ISSN 1330-3651 (Print), ISSN 1848-6339 (Online)
https://doi.org/10.17559/TV-20200224091110
Preliminary communication

Simulation of Academic Computer Networks Using Probability Distributions: A Case Study in A Campus Network

Mehmet Ali ALTUNCU, Fidan KAYA GÜLAĞIZ, Hikmetcan ÖZCAN*, Sümeyya İLKİN, Suhap ŞAHİN

Abstract: Computer networks are becoming more complex with the advances in technology. Hence, the installation of computer networks becomes more complicated and costly. Therefore, many parameters of the existing or planned networks, such as the requirements, limits and performance are modelled through simulators. Thus, it is possible to save both in terms of time and cost. Campus networks are networks that are established by consolidating many local area networks. The aim of this study is to model campus networks which have a general daily behaviour pattern, through simulators. The data used in the study are collected in real time from Siirt University. The daily behaviour of the network in working hours is divided into four separate time intervals according to the network traffic and in consideration of similar studies in the literature. The most appropriate distributions that model the transmission times of the incoming/outgoing packets at each time zone are identified. The results are presented in comparison with the previous studies conducted to model campus networks. At the same time, the most generic distributions that model the daily incoming / outgoing traffic of the network are identified. The distribution that best models the transmission times of the network packets was identified to be the lognormal distribution for TCP packets and the Generalized Pareto distribution for UDP packets. Compatibility of the distributions was determined through the use of Kolmogorov-Smirnov and Chi-Squared tests.

Keywords: local area network; network architecture; network statistical analysis; simulation; statistical distributions

1 INTRODUCTION

In order to accurately assess the effect of the protocols, applications, and users, used in network simulation, it is very important to generate a simulated traffic. There are two types of traffic to consider when modelling with simulators. The first is the application-specific traffic to be modelled for the target application, and the other is the background traffic generated by other applications on the network. Background traffic has a significant effect on the behaviour of the target application with regard to the use of network resources [1]. The dimension or extent of this effect was analysed by various researchers.

Venkatesh and Vahdat [2] conducted a behavioural analysis on the synthetic and real background traffic for different applications. According to the results they obtained, it was concluded that each application was affected by the intensity in traffic at certain levels, depending on the type of application.

In another study conducted by Venkatesh and Vahdat [3], it was demonstrated that structural traffic models specific to applications can be successfully established. When generating new traffic, the transmission frequency and time, distribution of packet sizes, characteristics of the flows and destination internet protocol (IP) and destination port addresses of the packets from the original traffic were taken into consideration. They demonstrated through their study that the traffic they generated with a different application, different network and user conditions was compatible with the real network traffic.

It was demonstrated by Nahum et al. [4] that WAN (Wide Area Network) conditions had a significant effect on network performance. In this study, parameters (file size, request transmission time, etc.) that might cause traffic density on the servers were identified.

In the study conducted by Eylen and Bazlamaçı [5], they needed background traffic in order to obtain a traffic similar to the real traffic conditions. For this purpose, background traffic was generated through Poisson distribution, in order to add random delays on the trial packages used in the study. The real traffic was modelled by generation of three different rates of traffic and the proposed method was analysed more accurately under background traffic.

When the conducted studies are examined, it can clearly be seen that the background traffic has a significant effect on both the applications and servers. The size of this effect varies according to many different parameters. For this reason, to ensure realistic analysis of the application, it is necessary to model the network traffic generated outside the application (background traffic).

So far, many different distributions were used to model the traffic that occurs on a network during the day. At earlier times, exponential modelling of packet transmission times by Paxson and Floyd [6] was accepted to be a convenient method. In later years, the Poisson distribution was shown to be accurate for designing a flow-based internet traffic model [7]. In 2008, Fras et al. [8] modelled the statistical processes of network traffic by using the probability density function. Histograms of the measured traffic were used to determine the parameters of the Pareto, Weibull and exponential distributions used in the study. The most fit distribution was evaluated through the use of Kolmogorov-Smirnov, Anderson-Darling and Chi-Squared statistical goodness of fit tests. In terms of package size, Weibull distribution was found to be more suitable than the other distributions, in all three tests.

Bhattacharjee and Nandi [9] compared the Log-Normal distribution and the Pareto distributions to model the transmission times of the academic network data. In the study, which was based on the statistical analysis of data in terms of location and time, it is concluded that the Log-Normal distribution is more suitable for its own data than Pareto distribution.

However, it was shown that the use of a single probability distribution was not suitable for the different behaviour of the network over different time periods [10, 11]. Garsva et al. [10] conducted a statistical analysis of the academic network data collected with Netflow. In this study, the network traffic was divided into eight time intervals. In general sense, it was seen that Pareto 2 distribution was suitable to model the packet transmission times during the more intensive (heavy tail) time intervals, and Weibull and Pareto2 distributions were more suitable to model the packet transmission times at the low-intensity traffic hours.

When the studies conducted until now are examined, fit distributions for modelling different types of networks were demonstrated, but the architecture of the modelled network was not included in the studies [9-11]. Meanwhile, there are only a few studies on networks with periodic behaviours throughout the day. In this study, both the architecture of the network from which the data is obtained and the probability distributions that model the packet transmission times within different time intervals are provided in comparison with the previous studies.

The next part of the study is organized as follows. The second section describes the modelled network architecture. The third section conducts a statistical analysis of the data on the network traffic. The fourth section briefly mentions the tests used to model package transmission times. Statistical analysis of the transmission times of the package is provided in the fifth section. We complete this paper with the conclusion and some guideless for future work.

2 MODELED NETWORK ARCHITECTURE

The network structure of the university is shown in detail in Fig. 1. Ulaknet provides the access of the University to the internet and the infrastructure is provided by the service provider (Türk Telekom). The bandwidth of the University is 500 Mbps [12].

