Network congestion analysis and traffic prediction using time series analysis

Ashutosh Shankhdhar
Assistant professor, Department of CEA
GLA University
Mathura, India
ashutosh.shankhdhar@gla.ac.in

Rohit Gupta
Department of CEA
GLA University
Mathura, India
rohit.gupta_cs18@gla.ac.in

***Abstract*—This Research paper will be looking at the various causes of network congestions with special emphasis to India as well as possible solutions approaches and work done in order to get better. This paper will also discuss about the causes of congestion and problems arises due to it and problems with some methods used to control congestion. To reduce congestion in the telecommunication network and to stop potential threats on internet we will be doing time series analysis on the network traffic data on the past dates and predict the traffic for future according to the future traffic network providers can increase the bandwidth required to reduce or remove the congestion in network, use Arima Model to get the approximate traffic on the network in the future and according to that predict the cost of bandwidth required.**

1. INTRODUCTION

The level of patronage being experienced in wired and wireless network communications in India is overwhelming. As reliance jio infocomm limited. was introduced in year 2016 it drastically reduced the charges of mobile network and other broadband connections too .As the result of reduced data rates the number of subscribers continues to increase at a very high rate', Average data consumption Was increased by around 4 times while voice consumption surged to 820 minutes per user from 400 minutes last year. Voice calls are becoming data calls as now LTE allows is and makes it cheaper. But industries only managed to install around 50,000 cell sites in 2018 compared with the need for over 100,000 cell sites to keep pace, there is a consequential increase in network congestion with this huge rise in usage. Mobile phone companies and network equipment vendors admit to the call mismatch especially in urban areas in demand and supply, but blame it on their inability to install telecom towers as needed. Some problems due to congestion are→ a) Packet Loss, b) High ping, c) performance degradation.

This high rise of internet usage also leaves use more vulnerable towards attack. An attack is defined as any actions that compromise the availability, confidentiality and integrity of a system. Some of the major attacks are a) DOS (Denial of Service), b) Probe, c) R2L (Remote to Local), d) U2R (User to Root).Aftermath of this is the reduction in the quality of services delivered to customers. Network Congestion is one of the primary cause of:

i) Performance degradation

ii) Performance variability

iii) Poor scaling in heavy parallel applications

iv) Low Bandwidth

We need new approaches to analyze, model and predict this critical behavior in order to improve the performance of large-scale parallel applications.

Anomaly in the network is one of the primary causes of:

i) Network Security break

ii) Packet drop

2. EXISTING METHODS

Some well-known techniques to reduce congestion in network are congestion control and congestion avoidance. These techniques include exponential backoff in protocols such as CSMA/CA in 802.11 and the similar CSMA/CD in the original Ethernet, "window reduction in TCP, and fair queueing in networking devices such as routers and network switches. Other techniques that address congestion include priority schemes which transmit some packets with higher priority ahead of others and the explicit allocation of network resources to specific flows through the use of admission control. There have been some problems with these techniques [6] Congestion Collapse, occurs when the network is increasingly busy, but little useful work is getting done. Problem: Classical congestion collapse: Paths clogged with unnecessarily-retransmitted packets .Fix: Modern TCP retransmit timer and congestion control algorithms.[7] Fragmentation-based congestion collapse Problem: Paths clogged with fragments of packets invalidated because another fragment (or cell) has been discarded along the path. [Kent and Mogul, 1987] Fix: MTU discovery [Mogul and Deering, 1990] Early Packet Discard in ATM networks [Romanow and Floyd, 1995].

3. RELATED WORK

[1] The accuracy of bandwidth prediction directly determine the quality of service in wireless network. When the available bandwidth estimation of the wireless network is high, more data will be injected to the network, which will reduce the quality of the network service and cause congestion. When available bandwidth estimation is low, the utilization of the channel will be reduced, and the throughput of the system will be reduced because the channel capacity of the wireless network is not effectively utilized. Obviously, if congestion control is implemented on the basis of inaccurate available bandwidth estimation, it will inevitably lead to congestion oscillation or a waste of resources. To increase the performance in a wireless network, researchers have presented many strategies of

congestion control protocols. These strategies can be categorized according to the technique they employ. In [3] the aim is to develop the uncontrolled model into a controllable model. Which reacts to the local build-up of queues with an objective of delaying onset of congestion at the network level when the sources are strongly long range dependent. In [4] before transfer of data between a source and destination, the source have to check the available queue size of the destination based on threshold. To understand the network congestion, RTT (Round-Trip Time) is calculated. Lesser the value of RTT, lesser is the congestion and vice versa.

