Machine Learning based Prediction of Internet Path Dynamics

Sarah Wassermann*, Pedro Casas†, Benoit Donnet*
* Université de Liège, † AIT Austrian Institute of Technology
* sarah.wassermann@student.ulg.ac.be, benoit.donnet@ulg.ac.be, † pedro.casas@ait.ac.at

ABSTRACT

We study the problem of predicting Internet path changes and path performance using `traceroute` and machine-learning techniques. Path changes are frequently linked to path inflation and performance degradation. Therefore, predicting their occurrence could improve the analysis of path dynamics using `traceroute`. By relying on neural networks and using empirical distribution based input features, we show that we are able to predict (i) the remaining life time of a path before it actually changes, and (ii) the number of path changes in a certain time slot with relatively high accuracy. We also show that it is possible to predict path performance in terms of latency, opening the door to novel, machine-learning-based approaches for RTT prediction.

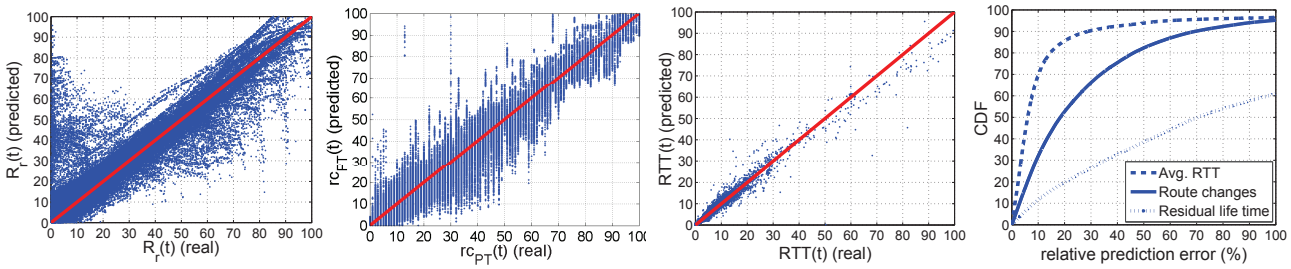
1. INTRODUCTION

Analyzing the performance of a certain path through active measurements requires to regularly measure or "sample" the path, by periodically launching traceroutes to retrieve relevant metrics. However, there is a constraint in how often measurements are performed, trading the accuracy of the analysis with the probing resource budget. As such, monitoring a large number of Internet paths through active measurements requires some smart ways to allocate a pre-defined probing budget. Internet paths change frequently due to inter/intra-domain routing changes, load balancing and even failures. Some of these changes can seriously disrupt performance, causing longer round-trip times, congestion, or even loss of connectivity [1]. For example, in [2], Google reports that inter-domain routing changes caused more than 40% of the cases in which clients experienced a latency increase of at least 100ms. We are currently extending our tool DisNETPerf [3] by adding an automatic approach to *dynamically* adapt the sampling rate of a path based on the remaining time until a next path change. In this paper, and similar to [4, 5], we propose to build a learning system which will be able to predict those path changes, relying on supervised machine learning algorithms. In addition, we explore the possibility of using the same approach to also predict path performance metrics such as RTT.

2. PREDICTING PATH DYNAMICS

Let us first introduce some basic definitions to formulate the corresponding learning and prediction problem. We define a path P as a sequence of links connecting a certain fixed source s to a fixed destination d. At any time t, path $P(t)$ is realized by a specific route r: this route consists of a specific sequence of links connecting s to d, and has an associated initial time t_0 when the route becomes active or in-place, and a final time t_f which corresponds to the time when r switches to another route realization, i.e., when the actual route changes. As such, a path $P(t)$ can be considered as a statistical time process, composed of a set of time-contiguous routes $r_i(t_0^i, t_f^i)$, each one with a duration $D(r_i) = t_f^i - t_0^i$, and with a total number of route changes $rc_P = i - 1$. We additionally define the duration of a route r as $D(r) = t_f - t_0$, its current life time at time t as $L_r(t) = t - t_0$, and its remaining life time at time t as $R_r(t) = t_f - t$. In our prediction problem, we want to estimate, at every time t, the remaining life time $R_r(t)$ of route r, namely $\hat{R}_r(t)$. As such, when $\hat{R}_r(t)$ becomes closer to 0, we would increase the sampling rate to better monitor the path performance in the event of a route change. In addition, we also want to predict – at every time t – the number of route changes a path experiences over a specific time window of length T, namely $\hat{rc}_{P_T}(t)$, which would allow to dynamically identify which paths are more prone to frequent changes. Besides path dynamics, we are also interested in predicting path performance, so we propose to predict the average RTT of a path at every time t, namely $\hat{RTT}_P(t)$.

To perform these estimations on a given path P at time t, we use as input a very rich set of features describing the statistical properties of the route duration $D(r)$ of P, its total number of route changes rc_P, the number of route changes rc_{P_T} for the target time window of length T, the current life time of the active route $L_r(t)$, as well as RTT, all of them computed on top of `traceroute` measurements. Note that we compute all these features by sampling their empirical distributions at multiple percentiles, assuming a predefined observation time T_{learn} of the monitored paths, during



(a) $\hat{R}_r(t)$ vs. $R_r(t)$. (b) $\hat{rc}_{P_T}(t)$ vs. $rc_{P_T}(t)$. (c) $\hat{RTT}_P(t)$ vs. $RTT_P(t)$. (d) Prediction relative error.