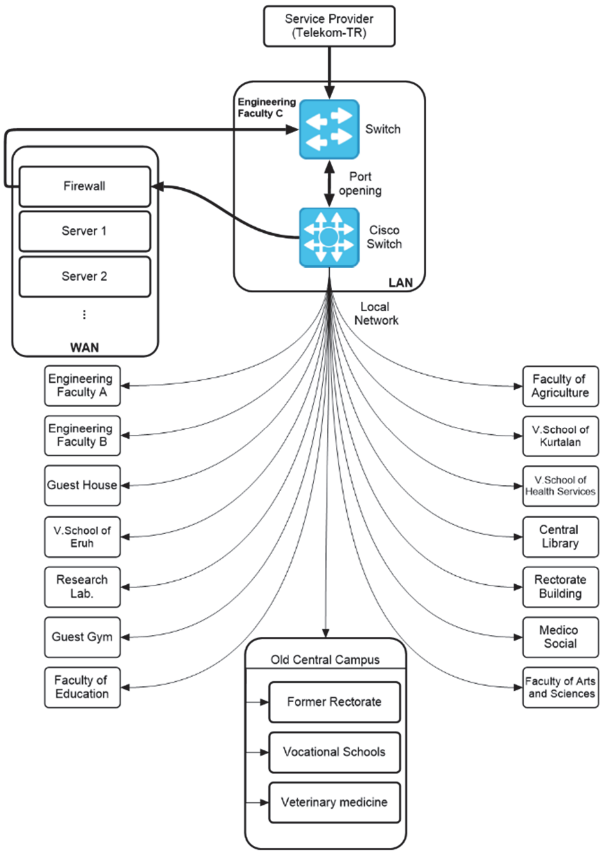


Figure 1 The network architecture of the university where daily network data is collected

There are local area networks between the faculties and vocational schools in the University. Each client connected to the network sends a request to open a port from the university to access the Internet. This request then passes through the firewall to reach Ulaknet and then internet access is provided. To access a server in the university network, again, a request is sent to open a port. However, direct access to the server is provided for this request without having to pass through the firewall. Communication between clients is provided through the Cisco Switch, without the need to open a port.

3 GENERAL STATISTICAL ANALYSIS OF NETWORK

In this part of the study, firstly, the Z-score method, which is used for eliminating outliers is mentioned. Then, the detailed statistical analysis of the modelled campus network is described.

3.1 Outlier Analysis

Various distributions are used in the literature to model different networks. The structure and overall behaviour of the network should be taken into consideration in determining the compatibility of the distributions. In our study, by reference to a prior study [13], modelling was performed by taking only weekdays into account when campus network data are obtained. The sample data belonging to the collected dataset is shown in Fig. 2. Traffic for one day is first classified as incoming and outgoing traffic. Each case is then divided into 4 different time intervals. Heavy network traffic conditions (working hours) are taken into consideration in determining these time intervals. In Tab. 1, the packets for incoming traffic are divided between time intervals 1 to 4 and the packets for outgoing traffic are divided between time intervals 5 to 8. When the table is analysed, it can be seen that the majority of the incoming and outgoing packets are transmitted in time intervals 2 (20.29%) and 6 (27.61%). A significant increase in network traffic was observed with the start of the workday and a significant decrease was observed with the end of the workday.

Table 1 Network traffic periods

Incoming	Packet / %	Time	Length / h	Packet / %	Outgoing
1	11.37	07:00-10:59	4	16.08	5
2	20.29	11:00-16:59	6	27.61	6
3	5.08	17:00-22:59	6	10.04	7
4	1.34	23:00-06:59	8	8.19	8

After the incoming and outgoing traffic is divided into time intervals and collected, before proceeding with the analysis of the data, the outlier values within the data need to be identified.

The observations which are numerically distant from the other data for some reason, are called outlier values. Outliers often lead to negative effects such as increasing error difference, influencing estimation results, and reducing the strength of statistical tests [14]. Therefore, outlier analysis methods are applied to data that do not have a normal distribution and have too many outlier values. [15].

In this study, Z-Score method, which is based on statistical approach, is used in determination of the outliers. In the Z-Score method, the average (μ) and standard deviation (σ) values are used to determine whether any value (z) is an outlier (Eq. (1)).

$$z = \frac{x - \mu}{\sigma} \tag{1}$$

The z value obtained in Eq. (1) is considered to be the normal value if it is within the (–3, 3) range. All values outside the defined values are outliers [16]. In our study, calculations are based on the number of packages that fall within each time interval. The average number of packets was calculated for each time interval and z values were calculated.

Table 2 Outlier analysis of daily data

	TCP (Transmission Control Protocol)		UDP (User Datagram Protocol)	
	Outliers / %	Packets / %	Outliers / %	Packets / %
1	0.26	78.34	0.15	21.61
2	0.78	75.76	0.24	24.21
3	2.77	80.08	0.35	19.90
4	0.35	67.60	0.40	32.36
5	0.18	88.45	0.39	11.49
6	0.58	85.42	0.28	14.54
7	1.19	89.63	0.45	10.31
8	0.03	83,57	0.68	16.31
Incoming	1.04	75.46	0.29	24.52
Outgoing	0.50	86.77	0.45	13.16
Average	0.77	81.12	0.37	18.84

The percentage of the outlier values obtained in the study conducted by Garsva et al. [13] based on TCP and UDP (TCP: 3.39 UDP: 3.77) is around 3 times the values obtained in our study (TCP: 0.77 UDP: 0.37). In Fig. 3, the data obtained by the interquartile range (IQR) technique that used by Garsva et al. [13] and the values obtained when the z score method is applied are shown comparatively. The results obtained by the IQR method on the left of the figure and the results obtained by the z-score method on the right are shown. When the IQR method is examined, there is an inconsistency in the data obtained. In Fig. 3a and Fig. 3b, the IOR method cut the data from the lower values than the z-score method, but in (c) it could not even eliminate very high values in the data. Thus, the z-score method was preferred because outlier values are eliminated more consistently in our study.

