Mathematical Theory and Modeling
ISSN 2224-5804 (Paper) ISSN 2225-0522 (Online)
Vol.8, No.6, 2018

MATHEMATICAL ANALYSIS OF TRAFFIC FLOW BEHAVIOUR OF THE INTERNET AT GHANA TECHNOLOGY UNIVERSITY COLLEGE

Emmanuel Kwame Mensah
Department of Computer and Telecommunication Engineering, Faculty of Engineering, Ghana Technology University College PMB 100 Accra-North, Ghana
Email: emensah@gtuc.edu.gh

Abstract
This paper analyses the internet traffic flows behaviour at Ghana Technology University College campus network. Internet traffic is a flow of data across the internet. This flow exhibit certain behaviour in accordance with Probability distribution. Statistical analysis was performed to understand the characteristic of traffic population. Empirical cumulative distribution process (CDF) together with essential statistical parameters were benchmarked. Goodness of Fit (GOF) test using Anderson-Darling (AD) estimation method was applied to establish the best probability distribution model which describes the situation alongside with Probability plot. The N-Flow traffic model was used to determine the burstiness of internet traffic in finite session. The Empirical and Theoretical CDF graph was used to determine the discrepancies' between them. At the end it was discovered that the data best fit normal distribution
Keywords: internet traffic, statistical analysis, cumulative distribution function, maximum probability plot, anderson-darling, normal distribution, poisson distribution, exponential distribution.

1. Introduction

One of the most important areas in human life is communication and engineering. It has helped to the growth and make life more comfortable across human race. Due the technological development of the world, communication and engineering has given birth to Telecommunication engineering which has its own traffic problems. Telecommunications traffic engineering (i.e. Tele -traffic engineering) is the application of traffic engineering theory to telecommunications. Tele-traffic engineers use their cognition of statistics including queuing theory, the nature of traffic, their practical fashion model, and their mensuration and computer simulation to make predictions and to programme telecommunication networks such as a telephone network or the Internet. These tools and knowledge help provide reliable service at lower cost. At a broader level, finger enabled development of electronic commerce department. Anyone with computer and internet approach will be in a spatial relation to become merchant and reaching out to customer across the globe, and any consumer will be able to shop the world for goods and services. In Ghana Technology University college internet usage is crucial to both academia and non-academic to the success of the institution, the nation and the world at large. Due to this reason, this paper seek to analysed traffic behaviour of the internet on campus. Moreover, determining which probability distribution best describes the behaviour is of great concern since this will help in a long way to make better plans for the academic growth of the university to make human life simpler. Moreover, it presents an analytic thinking of internet traffic in a campus meshwork for both inbound and outbound traffic flow. Internet traffic data was collected at the backbones network for 2 weeks. Statistical analysis was performed to understand the characteristic of traffic universe. Empirical cumulative distribution function (CDF) is evaluated and important parameters characterized. Anderson-Darling (AD) estimation method and probability plot were used to identify the best fitted distribution model.

In Networking, the determination of packet arrivals rate is used to determine traffic congestion. This in a long way either affect the companies and institutions positively or negatively. In most cases this congestion happens because the determination process has failed due to wrong use of traffic models. In this research, we seek to unearth which distribution best describe traffic behaviour in order to help solve the problem of congestion on campus from mathematical point of view.

2. Literature Review

This section discusses various probability distributions and its properties.

2.1 Probability Distributions

A probability distribution describes the set of all possible values in a range within which the probability of each event within a sample space occurs. The distribution is grouped into two depending the nature of the random value describing the distribution. The random variable in nature is said to be discrete or continuous. Due to this reason, the two groups of probability distribution is named; discrete and continuous probability distributions (Montgomery & Runger 2014).

2.2 Discrete Probability Distribution

2.2.1Poisson Probability Distribution

Let X be the random variable representing the number of occurrences of a Poisson experiment in some interval. Then, the probability distribution function for X is

$$\mathrm{p}(x) = \frac{(\lambda)^x e^{-\lambda}}{x!} \quad \forall x = 0,1,2,3,... \qquad \text{.......................................1}$$

and λ is the mean of the distribution[1].

2.2.2 Properties of Poisson Probability Distribution:

A random variable X has the Poisson probability distribution $\mathrm{P}(x)$ with parameter, λ then

$$E(X) = \text{Expected number of occurrences} = \lambda$$

and

$$Var(X) = \lambda$$

Furthermore, in Poisson distribution, mean =E(X) = =variance .

2.2.3 Poisson Process

Poisson process is a random process representing a discrete outcome of event taking place over continuous intervals of time or region. Examples of Poisson processes include:

- the arrival of telephone calls at a switchboard,
- the passing cars of an electric checking device.

All these examples involve a discrete random event. At any given small period of time, the probability that the event occurs is small; however, over a long time, the number of occurrence is large. Poisson distribution plays an extremely important role in science and engineering, since it represents an appropriate probabilistic model for a large number of observational phenomena (Gupta & Guttman 2013).

