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

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

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

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

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

Abstract— **End-End performance monitoring in the Internet, also called PingER is a part of SLAC National Accelerator Laboratory's research project. It was created to answer the growing need to monitor 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 financial channels. 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, namely 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.

The projects have been combined and have resulted in an exclusive end-to-end performance monitoring framework that is being setup with an extensive network probing method with a large set of tools to examine the data. PingER was thus formed to report end-to-end performance over pings. More particularly, the monitoring activities are forced by the two groups and the framework of the specific PingER project review in this report. The primary group is 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). The Standing Committee on Interregional Connectivity (SCIC) of the International Committee for Future Accelerators (ICFA) is the second group. 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 which the two groups faced was that the data was uncluttered and in a raw format. The data needs to be pre-processed and then correlated to be able to find a logical relation between the network performance of a country on a real-time basis and its economic growth and progress or to analyze 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 Linked Open Data was stored in CSV files. However due to difficulty in accessing the data, this method of storing data was replaced with relational databases. A system was devised to convert the flat files into relational database entries [2]. This system faces issues with scalability and efficiency. To resolve this issue, it was suggested that data be stored in the format of RDF Triples which belongs to the World Wide Web Consortium standard. Though this is the format currently in use, it too faces some challenges. Another issue is the amount of data. PingER generates enormous amounts of data and analyzing this data 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 [3]. The trends for various countries can be analyzed and compared and new conclusions can be drawn from them.

B. Framework

A 100-byte payload equipped pings are transferred 11 times by PingER at a time interval of 2 second. This is followed by 20 pings with a 1024-byte payload, again at 1 second time intervals, to each of a dataset of particularized remote nodes listed in an arrangement file. The initial ping is discarded since it is inactive due to priming caches. It is understood from UDP echo packets that the initial packet holds about 30% larger than subsequent packets [4]. The default ping timeout is 10 seconds. However, a weak link study indicates that is default time is too small to be considered since, the number repeating after 10 seconds but before 100 seconds is less than 0.1%. Dodge fragmentation is used to transfer small packets for every set of 10 pings which are called a 'representation'. 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 attempts 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 speed of transfer of these pings. The project presently monitors around 700 sites from approximately 160 countries across the globe [1]. It was developed by the IEPM group at SLAC National Accelerator Laboratory. The PingER data repository for network data is around a decade old and measurements from and to sites around the globe are found in it. PingER data monitors approximately 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.

E. Retrieval and Storage

Finally, this data is built form different reports at the analysis site using code is written in Perl and Java. All reports are available on a web page as HTML table from where they are 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, in rare situations, associations occur. 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, unpredictability, and quiescence.

A. Packet Loss

Packet loss is defined as the percentage of network packets lost while transmitting data from one host to another. Packet loss provides a good indication that a portion of the link is congested. Ideally, the performance of the application using TCP/IP will depreciate significantly after 4% packet loss [7] for the issue of packet resending administered by the TCP/IP algorithms. However, the consequence which the end user experiences will fluctuate according to the application. Video-conferencing will become unusable with moderate packet loss as it is highly interactive whereas e-mail which is non-interactive will work even with high packet loss.

B. Round-Trip Time (RTT)

The process of buffer queuing described beforehand also changes the Round-Trip Time (RTT). However, unlike packet loss, it is possible to depreciate losses to nil, it is never plausible to reduce the RTT to less than the time taken for light to travel the total distance along the optical fiber cable [8]. The minimum RTT shows that the length of the route adopted by the packets, he total number of hops counted, and the line speeds of the channel. Route change is indicated by the slightest mutation in the RTT [9].

C. Unreachability

Unreachability is the scenario where the remote node is discarded if the reply that is collected from all ten ping packets is nil. Network production is the matter of unreachability and it is necessary for correct network performance analysis. It is also difficult for the program analysis code to tell the difference [10]. Less than 2% of PingER specimens among nodes on these networks are completely lost; consequently, unreachability less than 2% is classified as reachable [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 [12].

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 [13]. 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

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

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 [14]. The primary aim to compare the datasets between various countries to try to analyze the internet performance among its hosts. The different steps in obtaining the correlation are listed below [15].

A. Collection of Data

The network data was obtained from [16]. 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:

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

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 significant advantages for continuous non-normal data which does not have obvious outliers. Pearson's correlation coefficient offers a substantial success in mathematical power even for distributions with moderate skewness or excess kurtosis. Hence, because of its known sensitivity to outliers, Pearson's correlation leads to a less powerful statistical test for distributions with maximum skewness or excess of kurtosis.

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.

