Network Performance of PingER data with respect to the growth of the Telecom industry in India

Naman Madan
Department of CSE, ASET
Amity University
Noida, UP, India
naman.madan25@gmail.com

Dr. A. Sai Sabitha
HOD, Department of IT, *ASET*
Amity University
Noida, UP, India
assabitha@amity.edu

Prof (Dr.) Abhay Bansal
HOD, Department of CSE, ASET
Amity University
Noida, UP, India
abhansal1@amity.edu

Prof (Dr.) Les Cottrell
SLAC National Accelerator
Laboratory
Stanford, CA, USA
cottrelll@slac.stanford.edu

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

Abstract: This paper aims towards analyzing the changes in the Internet Performance in India over a span of 3 years by using the clustering algorithm on PingER data. The paper analyzed the enhanced performance of the internet after major changes in the Telecom industry in India in the year 2016 and some insights into the level of improvement. PingER which is short for Ping End-to-end Reporting is a project started with an aim to monitor end-to-end performance of between Internet hosts by SLAC National Accelerator Laboratory, Stanford, California. And over the last decade or so they have collected a huge amount of data which is stored as space separated flat files along with sophisticated methods in order to enable searching for data fast.

Keywords - data analysis, PingER, Network Monitoring, clustering, Internet Performance

I. Introduction

PingER project, which was originally started in 1995 and at that time it was for the High Energy Physics community, but presently it's more focused towards Digital Divide measuring through internet performing [2]. As for now, this project has over 40 MAs (Measurement Agents) in 14 countries for measurements which are measuring around 700 sites in 160 countries. [9]

The data collected by pinging nodes are freely available and can be analyzed using different algorithms in order to find interesting patterns. In the current analysis, we are using a clustering algorithm in order to analyze the internet performance of India over the years. For this, we have collected three months data for each year that is 1st August to 31st October for years 2016, 2017 and 2018 (for 2018 till 3rd October). We particularly started collecting data from August 2016 because there was a major shift in the telecom sector after the introduction of Reliance Jio [1] which was publicly available on 5th September 2016 which majorly affected the Internet Performance. Through this analysis, we were able to study its effect in depth using interpretation and clustering analysis.

The sections with are further discussed in this paper are Literature survey, Methodology, Experimental Setup, and Output and Analysis.

II. LITERATURE SURVEY

A. History of PingER

Started back in 1995 by SLAC, which basically monitored links to a number of sites. Under the PingER project, there are many MAs short for Measurement Agents which ping different target sites periodically and the data collected by pinging these sites is stored by these agents.

Now, all these data collected by monitoring sites are used in different ways one of them is to study the progress of a country using its internet performance one such example of this is SLAC paper written in 2013 on quality of internet performance in Africa using this data [4].
There are several different parts of PingER like PingER operation, validation, analysis, deployment, Databases, deployment, data, and toolbox.

B. Data Mining Techniques

Data Mining Techniques refers to the process of discovering meaningful patterns from a large set of data also known as big data. Some of the Majorly used techniques in data science are classification, association, and clustering. We have used clustering to divide data into different groups for this analysis.

C. Clustering

Clustering is a statistical technique of dividing the large dataset into a smaller number of groups such that each data point in a group has some common properties or some common attributes. Clustering is subjective implies that there can multiple means to achieve the goal some of the most popular algorithms are Clustering based on Centroid, Distribution, Connectivity, and Density [7]. These are different sets of rules on the basis of which different data points are separated into groups and the two most popular clustering algorithms are K Means and Hierarchical Clustering [6].
For this analysis, we have used the K-means algorithm based on centroid. It is a repetitive or iterative algorithm and in each iteration, it locates local maxima. We first start by determining k which denotes groups using Davies–Bouldin index, which is an evaluation metric to examine how well are groups divided [2]. The algorithms can be defined in the following steps [8]:

1) Partitioning of the data points in k clusters.
2) Identification of the mean point of a group which is denoted as the seed point.