Mathematically, Poisson process can be described by the following formula:

$$\mathrm{p}(x, \lambda t) = \frac{(\lambda t)^x e^{-\lambda t}}{x!} \quad \forall x = 0,1,2,3,... \qquad \text{..2}$$

where, is the average number of outcomes per unit time or region. Hence, λt represents the number of outcomes.

Over here the cumulative distribution function (cdf) is given as

$$P_X(x;\lambda,t)=\sum_{x=0}^{n}\frac{(\lambda t)^x e^{-\lambda T}}{x!} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots 3$$

where $x \in [0,n]$, $,Time,$ t, and λ is mean of the event.

2.2.4 Properties of Poisson Process:

1. X represent the number of occurrences in a continuous interval.
   λ represent expected value of occurrences in this interval belong to a set of real numbers.
2. The probability of an occurrence is the same for any two intervals of equal length. The expected value of occurrences in an interval is proportional to the length of this interval.
3. The occurrence or nonoccurrence in any interval is independent of the occurrence or nonoccurrence in any other interval.
4. The probability of two or more occurrences in a very small interval is close to 0 (Gupta & Guttman 2013).

2.3 Continuous Probability Distribution

2.3.1 Normal Distribution

One of the most widely used probability distribution in most of the application fields is the normal distribution. A random variable X is said to be a normal distribution if it has the probability density function (pdf) given as

$$f(x;\lambda,\sigma^2)=\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-\lambda)^2}{2\sigma^2}\right) \quad \ldots\ldots\ldots\ldots 4$$

with parameters, mean λ $(-\infty < \lambda < \infty)$ and standard deviation, $\sigma > 0$ written as X ~ N (μ, σ^2). A normal distribution with $\mu = 0$ and $\sigma = 1$ is called the standard normal distribution. A random variable having the standard normal distribution is denoted by Z (Gupta & Guttman 2013).

2.3.2Properties of Normal Distribution

The normal random variable X has the following properties:

- if X has the normal distribution with parameters λ and σ then Y = α X + β is normally distributed with parameters $\alpha\lambda$ + β and $\alpha\sigma$. theoretically,
  $Z=(x-\lambda)/\sigma$ is a standard normal distribution.
- the normal pdf is a bell-shaped curve that is symmetric about λ and that attains its maximum value of:

$$\frac{1}{\sqrt{2\pi}\sigma}=\frac{0.399}{\sigma}$$

at x = λ . This is seen in figure 1 below.

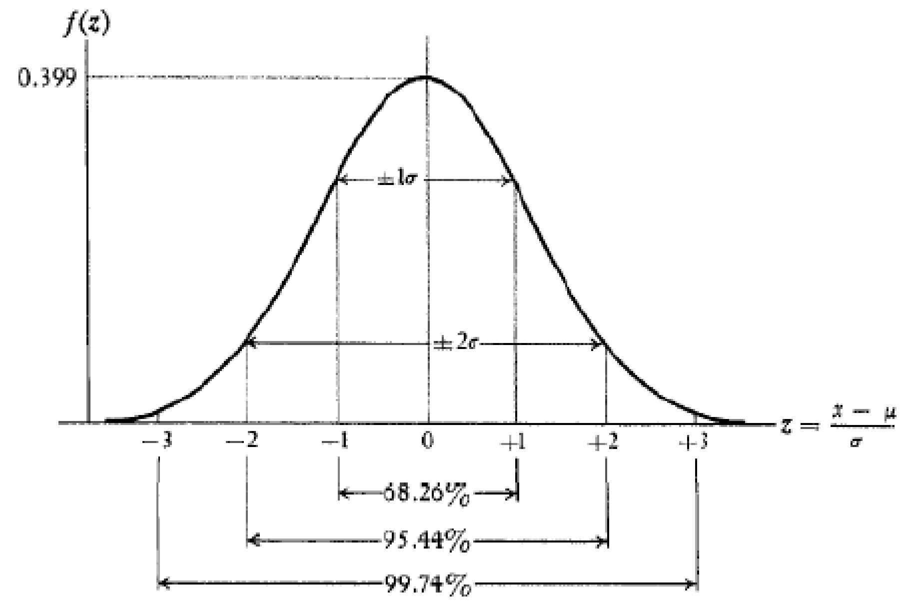

Figure 1 Normal Distribution

This shape is determined by the values of the mean and standard deviation.

