Received: Added at production | Revised: Added at production | Accepted: Added at production

DOI: xxx/xxxx

ARTICLE TYPE

Characterization of Probabilistic Structure of Network Traffic During COVID-19: A Study Based on MAWI Data

Anoushka Mittal[1] | Pranav Jain[2] | Karmeshu*[3] | Shachi Sharma[4]

[1]School of Computer Science and Statistics, Trinity College, Dublin, Ireland
[2]Viasat India Private Limited, Chennai, India
[3]School of Computer Science, University of Petroleum and Energy Studies, Uttarakhand, India
[4]Department of Computer Science, South Asian University, New Delhi, India

Correspondence
*Karmeshu, Email: karmeshu@gmail.com

Summary

The COVID-19 pandemic has greatly affected all aspects of human life including working of offices, businesses, industries, educational institutions etc. With more work load shifting online, changes in the network traffic are inevitable. The earlier investigations have generally focused on the qualitative aspects of network traffic data during COVID-19. In contrast, the paper presents a study based on MAWI data characterizing network traffic in terms of multimodal and unimodal probability distributions. It is found that a transition of multimodal Gaussian mixture model of byte and packet counts during normal period to that of unimodal Laplace distribution during COVID-19 period has emerged. Further it is observed that the probability distribution depicts the preponderance of small and large packets during normal period which changes to that of small sized packets during Covid-19 period. These findings are likely to be useful to the administrators to manage network during crisis periods.

KEYWORDS:
Internet traffic, COVID-19, mixture Gaussian distribution, statistical analysis, Laplace distribution, packet sizes

1 | INTRODUCTION

In the beginning of year 2020, the COVID-19 pandemic compelled governments in various countries to impose measures like lockdown leading to emergence of work-from-home culture. Consequently, a significant amount of workload switched to use networks from homes rather than work premises. This paradigm switch of work culture eventually leads to change in network traffic characteristics and its analysis becomes an obvious subject of study. Statistical analysis of network traffic has been an area of interest primarily because of its applications in dimensioning resources such as buffers and bandwidth, designing congestion and admission control algorithms and network management[1]. The results of network traffic analysis are also useful for service providers as they can design plans for customers accordingly. One such notable work conducted in the past highlights major changes in network traffic characteristics in a 14 year and 3 day longitudinal study[2].

A few studies have been conducted to quantify the impact of COVID-19 on network traffic. The analysis of network traffic at different vantage points reveals that the volume increased significantly soon after the governments imposed lockdown. Also, it is observed that the usage of applications such as gaming increased. A notable reduction in outgoing traffic and increase in incoming traffic is observed in a study carried out on university campus network traffic in Italy owing to the deployment of online teaching platform[3]. In another contemporary study on network traffic of European countries, the variation in latency is observed in paths in the network during COVID-19 peak periods. The latency is noted to be 3-4 times higher than in normal scenario.

Also, a rise in network outages is detected during lockdown period in United States communication network in a recent work[4]. There are many such analysis conducted in various counties and a detailed survey of this literature is provided in section 2.

The governments of various countries across the globe have enforced a variety of measures to contain the spread of COVID-19 pandemic. Hence, there are many studies adjudging the impact of the pandemic by focusing on diverse metrics such as latency and traffic volume. However, a detailed study with statistical characterization of network traffic highlighting the qualitative difference between pre-COVID and during COVID is lacking. The paper addresses this research gap by comparing the traffic behavior at three levels – byte counts, packet counts and packet sizes. The statistical analysis also provides insight into emergence of probability distributions characterizing pre and during COVID data. The results of our study become very relevant in the context of handling such situations in the future not only from a community perspective but also from a network service provider's perspective.

The paper is organized into eight sections. A discussion on the related literature on network traffic analysis and study of traffic patterns during Covid-19 pandemic is provided in section 2. Some preliminaries are outlined in section 3. The details of datasets used in the study are presented in section 4. The analysis of byte counts along with results are discussed in section 5. The behavior of packet counts and sizes are described in section 6 and 7 respectively. The last section 8 contains conclusion.

2 | RELATED WORK

The study of network traffic and its modeling have been an active area of research since the advent of telephone networks, mainly because of its importance in designing a quality network. The results of traffic studies are also used to compute performance measures such as loss probability, traffic demand, network capacity planning etc.[5,6]. In modern communication networks such as internet, traffic measurements and analysis also help Internet Service Providers (ISPs) both in network management as well as in developing appropriate usage plans for users[7]. The nature of the traffic carried on internet is undergoing change largely because of the more adoption and upgradtion of technologies like 4G/5G and optical communication[8]. There is considerable literature investigating the internet and other communication network traffic characteristics. One of the main works is by Borgnat *et al.*[9] in which 7 years and one day data was examined and the presence of Long Range Dependence (LRD) paradigm was established even amidst congestion or traffic restriction periods or anomaly occurrences. Recently, Fontugne *et al.*[10] analyze internet traffic data of 14 years and 3 days duration. They validate the existence of multifractals enabling practitioners to study traffic properties at different scales.

When the COVID-19 pandemic hit the world and the majority of workload shifted to home, the researchers also have become inquisitive to investigate its impact on communication network traffic. Feldmann *et al.*[11] examine the traffic data from three different sources – one ISP, three Internet Exchange Points (IXPs), and one metropolitan educational network. The IXPs are placed at the boundaries of networks and act as connectors by joining two networks, thus, allowing ISPs to send traffic out of their networks. As reported in[11], ISPs observe a 15% rise in traffic volume within one week of lockdown. It is further noted that the traffic has shifted to traditional non-peak hours and applications like gaming, video conferencing and VPN contribute more towards network traffic[11,12].

Favale *et al.*[3] investigate both incoming and outgoing network traffic at Politecnico di Torino campus network in Italy due to adoption of online teaching platform. They find an abrupt increase in outgoing traffic whereas reduction in incoming traffic. Candela *et al.*[13] investigate the impact of COVID-19 by analyzing network latency using data from network of Italy and other European countries. They note significant variation (approximately 3 to 4 times more) in network latency during COVID-19 lockdown. Simultaneously, 2 to 3 times more packet loss has been observed. A report on the network traffic from some Asian countries too find surge in traffic volume during lockdown period[14]. A similar report on internet traffic of Middle East and North Africa highlights the importance and dependency on internet during pandemic like COVID-19 and observe an increase in internet traffic volume, change in usage pattern as well as applications and around 30% rise in mobile data traffic[15].

In another study, Affinito *et al.*[16] focus on identification of popular internet applications during the pandemic and their analysis reveals that social networking applications like Facebook, communication applications like WhatsApp and Skype, entertainment applications like Youtube and Netflix have utilized most of the network bandwidth. After conducting an online survey between mid February 2020 to mid May 2020 in Japan, Yabe *et al.*[17] identify an inverse relationship between outings and internet usage. Lutu *et al.*[18] investigate data from a UK mobile operator. They find 140% rise in median voice traffic volume whereas a decrease in downlink data traffic volume and 10% reduction in throughput is observed. A survey of various papers on effect of COVID-19 on the internet traffic has been provided by Silva *et al.*[19]. In a more recent work, Bronzino *et al.*[20] conclude that even though

traffic volume increased in general during lockdown period leading to high utilization, some ISPs experience more latency than others.

It is worth noting that most of existing investigations relate to the study of impact of Covid-19 in terms of qualitative features of internet traffic viz. volume, latency, utilization, and its graphical representation. In order to gain more insight, it would be useful to carry out statistical analysis of the data enabling identification of appropriate probability distributions characterizing pre covid and during covid periods. Such an analysis allows network managers to examine the changing patterns of traffic in different periods.