[2] Congestion control is technique and mechanism that can either prevent congestion before it happens, or remove congestion, after it has happened. There is a congestion control mechanisms into two broad categories: congestion avoidance (open-loop congestion control) and congestion recovery (closed-loop congestion control). The strategy of congestion avoidance is preventive in nature" it is aimed to keep the operation of a network at or near the point of maximum power, so that congestion will never occur.

According to [5] we can predict bandwidth required if we have the details of data rate use on different days by using same model.

4. PROPOSED METHOD

By using the ARIMA model we will predict the data traffic for future and based of the traffic, service providers can increase the bandwidth to reduce congestion, it will be an easier way to reduce congestion. This method looks costly but it will ensure the least congestion on the network as providers will have an approximate value of traffic in future and existing methods wouldn't work for the traffic then increasing the bandwidth will be the best option, and according to the traffic cost of the bandwidth will be decided.

To predict the cost of bandwidth we can use

Cost=bandwidth required/traffic * cost of bandwidth

Predicting the exact cost of the bandwidth required is difficult because bandwidth required for traffic vary for different customer base and providers, the cost of bandwidth is also not fixed everywhere, this method will help network providers to get the amount of traffic in future and increase the bandwidth in advance to avoid congestion

A. Dataset

Dataset required for the prediction contains traffic on a network on different dates, it's unit is number of person on a day. It can also be gathered from different network providers to get the traffic on their network

B. Traffic analysis

To analyze and predict the data Time Series algorithm that is Arima Model Will be used ,by the help of this model we predict the traffic in the network and after finding the traffic we can predict the cost based on the required bandwidth which help as to analyse the congestion and the factors of traffic in the network. Time series data mostly arise when monitoring industrial process or tracking corporate business metrics.

Time series analysis accounts the fact that data points that are taken over time may have an internal structure (such as autocorrelation, trend or seasonal variation) that should be accounted for.

Before starting analysis on the data we have to check the data for non-stationary trends. If the data contains non stationary trends then the model may not work properly. To make the time series model work properly and give more accurate result we need stationary data.

C. Checking stationarity and transforming data

To check the stationarity of the data first we plot the graph of rolling mean and rolling standard of the data to check the stationarity with visual test. But this might not give the give the accurate result. To cross check the result we used ADF (Augmented Dickey fuller) test, it is one of the most useful and easy to perform statistical tests. Useful work of this test for us is that It can be used to determine the presence of unit root in the series, and hence it shows us if the series is stationary or not.

After using this test on our data we found that the series data is not stationary .So now we have to transform this data to make it stationary.

To make the series stationary first we used log transform and then difference the rolling mean and test stationarity again with ADF. As expected, the series was still non stationary, now we have to keep transforming the series and check with ADF until we get stationary series. Finally stationary series was found by subtracting log transformed series by itself one value shifted.

Final transformation of series will look like

$$X_n = ln(X_n / X_{n-1})$$

Here the Fig.1 shows the final stationary series Graph. With this data we are ready to make our prediction with Arima model.

Fig.1 Stationary data graph

D. ARIMA Model

'Auto Regressive Integrated Moving Average' also known as ARIMA is actually a class of models that explains a given time series based on its own past values, i.e it shows its own lags and the lagged forecast errors, so that equation can be used to predict future values.

