTICALLY ACCURATE INTERNET LOSS MEASUREMENT

2.1 Modelling the Measurement Process

The loss process on an Internet path is modelled as a continuous time binary stochastic process $I(t)$: $I(t) = 1$ if at time t an arriving packet would be dropped and $I(t) = 0$ otherwise. We assume that this process is wide-sense stationary, which means that its mean, variance and auto-covariance are all constant with respect to t for all $t \geq 0$. The mean of I is just the loss rate which we denote p, and can consequently be written $p = \mathbb{E}[I(t)]$, and in addition we use the notation $\text{VAR}(I(t)) = \sigma^2$, and $\text{Cov}(I(t), I(t+s)) = R(s)$.

In active probing, $I(t)$ is not observed directly. Instead, we shall send N probes at times $t_1, \ldots, t_N$, resulting in samples $I_i = I(t_i)$ of the underlying process. Given N samples, $I_1, \ldots, I_N$, the standard estimator for the loss rate is

$$\hat{p} = \frac{1}{N} \sum_{i=1}^{N} I_i. \tag{1}$$

When the loss process is stationary and ergodic, and we use Poisson probes, the above estimator is unbiased.

The variance of $\hat{p}$ based on N samples of $I(t)$ is

$$\text{VAR}(\hat{p}) \;=\; \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \mathbb{E}[I_i I_j] - p^2 = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} R(\tau_{ij}),$$

where $R(\tau)$ is the autocovariance function of $I(t)$, and τ_{ij} is the time between the i and jth probes.

We know the sample times t_i, and so if we knew the autocovariance function of $I(t)$, we could easily compute the above sum. The standard statistical approach to compute $R(\tau)$ is to build a model of the data, estimate the model parameters, and use this model to derive the autocovariance. To this end, we model loss process as alternating periods of consecutive losses and periods without loss, i.e., $I(t) \equiv \{(A_i, B_i)\}_{i=1}^{\infty}$ where A_i is the length of the ith loss period and B_i is the length of the ith no-loss period. This process is an alternating *renewal* process iff the A_i are IID (Independent, and Identically Distributed), the B_i are IID, and A_i and B_j are independent for all i and j. The alternating renewal process is described by the density functions of A_i and B_i, which we denote by $f_A(\cdot)$ and $f_B(\cdot)$ respectively. We use the Gamma distribution to model the ON/OFF periods with four parameters $w = \{(k_0, \theta_0), (k_1, \theta_1)\}$.

The model parameters are inferred from the measurements. Given these parameters, we can compute the autocovariance function by using a Laplace transform method or simulations [3]. Once we know the autocovariance function $R(\tau)$, we can then compute Gaussian confidence intervals for $\hat{p}$ (we use 95th percentile intervals in our results), i.e.,

$$\hat{p} \pm z_{\beta/2}\sqrt{\mathrm{VAR}(\hat{p})/N}, \qquad (2)$$

where $z_{\beta/2}$ are such that $\mathbb{P}(-z_{\beta/2} < N(0,1) < z_{\beta/2}) = 1-\beta$. For example, to compute the 95% confidence intervals, $z_{2.5} = 1.96$ is used and the confidence intervals would be $\hat{p} \pm 1.96\sqrt{\mathrm{VAR}(\hat{p})/N}$.

2.2 Inferring the Loss Rate from Measurement

Note that we do not observe the process $I(t)$ directly. Instead, we only observe $I(t)$ through our probes $I_1, \ldots, I_N$. We need to infer the model parameters from these probes. One standard approach for estimating parameters of an underlying renewal process from indirect observations is to use a Hidden Semi-Markov Model (HSMM) [6]. Our inference technique is similar to the inference of a HSMM with missing observations in [7]. Detail of the inference algorithm, SAIL (statistically accurate Internet loss measurement), is provided in [3]. Once the model parameters are inferred, SAIL uses (1) and (2) to compute the loss rate and its 95% confidence interval.

3. VALIDATION

We have performed extensive validation of our SAIL technique both in simulations and with real traffic traces [3]. We present here only results for one set of traces.

