Correlation Analysis on Real-Time Tab-Delimited Network Monitoring Data

Aditya Pan
Department of CSE, ASET
Amity University
Noida, India
pan.aditya93@gmail.com

Jahin Majumdar
Department of CSE, ASET
Amity University
Noida, India
jahin07@gmail.com

Prof. (Dr.) Abhay Bansal
Department of CSE, ASET
Amity University
Noida, India
abansal1@amity.edu

Prof. (Dr.) Bebo White
SLAC National Accelerator Laboratory
Stanford, CA, USA
bebo@slac.stanford.edu

Prof. (Dr.) Roger Leslie Anderton Cottrell
SLAC National Accelerator Laboratory
Stanford, CA, USA
cottrell@slac.stanford.edu

Abstract— **The PingER End-End performance monitoring of the Internet, is led by the SLAC National Accelerator Laboratory. It was created to answer the growing need to monitor the network both to analyze current performance and to designate resources to optimize execution between research centers, and the universities and institutes co-operating on present and future operations. The monitoring support reflects the broad geographical area of the collaborations and requires a comprehensive number of research and resources. The data architecture retrieval and methodology of the interpretation have emerged over numerous years. Analyzing this data is the main challenge due to its high volume. By using correlation analysis, we can make crucial conclusions about how the network data affects the performance of the hosts and how it depends from countries to countries.**

Keywords—Correlation; Network Monitoring; Ping; Tab-Delimited; PingER

I. INTRODUCTION

A significant challenge is presented at laboratories located worldwide that are mainly concerned with modern high-energy nuclear and particle physics research. SLAC National Accelerator Laboratory at Stanford, CA, USA has collaborated with a number of research laboratories worldwide, for example the Brookhaven National Laboratory (BNL), Relativistic Heavy Ion Collider (RHIC) and the Large Hadron Collider (LHC) at the European Center for Particle Physics (CERN). The research laboratories form petabytes (10^{15} bytes) or even Exabytes (10^{18} bytes) [1] of data while performing the research and recording it. A larger chunk of this data is delivered via the Internet for analysis to the experiments' collaborators at universities and research institutes everywhere in the world.

PingER was formed to report end-to-end performance over pings. The monitoring activities were originally sponsored by the Network Monitoring Task Force (NMTF) of the Energy Sciences Network (ESnet). This group takes particular interest in performance between laboratories funded by the universities and institutes involved in research at these laboratories and the U.S. Department of Energy (DoE). More recently it has been sponsored by the Standing Committee on Interregional Connectivity (SCIC) of the International Committee for Future Accelerators (ICFA). This second group addresses problem areas of international and especially variant performance in multiple networks connecting research institutes and universities performing high energy physics research.

An issue regarding the usage of the network data is that given the volumes of the data and the limited compute power, the data needs to be pre-processed to provide useful reports and then correlated to be able to find logical relations. Such logical relations include the network performance of a country and its economic growth and progress, or to analyze the impact of natural disasters. Since such data is of high volume, correlation analysis needs to be performed on multiple tuples to find an accurate relation between a country's economic factor and internet performance.

The rest of the paper is organized as follows: Section II discusses the background of the PingER project. Section III discusses the data analysis metrics used in the analysis of network data obtained from PingER. Section IV presents the methodology using Pearson's correlation analysis. Section V provides the results of the analyzed data on three crucial network metrics. Section VI concludes the paper.

II. BACKGROUND

The PingER project was initially started in 1995 with the objective of aiding the High End Physics Research. However, in the recent years it has shifted its focus to measuring the performance in the digital world from the point of view of Internet pings. It is an end to end internet monitoring tool, started by SLAC National Accelerator Laboratory, Stanford, CA. SLAC has collaborated with various other institutions to set up network monitoring sites all across the globe.