3) Assigning each data point the closet mean point.

4) Repeating Step 2 until there is no data point left for assignment.

III. METHODOLOGY

Research strategy applied to the particular field of study is as follows:

1) Collecting Data: The data are collected from the Stanford SLAC website [3] and this data is freely available. For this particular analysis, we have considered data collected by EDU.SLAC.STANFORD.PINGER by pinging nodes in India over the period of three months for three years, which are 1st August to 31st October for years 2016, 2017 and 2018 (for 2018 till 3rd October). This data is stored in the ”.Tsv” format. The dataset is further explained in the Experimental Setup section.

2) Data Cleansing: There were many missing values in these datasets and for our analysis columns containing missing values were not considered.

3) Clustering and Analysis: The K means algorithm is applied using Python along with scikit-learn library which is free machine learning software for clustering the data and then the results were analyzed.

IV. EXPERIMENTAL SETUP

This section discusses Metadata about Data, Datasets, Tools, and Scripts Used, and Identification of 'K' using Davies–Bouldin index.

A. Metadata about Data

The data is collected from Stanford SLAC website [3] considering 'pinger.slac.stanford.edu' source hostname. The attributes considered by SLAC for data are source_host, source_host_address, size, destination_host_name,destination_host_address, unix_epoch_time, sent, rcvd, min, max, avg, seq_rev, and Rtt_rcv. For this particular analysis, we have considered
destination_host_name,destination_host_address,

unix_epoch_time, min, max, avg because for analyzing Internet Performance we have used Round-trip delay time (RTT) which measure network latency. Here min, avg, max represents min RTT, avg RTT, and max RTT. Moreover, source host and address remain same so no need to include them.

Of total 71503 tuples, 11668 tuples were collected in 2016, 35705 tuples were collected in 2017 and 24131 were collected in 2018.

There were many missing values and the tuples containing missing values were removed and after removing these tuples 41450 values left. The nodes which are set up in India were considered as the destination host which are "pingeramity.in", "www.mitpune.in", "speedtest.hns.net.in", "mail.prl.res.in" having IP addresses 202.12.103.71, 203.199.134.21, 111.91.122.166, 210.212.155.234.

Of total 41450 tuples, 8679 tuples were collected in 2016, 18628 tuples were collected in 2017, and 14143 tuples were collected in 2018.

B. Datasets

The data collected from the website was divided into two datasets for analysis. The first dataset contains three attributes min, avg, and max and the second dataset contains four attributes unix_epoch_time, min, avg, max. Figure 4.1 represents two datasets where the left shot is dataset 1 and the right shot is dataset 2. Two case studies discussed in Output and Analysis section are based on these two datasets.

C. Tools and Scripts Used

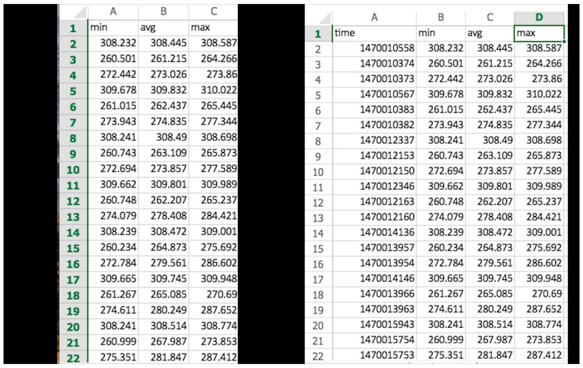
For clustering and analysis of data, python scripts were used. For analyzing the data pandas(0.23.4) library was used and for clustering scikit-learn(1.1.0). Figure 4.2 is a snippet of python script used for clustering.

D. Identification of 'K' using Davies–Bouldin index

Davies–Bouldin index was considered to find the optimum value of 'k'. It determines how well clustering is done. We first calculated different values of 'k' and then determine Davies–Bouldin index for each value of 'k' and then the value of 'k' for which Davies–Bouldin index is closest to 0 is considered optimum.