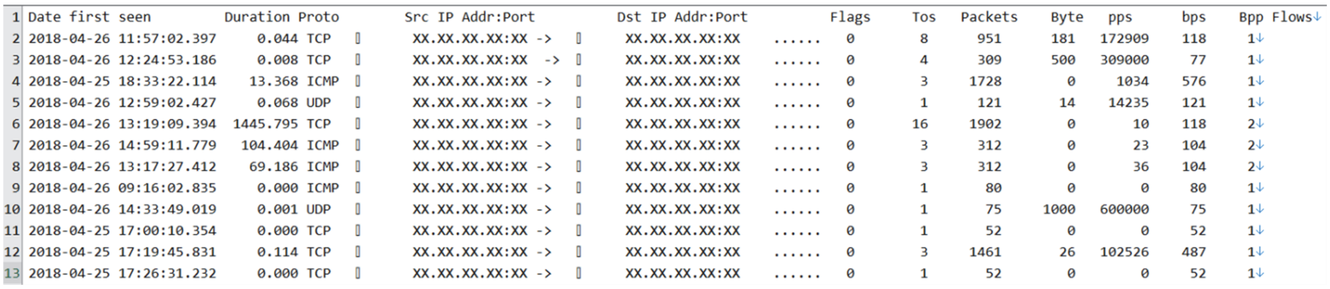
When the left side of Tab. 2 is examined, for example, time zone 1 represents the incoming packets within the interval of 07: 00-10: 59. While 78.34% of these packets are TCP protocol packets, 21.61% of them are UDP protocol packets. 0.26% of the incoming TCP packets within this interval contain outlier values, while 0.15% of the UDP packets contain outlier values, which were eliminated. When the table is broadly analysed, it is observed that the outliers are higher in the periods when the network traffic is intense (2-3 for incoming traffic, 6-7 for outgoing traffic). At the same time, an average outlier ratio of 0.77% is observed in the TCP protocol, while the average value is 0.37% in the UDP protocol. The majority of outliers for the TCP protocol are observed in the incoming packets (1.04%), whereas for the UDP protocol, more outliers are observed in the outgoing packets (0.45%).

3.2 Statistical Analysis of Network

After eliminating the outliers in the data, graphical distribution of the daily traffic is obtained. The graphs are given in Fig. 3. Both the protocol-based and graphs containing all the protocols are given in detail. The x-axis of the graphs represents the clock and the y-axis represents the number of flows within the relevant time interval. For the representation of the graphs, Garsva et al. [13] are taken as the reference. It can be seen that the distribution of network traffic in the graphs is consistent with the distribution in the time intervals presented in Tab. 1.

In Tab. 3 and Tab. 4, packet and flow information on incoming and outgoing traffic are given in detail, respectively. In terms of incoming traffic, TCP traffic is 7.9 times that of UDP traffic. The traffic generated by ICMP protocol traffic is very low compared to TCP and UDP traffic. The number of TCP flows is 3.4 times the number of UDP flows, but the number of TCP and UDP packets per flow is almost the same for incoming traffic. Again, the average size of TCP packets is higher than UDP packets (around 2.4 times).

In terms of outgoing traffic, TCP traffic is 6.3 times the UDP traffic. The number of TCP flows is 4.3 times the number of UDP flows, but unlike incoming traffic, the number of packets per flow is 1.6 times for TCP than that of UDP. When the average size of the packets is compared on the basis of outgoing traffic, it is seen that there is no significant difference.