3 | PRELIMINARIES

The focus of our study is to analyze and compare the network traffic patterns by computing basic statistical measures like mean, median, variance, coefficient of variation (CV), kurtosis, skewness. We also compute measure of uncertainty viz. entropy.

3.1 | Entropy

The entropy of a random variable X is defined in terms of amount of average uncertainty being contained in it as[21]

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log p(x_i), \qquad 0 \log 0 = 0, \tag{1}$$

where $p(x_i)$ denotes the probability that X takes value x_i from n possible alternatives.

We next present some probability distributions utilized in the study.

3.2 | Probability Distributions

In this subsection, the Laplace and mixture Gaussian distributions are discussed to characterize different periods of study.

3.2.1 | Laplace Distribution

The Laplace distribution is a continuous probability distribution with probability density function (PDF)[22]

$$f(x; \theta, s) = \frac{1}{2s} e^{\frac{-|x-\theta|}{s}}, \quad -\infty < x < \infty, \tag{2}$$

where $\theta \in (-\infty, \infty)$ and $s > 0$ are location and scale parameters respectively. The cumulative distribution function (CDF) corresponding to (2) is

$$F(x; \theta, s) \begin{cases} \frac{1}{2} e^{\frac{-|x-\theta|}{s}} & \text{if} x < \theta, \\ 1 - \frac{1}{2} e^{\frac{-|x-\theta|}{s}} & \text{if} x \geq \theta. \end{cases} \tag{3}$$

The Laplace distribution is symmetric around θ, i.e. for any real x,

$$f(\theta - x; \theta, \sigma) = f(\theta + x; \theta, \sigma) \tag{4}$$

and

$$F(\theta - x; \theta, \sigma) = 1 - F(\theta + x; \theta, \sigma). \tag{5}$$

The mean, median, and mode of the Laplace distribution are all equal to θ. The Laplace distribution can be derived from the difference of two exponential random variables. Hence, it is also known as double exponential distribution or two-tailed exponential distribution or bilateral exponential law. The key feature of this distribution is in measuring the deviation from a value (usually mean), often called errors. The Laplace distribution has been applied in image compression, image and speech recognition, ocean engineering, hydrology, finance, and can be obtained via scale mixture of normals[22].

3.2.2 | Gaussian Mixture Model

Gaussian Mixture Model (GMM) is a multivariate distribution which is a mixture of component random variables where each random variable again follows multivariate Gaussian distribution. The components are also assigned weights based on their contribution to the density function.

For a GMM with N components such that the k^{th} component has mean μ_k and variance σ_k in the univariate case or mean $\boldsymbol{\mu}_k$ and covariance matrix $\boldsymbol{\Sigma}_k$ in the multivariate case; component weight π_k such that $\sum_{k=1}^{N} \pi_k = 1$, the pdf is[23]

$$p\left(\mathbf{x}|\pi, \boldsymbol{\mu}, \boldsymbol{\Sigma}\right) = \sum_{k=1}^{N} \pi_k \phi\left(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right), \tag{6}$$

where $\phi(.)$ is multivariate normal density

$$\phi\left(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right) = (2\pi)^{\frac{-d}{2}} \left|\boldsymbol{\Sigma}_k\right|^{\frac{-1}{2}} \exp\left\{\frac{-1}{2}\left(\mathbf{x} - \boldsymbol{\mu}_k\right)^T {\boldsymbol{\Sigma}_k}^{-1} \left(\mathbf{x} - \boldsymbol{\mu}_k\right)\right\}, \tag{7}$$

where d is the number of random variables. For univariate GMM $d = 1$, simplifying (6) and (7) into

$$p(x) = \sum_{k=1}^{N} \pi_k \phi\left(x|\mu_k, \sigma_k\right), \tag{8}$$

and

$$\phi\left(x|\mu_k, \sigma_k\right) = \frac{1}{\sigma_k\sqrt{2\pi}} \exp\left\{\frac{-\left(x - \mu_k\right)^2}{2{\sigma_k}^2}\right\} \tag{9}$$

respectively. GMMs are widely used because of their flexibility to handle both univariate and multivariate data. Also, GMMs are used for segregating different normally distributed sub-populations within a given population. Alternatively, the various homogeneous parts of the heterogeneous data can be modelled well by mixture Gaussian distribution. This is also the prime reason, we have included GMM in this study. The packets data is a heterogeneous data which is a mixture of various applications and protocols. GMM thus helps in identifying these homogeneous components in the data.

3.3 | KL divergence

The Kullback Leibler Divergence, also called cross-entropy or relative entropy is used to quantify closeness of two probability distributions. For two given probability distributions P an Q, the KL divergence[21] is

$$KL\left(P||Q\right) = \sum_{x} P\left(x\right) \log \frac{P(x)}{Q(x)}. \tag{10}$$

Lower the value of KL metric, closer the two distributions are. It may be noted that KL measure becomes zero when the two distributions are identical.

4 | DATASETS

For this study, we have chosen Measurement and Analysis on the WIDE Internet working group traffic archive (MAWI) dataset[24]. The MAWI dataset has been used in several internet analysis studies and has become de facto for internet traffic studies[10]. The WIDE project is a research consortium, established in Japan in 1987, with a focus on the empirical study of the live large-scale internet. WIDE runs its own internet testbed carrying both commodity traffic and research experiments. The testbed is used by network researchers, engineers and students of universities, industries and government. The WIDE testbed is connected to Internet2 via TransPAC so that traffic to/from US/EU/Asia Pacific academic sites traverse on this path. WIDE is also responsible for various internet operations including the M-root name server, NSPIXP (Network Service Provider internet eXchange Point), AI3 (Asian Internet Interconnection Initiatives), and 6Bone in Japan. The MAWI working group has carried out network traffic measurement, analysis, evaluation, and verification from the beginning of the WIDE project.

The MAWI dataset has been used extensively in many studies on internet traffic, to the best of our knowledge in 261 studies so far[25,26,10]. One seminal work is by Fontugne *et al.*[10] where they have analyzed statistically the internet traffic of 14 year 3 day and establish the presence of two time-scales in the traffic. We have used the data from the Agurim site which is a network traffic monitor based on flexible multi-dimensional flow aggregation in order to identify significant aggregate flows in traffic, on the Sample Point-F of the WIDE backbone. The Sample Point -F monitors the transit link of WIDE to the upstream ISP, in operation since 2006. The tool aggregates the traffic based on the total duration, with the lowest resolution being 5 minutes and all the higher resolutions are recursively aggregated from the 5-minute flows. We downloaded one day resolution data for the years of 2014, 2015, 2016, 2017, 2018, 2019, 2020 and a few months of 2021. Apart from the aggregate traffic flow data, we have used the daily trace files collected on Sample Point-F, each sampled for a duration of 15 minutes at 14:00 JST each day.

Table 1 Statistical measures for byte counts from 2014 to 2020.

Year	Mean	Variance	Skewness	Kurtosis	Coefficient of variation	Geometric Mean	Median	Entropy
2014	3.46E+12	6.42E+23	1.349302	8.678358	0.231257	3.38E+12	3.4E+12	2.1737
2015	5.23E+12	3.69E+24	0.495788	2.011367	0.367237	4.89E+12	4.78E+12	2.9653
2016	3.78E+12	1.11E+24	0.167707	3.788149	0.278637	3.56E+12	3.82E+12	2.4699
2017	5.35E+12	1.23E+24	0.161692	2.749022	0.207299	5.23E+12	5.25E+12	2.7067
2018	4.94E+12	1.35E+24	0.085003	2.465907	0.235056	4.8E+12	4.91E+12	2.4838
2019	4.86E+12	1.95E+24	0.010691	2.346308	0.287311	4.64E+12	4.96E+12	3.1598
2020	3.42E+12	9.67E+23	0.529714	3.937503	0.287164	3.28E+12	3.36E+12	2.6872

Table 2 KL divergence for byte counts from 2014 to 2020.