- 68.26% of the total area bounded by the curve lies between $\lambda \mp 1$.
- 95.44% is between $\lambda \mp 2$.
- 99.74% is between $\lambda \mp 3$.
- The expected value is $E(X) = \lambda$.
- The variance is $Var(X) = \mathrm{E}(\mathrm{X} - \lambda)^2 = \sigma^2$
- The moment generating function (mgf) is given as

$$M_X(t) = \exp\left(\lambda t + \frac{\sigma^2 t^2}{2}\right) \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad \ldots.5$$

- The cumulative distribution function (cdf) for the normal distribution random variable X is

$$F_X(x;\lambda,\sigma^2) = \Pr\left(X \le x\right) = \frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{x} \exp\left(-\frac{(\mathrm{y}-\lambda)^2}{2\sigma^2}\right)dy \quad \ldots\ldots\ldots.6$$

which can be rewritten as

$$F_X(x;\lambda,\sigma^2) = \Pr\left(\frac{X-\lambda}{\sigma} < \frac{x-\lambda}{\sigma}\right) = \Pr\left(Z < \frac{x-\lambda}{\sigma}\right) = \Phi\left(\frac{x-\lambda}{\sigma}\right) \ldots\ldots \quad 7$$

where $\Phi(.)$ is the cdf for the standard normal random variable:

$$\Phi(z) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{z} \exp\left(-\frac{y^2}{2}\right)dy$$

- The $100(1-\alpha)\%$ percentile of the normal variable X is defined by the formula

$$x_\alpha = \lambda + \sigma z_\alpha$$

Where z_α is the $100(1-\alpha)\%$ percentile of the standard normal random variable Z .

- $z_\alpha = -z_{1-\alpha}$.
- if $X_1 \sim N(\lambda_1,\sigma_1^{\,2})$ and $X_2 \sim N(\lambda_2,\sigma_2^{\,2})$ are independent then,

$$aX_1 + bX_2 + c \sim N(a\lambda_1 + b\lambda_2 + c, a^2\sigma_1^{\,2} + b^2\sigma_2^{\,2}).$$

- if $X_i \sim N(\lambda,\sigma^2), \forall \mathrm{i} = 1,2,\ldots,\mathrm{n}$ are independent identical distribution then

$\overline{X} \sim N\left(\lambda, \sigma^2/n\right)$ where $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is the sample mean.

- if $X_i \sim N(\lambda, \sigma^2), \forall \mathrm{i} = 1, 2, ..., \mathrm{n}$ are independent identical distribution then $\overline{X}$ and S^2 are independently distributed where

$$S = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 \quad \text{is the sample variance.} \quad \ldots\ldots\ldots\ldots..8$$

(Gupta & Guttman 2013) .

2.4 Exponential Distribution

A random variable X is said to be exponential distributed if it has the probability density function (pdf) defined as

$$f(x;\lambda) = \begin{cases} \lambda \exp(-\lambda x) & x \geq 0, \lambda > 0 \\ 0 & x < 0 \end{cases} \quad \ldots\ldots\ldots\ldots\ldots\ldots..9$$

with parameter λ which represents the mean number of events per unit time. Hence, it is written as $X \sim Exp(\lambda)$. An example of λ is the rate of arrivals or the rate of failure (Gupta & Guttman 2013).

2.4.1Properties of Exponential random variable

An exponential random variable X has the following properties:

- It is closely related to the Poisson distribution – if X describes for example the time between two failures then the number of failures per unit time has the Poisson distribution with parameter λ
- The cdf is

$$F_X(x:\lambda) = \lambda\int_0^x \exp(-\lambda y)dy = 1 - \exp(-\lambda x) \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad .10$$

- The 100(1 − α)% percentile is

$$x_\alpha = -\frac{1}{\lambda}Log(\alpha) \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad \ldots\ldots11$$

- The expected value is:

$$E(X) = \frac{1}{\lambda}$$

- The variance is

$$Var(X) = \frac{1}{\lambda^2}$$

-
- The moment generating function (mgf) is

$$M_X(t) = \frac{\lambda}{\lambda - t} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \ldots12$$

(Gupta & Guttman 2013).

3.0 METHODOLOGY

3.1 Internet Data Traffic Modelling

A widely used probability distribution that has comes out with the idea that events happen randomly in a period of time (or space) is the Poisson distribution. It is one of the most widely used and oldest

traffic model. The memoryless Poisson distribution is the most important model used for analysing traffic in traditional telephony networks.
The two key assumptions of the Poisson model are:

- the number of sources is infinite and independent
- the traffic interarrival pattern is random and exponential

A random variable X that is equal to number of events in Poisson process is a Poisson random variable if it's probability density function (pdf) and cumulative distribution function (cdf) which is given as defined in equations (2) and (3) where λ, the expected number of internet data transmitted per unit time, t .

Determination of Expected Queuing length Using Mathematical Modelling

Queues always arise in many situations. In this paper, consideration was made of situation of arrival and departure of internet data in and out Ghana Technology University College router. Consider a situation where a data arrives in a queue in time δt is $\lambda\delta t$ and probability that a data depart from a queue in time, δt is $\mu\delta t$. Let $P_i(t)$ denotes the probability of *ith* data in the queue at time, t $\forall i = 0,1,2,\ldots$.