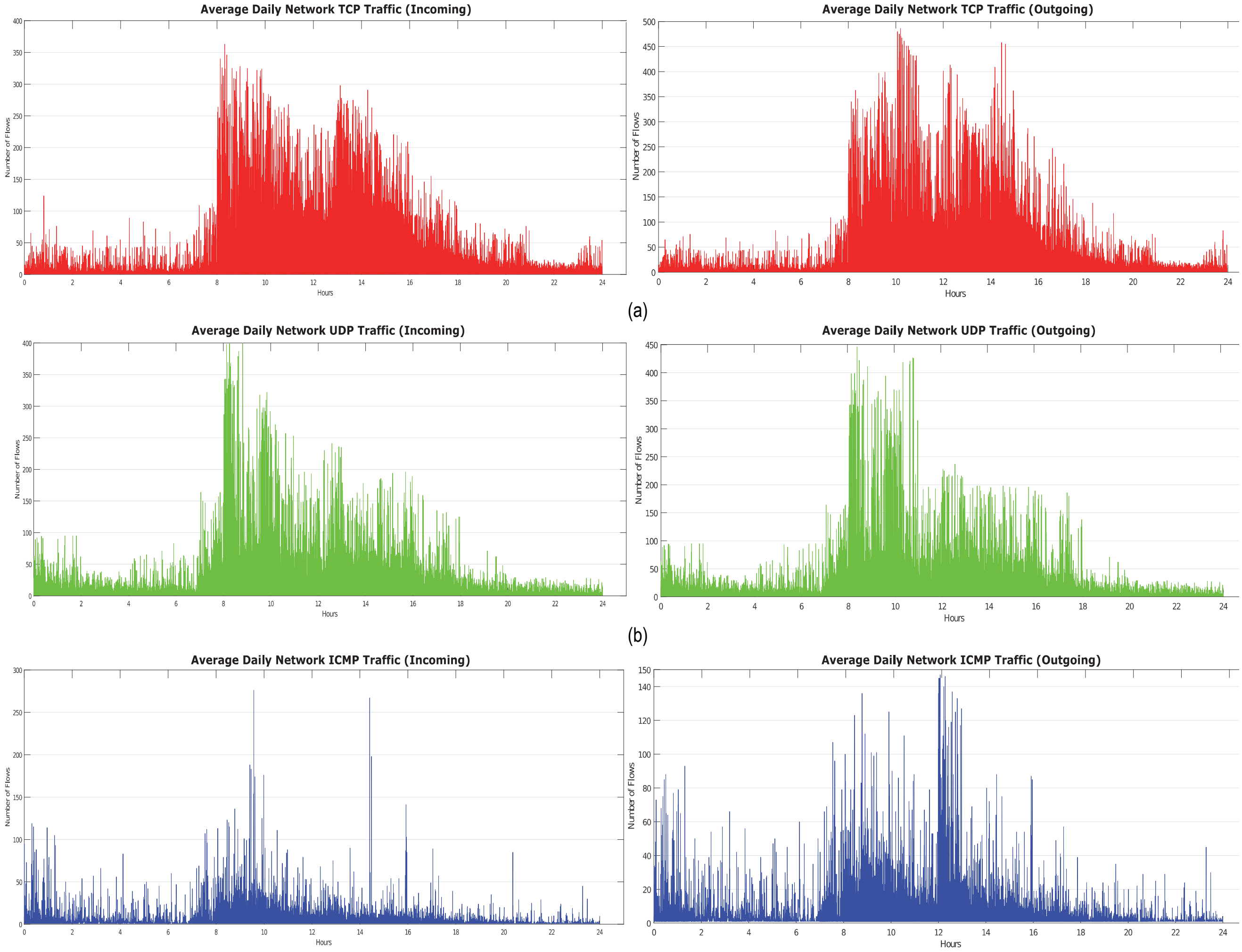


```
Date first seen          Duration Proto       Src IP Addr:Port          Dst IP Addr:Port          Flags     Tos   Packets    Byte    pps      bps   Bpp Flows↓
2018-04-26 11:57:02.397     0.044 TCP   ▯    XX.XX.XX.XX:XX ->   ▯    XX.XX.XX.XX:XX   ......   0        8      951     181   172909    118    1↓
2018-04-26 12:24:53.186     0.008 TCP   ▯    XX.XX.XX.XX:XX   -> ▯    XX.XX.XX.XX:XX   ......   0        4      309     500   309000     77    1↓
2018-04-25 18:33:22.114    13.368 ICMP  ▯    XX.XX.XX.XX:XX ->   ▯    XX.XX.XX.XX:XX   ......   0        3     1728       0     1034    576    1↓
2018-04-26 12:59:02.427     0.068 UDP   ▯    XX.XX.XX.XX:XX ->   ▯    XX.XX.XX.XX:XX   ......   0        1      121      14    14235    121    1↓
2018-04-26 13:19:09.394  1445.795 TCP   ▯    XX.XX.XX.XX:XX ->   ▯    XX.XX.XX.XX:XX   ......   0       16     1902       0       10    118    2↓
2018-04-26 14:59:11.779   104.404 ICMP  ▯    XX.XX.XX.XX:XX ->   ▯    XX.XX.XX.XX:XX   ......   0        3      312       0       23    104    2↓
2018-04-26 13:17:27.412    69.186 ICMP  ▯    XX.XX.XX.XX:XX ->   ▯    XX.XX.XX.XX:XX   ......   0        3      312       0       36    104    2↓
2018-04-26 09:16:02.835     0.000 ICMP  ▯    XX.XX.XX.XX:XX ->   ▯    XX.XX.XX.XX:XX   ......   0        1       80       0        0     80    1↓
2018-04-26 14:33:49.019     0.001 UDP   ▯    XX.XX.XX.XX:XX ->   ▯    XX.XX.XX.XX:XX   ......   0        1       75    1000   600000     75    1↓
2018-04-25 17:00:10.354     0.000 TCP   ▯    XX.XX.XX.XX:XX ->   ▯    XX.XX.XX.XX:XX   ......   0        1       52       0        0     52    1↓
2018-04-25 17:19:45.831     0.114 TCP   ▯    XX.XX.XX.XX:XX ->   ▯    XX.XX.XX.XX:XX   ......   0        3     1461      26   102526    487    1↓
2018-04-25 17:26:31.232     0.000 TCP   ▯    XX.XX.XX.XX:XX ->   ▯    XX.XX.XX.XX:XX   ......   0        1       52       0        0     52    1↓
```

Figure 2 Sample dataset

Again, the right side of the tables presents the data for the incoming and outgoing traffic of the network modelled by Garsva et al. [13]. When compared in terms of total traffic, the traffic values in the study [13] and the traffic in our study are close. However, if the number of flows on the basis of protocol is compared, the number of TCP flows in our study is approximately 3 times that of the flows in study no [13], while the UDP and ICMP packet flows in study no [13] are higher than those in the traffic we modelled. The average number of packets per flow and the average size of the packets are higher than the values of the network in our study.

Incoming and outgoing data traffic is also analysed with regard to some known ports. The number of transmitted packets and the packet size information by port types are presented in Tab. 5 and Tab. 6.



(a)

(b)

(c)

Figure 3 Daily network traffic based on number of flows (a) TCP; (b) UDP; (c) ICMP

Table 3 Incoming traffic data

	Our Study			(Garsva et. al) [13]		
Parameter\Protocol	TCP	UDP	ICMP	TCP	UDP	ICMP
Traffic Total, B	1425×10^8	179×10^8	0.22×10^8	843×10^9	257×10^9	0.158×10^9
Flows Total	20831009	6116872	110752	12954137	11733655	438532
Packets Total	5065×10^5	1534×10^5	2×10^5	1007×10^6	265x10^6	2×10^6
Average Packets in Flow	24	25	2	78	23	4
Average Size of Packets, B	281	117	95	838	970	96

Table 4 Outgoing traffic data

	Our Study			(Garsva et. al) [13]		
Parameter\Protocol	TCP	UDP	ICMP	TCP	UDP	ICMP
Traffic Total, B	495×10^8	78×10^8	0.54×10^8	131×10^9	91×10^9	0.358×10^9
Flows Total	32806929	7552372	51814	12072179	10958168	882523
Packets Total	4189×10^5	635×10^5	2×10^5	464×10^6	169×10^6	3×10^6
Average Packets in Flow	13	8	5	38	15	3
Average Size of Packets, B	118	123	207	282	539	125

There are 65536 ports available for use in TCP or UDP. They are divided into three ranges [10, 17];

- 0-1023: Well known ports
- 1024-49151: Registered ports
- 49152-65535: Dynamic ports.

Well known ports such as HTTP, FTP, SMTP, contain port numbers used for standard pre-defined operations. Registered ports can be used by common user operations or programs executed by common users in most systems whereas dynamic ports can be used dynamically by any application [10, 17].