Figure 4.3 represents values of 'k' for dataset 1 (with three attributes) and the minimum value of DB index is at k=3.

Figure 4.4 represents values of 'k' for dataset 2 (with four attributes) and the minimum value of DB index is at k=3.

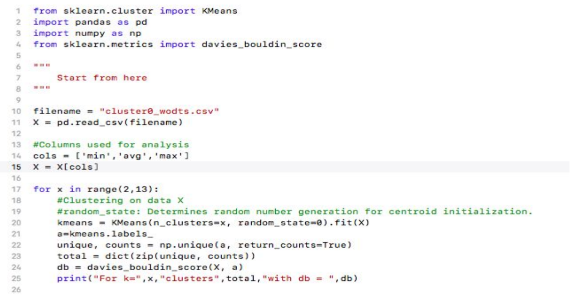


	A	B	C
1	min	avg	max
2	308.232	308.445	308.587
3	260.501	261.215	264.266
4	272.442	273.026	273.86
5	309.678	309.832	310.022
6	261.015	262.437	265.445
7	273.943	274.835	277.344
8	308.241	308.49	308.698
9	260.743	263.109	265.873
10	272.694	273.857	277.589
11	309.662	309.801	309.989
12	260.748	262.207	265.237
13	274.079	278.408	284.421
14	308.239	308.472	309.001
15	260.234	264.873	275.692
16	272.784	279.561	286.602
17	309.665	309.745	309.948
18	261.267	265.085	270.69
19	274.611	280.249	287.652
20	308.241	308.514	308.774
21	260.999	267.987	273.853
22	275.351	281.847	287.412

	A	B	C	D
1	time	min	avg	max
2	1470010558	308.232	308.445	308.587
3	1470010374	260.501	261.215	264.266
4	1470010373	272.442	273.026	273.86
5	1470010567	309.678	309.832	310.022
6	1470010383	261.015	262.437	265.445
7	1470010382	273.943	274.835	277.344
8	1470012337	308.241	308.49	308.698
9	1470012153	260.743	263.109	265.873
10	1470012150	272.694	273.857	277.589
11	1470012346	309.662	309.801	309.989
12	1470012163	260.748	262.207	265.237
13	1470012160	274.079	278.408	284.421
14	1470014136	308.239	308.472	309.001
15	1470013957	260.234	264.873	275.692
16	1470013954	272.784	279.561	286.602
17	1470014146	309.665	309.745	309.948
18	1470013966	261.267	265.085	270.69
19	1470013963	274.611	280.249	287.652
20	1470015943	308.241	308.514	308.774
21	1470015754	260.999	267.987	273.853
22	1470015753	275.351	281.847	287.412

Fig 4.1: Datasets used.

As the datasets were quite big so these clusters were further subdivided into clusters and the same steps were taken to find the value of 'k' for them.



```python
from sklearn.cluster import KMeans
import pandas as pd
import numpy as np
from sklearn.metrics import davies_bouldin_score

"""
    Start from here
"""

filename = "cluster0_wodts.csv"
X = pd.read_csv(filename)

#Columns used for analysis
cols = ['min','avg','max']
X = X[cols]

for x in range(2,13):
    #Clustering on data X
    #random_state: Determines random number generation for centroid initialization.
    kmeans = KMeans(n_clusters=x, random_state=0).fit(X)
    a=kmeans.labels_
    unique, counts = np.unique(a, return_counts=True)
    total = dict(zip(unique, counts))
    db = davies_bouldin_score(X, a)
    print("For k=",x,"clusters",total,"with db = ",db)

```

Fig 4.2: Python Script.

The optimum value of 'k' for cluster_0 of dataset 1 is 3 and similarly, for cluster_1 and cluster_2 it is also 3.