	2014	2015	2016	2017	2018	2019
2015	0.6447					
2016	1.5947	0.3221				
2017	1.1725	0.4262	0.5589			
2018	0.2811	0.3648	1.8271	0.8633		
2019	0.6977	0.3426	0.8790	0.2083	0.3654	
2020	0.5238	0.8020	2.1434	0.6528	0.1947	0.3447

Table 3 p-values obtained by fitting GMMs and Laplace distributions in byte counts.

Year	GMM with one component	GMM with two components	GMM with three components	Laplace distribution
2014	0.7563	0.9472	0.9998	0.1620
2015	0.0230	0.6947	0.9999	6.1548E-5
2016	0.9467	0.9467	0.9877	0.0948
2017	0.4530	0.8159	0.8683	0.0091
2018	0.1631	0.9478	0.9991	0.0011
2019	0.6335	0.9727	0.9998	0.0800
2020	0.3507	0.8693	0.9991	0.5444

We have used 14 traces, 7 traces for the duration 8th to 14th April 2020, the first week of the state of emergency in Japan, 7 traces for the same duration in 2019. The reason for choosing this particular week of April is because a state of emergency was declared in Japan on 8th April 2020 to curb the spread of COVID-19 pandemic.

The results of the study are presented in subsequent sections.

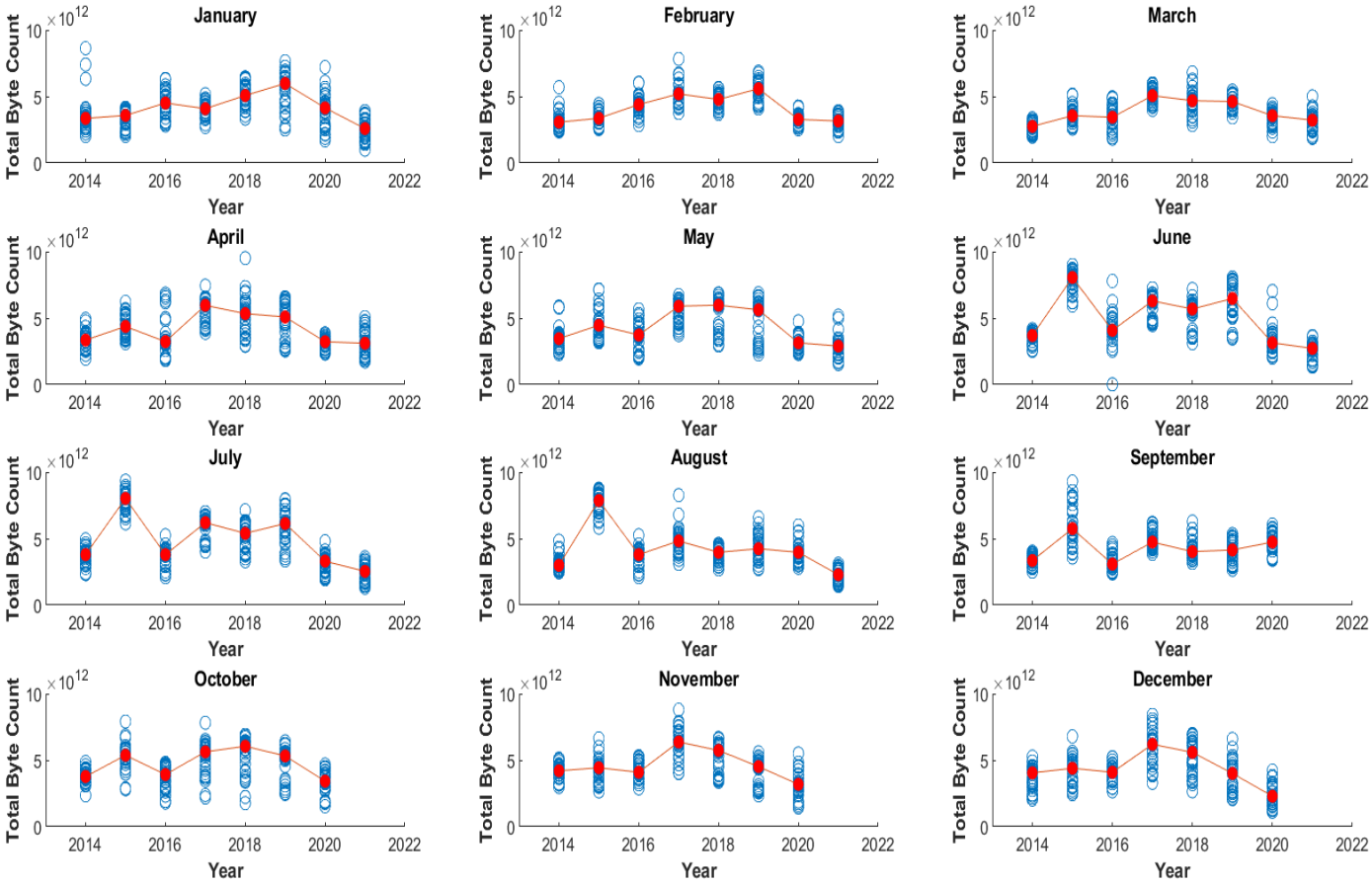
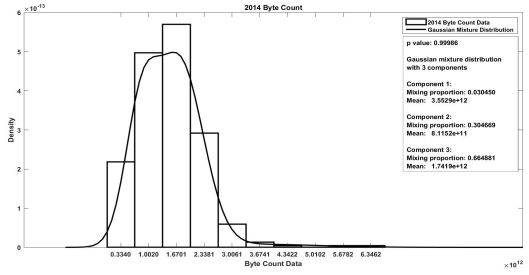


Figure 1 Total byte counts categorized by months from 2014 to 2020.

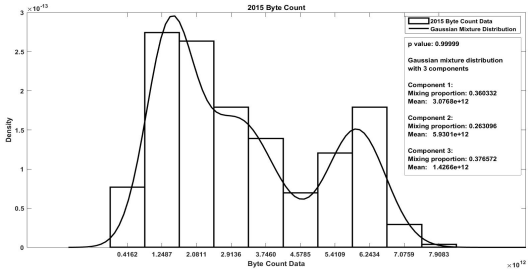
5 | ANALYSIS OF BYTE COUNTS

Both byte counts and packet counts transferred on a network are measures of traffic volume. This section contains the results of the analysis of the byte counts data for the years 2014 to 2020. The basic statistics, that is, mean, variance, geometric mean, median, skewness, kurtosis, coefficient of variation and entropy are computed for all the years and are presented in Table 1. It is easy to note a decrease in mean, variance, geometric mean and median for 2020 compared to 2019 whereas the coefficient of variation reduces only marginally and the skewness as well as kurtosis show a rise. The Table 2 shows the KL-divergence values. The comparison of 2020 with 2019 and 2018 indicates that the traffic in 2020 is more closer to that of 2018 than 2019. This may throw light on evolutionary aspects of traffic.