Table 5 Incoming traffic statistics according to port number

Port Number	Packet Count	Packet Size, B	Average Packet Size, B
443	4.86E+08	1.29E+11	266.22
53	5500595	7.16E+08	130.19
80	1.44E+08	2.73E+10	189.48
21	16174	932176	57.63
22	645984	7.07E+07	109.41
0-1023	6.38E+08	1.58E+11	247.14
1024-49151	1.63E+07	2.27E+09	139.75
49152-65535	5967911	5.78E+08	96.78

Table 6 Outgoing traffic statistics according to port number

Port Number	Packet Count	Packet Size, B	Average Packet Size, B
443	2.84E+06	2.91E+09	1024.64
53	5.09E+05	1.48E+08	290.55
80	7.56E+06	7.20E+09	951.76
21	2450	1.79E+05	72.91
22	27075	5.45E+06	201.38
0-1023	1.14E+07	1.03E+10	906.31
1024-49151	1.49E+08	1.76E+10	118.06
49152-65535	3.22E+08	2.95E+10	91.63

When Tab. 5 and Tab. 6 are analysed, it can be concluded that the highest number of packets are transmitted from ports in the range of 0-1023, and the port with the largest average packet size per packet is 443. With regard to outgoing packets, it is seen that the highest number of packets are transmitted from ports in the range of 49152-65535, which are used by dynamic applications; however, the port with the largest average packet size per packet is port 80.

4 GOODNESS OF FIT TESTS

In this section, the probability distributions, which would most efficiently model the time intervals and the general traffic of one day presented in Tab. 1, are identified. Kolmogorov-Smirnov and Chi-Squared tests are performed to determine the most suitable distribution. Kolmogorov-Smirnov and Chi-Squared tests are nonparametric tests. Nonparametric tests are widely used when knowledge of the data to be modeled is not available. They can also process limited number of data. In general, they process data faster according to parametric tests. In scope of the study nonparametric tests were preferred because there was data sparsity in some time intervals to be modeled. Also, since it would not be correct to determine suitability over a single test, the most commonly used Kolmogorov-Smirnov and Chi-Squared tests were preferred among non-parametric tests. The descriptions of these tests and the parameters of the fit distributions are presented in detail in the subsections.

4.1 Kolmogorov-Smirnov Test

The Kolmogorov-Smirnov (K-S) test is a non-parametric goodness of fit test used to differentiate the changes in the data. In this way, it provides more successful results than parametric data in cases where the assumptions about the data are insufficient. K-S test is applied in the modelling of the internet network as well as fields such as astronomy and wireless sensor networks. The aim of the K-S test is to compare the Cumulative Distribution Function (CDF) of the data with the recommended CDF [18, 19]. This comparison process is performed by following the steps below [19, 20].

Step 1: If the observed frequencies are equal to the expected frequencies Hypothesis 0 is accepted, if not Hypothesis 1 is accepted.

Step 2: $D = \max | F_o - F_e |$ (2)

The test statistic value is calculated with the formula in Eq. (2). In Eq. (2), D represents the test statistic, F_o represents the observed cumulative frequency and F_e represents the expected cumulative frequency.

Step 3: $k(\alpha, N) = \sqrt{\frac{1}{2N} \ln\left(\frac{2}{\alpha}\right)}$ (3)

The critical value is calculated with the formula in Eq. (3). In Eq. (3), N represents the number of observations. If the test statistic value is greater than the critical value, Hypothesis 1 is assumed to be α significant. Otherwise, Hypothesis 0 is valid.

4.2 Chi-Squared Test

The Chi-Squared distribution is also often used to test two independent qualitative criteria. The process steps are almost identical to the Kolmogorov-Smirnov (K-S) test. Hypothesis 0 indicates that the two criteria are independent and Hypothesis 1 indicates that there is a correlation between the two criteria. The only difference with the K-S test is that the test statistic value is calculated as shown in Eq. (4). Nevertheless, for the Chi-Squared test to be performed, the expected frequencies must be greater than 5, [21]. This seems to be a significant disadvantage compared to the Kolmogorov-Smirnov test.

$$\chi^2 = \frac{(O - E)^2}{E} \tag{4}$$

In Eq. (4), O represents the observed frequency, E represents the expected frequency and χ^2 represents the chi-square value.

5 PACKET INTER ARRIVAL TIME STATISTICAL ANALYSIS

In Tab. 7, the distributions fit for packet transmission times for each section are determined for both TCP and UDP packets. When the table is examined, it can be observed that Pareto 2 distribution is prominently fit for modelling the transmission time of the packet for both protocols. Other remarkable situations in the table are the compatibility of the Log Logistic distribution for the TCP protocol and the Weibull distribution for the UDP packets during the low traffic time intervals 4-8. The parameters of the distributions listed in the table and the feasibility values according to Kolmogorov-Smirnov and Chi-Squared tests are listed in detail. In Tab. 7 parameter 1 column, α and k symbols represent the shape parameter, σ symbol represents the standard deviation value of the Lognormal distribution, β and σ symbols in the Parameter 2 column represent the scale parameter, and finally μ symbol, which is the third parameter, represents the location parameter for Generalized Extreme Value distribution and the mean value for the Lognormal distribution.