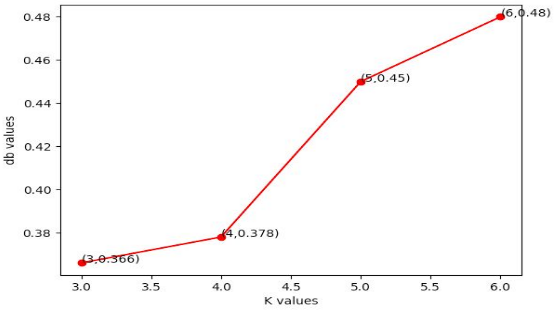


Fig. 4.3: Values of 'k' for dataset 1.

The optimum value of 'k' for cluster_0 of dataset 2 is 6 and similarly, for cluster_2 it is 3.

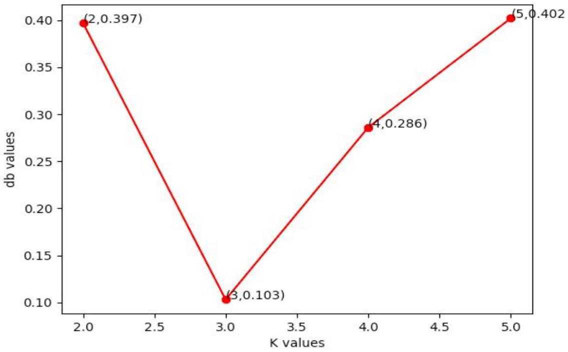


Fig. 4.4: Values of 'k' for dataset 2

V. OUTPUT AND ANALYSIS

Two datasets shown in Fig 4.1 were divided into three clusters each by applying the k-means clustering algorithm. Since the datasets were too large so we further divided these clusters into 'k' clusters by following the same steps as followed previously. And the output is shown in Appendix 1 and 2.

Dataset 2 was divided into three clusters on the basis of year. From the output of clustering of dataset 2, the following observations were made - node named 'speedtest.hns.net.in' was not active in the year 2016 and 2107 but it was up and running in 2018. whereas 'mail.prl.res.in' was working in 2016 and in 2017 it contributed for almost around 50% but in 2018 it was not active for the months analyzed. Nodes 'www.mitpune.com' and 'pingeramity.in' showed consistent performance but 'www.mitpune.com' performance was better than all.

Using the output from dataset 1 we were able to compare the network performance over the years. Cluster1 is an outlier as it contains one objects having a very High Response time which account for poor connection at that time. Cluster2 contains objects with less response time than cluster0. Further analyzing cluster2 the cluster2_0 has the least response time of all and 15.3% of pings are of the year 2016 and cluster2_1 having bit higher response time contains 17.158% pings of the year 2016 and cluster2_2 having the maximum response time of all contains 43.24% pings of the year 2016 whereas cluster2_0 contains 52.429% and 32.26% pings of the year 2017 and 2018. 39.48% and 43.35% pings of the year 2017 and 2018 are in cluster2_1. And cluster2_2 only contains 7.7% pings from 2018 hence over the year the response time of host decreased and become consistent which can be accounted for enhancement of network performance. In September 2016 Reliance Jio was launched in India which is an Indian network operator and this resulted in the competition in Telecom industry to provide a high-speed network connection and many of operators providing a low-speed network connection were shut like Tata Docomo. This is one of the reasons for better network connection in India.

VI. CONCLUSION AND FUTURE WORK

There was a drastic performance improvement of the network in India after major changes in Telecom industry in India in the year 2016 as shown in data above. The data which is available through the PingER project can further be used for let's say to compare the performance from (let's say) India and Pakistan to the rest of the world. It can also be used to compare ping performance with other metrics such as Digital Opportunity Index, the Human Development Index etc and it can also be used to study the effect of natural disaster on internet performance.