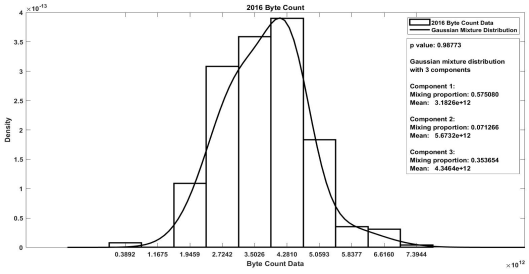
The plot of total bytes counts for every month of years 2014-2020 is shown in Figure 1. The red line represents the median values of the byte counts. It can be noted that for most of the months (except August and September), the median byte value for 2020 is lower than that for 2019. In August and September 2020, the COVID-19 cases started declining after second wave leading to higher byte counts in these months. To understand the change in dynamics of byte counts further, we fit the GMMs with one, two and three components and also the Laplace distribution to the data. The goodness of fit has been tested by KS-statistic[27]. The p-values of the fits are shown in Table 3. The GMM with three components provide the best fit for all the years. The plots of best fits are presented in Figure 2. An interesting finding is that the Laplace distribution only fits to the year 2020 byte data with p value of 0.5444 whereas it fails to fit for all other years. Even though, fitting GMM with 3 components to 2020 byte counts results in p value of 0.9991, Laplace distribution will be favored over GMM with 2 and 3 components using Akaike



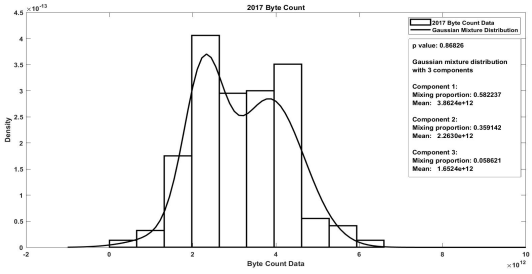
(a) 2014: GMM with 3 components



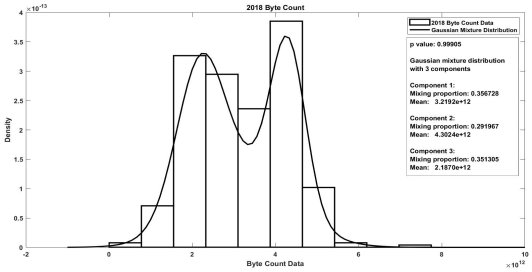
(b) 2015: GMM with 3 components



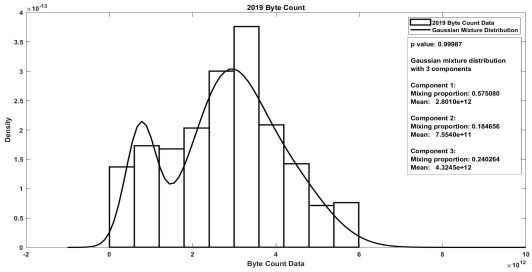
(c) 2016: GMM with 3 components



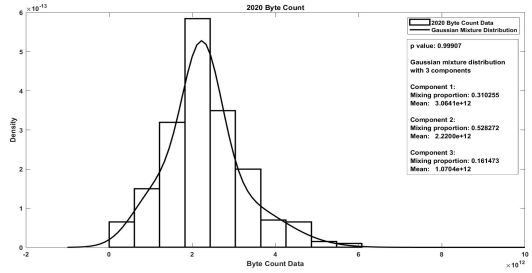
(d) 2017: GMM with 3 components



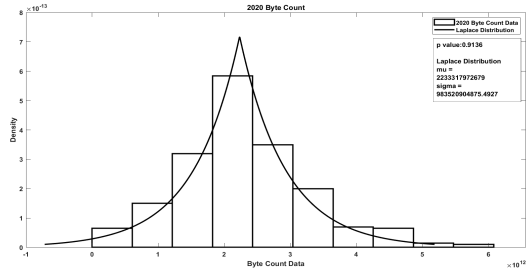
(e) 2018: GMM with 3 components



(f) 2019: GMM with 3 components



(g) 2020: GMM with 3 components



(h) 2020: Laplace distribution

Figure 2 Fitting Gaussian mixture distribution with 3 components to the byte counts from 2014 to 2020 and Laplace distribution to 2020 byte counts.

Information Criteria (AIC)[28] as it penalizes GMMs for increased number of parameters. It is further noted from Figure 2 that the histogram of 2020 is unimodal while for preceding years, it is bimodal. This signifies major shift in bytes count data of year 2020 due to COVID-19 pandemic and underlying reason may be due to the scale mixture of normals.

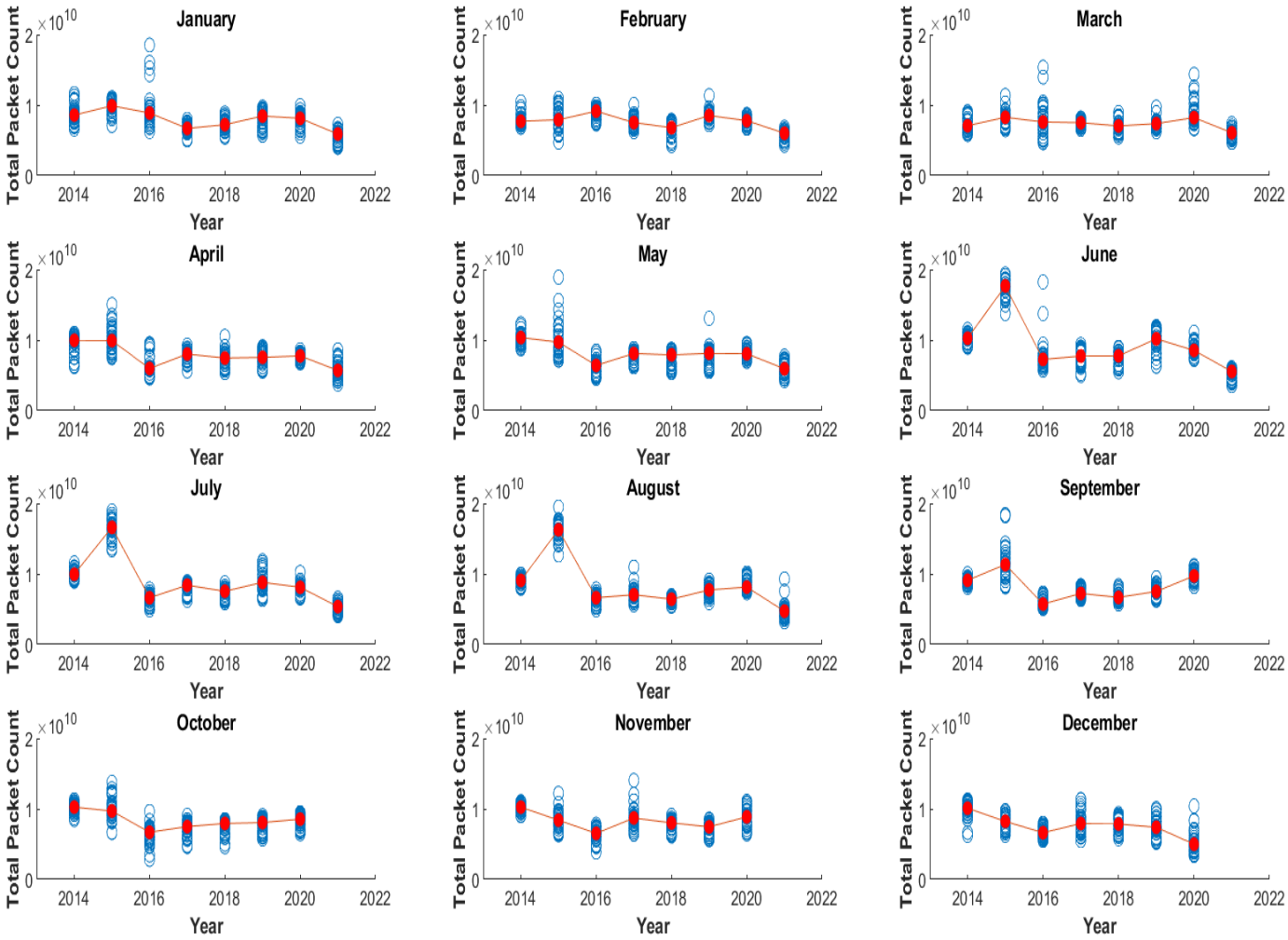


Figure 3 Total packet counts categorised per month from 2014 to 2020.