Table 7 Probability distributions appropriate to packet inter arrival times

	Section	Distribution	Parameter 1	Parameter 2	μ	K-S	Chi-Squared	(Garsva et al.) [13]
TCP	1	Pareto 2	1.4203 (α)	20607.0 (β)	-	0.18561 (Rank 5)	10.299	Pareto 2
	2	Pareto 2	0.30614 (α)	10.209 (β)	-	0.09153 (Rank 6)	2.8005	Weibull
	3	Pareto 2	0.50204 (α)	32.246 (β)	-	0.15484 (Rank 6)	19.065	Pareto 2
	4	Log Logistic	1.2836 (α)	106.47 (β)	-	0.10435 (Rank 9)	5.329	Pareto 2
	5	Pareto 2	1.3191 (α)	18906.0 (β)	-	0.18487 (Rank 7)	10.297	Weibull
	6	Pareto 2	0.31418 (α)	12.011 (β)	-	0.09128 (Rank 8)	0.5196	Gamma
	7	Pareto 2	0.51974 (α)	39.156 (β)	-	0.17342 (Rank 4)	14.523	Weibull
	8	Log Logistic	1.296 (α)	116.35 (β)	-	0.0935 (Rank 7)	4.9089	Weibull
UDP	1	Pareto 2	1.4182 (α)	3979.9 (β)	-	0.2305 (Rank 7)	35.253	Lognormal
	2	Pareto 2	1.67 (α)	719.48 (β)	-	0.22721 (Rank 10)	45.684	Pareto 2
	3	Generalized Extreme Value	0.99043 (k)	110.19 (σ)	263.45	0.22836 (Rank 11)	-	Pareto 2
	4	Weibull	1.375 (α)	634.58 (β)	-	0.18084 (Rank 9)	-	Pareto 2
	5	Pareto 2	1.4034 (α)	4884.9 (β)		0.23731(Rank 10)	5.9382	Weibull
	6	Pareto 2	0.90471 (α)	571.54 (β)	-	0.2385 (Rank 6)	73.654	Pareto 2
	7	Lognormal	1.3739 (σ)	-	6.5643	0.26144 (Rank 10)	61.41	Pareto 2
	8	Weibull	1.5775 (α)	1118.9 (β)	-	0.20075 (Rank 4)	-	Pareto 2

The fit distributions obtained specifically for the traffic sections in the study conducted by Garsva et al. [13] are compared with the distributions obtained in our study, according to Tab. 7. Garsva et al. concluded that Pareto 2 and Weibull distributions were fit in general, likewise, Pareto 2 distribution was fit for various sections in our study. When the studies are compared specifically on a section basis, it can be concluded that the same protocol is fit for sections 1 and 3 for the TCP protocol and sections 2 and 6 for the UDP protocol. In our study, unlike Garsva et al., there are also instances where Log Logistic and Generalized Extreme Value distributions are also suitable. (Sections 4 and 8 for TCP protocol and sections 3 for UDP protocol).

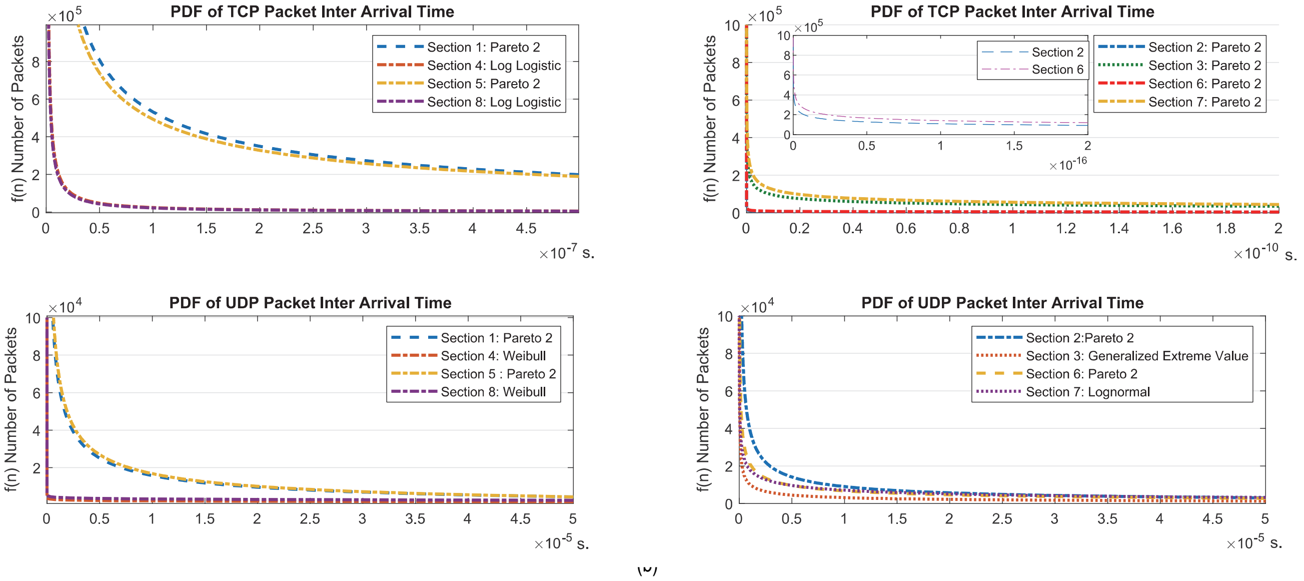
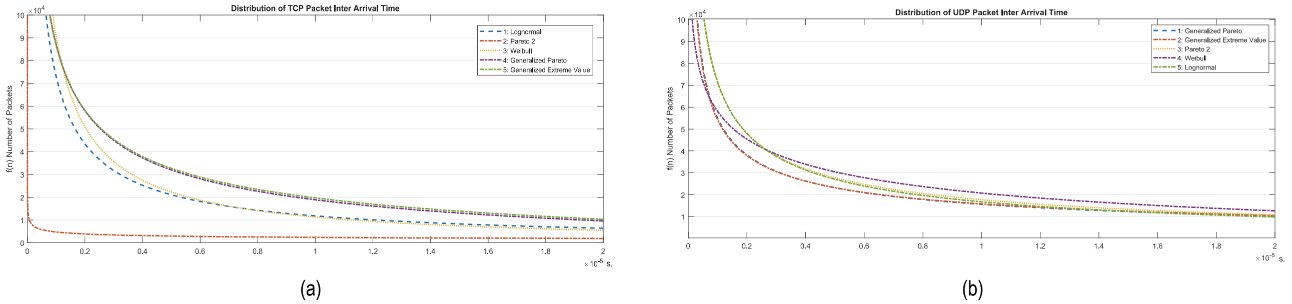


Figure 4 Pdf of packet inter arrival time graphics (a) TCP; (b) UDP

Fig. 4 also presents the pdf (probability distribution function) of the fit distributions. In the graphs in Fig. 4, the y-axis represents the pdf function and the x-axis represents the transmission time of the packets in seconds. When Fig.4 is examined, it is understood that the number of packets sent for the TCP protocol during time intervals 2-6, 3-7 is higher and in these time intervals the transmission time between packets is shorter. The specified time intervals correspond to working hours with intensive traffic. The frequency of packet transmission decreases with the decrease in traffic. As can be seen in Fig. 4a, in the distribution graphs of time intervals 1-5, 4-8, the frequency of packet transmission decreases notably whereas in Fig. 4b the packet transmission time distributions for UDP traffic are presented. The time interval with the lowest UDP traffic is observed to be 1-5. However, no significant difference in the packet transmission frequency is observed in other time intervals.