3.1 Data Sources

We apply SAIL to loss measurements between our dedicated measurement machines and between PlanetLab hosts. For each pair of source and destination, we send 10000 Poisson probes between the two machines with a fixed rate $\lambda = 10$ packets/second. The probes are UDP packets with a payload of 40 bytes. At the sender, we keep track of the sending times of each of the probes. The probe outcomes and probe sending times are used as input to the SAIL algorithm. We collected in total 1100 stationary packet traces.

3.2 Results

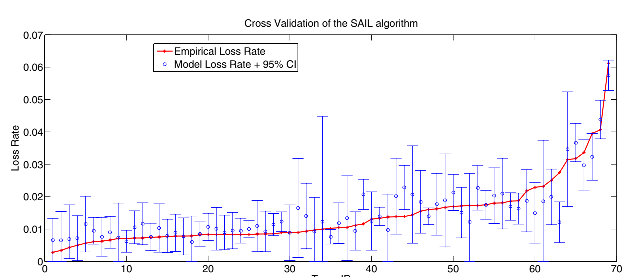

Figure 1: Cross validation for loss rate estimated using SAIL for network losses.

We validate SAIL by breaking each trace into two subsamples so that each sample has an even probability of being put into either subsample: one called the *inference trace* and the other the *cross validation trace*. The subsampling property of the Poisson process means that each of these subsamples forms a Poisson process as well, with rate $\lambda/2$. The PASTA property suggests that the two traces should (on average) report the same loss rate.

We plot in Figure 1 the loss rate and its 95% confidence interval provided by the SAIL for the inference trace. We also plot on the same figure the empirical loss rate of the validation trace. In order to make the figure comprehensible we report on 69 randomly chosen stationary traces (similar results are seen in the other traces).

Two important observations can be made. Firstly, the figure clearly shows that our model gives very accurate estimates of the loss rate variance. In 65 out of 69 traces (94%), the validation loss rate falls in the 95% confidence interval derived using SAIL. In the complete dataset 1012 out of 1100 (92%) fall in the interval. Note the datasets used for inference and validation are completely separate, so we have satisfied our chief goal of being able to estimate the accuracy of loss estimates to a reasonable degree of fidelity.

Secondly, the width of the confidence intervals is large. In many of the traces, the interval is larger than the loss rate itself. This is a fundamental property of loss rate measurements, not a failure of SAIL. The intervals must be this wide to encompass the intrinsic variation of loss rate estimates. The implication is that there is a large uncertainty involved in loss rate measurements and extreme care need to be taken in their interpretation. In fact, it is not overstating the case to say that loss rate measurements are much less useful than typically assumed. Without accurate confidence intervals it would be easy to come to erroneous conclusions, and this could have dramatic consequences if these measurements were, for instance, used as evidence in contractual litigation (say for purported failure to satisfy an SLA).

4. CONCLUSION

We have developed SAIL, a statistically accurate Internet loss measurement technique. SAIL employs sophisticated modelling methods from the Hidden Semi-Markov Model to accurately infer the parameters of the underlying loss process and then uses them to compute the loss rate and its confidence interval. We show that SAIL outperforms alternative methods in both simulations and real packet traces.

5. REFERENCES

[1] M. Yajnik et al. Measurement and modelling of the temporal dependence in packet loss. In *Proc. of Infocom'99*, 1999.
[2] R. Koodli and R. Ravikanth. *One-way Loss Pattern Sample Metrics*. IETF, RFC 3357, 2002.
[3] H. X. Nguyen and M. Roughan. Rigorous statistical analysis of Internet loss measurements. In *University of Adelaide Technical Report*, *http://adelaide.edu.au/directory/hung.nguyen* 2010.
[4] V. Paxson, G. Almes, J. Mahdavi, and M. Mathis. *Framework for IP performance metrics*. IETF, RFC 2330, 1998.
[5] K. Salamatian and S. Vaton. Hidden Markov Model for network communcation channels. *Proc. of Sigmetrics*, 2001.
[6] S.-Z. Yu. Hidden Semi-Markov model. *Artifical Intelligence*, 174:215–243, November 2009.
[7] S.-Z. Yu and H. Kobayashi. A Hidden Semi-Markov Model with missing data and mulitple observation sequences for mobility tracking. *Elsevier Transations on Signal Processing*, 83:235–250, December 2003.