6 | ANALYSIS OF PACKET COUNTS

This section contains the results of analysis done on packet counts data from 2014 to 2020. The values of basic statistics and KL divergence are shown in Tables 4 and 5 respectively. The mean, geometric mean, median, kurtosis, variance and coefficient of variation of packet counts in 2020 have increased in comparison to past few years where as skewness and entropy have reduced. KL-divergence values show the closeness of packet counts in 2020 with 2017. Again, this shows a pointer to the evolutionary aspect of network traffic. The total packet counts are plotted in Figure 3 month-wise. Again, the red color line represents the median value of packet count for the specified month. The packet counts also fall in year 2020 till August 2020, thereafter the COVID-19 situation improved in Japan and packet counts started increasing. However, in month of December, again the pandemic started spreading and packet counts show a dip.

Like bytes count analysis, we fit the GMMs and Laplace distributions to the data. The p-values obtained from the fits are shown in Table 6. The three component GMM provides best fit for most of years. Surprisingly, the Laplace distribution provides a very good fit for year 2020. This is a significant finding as Laplace distribution fits to byte count of 2020 as well. The plots of

Table 4 Statistical measures for the packet counts from 2014 to 2020.

Year	Mean	Variance	Skewness	Kurtosis	Coefficient of variation	Geometric Mean	Median	Entropy
2014	9.35E+09	1.66E+18	-0.61735	2.825897	0.13765	9.25E+09	9.45E+09	2.9079
2015	1.13E+10	1.51E+19	0.753131	2.417621	0.343416	1.07E+10	9.92E+09	2.9079
2016	7.05E+09	4.36E+18	2.819101	16.62382	0.29647	6.81E+09	6.68E+09	1.8680
2017	7.6E+09	1.26E+18	0.972983	6.693611	0.147854	7.52E+09	7.52E+09	2.2445
2018	7.18E+09	1.02E+18	-0.06651	2.555822	0.140442	7.11E+09	7.17E+09	2.6980
2019	7.94E+09	1.99E+18	0.655011	3.695782	0.177764	7.82E+09	7.96E+09	2.7938
2020	8.14E+09	2.12E+18	-0.27027	4.92724	0.179019	7.99E+09	8.14E+09	2.3732

Table 5 KL divergence for the packets count data from 2014-2020.

	2014	2015	2016	2017	2018	2019
2015	0.8760					
2016	1.9760	0.6527				
2017	1.8744	0.4054	0.1241			
2018	1.3308	0.3967	0.2580	0.1541		
2019	1.7152	0.2683	0.6690	0.2329	0.5299	
2020	1.1678	0.3538	0.2145	0.1296	0.1541	0.2577

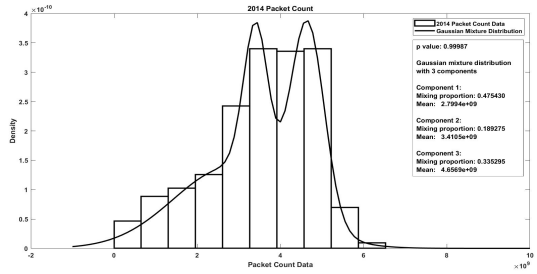
Table 6 p-values obtained by fitting GMMs and Laplace distribution to packet counts.

Year	GMM with one component	GMM with 2 components	GMM with 3 components	Laplace distribution
2014	0.2622	0.7578	0.99987	0.0019
2015	8.8711E-5	0.9990	0.99987	8.7977E-08
2016	0.00081995	0.4007	0.98829	0.0149
2017	0.4530	0.6964	1.0000	0.0442
2018	0.4530	0.9990	0.99999	0.0055
2019	0.5106	0.5711	0.99987	0.3035
2020	0.1934	0.9883	0.99987	0.9986

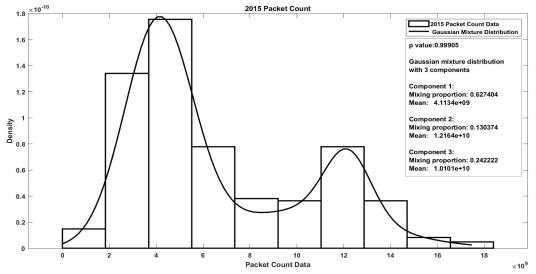
good fits for all the years are shown in Figure 4. The GMM plots again highlight that, for most of the years, the packet counts depict characteristic bimodal behaviour whereas 2020 displays a unimodal behaviour. This interesting finding may be due to the fact that probability mass shifted towards small packet sizes in contrast to previous years. Even though, the bytes and packet counts show some decline in the total volume, the qualitative behavior of the traffic has undergone major shift as characterized by the fact that Laplace distribution explains 2020 traffic well.

7 | ANALYSIS OF PACKET SIZES

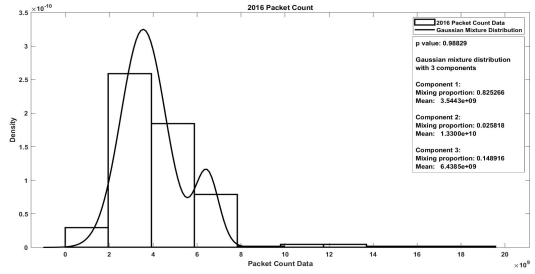
Unlike other studies conducted to understand the impact of COVID-19 on network traffic where the focus is on traffic volume, we include analysis of packet sizes in our study. For this purpose, we use the 15-minute packets trace data from January 2017 to



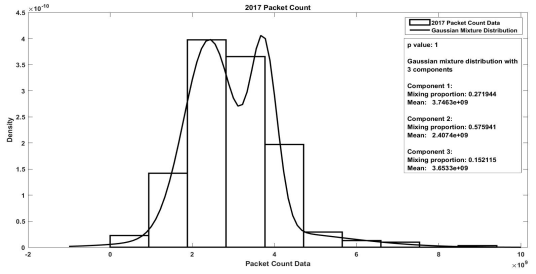
(a) 2014: GMM with 3 components



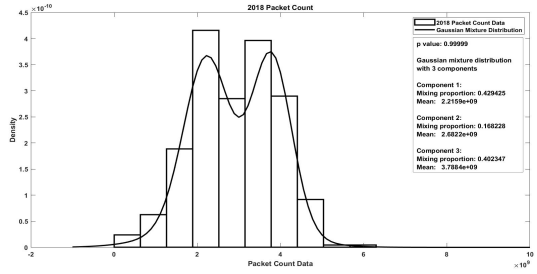
(b) 2015: GMM with 3 components



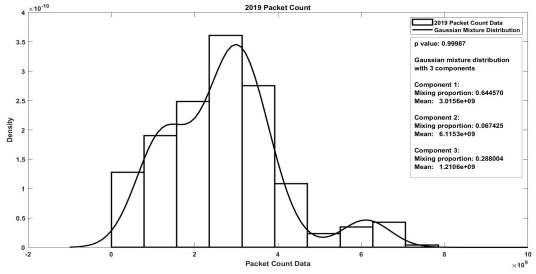
(c) 2016: GMM with 3 components



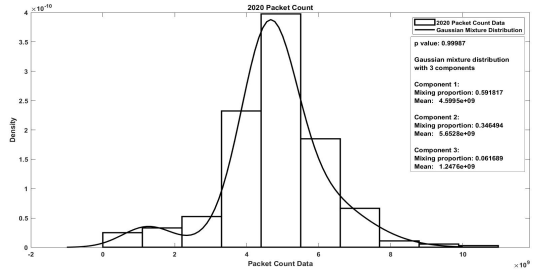
(d) 2017: GMM with 3 components



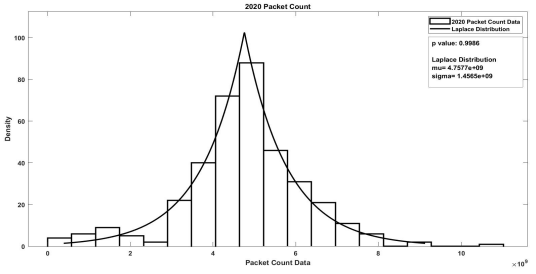
(e) 2018: GMM with 3 components



(f) 2019: GMM with 3 components



(g) 2020: GMM with 3 components



(h) 2020: Laplace Distribution

Figure 4 Fitting Gaussian mixture distribution with 3 components to the packet counts from 2014 to 2020 and Laplace distribution to 2020 packet count.