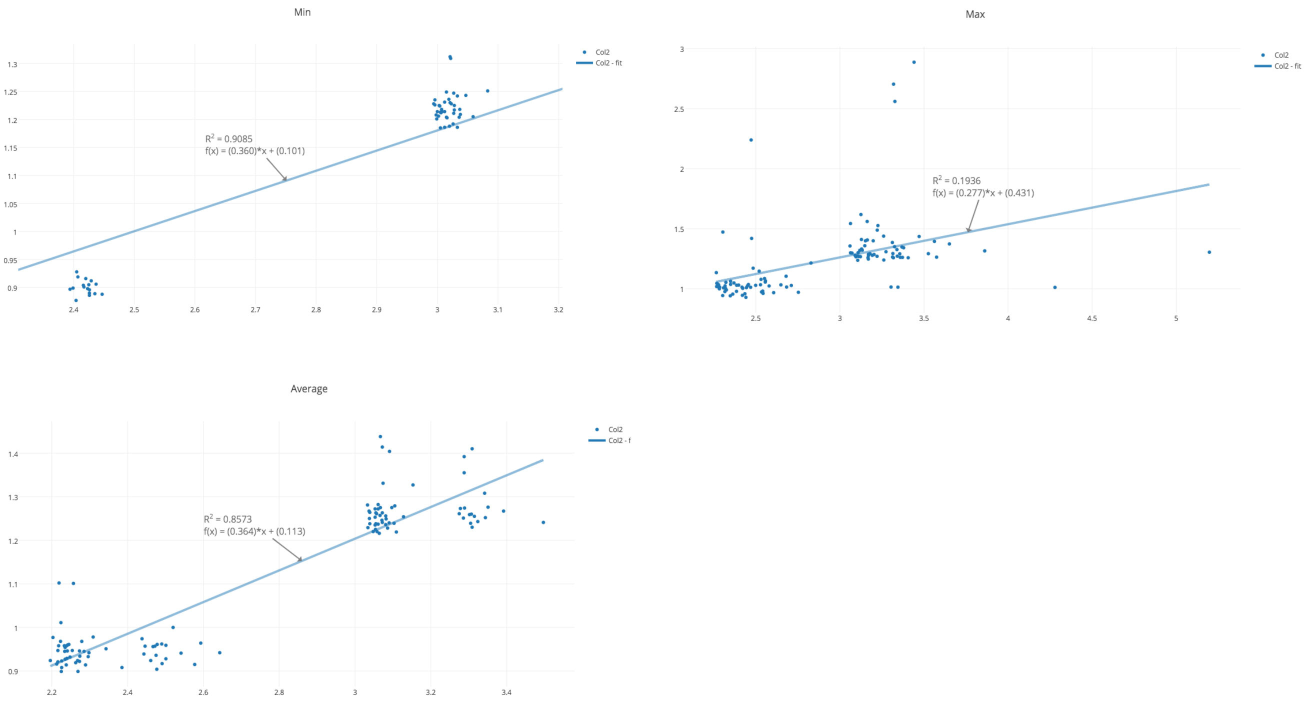

Figure 1: Correlation Graph of Min, Max and Average RTT

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.

References

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

[2] R. L. A. Cottrell and J. Halperin, “Effects of Internet Performance on Web Response Times,” http://www.slac.stanford.edu/comp/net/wan-mon/ping/correlation.html, Dec. 1996.

[3] M. Mathis et al., “The Macroscopic Behavior of the TCP Congestion Avoidance Algorithm,” Comp. Commun. Rev., vol. 27, no. 3, July 1997.

[4] M. Horneffer, IPPM mailing list, http://www.advanced.org/IPPM/archive/0246.html, Jan. 1997.

[5] V. Paxson et al., “Framework for IP Performance Metrics,” RFC 2330, ftp://ftp.isi.edu/in-notes/rfc2330.txt, May 1998.

[6] The Surveyor Project Advanced Networks, http://www.advanced.org/surveyor

[7] Abilene Network Operations Center Web site, http://www.abilene.iu.edu

[8] The National Internet Measurement Infrastructure (NIMI) Project, http://www.psc.edu/networking/nimi

[9] The Active Measurement Project (AMP), http://amp.nlanr.net/AMP

[10] The RIPE Test Traffic Project, http://www.ripe.net/test-traffic

[11] Piatetsky-Shapiro, Gregory (1991), Discovery, analysis, and presentation of strong rules, in Piatetsky-Shapiro, Gregory; and Frawley, William J.; eds., Knowledge Discovery in Databases, AAAI/MIT Press, Cambridge, MA.

[12] Brin, Sergey; Motwani, Rajeev; Ullman, Jeffrey D.; and Tsur, Shalom; Dynamic itemset counting and implication rules for market basket data, in SIGMOD 1997, Proceedings of the ACM SIGMOD International Conference on Management of Data (SIGMOD 1997), Tucson, Arizona, USA, May 1997, pp. 265-276

[13] Webb, Geoffrey I. (2000); Efficient Search for Association Rules, in Ramakrishnan, Raghu; and Stolfo, Sal; eds.; Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2000), Boston, MA, New York, NY: pp. 99-107

[14] 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

[15] 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.

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