Then,

$$P_0(t+\delta t) = P_0(t)\times \Pr obability\ of\ no\ arrival\ in\ \delta t + P_1(t)\times \Pr obability\ of\ one\ departure\ in\ \delta t$$

The probability of departure in δt is $\mu\delta t$. Assume that δt is sufficiently small so that only one departure is possible. Since the probability of an arrival is $\lambda\delta t$ then the probability of no arrival is $1-\lambda\delta t$

Hence, $$P_0(t+\delta t) = P_0(t)(1-\lambda\delta \mathrm{t}) + \mathrm{P}_1(t)\mu\delta t$$

$$\frac{P_0(t+\delta t)-P_0(t)}{\delta t} = \mu P_1(t) - \lambda P_0(t)$$

As $\delta t \to 0$, $$\frac{dP_0(t)}{dt} = \mu P_1(t) - \lambda P_0(t) \quad \ldots\ldots 13$$

Similarly.

$$\begin{aligned} P_x(t+\delta t) = P_x(\mathrm{t})\times \Pr obability\ of\ no\ arrival\ or\ departure\ in\ \delta t + \\ P_{x+1}(\mathrm{t})\times \Pr obability\ of\ one\ departure\ in\ \delta t + \\ P_{x-1}(\mathrm{t})\times \Pr obability\ of\ one\ arrival\ in\ \delta t \end{aligned}$$

Which gives,

$$\begin{aligned} P_x(t+\delta t) &= P_x(\mathrm{t})(1-\lambda\delta\,\mathrm{t})(1-\mu\delta\,\mathrm{t}) + P_{x+1}(\mathrm{t})\mu\delta\,\mathrm{t} + P_{x-1}(\mathrm{t})\lambda\delta\,\mathrm{t} \\ &= \mathrm{P}_x(t)[1-\lambda\delta\,\mathrm{t} - \mu\delta\,\mathrm{t} + \lambda\mu(\delta t)^2] + P_{x+1}(\mathrm{t})\mu\delta\,\mathrm{t} + P_{x-1}(\mathrm{t})\lambda\delta\,\mathrm{t} \end{aligned}$$

Rearranging,

$$\frac{P_x(t+\delta t)-P_x(\mathrm{t})}{\delta t} = -(\lambda+\mu-\lambda\mu\delta t)P_x(t) + \mu P_{x+1}(t) + \lambda P_{x-1}(t)$$

As $\delta t \to 0$,

$$\frac{dP_x(t)}{dt} = -(\lambda+\mu)P_x(t) + \mu P_{x+1}(t) + \lambda P_{x-1}(t)$$
…………………………………..14

Condition I (i.e. $\mu > \lambda$)

If $\mu > \lambda$ then the values of $P_0(t), P_1(t),$ etc tend to particular value as $t \to \infty$. This is known as the steady state. Hence steady state of the system has been achieved. Also, at equilibrium, $dP_x(t)/dt = 0$ $\forall$ values of x and equations 13 and 14 becomes

$$\mu P_1 = \lambda P_0 \quad \text{……………………………………..15}$$

and

$$\mu(P_{x+1} - P_x) = \lambda(P_x - P_{x-1}) \quad \text{………………………..16}$$

where $P_0(t), P_1(t),$ etc denotes the steady state values.

Moreover, Equations 15 and 16 can be used to find $P_0, P_1, \mathrm{P}_3,$ etc since equation 16 can be written in descending order we h,

$$\left.\begin{array}{l} \lambda(P_x - P_{x-1}) = \mu(P_{x+1} - P_x) \\ \lambda(P_{x-1} - P_{x-2}) = \mu(P_x - P_{x-1}) \\ \quad \bullet \qquad\qquad \bullet \\ \quad \bullet \qquad etc \qquad \bullet \\ \quad \bullet \qquad\qquad \bullet \\ \lambda(\mathrm{P}_1 - P_0) = \mu(P_2 - P_1) \\ \lambda P_0 = \mu P_1 \end{array}\right\} \quad \text{………………… ….17}$$

Adding equations in equation 17 up vertically, most of the terms cancel out, hence

$$\lambda P_x = \mu P_{x+1}$$

Therefore, the successive probabilities form a geometric progression with a common ratio of λ/μ and a first term, P_0. This statement confirms condition I which states for steady state $\mu > \lambda$. In general

$$P_x = P_0\left(\frac{\lambda}{\mu}\right)^x \quad \forall x = 0,1,2,3,\ldots \qquad \text{……………………18}$$

To find P_0, we use the condition that the sum of all probabilities under the sample space in 1. Hence,

$$P_0\left(1+\frac{\lambda}{\mu}+\left(\frac{\lambda}{\mu}\right)^2+\dots\right)=1$$

$$P_0\left(\frac{1}{1-\lambda/\mu}\right)=1$$

$$P_0=1-\left(\lambda/\mu\right)\dots\dots 19$$

Therefore,

$$P_x=\left(1-\lambda/\mu\right)\left(\frac{\lambda}{\mu}\right)^x \quad \forall x=0,1,2,3,\dots \dots\dots 20$$