Clusters	**Sub-Clusters**	**Objects**	**Composition**	**Avg RTT**
Cluster0	Cluster0_0	3530	www.mitpune.com - 88.5% mail.prl.res.in - 8.866% speedtest.hns.net.in - 2.436% pingeramity.in - 88.5%	312.28
	Cluster0_1	1	pingeramity.in - 100%	759.8
	Cluster0_2	75	www.mitpune.com - 2.6% mail.prl.res.in - 25.33% pingeramity.in - 72%	404.24
	Cluster0_3	3	mail.prl.res.in - 33.33% pingeramity.in - 66.66%	602.8
	Cluster0_4	1	pingeramity.in - 100%	2663.8
	Cluster0_5	10	mail.prl.res.in - 100%	741.98
Cluster1	-	1	pingeramity.in - 100%	14318.75
Cluster2	Cluster2_0	18057	www.mitpune.com - 86.9% mail.prl.res.in - 12.6599% speedtest.hns.net.in - 0.199% pingeramity.in - 0.238%	257.49
	Cluster2_1	13142	www.mitpune.com - 16.154% mail.prl.res.in - 40.7% speedtest.hns.net.in - 41.55% pingeramity.in - 1.575%	270.117
	Cluster2_2	6630	www.mitpune.com - 0.7% mail.prl.res.in - 57.6% speedtest.hns.net.in - 6.77% pingeramity.in - 34.9%	280.77

Appendix 1: Clustering of dataset 1

Clusters	**Sub-Clusters**	**Objects**	**Composition**	**Duration**
Cluster0	Cluster0_0	2897	www.mitpune.com - 33.4% mail.prl.res.in - 33.39% speedtest.hns.net.in - 0% pingeramity.in - 33.18%	21\08\16 - 31\08\16
	Cluster0_1	2878	www.mitpune.com - 33.759% mail.prl.res.in - 33.4138% speedtest.hns.net.in - 0% pingeramity.in - 32.827%	1\08\16 - 11\08\16

	Cluster0_2	2904	www.mitpune.com - 33.5399% mail.prl.res.in - 33.5399% speedtest.hns.net.in - 0% pingeramity.in - 32.92%	11\08\16 - 21\08\16
Cluster1	Cluster1_0	4486	www.mitpune.com - 48.59% mail.prl.res.in - 0% speedtest.hns.net.in - 48.59% pingeramity.in - 2.8%	21/08/18 - 13/09/18
	Cluster1_1	4690	www.mitpune.com - 40.04% mail.prl.res.in - 0% speedtest.hns.net.in - 40.085% pingeramity.in - 19.872%	13/09/18 - 03/10/18
	Cluster1_2	4967	www.mitpune.com - 39.7% mail.prl.res.in - 0% speedtest.hns.net.in - 39.7% pingeramity.in - 20.5959	01/08/18 - 21/08/18
Cluster2	Cluster2_0	6320	www.mitpune.com - 47.15 mail.prl.res.in - 47.1% speedtest.hns.net.in - 0% pingeramity.in - 5.74%	02/10/17 - 02/11/17
	Cluster2_1	6063	www.mitpune.com - 48.5898% mail.prl.res.in - 48.5% speedtest.hns.net.in - 0% pingeramity.in - 2.9%	01/08/17 - 31/08/17
	Cluster2_2	6245	www.mitpune.com - 48.038% mail.prl.res.in - 47.7% speedtest.hns.net.in - 0% pingeramity.in - 4.259%	31/08/17 - 02/10/17

Appendix 2: Clustering of dataset 2

References

[1] Haq, N. (2017). Impact of Reliance JIO on Indian Telecom Industry. International Journal of Engineering and Management Research, 7 (3),259-263

[2] http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.102.4114&rep=rep1&type=pdf

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

[4] http://www.slac.stanford.edu/cgi-wrap/getdoc/slac-pub-15333.pdf

[5] http://www.slac.stanford.edu/pubs/slacpubs/10000/slac-pub-10186.pdf

[6] https://dl.acm.org/citation.cfm?id=46712

[7] https://www.cc.gatech.edu/~isbell/reading/papers/berkhin02survey.pdf

[8] https://www.jstor.org/stable/2346830?seq=1#page_scan_tab_contents

[9] www.slac.stanford.edu/xorg/icfa/icfa-net-paper-jan10/report-jan10.doc