Table 8 Conformity results of the distributions commonly used in modelling computer networks to the modelled network

Protocol	Distribution	Parameter 1	Parameter 2	μ	K-S	Chi-Squared	Distribution (Garsva et. al) [13]
TCP	Lognormal	2.9991 (σ)	-	7.9088	0.11763	10.494	Pareto 2
	Pareto 2	0.33705 (α)	216.48 (β)	-	0.14342	7.2422	Weibull
	Weibull	0.40333 (α)	9326.8 (β)	-	0.14348	27.267	Lognormal
	Generalized Pareto	0.98227 (k)	10078.0 (σ)	-2839.6	0.23273	23.172	Exponential
	Generalized Extreme Value	0.98314 (k)	9568.8 (σ)	3508.4	0.25293	29.091	Gamma
UDP	Generalized Pareto	0.98035 (k)	3406.1 (σ)	603.55	0.0896	-	Pareto 2
	Generalized Extreme Value	0.98133 (k)	3231.7 (σ)	2748.6	0.09098	4.0479	Weibull
	Pareto 2	1.13 (α)	5432.4 (β)	-	0.09439	7.5846	Lognormal
	Weibull	0.71509 (α)	8775.1 (β)		0.11812	7.4368	Gamma
	Lognormal	1.8545 (σ)	-	8.4215	0.12859	10.855	Exponential



(a)

(b)

Figure 5 Pdf of packet inter arrival time graphics according to distributions

Table 9 Detection of appropriate distributions in the lower tail region

Distribution	Parameter 1	Parameter 2	μ	K-S	Chi-Squared	Distribution (Bhattacharjee and Nandi) [9]
Generalized Pareto	0,9891 (k)	613,04 (σ)	150,62	0.1481	-	Log-Normal (Rank 1)
Lognormal	1,5946 (σ)	-	6,9421	0,22836	11,987	Pareto (Rank 2)

In the literature, the distributions commonly used for modelling computer networks are listed [13] as Weibull, Pareto, Gamma, Exponential and Lognormal. By taking these distributions into consideration, the distributions that best represent the modelled general traffic (incoming and outgoing) are listed in Tab. 8. Since Generalized Extreme Value and Generalized Pareto distributions model data better than Exponential and Gamma distributions, the results of these distributions are not included in the table. Unlike the results in Tab. 7, in terms of general traffic, the distribution that best modelled the network TCP protocol was the lognormal distribution, and the distribution that best modelled the UDP protocol was the Generalized Pareto distribution. Whereas in the study conducted by Garsva et al. [13], Pareto 2 distribution was the fit distributional for modelling an academic network for both TCP and UDP protocol. The graphs for the distributions are presented in detail in Fig. 5.

In our study, the statistical modelling of an academic network is performed comparatively by reference to the study conducted by Garsva et al. [13]. It is possible to model all campus networks around the world with the same physical conditions and working hours as the network modelled by the study performed. In addition, fit distributions to model the data collected by Bhattacharjee and Nandi [9] between 16:15-17:30 hours for modelling an academic network are identified. In the result of the study, it is concluded that Log Normal distribution models the data better than Pareto distribution in the specified time interval. The data collected in our study is the daily data collected from 15:00 to 00:00 on 25.04.2018 and from 00:00 to 15:00 on 26.04.2018. In order to compare with the study conducted by Bhattacharjee and Nandi [9], data section of the 16:15-17:30 interval is selected and the fit distributions for this section are obtained. The results are shown in Tab. 9 in detail.

Bhattacharjee and Nandi stated that lognormal distribution in lower tail regions showed better compatibility than Pareto distribution [9]. However, according to the table obtained in our study, Generalized Pareto distribution (0.1481) produced more consistent results than lognormal distribution (0.22836).

6 CONCLUSION

This study is conducted with the aim of modelling a campus network with periodic behaviour with simulators, by using real time data collected from Siirt University Campus Network. The data collected are analysed statistically on both incoming/outgoing traffic and port basis. The results are presented in comparison with the results of prior studies conducted by Garsva et al. [13] with the aim of analysing a campus network. The comparison results with the study conducted by Bhattacharjee and Nandi [9] on academic networks are also presented.

When the obtained results are analysed on the basis of general traffic, the distribution that best models the transmission time of TCP packets in the network is the Lognormal distribution and the distribution that best models the arrival time of UDP packets is the Generalized Pareto distribution. In terms of daily periods, Pareto 2, Weibull, Logistic, Lognormal and Generalized Extreme Value distributions are found to be the fit distributions. The results are given in detail along with the network architecture to enable modelling with simulators.

7 REFERENCES

[1] Ting, L. (2014). *Background Traffic Modeling for Large-Scale Network Simulation*. PhD dissertation, Dept. Elect. and Comp. Eng. Florida Inernational University., Florida, USA. https://doi.org//10.25148/etd.FI14040803

[2] Venkatesh, K. V. & Amin, V. (2008). Evaluating Distributed Systems: Does Background Traffic matter? *USENIX Annual Technology Conference*, 227-240.