The expected (average) queue length is defined by

$$E(X)=\sum_{x=0}^{\infty}xP_x$$

$$=\sum_{x=0}^{\infty}x\left(1-\lambda/\mu\right)\left(\frac{\lambda}{\mu}\right)^x$$

$$=\left(1-\lambda/\mu\right)\left(\frac{\lambda}{\mu}\right)\sum_{x=1}^{\infty}x\left(\frac{\lambda}{\mu}\right)^{x-1}$$

$$=\left(\frac{\lambda}{\mu}\right)\left(1-\lambda/\mu\right)\left[1+2\left(\frac{\lambda}{\mu}\right)+3\left(\frac{\lambda}{\mu}\right)^2+\dots\right]=\left(\frac{\lambda}{\mu}\right)\left(\frac{\mu-\lambda}{\mu}\right)\left(\frac{\mu}{\mu-\lambda}\right)^2$$

$$=\frac{\lambda}{\mu-\lambda}\dots\dots 21$$

Condition II (i.e. $\mu<\lambda$)

If $\mu<\lambda$, the system will become unstable and the queue length will increase indefinitely.

3.2Method of Data Analysis

In this section, data was analysed in two approaches i.e. experimental and theoretical approaches.

3.2.1 Experimental Approach

In the experimental approach, data was captured from Computer Department from Ghana Technology University College using Wire shark software. The data captured was arrival rate and departure rate of data packets. The data was exported to excel where it was analysis using hypothesis.

3.2.2 Goodness of Fit Technique:

Goodness of fit way of doing things means the method of examining how well a sample data agrees with a given distribution as it population. The goodness of fit technique are:

- Test of chi-square types
- Moment ratio technique
- Test based on correlation
- Test based on empirical distribution function.

- Maximum Likelihood Estimates (MLE)

Most of these test statistics suffer from serious limitation. In general, test of chi-square type have less power due to loss of information caused by grouping. The statistical distribution explanation of Chi-square statistics is large sample theory (it explains why some event works or happens the way it does). The higher moment usually under guessed a number and this fact prevent the use of moment ratio method been used in analysis and so would be the case with correlation type test.
According to Stephen (1974, 1977) several powerful studies have revealed Empirical Distribution Function (EDF) test to be the more powerful than other test of fit for a wide range of sample size into a given Probability distribution. Due to this reason, EDF approach will be used in this research to investigate whether the traffic pattern behaviour in Ghana Technology University college follows Poisson distribution or not. If not, feather investigation will done to unearth reasons why Poisson failed and detect which probability distribution best fit the traffic flow.

From theory, Poisson arrivals have two key characteristics: the interarrival times are both exponentially distributed, and independent. We discuss testing for each in turn.

Hypothesis

Over here hypothesis was done using probability plot. A probability plot is a graphical method that is used to evaluate whether a particular data fit a particular probability distribution. It estimates percentiles and compare with different sample distributions. Below is the flow chart for the hypothesis.

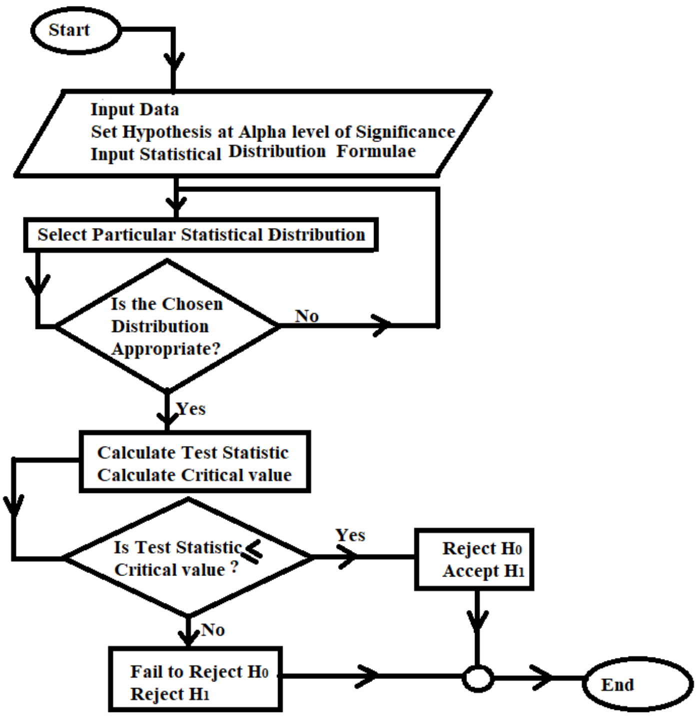

Figure 2 Flow Chart for Hypothesis

Justification for using Anderson-Darling Goodness -of -Fit Tests of Fitting a data into a Probability Distribution

In statistics testing, determining whether a particular data fit a particular probability distribution using hypothesis, the following tests Chi square test, Kolmogorov- Smirnow (K-S) test and Anderson-Darling test are widely used. Chi square test is used when data measurements are ordinal.