June 2021 which is collected at daily intervals and present in the MAWI data archive. We have chosen 15-minutes trace because of two reasons: (i) the size of the entire day trace remains huge (ii) the availability of 15 minutes complete traffic trace. It is worth highlighting that 15 minutes traffic trace has been used to study daily traffic characteristics in the past[25,26,10] In the traces, small packets are the packets whose size is less than or equal to 144 bytes. Usually signalling packets are of small size. The large packets are greater than or equal to 1400 bytes in size (usually data frames) and in between are medium size packets.

Table 7 Total count of different packet sizes in daily 15 minutes interval data.

Year	Size	Mean	Ratio	Standard deviation	Coefficient of Variation
	small	47653326.11	0.45	6375602.06	0.13
2017	medium	9029031.20	0.09	5311082.06	0.59
	large	48332347.62	0.46	13540758.62	0.28
	small	43876678.84	0.45	6020184.58	0.14
2018	medium	10868363.67	0.11	2543274.74	0.23
	large	43749220.05	0.44	20247473.19	0.46
	small	52134564.62	0.48	8617465.86	0.17
2019	medium	11384371.74	0.10	7017302.89	0.62
	large	45610198.45	0.42	20721761.15	0.45
	small	**65624192.62**	**0.64**	**14801722.61**	**0.23**
2020	**medium**	**12493349.03**	**0.12**	**7704208.47**	**0.62**
	large	**23917444.52**	**0.23**	**12403570.48**	**0.52**
	small	**43557406.99**	**0.53**	**11386657.39**	**0.26**
2021	**medium**	**17411608.38**	**0.21**	**8418423.96**	**0.48**
	large	**20563082.01**	**0.25**	**8950038.92**	**0.44**

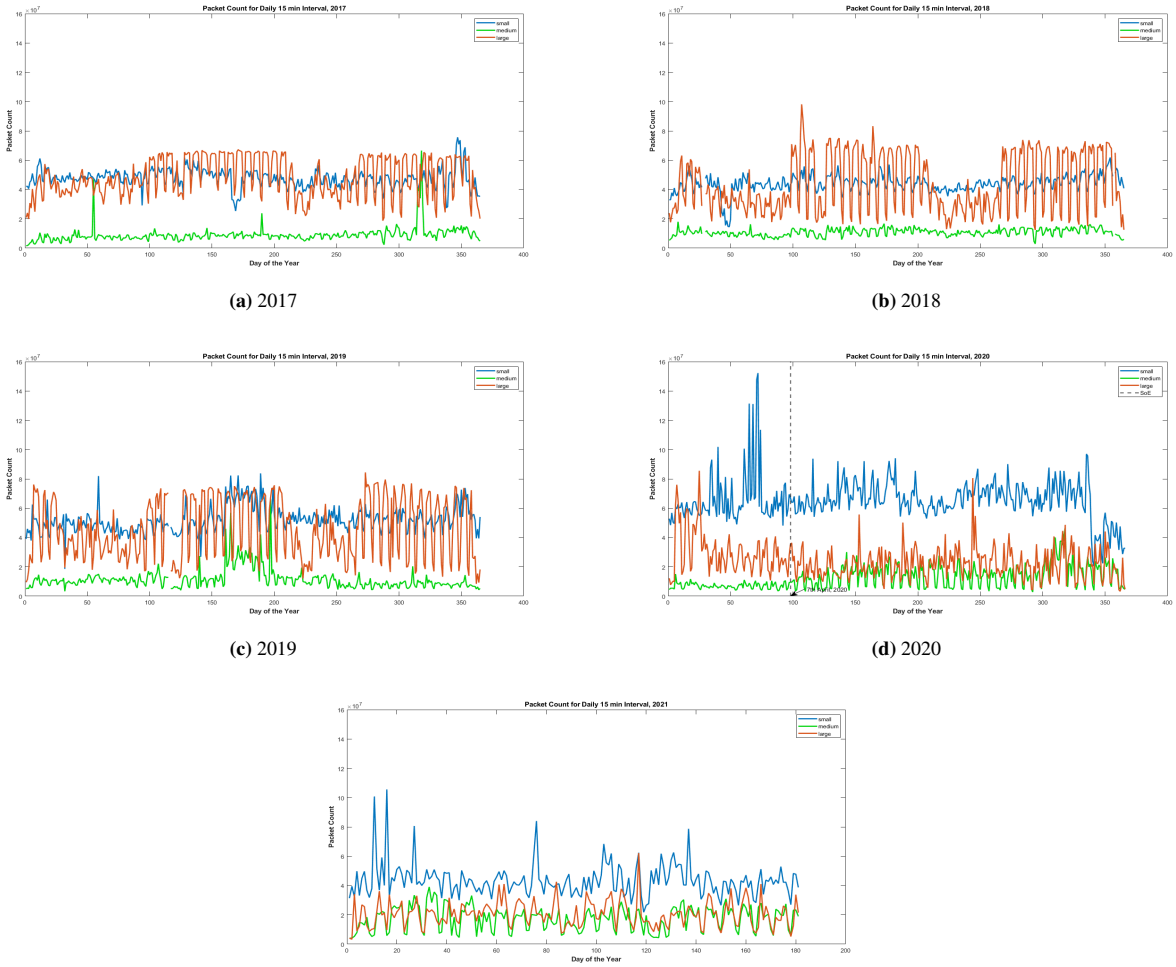
The year wise count of small, medium and large size packets is collated in Table 7. For the period 2017 to 2019, we observe that the number of small packets is comparable to the number of large packets in the network. But, the year 2020 shows a sharp increase in number of small packets and a decline in large packets. The proportion of small packets increases from 48% in 2019 to 64% in 2020. In the first 6 months of 2021, the COVID-19 situation has improved and the proportion of small packets returns to 53% in 2021. This shift of pattern is more clear from Figure 5. The blue, green and brown colors represent small, medium and large packet sizes respectively in the figure. In years 2017 to 2019, the medium size packets remain clearly distinguishable from small and large packets whereas small packets are distinguishable from medium and large packet sizes in year 2020 and first half of 2021 too. This analysis implies that applications, such as voice data calls which generates small size packets, have been used more during the state of emergency period started from 7^{th} April 2020 in Japan.

The plots of other statistics, that is, entropy, mean, variance and coefficient of variation of packet sizes are shown in Figure 6. As we observed earlier, due to an increase in small and medium size packets but a decrease in large size packets, we also observe a significant decrease in the average packet size. There is a 35.3% decrease in the 2020 and a 25.3% decrease in average packet size in 2021 compared to 2019. The daily variation in packet sizes shows an overall reduction of 16.2% in 2020 and 15% in 2021 compared to 2019.Thus, with a decrease in variance but a more significant decrease in mean, we observe an increase in the daily coefficient of variation as well.