[3] Venkatesh, K. V. & Amin, V. (2009). Swing: Realistic and Responsive Network Traffic Generation. *IEEE/ACM Transactions on Networking, 17*(3), 712-725. https://doi.org//10.1109/TNET.2009.2020830

[4] Nahum, E. M., Roşu, M. C., Seshan, S., & Almeida, J. (2001). The Effects of Wide-Area Conditions on WWW Server Performance. *ACM Sigmetrics Performance Evaluation Review*, 257-267, Cambridge, USA. https://doi.org//10.1145/384268.378790

[5] Eylen, T. & Bazlamaçcı, C. F. (2015). One-Way Active Delay Measurement with Error Bounds. *IEEE Transactions on Instrumentation and Measurement, 64*(12), 3476-3489. https://doi.org//10.1109/tim.2015.2469431

[6] Paxson, V. & Floyd, S. (1995). Wide Area Traffic: The Failure of Poisson Modeling. *IEEE/ACM Transactions on Networking*, *3*(3), 226-244. https://doi.org//10.1109/90.392383

[7] Barakat, C., Thiran, P., Iannaccone, G., Diot, C., & Owezarski, P. (2003). Modeling internet backbone traffic at the flow level. *IEEE Transactions on Signal Processing*, *51*(8), 2111-2124. https://doi.org//10.1109/TSP.2003.814521

[8] Fras, M., Mohorko, J., & Cucej, Z. (2008). A new goodness of fit test forhistograms regarding network traffic packet size process. *IEEE International Conference on Advanced Technologies for Communications (ATC)*, 345-348, Hanoi, Vietnam. https://doi.org//10.1109/ATC.2008.4760593

[9] Bhattacharjee, A. & Nandi, S. (2010). Statistical analysis of network traffic inter-arrival. *IEEE 12th International Conference on Advanced Communication Technology (ICACT)*, 1052-1057, Phoenix Park, South Korea.

[10] Garsva, E., Paulauskas, N., Grazulevicius, G., & Gulbinovic, L. (2012). Academic Computer Network Traffic Statistical Analysis. *IEEE 2nd Baltic Congress on Future Internet Communications (BCFIC)*, 100-105, Vilnius, Lithuania. https://doi.org//10.1109/BCFIC.2012.6217987

[11] Garsva, E., Paulauskas, N., & Grazulevicius, G. (2015). Packet size distribution tendencies in computer network flows. *2015 Open Conference of Electrical, Electronic and Information Sciences (eStream)*, 1-6, Vilnius, Lithuania. https://doi.org//10.1109/eStream.2015.7119483

[12] Siirt University, (2017, November 23). Bant Genişliği. Retrieved from http://bidb.siirt.edu.tr/detay/bant-genisligi/973345.html

[13] Garsva, E., Paulauskas, N., Grazulevicius, G., & Gulbinovic, L. (2014). Packet Inter Arrival Time Distribution in Academic Computer Network. *Elektronika ir Elektrotechnika*, *20*(3), 87-90. https://doi.org//10.5755/j01.eee.20.3.6683

[14] Olewuezi, N. P. (2011). Note on the comparison of some outlier labeling techniques. *Journal of Mathematics and Statistics*, *7*(4), 353-355. https://doi.org//10.3844/jmssp.2011.353.355

[15] Ovla, H. D. & Taşdelen, B. (2012). Aykırı Değer Yöntemi. *Mersin Üniversitesi Sağlık Bilimleri Dergisi*, *5*(3), 1-8.

[16] Seo, S. (2006). *A Review and Comparison of Methods for Detecting Outliers in Univariate Data Sets.* PhD dissertation, Graduate Faculty of Public Health Pittsburgh University, Pittsburgh, USA.

[17] Cotton, M., Eggert, L., Mankin, A., & Westerlund, M. (2008, August 21). IANA Allocation Guidelines for TCP and UDP Port Numbers draft-cotton-tsvwg-iana-ports-00. Retrieved from https://tools.ietf.org/id/draft-cotton-tsvwg-iana-ports-00.html

[18] Lall, A. (2015). Data streaming algorithms for the Kolmogorov-Smirnov test. *2015 IEEE International Conference on Big Data*, 95-104, Santa Clara, USA. https://doi.org//10.1109/BigData.2015.7363746

[19] Kashef, S. S. & Azmi, P. (2013). Cumulant based Kolmogorov-Smirnov spectrum sensing. *2013 IEEE 11th Malaysia International Conference on Communications (MICC)*, 104-109, Kuala Lumpur, Malaysia. https://doi.org//10.1109/MICC.2013.6805808

[20] Bircan, H., Karagöz, Y., & Kasapoğlu, Y. (2003). Ki-Kare ve Kolmogorov Smirnov Uygunluk Testlerinin Simülasyon ile Elde Edilen Veriler Üzerinde Karşılaştırılması. *Çukurova Üniversitesi. İktisadi ve İdari Bilimler Dergisi*, *4*(1), 69-80.

[21] Sharma, D. (2015). Implementing Chi-Square method and even mirroring for cryptography of speech signal using Matlab. *1st International Conference on Next Generation Computing Technologies (*NGCT), 3, 94-397, Dehradun, India. https://doi.org//10.1109/NGCT.2015.7375148

Contact information:

Mehmet Ali ALTUNCU, MSc
Kocaeli University,
Computer Engineering Department,
41100, Umuttepe, Kocaeli, Turkey
E-mail: mehmetali.altuncu@kocaeli.edu.tr

Fidan KAYA GÜLAĞIZ, PhD
Kocaeli University,
Computer Engineering Department,
41100, Umuttepe, Kocaeli, Turkey
E-mail: fidan.kaya@kocaeli.edu.tr

Hikmetcan ÖZCAN, MSc
(Corresponding author)
Kocaeli University,
Computer Engineering Department,
41100, Umuttepe, Kocaeli, Turkey
E-mail: hikmetcan.ozcan@kocaeli.edu.tr

Sümeyya İLKİN, MSc
Kocaeli University,
Computer Engineering Department,
41100, Umuttepe, Kocaeli, Turkey
E-mail: sumeyya.ilkin@kocaeli.edu.tr

Suhap ŞAHİN, Associate Professor
Kocaeli University,
Computer Engineering Department,
41100, Umuttepe, Kocaeli, Turkey
E-mail: suhapsahin@kocaeli.edu.tr