A. History

When PingER originally started the data was stored in separated space files. To assist in searches a fast binary search algorithm was applied to the data. This system, however, faces flexibility in real time data selection. To resolve this issue, it was suggested that data be stored in a relational database. The National University of Science and Technology in Pakistan attempted to provide this capability [2]. However it did not scale well and was too slow. More recently the format of RDF

978-1-4673-8203-8/16/$31.00

Triples which belongs to the World Wide Web Consortium standard was proposed and a Proof of Concept attempted. However it was too slow and did not scale as implemented. An issue is the amount of data. PingER generates enormous amounts of data and analyzing this data in real-time becomes a challenging task. The objective is to find interesting and undiscovered patterns in the data by clustering data based on different parameters such as the country it belongs to. The trends for various countries can be analyzed and compared and new conclusions can be drawn from them [6].

B. Framework

For each remote node in a list, every 30 minutes PingER sends a single 100byte ping [3] to prime the caches [4]. It follows this by sending up to 30 100byte pings with a separation with a separation of 1 second and a default timeout (20 seconds) to the remote node. PingER stops sending the pings to the remote node when/if it receives 10 responses. This is followed for a similar set of 1000byte pings. Every monitoring node-remote node bundle is called a pair.

C. Resource Description Framework (RDF)

Resource Description Framework (RDF) Triples are a format of data representation comprising of three parts which are a subject, a predicate and an object. The subject contains either a blank node or an URI reference. The predicate contains an RDF URI reference. The object is also a reference which could either be a blank node or a literal. It belongs to the family of World Wide Web consortium specifications. RDF triples are used because it makes data representation in a semantic web much simpler and organized. It puts all data into a common format which makes it simpler to integrate and combine the data. Hence, SLAC also attempted to put all the PingER data into RDF triple format [5].

D. Objective

The basic objective is to monitor end to end pings in a network and study internet performance based on the Round Trip Time of these pings. The project presently monitors around 700 sites from approximately 160 countries across the globe [6]. It was developed by the IEPM group at SLAC National Accelerator Laboratory. The PingER data repository for network data is around 18 years old, and measurements to and from sites around the globe are found in it. PingER data monitors countries that contain over 99% of the internet population data. PingER had 20 monitoring sites all around the world in December 1999. Eight of which were in the United States. Monitoring sites in Asia were located in Japan and China. Japan had two and China had one. Now it has about 50 working MAs in 20 countries.

E. Retrieval and Storage

Finally, this data is built to form different reports at the analysis site using code that is written in Perl. All reports are available on a web page as HTML table from where data may be retrieved in Comma Separated Values (CSV) or tab-separated values (TSV) format and imported for correlation analysis.

III. Metrics used in PingER Measurement

In ideal circumstances, network traffic should cross the Internet at the highest speed for the medium. However queuing in routers etc, can often add extra delays. Five known metrics are represented to design and to appear as the effect of this queuing to judge network performance in PingER. The five known metrics are known to be packet loss, Round Trip Time (RTT), unreachability and jitter.

A. Packet Loss

Packet loss is defined as the percentage of network packets lost while transmitting data from one host to another. Packet loss indicates well that the link is congested enough for packets to get discarded in transit. If a 4% packet loss is incurred, it ideally means that the application using TCP/IP will depreciate to a great extent [7]. This is mainly due to the effect of resending a packet administered by TCP/IP algorithms. However, the end-user experiences will fluctuate to a great extent in the application. Heavy tasks like video-conference will become unusable with moderate packet loss due to high interactivity whereas e-mail which is non-interactive will work even with high packet loss [8].

B. Round-Trip Time (RTT)

The process of buffer queuing described beforehand also changes the Round-Trip Time (RTT). It is never plausible to reduce the RTT to less than the time taken for light to travel the total distance along the medium (e.g. optical fiber cable) [9]. The minimum RTT as an indicator typically eliminates the effect of queueing and hence is determined by shows that the length of the route adopted by the packets, the total number of hops counted, and the line speeds of the links. Route change is often thus indicated by marked change in the minimum RTT [10].

C. Unreachability

Unreachability is the scenario where the remote node is discarded if the reply that is collected from all sent ping packets is nil. Measuring Unreachability is necessary for correct network performance analysis [11].

D. Quiescence

If a reply is received by all 10 packets sent to a remote node, the network is deemed to be non-busy or quiescent. The incidence of the zero packet loss is an indication to use the system. An 8 work hours per weekday occupied network and quiescent at other times, is said to have a dormant percent of about 85%. If the system is non-quiescent all during the day, it is considered to be poor and needs upgrading [11].