Kolmogorov- Smirnow test has an advantage over Chi square test. This is because with the K-S test, it is not important to divide observed data into intervals, hence the problem associated with small errors or small number of interval in the Chi-Square test would not be an issue with the K-S test.

Moreover, in fitting the data, we normally, propose observed theoretical and empirical cumulative distributive functions (i.e. CDF). In K-S test, both proposed theoretical and empirical CDF are relatively flat at the tail of probability distribution. Hence maximum deviation seldom occur in the tail of the distribution whereas the chi-square test, empirical frequencies at the tail must generally be grouped together. As a result both tests (i.e. K-S, Chi square) would not reveal any discrepancy between the empirical and theoretical frequencies at the tails of the proposed distribution. It is the Anderson-Darling (AD) test that reveals this discrepancies if available when comparing theoretical and empirical CDF's in both the upper and lower tails. Hence it is preferred in data fitting analysis than Chi square and K-S test.

Anderson-Darling (AD) test

The Anderson-Darling Goodness of fit test places more weight or discriminating powers at the tail of the distribution. This can be important when the tails of a selected theoretical distribution are of practical significance. Below is the procedure for applying the Anderson- Darling method:

1. Arrange the observed data in increasing order: $x_1, x_2, x_3, ..., \mathrm{x}_i, ..., x_n$ with x_n as the largest value.
2. Evaluate The CDF of the proposed distribution (in this case is Poisson) $F_X(P(X_i))$ at x_i for $\forall i = 1, 2, ..., n.$
3. Calculate the Anderson-Darling statistic

$$A^2 = -\sum_{i=1}^{n}[(2i-1)\{\ln F_X(\mathrm{P}(X_i)) - \ln[1 - F_X(\mathrm{P}(X_{n+1-i}))]\} / \mathrm{n}] - n \ldots\ldots \quad \ldots \quad .22$$

where $i \in [1, n]$, $P(X_i)$ is the Probability density is function for Poisson distribution of a random

variable, X and $F_X(\mathrm{P}(X_{N+1-i}))$ is cumulative probability function of a random variable, X .

4. Compute the adjusted test statistic A^* to account for the effect of sample size, n . The Adjustment depends on the selected form of the distribution. For exponential distribution, the test statistic is given as

$$\mathrm{A}^* = A^2\left(1 + \frac{0.6}{\sqrt{n}}\right) \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad ..23$$

Moreover, the test statistic for normal distribution is

$$A^* = A^2\left(1.0 + \frac{0.75}{n} + \frac{2.25}{n^2}\right) \ldots\ldots\ldots\ldots\ldots\ldots\ldots 24$$

where n is the sample size.

5. Select a significance level α and determine the corresponding critical value C_α for the appropriate distribution type. This Critical value C_α is read from Anderson-Darling statistical table under the exponential distribution section when dealing with exponential distribution. For normal distribution, critical value, c_α is given by

$$c_{\alpha} = a_{\alpha}\left(1.0 + \frac{b_1}{n} + \frac{b_2}{n^2}\right) \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots 25$$

where a_{α}, and $b_i \ \forall i = 0,1$ are values determine from critical values at significance level α of the AD

statistical table under Normal distribution.

6. For a given distribution, compare A^* with the appropriate critical value C_{α} . If $A^* < C_{\alpha}$ then the proposed distribution (i.e. Null hypothesis) is accepted at α significance level. Otherwise the alternate Hypothesis is accepted.

Theoretical Approach

This approach uses information based on statistical theory to fit a data into a model.

The *N*-Flow traffic model

The N-Flow model is a good model to represent real telecommunication traffic. N-Flow model with heavy-tailed distribution for ON durations produces self-similar traffic. In order to understand behaviour of the *N*-Flow model, consider the simplest case when $N = 1$. We will use the following to present 1-flow model.

$\overline{OFF}$ =mean time of OFF period

$\overline{n_p}$ =mean number of packets during ON period

$\lambda_p =$ peak transmission rate during ON period

$\overline{ON} = \overline{n_p}/\lambda_p$ =mean length of ON period

$\kappa =$ average transmission rate

$\rho = \kappa/\upsilon$

By definition of ON and OFF period

$$\kappa = \frac{\overline{ON}}{\overline{ON} + \overline{OFF}}\lambda_p \ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots 26$$

The 1-Flow model depends on four separate distributions. Each one governs a different sub-process which all together describe characteristics of the 1-Flow model. They are:

• SV: **Packet Service Time Distribution** with mean $1/\upsilon$. This distribution depends on packet size distribution and router speed.