Figure 1: Real and predicted values.

which we collect these statistics for learning purposes. A worth-mentioning observation is that the set of features we are considering in this work is richer in terms of statistical properties than the one considered in previous work [4, 5], as, in particular, we are using as input empirical distributions and not only single metric values. Finally, we build a predictor based on neural networks, as these have shown good learning performance in a large number of applications. More precisely, we consider a standard multi-layer perceptron (MLP) regressor.

3. PRELIMINARY RESULTS – M-LAB

We study the performance of the proposed predictor, analyzing a full week of Paris `traceroute` measurements performed through M-Lab. The dataset corresponds to the first week of January 2016. During this week, we observe more than 450,000 different paths, sampled through Paris `traceroute` measurements from more than 100 geo-distributed servers. Unfortunately, not all of these paths are periodically sampled during this week; indeed, when analyzing the number of `traceroute` measurements for each of these paths, we found that only 15.725 paths have been sampled more than 10 times, and only 2346 paths have at least 100 `traceroute` associated measurements during the analyzed week. We use 100 as threshold to avoid reducing the useful dataset even more. Nevertheless, the more `traceroute` measurements we have for a path, the higher visibility on potential route changes. Having 100 samples in a week means a minimum path sampling rate of one `traceroute` every 100 minutes. For each of these 2346 paths P, we compute the distribution of the aforementioned input features during an observation period $T_{learn} = 1$ week. Note that while we use the full week of measurements to compute the input features of our predictor, performed evaluations are done on a 10-fold cross-validation basis, to avoid biased results. A large fraction of paths are rather stable, with about 40% of the observed routes lasting more than one hour, and with about 40% of the paths showing no path changes during the whole week. There is also an important share of dynamic paths, and in fact more than 30% of them have at least 20 route changes during the week, with routes lasting only a couple of minutes.

Figures 1(a), 1(b) and 1(c) report the obtained prediction results for $\hat{R}_r(t)$, $\hat{rc}_{P_T}(t)$ and $\hat{RTT}_P(t)$ respectively. We take $T = 24$ hours, meaning that we are interested in predicting the daily number of route changes of a path. Before commenting on the results, it is worth mentioning that previous work [4, 5] trying to predict $R_r(t)$ using `traceroute` data has already acknowledged that providing high prediction accuracy is very challenging. The figures plot both the real value of the corresponding metric as well as the predicted value, for the complete set of more than 350,000 `traceroute` measurements. Results are normalized to the maximum of each metric observed in T_{learn}. The straight line represents the ideal prediction scenario, in which $\hat{X} = X$.

The first observation in the three scenarios is that the correlation between predicted and real values are very high. Indeed, most of the values lie around the diagonal, and the computed correlation factors are always above 0.9. However, it is clear that the predictor results in important errors, especially for $R_r(t)$, both by underestimating as well as overestimating the real values. A second observation is that errors are higher for smaller values of $R_r(t)$ as well as for bigger values of $RTT_P(t)$, as they denote highly dynamic paths.

Figure 1(d) depicts the distribution of the relative prediction errors, defined as $\frac{|X-\hat{X}|}{X} \times 100$ (note that we consider the absolute prediction error when a sample $X = 0$). The relative prediction error for $R_r(t)$ is above 50% for about 60% of the samples, confirming the challenges already found in [4, 5] to predict the residual life time of a route. However, results are much more encouraging when considering the prediction of $rc_{P_T}(t)$ and the prediction of the average RTT, $RTT_P(t)$. For example, relative errors for the prediction of $rc_{P_T}(t)$ are below 50% for more than 80% of the `traceroute` measurements, and below 30% for more than 90% of the measurements when predicting RTTs. To conclude, we are currently working on an extensive comparison of our proposal to previous work [4]. We believe that the input features used by our predictor have better predicting power than those in [4], due to the inclusion of significant statistical properties.

4. REFERENCES

[1] U. Javed, I. Cunha, D. Choffnes, E. Katz-Bassett, T. Anderson, and A. Krishnamurthy, "Poiroot: Investigating the root cause of interdomain path changes," in *Proc. of the ACM SIGCOMM 2013*, August 2013, pp. 183–194.

[2] Y. Zhu, B. Helsley, J. Rexford, A. Siganporia, and S. Srinivasan, "Latlong: Diagnosing wide-area latency changes for CDNs," *IEEE Transactions on Network and Service Management*, vol. 9, no. 3, pp. 333–345, September 2012.

[3] S. Wassermann, P. Casas, B. Donnet, G. Leduc, and M. Mellia, "On the Analysis of Internet Paths with DisNETPerf, a Distributed Paths Performance Analyzer," *Proc. 10th IEEE Workshop on Network Measurements (WNM)*, November 2016.

[4] I. Cunha, R. Teixeira, D. Veitch, and C. Diot, "Predicting and tracking internet path changes," in *Proceedings of the ACM SIGCOMM 2011 Conference*, August 2011, pp. 122–133.

[5] ——, "Dtrack: A system to predict and track internet path changes," *IEEE/ACM Transations on Networking*, vol. 22, no. 4, pp. 1025–1038, August 2014.