As pointed out in earlier study[17] where outings and internet usage are found to have inverse relationship resulting in more usage of internet at home, it would be interesting to compare the traffic in 2020 with that of 2019 on Sunday as shown in Figure 7. The preponderance of small sized packets in 2020 is on account of compound effect of two factors viz. entertainment at home and work from home. Further, there is a reduction in probability mass of large packet sizes in 2020 in relation to 2019. This brings out emergence of qualitative changes in the traffic behavior during state of emergency.

8 | CONCLUSION

The aim of the present study is to investigate whether probabilistic structure of traffic characteristics in terms of byte and packet counts besides packet sizes in pre-Covid period 2014-2019 remained robust and continued to capture the characteristics during Covid period in 2020. Based on MAWI data, statistical analysis reveals that network traffic behavior exhibits a qualitative change. An important finding relates to the fact that generally bimodal traffic characteristics during pre-Covid, captured through GMMs, undergo a qualitative change resulting in unimodal Laplace distribution in 2020. This raises a question as to the reason for emergence of Laplace distribution which can be justified via a scale mixture of Gaussian or normal variates. This implies a significant change in traffic behavior which reveals shifting of probability mass from preponderance of large to small packet



(a) 2017

(b) 2018

(c) 2019

(d) 2020

(e) 2021

Figure 5 Packet size versus day of the year, from 2017 to 2021.

sizes in 2020 indicating that application like voice telephony, WhatsApp become more popular during lockdown or state of emergency enforced by the government.

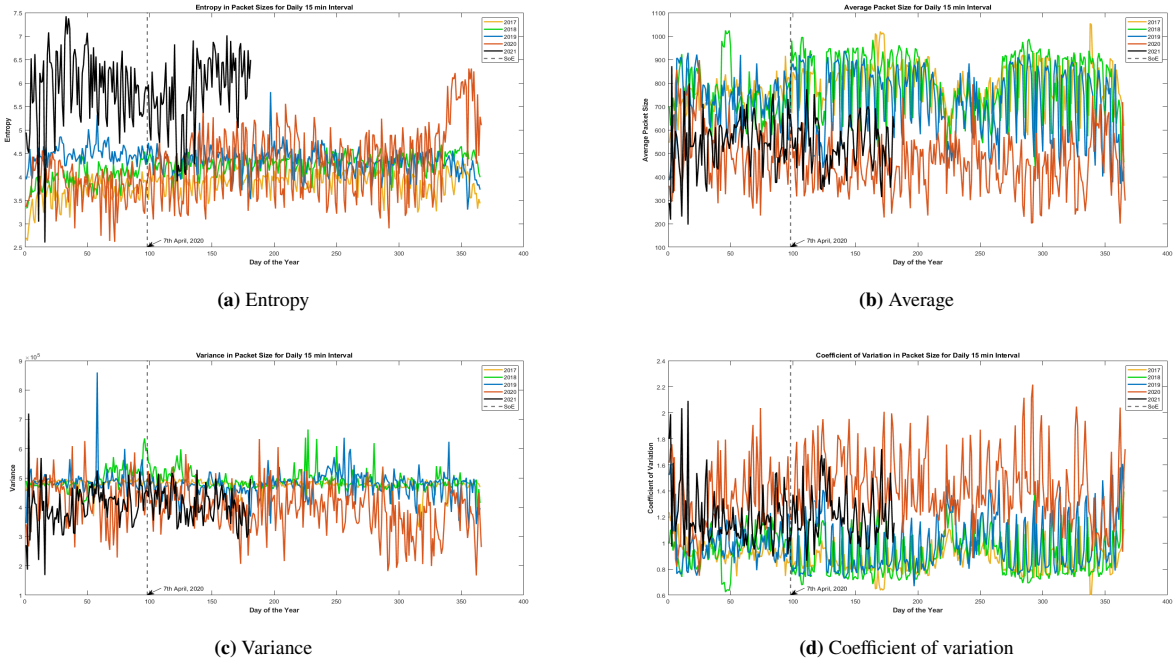
A pertinent question as part of future enquiry is to examine the unfolding of network traffic dynamics from pre-Covid to during Covid period. It would also be interesting to examine the emergence of Laplace distribution in network traffic between Ukraine and Russia war.

9 | ACKNOWLEDGEMENT

The authors would like to thank Dr. Kenjiro Cho, Director, IIJ Research Laboratory, Japan for promptly responding to their queries.

Author contributions

All authors contributed equally to this study.



(a) Entropy

(b) Average

(c) Variance

(d) Coefficient of variation

Figure 6 Statistical measures of packet sizes for daily 15 minute interval data from 2017 to 2021.

Financial disclosure

None reported.

Conflict of interest

The authors declare no potential conflict of interests.

References

1. Iversen V. *Teletraffic Engineering Handbook*. ITU-T . 2001.

2. Fontugne R, Abry P, Fukuda K, et al. Scaling in Internet Traffic: A 14 Year and 3 Day Longitudinal Study, With Multiscale Analyses and Random Projections. *IEEE/ACM Transactions on Networking* 2017; 25(4): 2152-2165. doi: 10.1109/TNET.2017.2675450

3. Favale T, Soro F, Trevisan M, Drago I, Mellia M. Campus Traffic and e-Learning during COVID-19 Pandemic. *Computer Networks* 2020; 176.

4. Alotibi F, Velagapudi A, Madan A, et al. The Impact of COVID-19 on Communication Network Outages. In: IEEE. ; 2022: 1-8

5. Christensen K, Javagal N. Prediction of future world wide web traffic characteristics for capacity planning. *International Journal of Network Management* 1997; 7(5): 264–276.

6. Roberts J. Traffic theory and the Internet. *IEEE Communications Magazine* 2001; 39(1): 94–99. doi: 10.1109/35.894382

7. Tammaro D, Valenti S, Rossi D, Pescape A. Exploiting packet-sampling measurements for traffic characterization and classification. *International Journal of Network Management* 2012; 22(6): 451–476. doi: 10.1002/nem.1802