• OFF: **OFF Time Distribution** with mean $\overline{OFF}$. This distribution depends on how bursts are generated and how often.

• ON**: ON Time Distribution** with mean $\overline{ON}$. This distribution depends on packet size distribution and nature of an application. For example, the ON-time distribution of file transmission depends on the distribution of file sizes. While the ON-time distribution of voice transmission depends on the talk period of a speaker between breathing pauses. $\overline{ON}$ can be computed from $\overline{n_p}/\lambda_p$

IN: **Inter-Packet Time Distribution** during a burst with mean $1/\lambda_p$. This distribution depends on how packets are generated and how often.

A useful shape parameter for describing a 1-Flow arrival process is the so-called Burstiness parameter, b is defined as follows.

$$b = \frac{\overline{ON}}{\overline{ON} + \overline{OFF}} = 1 - \frac{\kappa}{\lambda_p} \quad \text{.....................................}27$$

Equation (18) can be rewritten as

$$\kappa = \lambda_p (1 - b) \quad \text{... } 28$$

The value of b varies over a range of [0, 1] inclusively. If b = 0, then $\overline{OFF}$ = 0 which means that bursts abut each other and packets are continuously transmitted with rate κ . The 1-Flow (ON/OFF) process is then reduced to a renewal process. On the other hand, if b = 1, then $\overline{ON}$ = 0, which means that all packets in a burst are transmitted simultaneously with infinite rate which is a self-similar process, λ_p = ∞. The 1-Flow (ON/OFF) process becomes a bulk arrival process. When packets are produced by more than one ON/OFF source (i.e., N > 1), an effective arrival rate, λ , of the *N*-Flow model is the sum of individual mean rates, i.e., $\lambda = N\kappa$ (Siriwong, Lipsky, & Ammar 2007)

DATA ANALYSIS

Over here the analysis of the investigation of traffic behaviour is presented in two different ways of which the methods (i.e. experimental and theoretical) are interrelated.

Experimental Analysis

Hypothesis test for Anderson-Darling Goodness -of -Fit Test

First Hypothesis

H_0 : Population from which the interarrival sample is drawn is Exponentially Distributed.

H_1 : Population from which the interarrival sample is drawn is not Exponentially Distributed

at α = 5% level of significance

Test Statistic:

Minitab uses the Anderson-Darling statistic to calculate the p-value. Over here, the P-value play the role of test statistic in decision taking. Below is figure 3 illustrate probability plot of interarrival packets which gives information about the P-value. Brown lines are theoretical lines whereas blue is empirical lines.