E. Unpredictability

Unpredictability is obtained from a formula which is based on the variation of packet loss and RTT. The success rate of the ping is the proportion of data responses obtained from the amount of packets sent, and the ping ratio is twice that of the ping payload as compared to the average RTT [12]. In any period of time, ‘st’ is the ratio of the mean and maximum ping success, ‘rp’ is the average and highest ping rate. They are linked to produce the unpredictability, ‘un’, where

$$un = \frac{1}{\sqrt{2}}\sqrt{(1-rp)^2 + (1-st)^2} \quad (1)$$

IV. METHODOLOGY

The task was to use Pearson's correlation analysis on two separate datasets from different countries and try to obtain a relation between them. The first dataset represents the hosts in SLAC, CA, USA and the second dataset was from Europe. Pearson's Correlation Analysis was used and compared with two datasets with the values of min, max and average time taken for the ping to reach [13]. The primary aim is to compare the datasets between various countries to try to analyze the internet performance among its hosts [14]. The different steps in obtaining the correlation are listed below.

A. Collection of Data

The network data was obtained from [15]. A total of 155 datasets were collected and analyzed. Out of that, two datasets were chosen, and Pearson's correlation was applied to them. The data sets had the following attributes: source_host_name, source_host_address, destination_host_name, destination_ host_address, size, unix_epoc_time, snt, rcv, min, avg, max, seq_rcv (i=1, rcvd=10) and rtt_rcv (i=1, rcvd=10). The most current data was obtained to the current date. The data had to analyzed well because applying a correlation function would require the data to have a same number of pings. Hence, pre-processing was required for the data to be available and ready for statistical analysis. The excess data was ignored for sites with the larger number of pings. A few excerpt from the data is shown below in the following two tables:

Table 1: First Dataset

source_host_name	source_host_address	destination_host_name	destination_host_address	size	unix_epoc_time	snt	rcv	min	avg	max
pinger.slac.stanford.edu	134.79.104.80	netgate.net	205.214.169.4	100	1444090135	10	10	2.224	2.343	2.451
pinger.slac.stanford.edu	134.79.104.80	netgate.net	205.214.169.4	1000	1444090144	10	10	3.028	3.09	3.167
pinger.slac.stanford.edu	134.79.104.80	netgate.net	205.214.169.4	100	1444090156	10	10	3.001	2.674	2.863

Table 2: Second Dataset

source_host_name	source_host_address	destination_host_name	destination_host_address	size	unix_epoc_time	snt	rcv	min	avg	max
pinger.slac.stanford.edu	134.79.104.80	pinger.stanford.edu	171.66.6.39	100	1446336998	10	10	0.904	0.951	1.025
pinger.slac.stanford.edu	134.79.104.80	pinger.stanford.edu	171.66.6.39	1000	1446337007	10	10	1.225	1.24	1.259
pinger.slac.stanford.edu	134.79.104.80	pinger.stanford.edu	171.66.6.39	100	1446337723	10	10	1.234	0.934	1.044

B. *Applying Pearson's Correlation*

Pearson's correlation analysis is given by the following formula:

$$r = \frac{\sum XY - \frac{\sum X \sum Y}{N}}{\sqrt{\left(\sum X^2 - \frac{(\sum X)^2}{N}\right)\sqrt{\left(\sum Y^2 - \frac{(\sum Y)^2}{N}\right)}}} \quad (2)$$

The dataset with negate.net as destination_host_name was chosen to be the first dataset (X). The other dataset with pinger.stanford.edu as destination_host_name was selected to be the second dataset (Y). A linear regression function according to y = ax + b was applied first to the min time. The attributes of max time and the average time was used later on. The first dataset has 115 ping data while the second dataset had 194 ping data. The second dataset therefore, had to be pre-processed to allow the correlation function to run smoothly and correctly. Pearson's correlation analysis was chosen because Pearson's correlation coefficient has various advantages for continuous non-normal data having no obvious outliers. Pearson's correlation coefficient offers a major and variable success in mathematical power even for distributions with moderate to high skewness or excess kurtosis. Hence, Pearson's correlation leads to a less powerful statistical and methodological test for distributions with maximum skewness or excess of kurtosis because of its known sensitivity to outliers for continuous non-normal data. Hence, Pearson's correlation was used as a preferred method for correlation analysis of network data.