In statistics and econometrics, and in particular in time series analysis, an autoregressive integrated moving average model is a generalization of an autoregressive moving average model. Both of these models are fitted to time series data either to better understand the data or to predict future points in the series.
AR: Autoregression. This is a model that uses the dependent relationship between a value and some number of lagged forecast.
MA model: Moving Average model is a model that uses the dependency between a value and a residual error from a MA model applied to lagged observations. An ARIMA model is one where the time series was differenced at least one time to make it stationary and it is combination of the AR and the MA models.
So the equation of model is

$$\gamma_t = \alpha + \beta_1 \gamma_{t-1} + \beta_2 \gamma_{t-2} + \ldots + \beta_p \gamma_{t-p} \varepsilon_t + \Phi_1 \varepsilon_{t-1} + \Phi_2 \varepsilon_{t-2} + \ldots \Phi_q \varepsilon_{t-q}$$

ARIMA model equation in words:
Predicted γ_t = constant + Linear combination Lags of γ + Linear Combination of Lagged forecast.

5. RESULT

After all the above process we can apply ARIMA model to get result

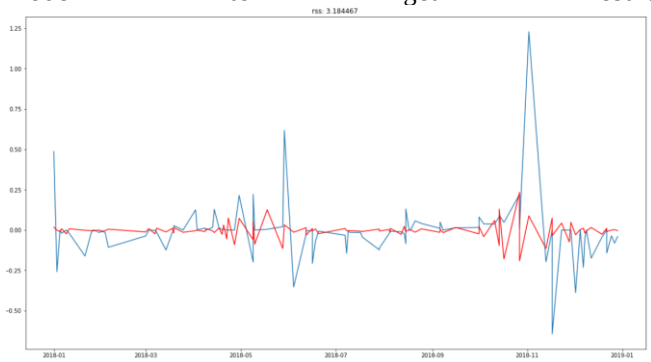

Fig.2 Plot of fitted values in model

Here Fig.2 shows the graph of fitted values in the ARIMA model with red line and blue line shows the dataset. It shows the analysis of the traffic of the values in the past dates.

E. Graphical summarization

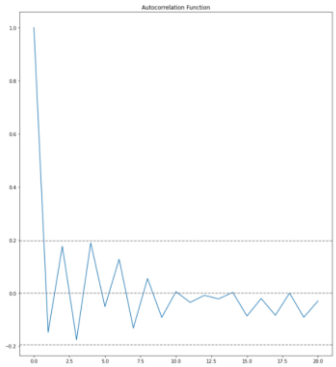

Fig.3a
Autocorrealation funtion

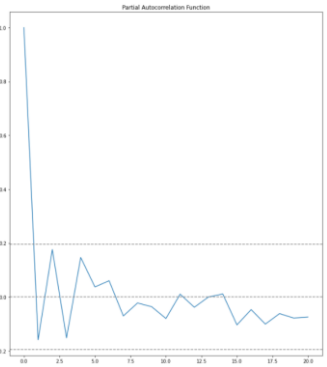

Fig.3b
partial autocorrealation function

Fig.3 graphically summarize the strength of relation with an observations in time series with some lags. These two plots show the Autocorrelation and and partial autocorrelation of the final transformed stationary data with number of lags 20.

F. Predictions

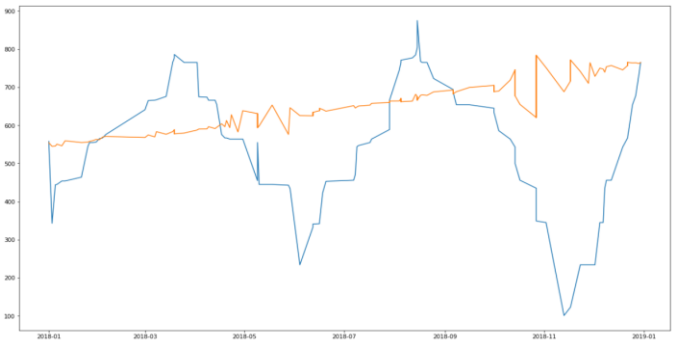

Fig.4 graph of predicted log values and dataset

This fig.4 shows the relation between our original dataset (blue) and the prediction by ARIMA model (yellow). This will help model to predict for the future values of traffic in the network .Surely it shows some difference from original data that is because of all the transformation on the data that won't show much effect on the predictions.