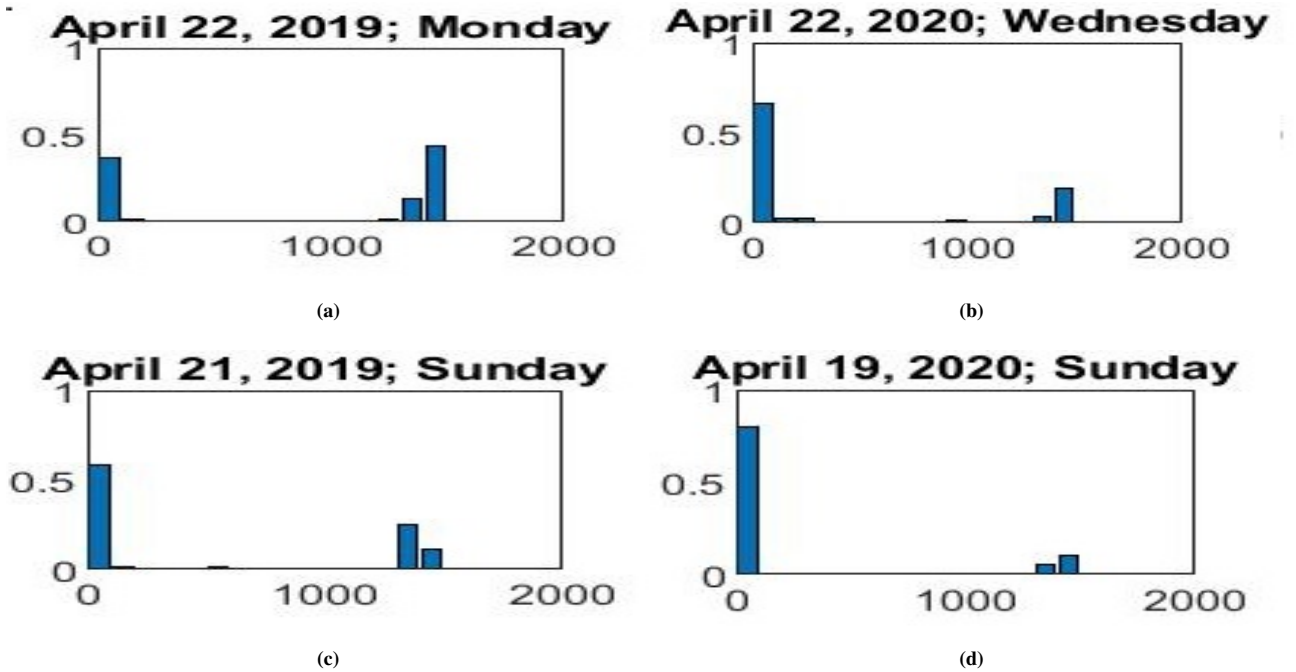


Figure 7 Comparison of weekday and weekend packet sizes.

8. Boussada M, Garcia J, Frikha M. New fluid approaches for studying the performance of elastic traffic under class based weighed fair queuing system. *International Journal of Network Management* 2018; 28(6). doi: 10.1002/nem.2037

9. Borgnat P, Dewaele G, Fukuda K, Abry P, Cho K. Seven years and one day: Sketching the evolution of internet traffic. In: IEEE. ; 2009: 711-719

10. Fontugne R, Abry P, Fukuda K, et al. Scaling in Internet traffic: A 14 year and 3 day longitudinal study, with multiscale analyses and random projections. *IEEE/ACM Transactions on Networking* 2017; 25(4): 2152–2165. doi: 10.1109/T-NET.2017.2675450

11. Feldmann A, Gasser O, Lichtblau F, et al. The Lockdown Effect: Implications of the COVID-19 Pandemic on Internet Traffic. In: ACM. ; 2020: 1–18

12. Feldmann A, Gasser O, Lichtblau F, et al. A year in lockdown: How the waves of COVID-19 impact internet traffic. *Communications of the ACM* 2021; 64(7): 101–108. doi: 10.1145/3465212

13. Candela M, Luconi V, Vecchio A. Impact of the COVID-19 pandemic on the Internet latency: a large-scale study. *Computer Networks* 2020; 182.

14. Society I. COVID-19 Impact on Internet performance - case study of Afghanistan, Nepal, and Sri Lanka. tech. rep., Internet Society; 11710 Plaza America Drive, Suite 400 Reston, VA U.S.A: 2021.

15. Kende M. Impact of COVID-19 on the Internet ecosystem in the Middle East and North Africa. tech. rep., Internet Society; 11710 Plaza America Drive, Suite 400 Reston, VA U.S.A: 2020.

16. Affinito A, Botta A, Ventre G. The impact of covid on network utilization: an analysis on domain popularity. In: IEEE. ; 2020: 1-6

17. Yabe N, Hanibuchi T, Adachi H, Nagata S, Nakaya T. Relationship between Internet use and out-of-home activities during the first wave of the COVID-19 outbreak in Japan. *Transportation Research Interdisciplinary Perspectives* 2021; 10. doi: https://doi.org/10.1016/j.trip.2021.100343

18. Lutu A, Perino D, Bagnulo M, Frías-Martínez E, Khangosstar J. A Characterization of the COVID-19 Pandemic Impact on a Mobile Network Operator Traffic. In: ACM. ; 2020: 19–33

19. Silva G. dC, Ferrari A, Osinski C, Pelacini D. The behavior of Internet traffic for Internet services during COVID-19 pandemic scenario. arXiv; 2021

20. Bronzino F, Feamster N, Liu S, Saxon J, Schmitt P. Mapping the digital divide: Before, during, and after COVID-19. In: SSRN. ; 2021

21. Karmeshu ., ed.*Entropy Measures, Maximum Entropy Principle and Emerging Applications*. Studies in Fuzziness and Soft ComputingSpringer Berlin, Heidelberg. first ed. 2003.

22. Geraci M, Borja M. Notebook: The Laplace distribution. *Significance* 2018; 15(5): 10–11. doi: 10.1111/j.1740-9713.2018.01185.x

23. McLachlan G, Rathnayake S. On the number of components in a Gaussian mixture model. *WIREs Data Mining Knowledge Discovery* 2014; 4(5): 341–355. doi: 10.1002/widm.1135

24. Cho K, Mitsuya K, Kato A. Traffic Data Repository at the WIDE Project. USENIX 2000 FREENIX Track, San Diego, CA; 2000.

25. Borgnat P, Dewaele G, Fukuda K, Abry P, Cho K. Seven years and one day: Sketching the evolution of Internet traffic. In: IEEE. ; 2009: 711-719

26. Kato M, Cho K, Honda M, Tokuda H. Monitoring the Dynamics of Network Traffic by Recursive Multi-Dimensional Aggregation. In: USENIX Association. ; 2012; Hollywood, CA.

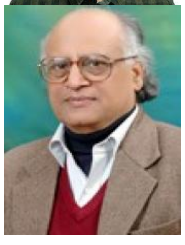
27. Massey F. The Kolmogrov-Smirnov test for goodness of fit. *Journal of the American Statistical Association* 1951; 46(253): 68–78.

28. Akaike H. A new look at the statistical model identification. *IEEE Transactions on Automatic Control* 1974; 19(6): 716-723. doi: 10.1109/TAC.1974.1100705

AUTHORS BIOGRAPHY

Anoushka Mittal. Anoushka Mittal obtained a Bachelor in Technology degree in Computer Science and Engineering from the School of Engineering, Shiv Nadar University, Delhi NCR, India, in 2021. She is pursuing her M.Sc. in Computer Science from Trinity College Dublin, Ireland. Her current research interests include data mining, computer vision, and machine learning.

Pranav Jain. Pranav Jain is working as software engineer at ViaSat, Chennai, India. He completed his Bachelor's degree in Computer Science and Engineering from Shiv Nadar University, India, with a minor in Big Data Analytics. At ViaSat, he has been a part of the Global Network and Technology segment where he has worked on developing various automations with the information security team.



Karmeshu. Dr. Karmeshu is a distinguished professor in the School of Computer Science, University of Petroleum and Energy Studies, Dehradun, Uttarakhand, India. He is also holdinh the position of honorary professor at Department of Computer Science and Engineering at Shiv Nadar University, India. In January 2021, he received D.Sc. degree (Honoris Causa) from the Dayalbagh Educational Institute, Agra, India in recognition of his outstanding pioneering contribution the field of mathematical modeling. He is currently chairman of the scientific computing panel of Naval Research Board (NRB) of DRDO, Government of India. Dr. Karmeshu has been working in the area of modeling and simulation of nonlinear stochastic systems and maximum entropy framework. He has published more than 100 research papers in international journals.

Shachi Sharma. Dr. Shachi Sharma has received her Ph.D. degree from School of Computer and System Sciences, Jawaharlal Nehru University, New Delhi, India in 2007. She worked as Research Scientist at IBM Research Laboratory, New Delhi, India from 2008-2017. She was a visiting faculty at Indraprastha Institute of Information Technology (IIIT) Delhi, India in 2018 and subsequently joined South Asian University, New Delhi, India. Her research work is substantiated by many international research publications and US patents. Her research interests include data analysis, stochastic modeling and applications of maximum entropy framework.