V. RESULTS

The maximum, minimum and the average RTT was extracted from the tab-delimited network data and then analyzed with the help of Pearson's correlation. The correlation coefficient was found out followed by the equation of the straight line through the graph. Thereafter, the RTT is analyzed on the graph for each parameter. The following results were obtained for the maximum, minimum and the average time taken for the hosts to transmit a ping from one country to another. Pearson's correlation provides the statistical analysis to find the relationship between a country's internet performance and economic growth. The results were graphed using a simple graphical tool and the data is shown as two clusters each belonging to two datasets.

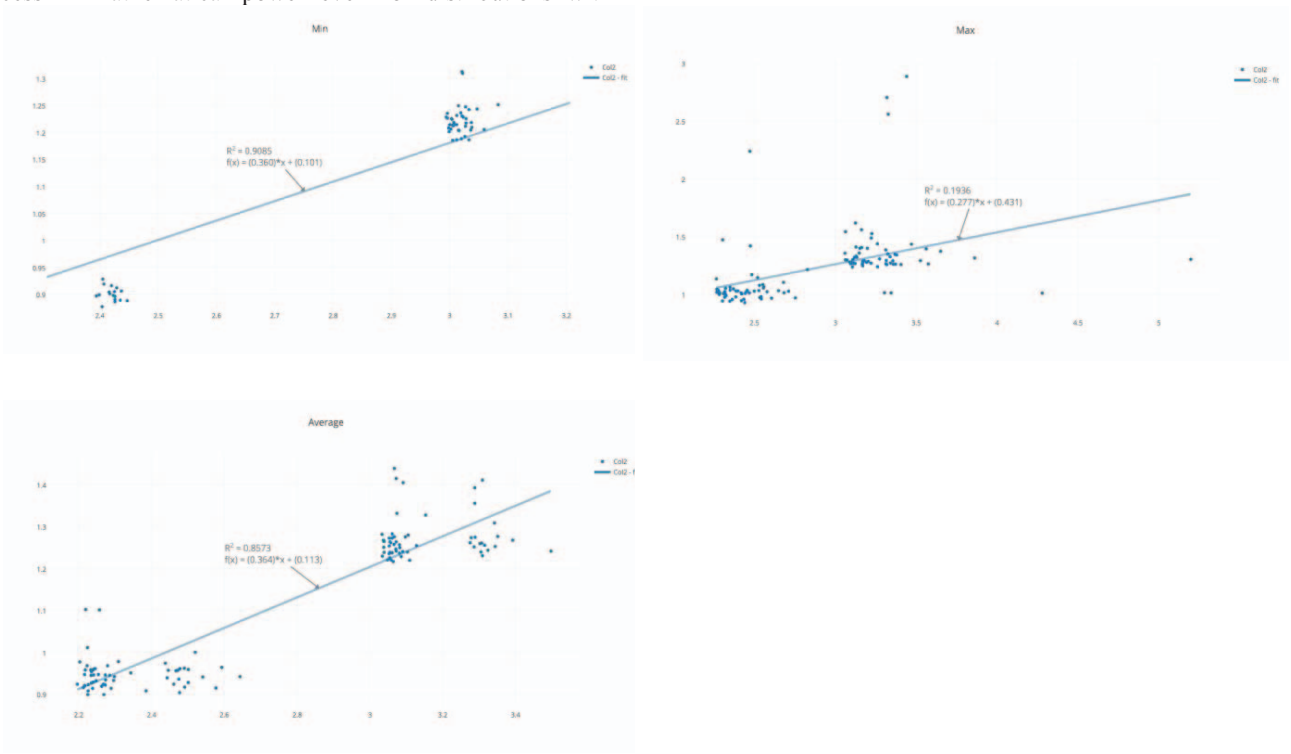

Figure 1: Correlation Graph of Min, Max and Average RTT

a) The min time was first analyzed with Pearson's correlation. It showed a very strong correlation (R2 = 0.9085). The linear equation which fit the curve is: y = (0.360) * x + (0.101).

b) The max time was then analyzed with Pearson's correlation. It showed a very weak correlation (R2 = 0.1936). The linear equation which fit the curve is y = (0.277) * x + (0.431).

c) The average time was finally analyzed with Pearson's correlation. It showed a moderately strong correlation (R2 = 0.8573). The linear equation which fit the curve is y = (0.364) * x + (0.113).