We can't get the values of traffic correctly still now .In order to get the accurate prediction values we have to reverse all the transformations which we have performed in order to get the stationary series on the obtained predicted values to get the original predicted traffic

$$X_n = e^{(X_n + X_{n-1})}$$

Here X is the traffic and n is the day and n

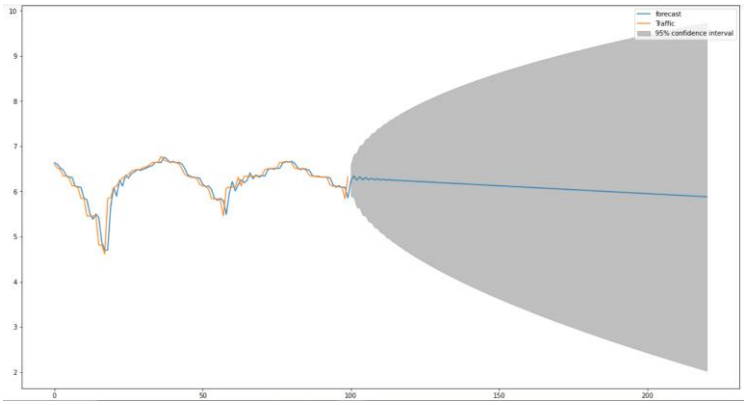

Fig.5 ARIMA traffic prediction Plot

Fig.5 is the plot of the predictions by the model after this we can find the maximum traffic on a network in a time period, So according to that the network provider can increase bandwidth .We are finding maximum traffic in a time period because we don't want to change bandwidth again and again, provider need to change the bandwidth once and most probably traffic won't go more than that value and network providers won't have to increase more bandwidth than that. [5] It will also work almost as same as predicting bandwidth with the benefit that bandwidth requirement changes for

different places and hence predicting traffic is better than predicting bandwidth directly.

6. Conclusion

In this paper we have discussed about the problems arises because of the congestion in network and some causes of the congestion in network. Here we have also discussed about some methods used to control congestion in network and some works done to reduce or control congestion, and at last our proposed method to reduce congestion by predicting traffic in the network so that providers can increase the bandwidth accordingly to avoid congestion.

References

[1] Wang, Hui, Junyong Tang, and Bo Hong. "Research of Wireless Congestion Control Algorithm Based on EKF." *Symmetry* 12.4 (2020): 646.

[2] Kushwaha, Vandana, and Ratneshwer Gupta. "Congestion control for high-speed wired network: A systematic literature review." *Journal of Network and Computer Applications* 45 (2014): 62-78.

[3] Arrowsmith, D. K. A., et al. "Internet packet traffic congestion." *2003 IEEE International Symposium on Circuits and Systems (ISCAS)*. Vol. 3. IEEE, 2003.

[4] Abirami, Sivaprasad, and Kumudhan Cherarajan. "Analysis on Congestion Control Mechanism in Wireless Networks." *Progress in Advanced Computing and Intelligent Engineering*. Springer, Singapore, 2018. 469-480.

[5] Rana, Nandalal, Krishna P. Bhandari, and Surendra Shrestha. "Network Bandwidth Utilization Prediction Based on Observed SNMP Data." *Journal of the Institute of Engineering* 13.1 (2017): 160-168.

[6] Constantinescu, Doru, et al. *Congestion and error control in overlay networks*. 2007.

[7] Katabi, Dina, Mark Handley, and Charlie Rohrs. "Congestion control for high bandwidth-delay product networks." *Proceedings of the 2002 conference on Applications, technologies, architectures, and protocols for computer communications*. 2002.

[8] Allman, Mark, Vern Paxson, and Wright Stevens. "TCP congestion control." (1999).

[9] Sarolahti, P. "Network Working Group S. Floyd Request for Comments: 4782 M. Allman Category: Experimental ICIR A. Jain F5 Networks." (2007).

[10] Kurdoglu, Eymen, et al. "Real-time bandwidth prediction and rate adaptation for video calls over cellular networks." *Proceedings of the 7th International Conference on Multimedia Systems*. 2016.

[11] Oumar, O. A., S. Dyllon, and Perry Xiao. "Educational Network Bandwidth Analysis and Prediction." *the 14th International Conference on Machine Learning and Data Mining (MLDM'2018) July 14-19, 2018*. 2018.

[12] Ramakrishnan, Nipun, and Tarun Soni. "Network traffic prediction using recurrent neural networks." *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*. IEEE, 2018.