VI. Conclusion

The above results show that the min time and the average time are strongly correlated among the datasets. The max time on the other hand has a weak correlation. By analyzing further datasets among hosts belonging to separate countries, we can find out and conclude much more interesting results. The datasets considered in this paper includes hosts situated in different developed countries. Hence, the minimum and the average time showed a string correlation indicating that the countries internet performance is on the higher percentile. The results show that two developed countries have a higher correlation in minimum and average time. Developing or third world countries would be expected to have a lower R2 value for minimum or average time. The maximum time was found out to be similar for all type of hosts belonging to different countries. Future research include using correlation analysis for hosts belonging to different research universities and using a different correlation parameter.

Acknowledgment

The dataset has been obtained from SLAC National Accelerator Laboratory and the authors would like to thank SLAC National Accelerator Laboratory for the same.

References

[1] Antcheva, I., Ballintijn, M., Bellenot, B., Biskup, M., Brun, R., Buncic, N., ... & Franco, L. (2009). ROOT—A C++ framework for petabyte data storage, statistical analysis and visualization. Computer Physics Communications, 180(12), 2499-2512.

[2] Nabi, G., Maqsood, M. F., & Naseer, B. (2011). Analysis and Implementation of Relational Archive Site for PingER and CBG Integration with TULIP.

[3] J. Postel, "Internet Control Message Protocol," RFC 792, ftp://ftp.isi.edu/in-notes/rfc792.txt, Sept. 1981.

[4] Horneffer, M. reported in http://www.advanced.org/IPPM/archive.2/0246.html (no longer available) that using UDP-echo packets and an inter-arrival-time of about 12.5 seconds the first packet takes about 20% more time to return.

[5] Souza, R., Cottrell, L., White, B., Campos, M., & Mattoso, M. (2014). Linked open data publication strategies: Application in networking performance measurement data.

[6] R Les Cottrell. Shawn McKee, "ICFA SCIC Network Monitoring Report 2015", http://www.slac.stanford.edu/xorg/icfa/icfa-net-paper-jan15/report-jan15.docx, Jan. 2015.

[7] Lakshman, T. V., & Madhow, U. (1997). The performance of TCP/IP for networks with high bandwidth-delay products and random loss. Networking, IEEE/ACM Transactions on, 5(3), 336-350.

[8] Yu, T., & Lin, K. J. (2005). Service selection algorithms for Web services with end-to-end QoS constraints. Information Systems and E-Business Management, 3(2), 103-126.

[9] Wille, E. C., Mellia, M., Leonardi, E., & Marsan, M. A. (2006). Algorithms for IP network design with end-to-end QoS constraints. Computer Networks, 50(8), 1086-1103.

[10] Karagiannis, T., Faloutsos, M., & Riedi, R. H. (2002, November). Long-range dependence: now you see it, now you don't!. In Global Telecommunications Conference, 2002. GLOBECOM'02. IEEE (Vol. 3, pp. 2165-2169). IEEE.

[11] Matthews, W., & Cottrell, L. (2000). The PingER project: active Internet performance monitoring for the HENP community. Communications Magazine, IEEE, 38(5), 130-136.

[12] Cottrell, R. L. A., Logg, C. A., & Martin, D. E. (1998). What is the Internet doing? Performance and reliability monitoring for the HEP Community. Computer physics communications, 110(1), 142-148.

[13] Shaheen, M; Shahbaz, M; and Guergachi, A; Context Based Positive and Negative Spatio Temporal Association Rule Mining, Elsevier Knowledge-Based Systems, Jan 2013, pp. 261-273.

[14] Wong, Andrew K.C.; Wang, Yang (1997). "High-order pattern discovery from discrete-valued data". IEEE Transactions on Knowledge and Data Engineering (TKDE): 877–893.

[15] http://slac.stanford.edu/cgi-wrap/ping_data.pl