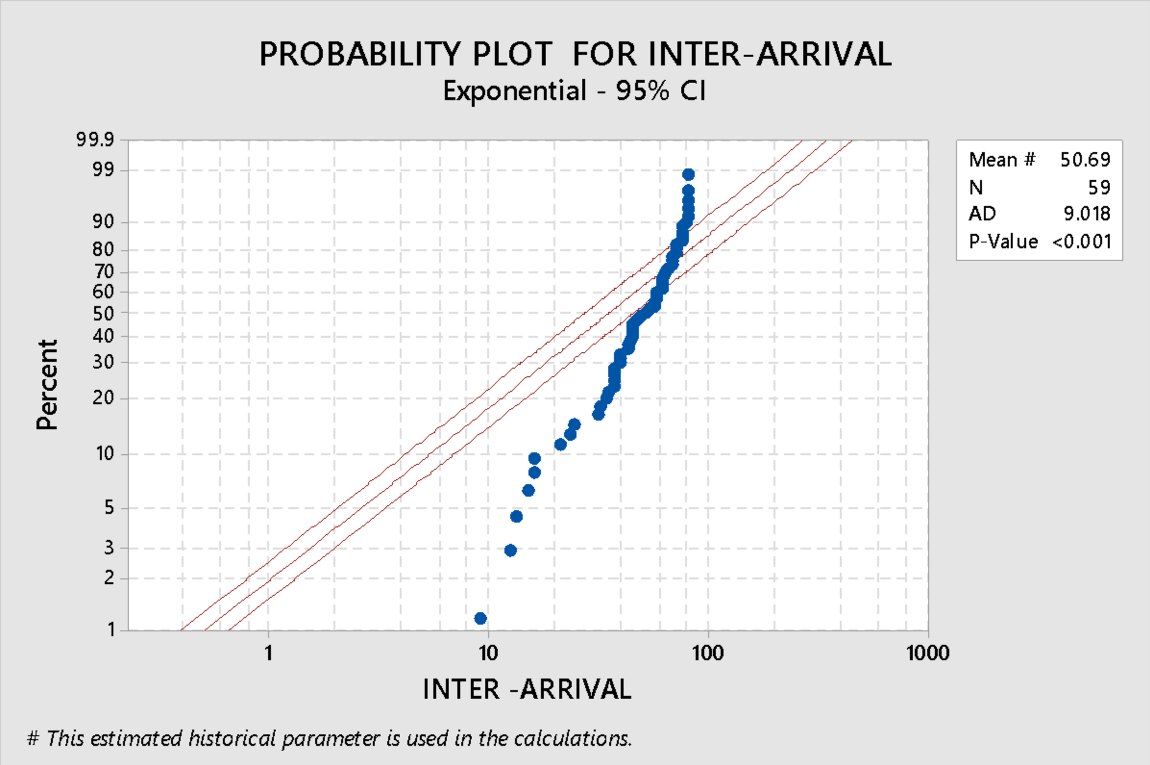

Figure 3: Probability Plot of Inter- arrival Packets for Exponential

Critical Value

For Minitab, Critical Value =level of significance, α =5%

Decision

Since test statistic is less the critical value i.e. 0.001< 0.05. Hence we reject H_0 and accept H_1. Moreover, the P-value is a probability that measures the evidence against the null hypothesis. Smaller P-values provide stronger evidence against the null hypothesis.

Conclusion of the first Hypothesis

From Figure 3, theoretically larger values of the Anderson-Darling statistic indicate that the data do not follow an exponential distribution. Hence Anderson-Darling statistic of 9.018 is an indication that the data do not follow exponential distribution. Moreover, since the Empirical data points does not lies along the straight line, the proposed Empirical data point does not fit Exponential distribution. Since the interarrival packets are not exponentially distributed. Therefore, the population from which the interarrival sample was drawn is not Exponentially Distributed

Second Hypothesis

Hypothesis:

H_0 : Population from which the inter-departure sample is drawn is Exponentially Distributed.

H_1 : Population from which the inter- departure sample is drawn is not Exponentially Distributed

at α = 5% level of significance

Test Statistic:

Minitab uses the Anderson-Darling statistic to calculate the p-value. The P-value play the role of test statistic in decision taking. Figure 4 below is the probability plot of inter-departure packets and it gives information about the P-value.

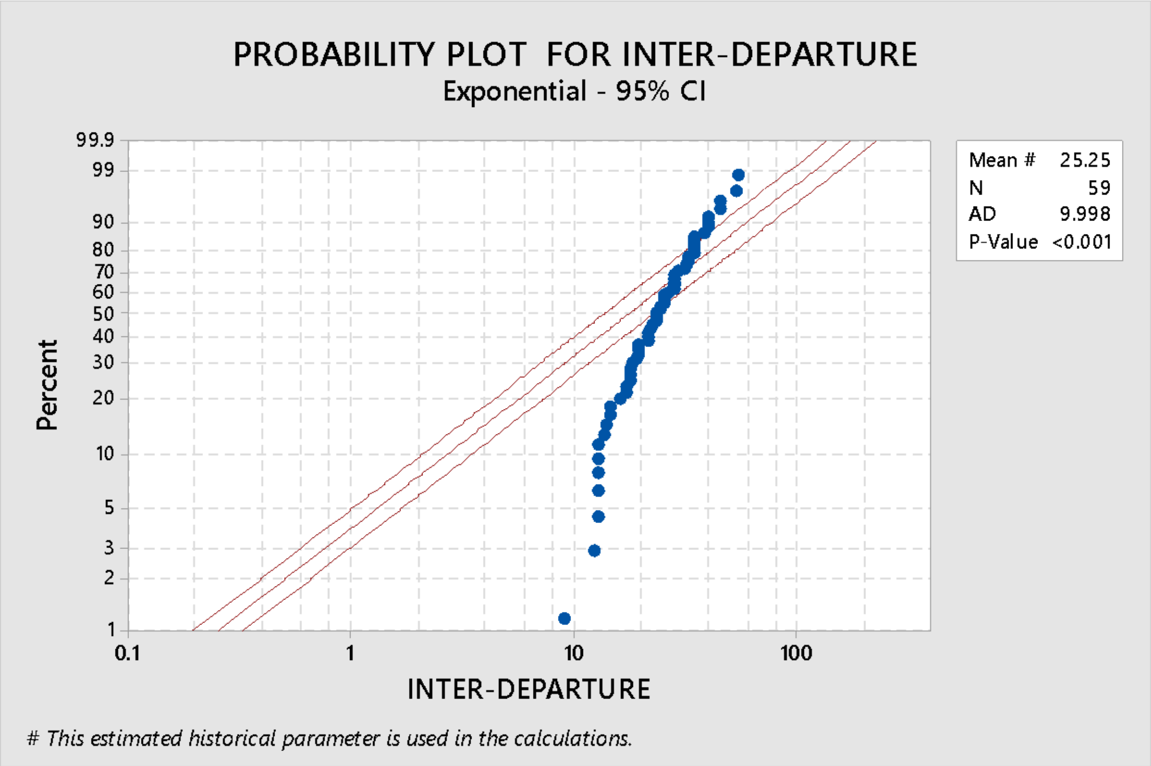

Figure 4: Probability Plot of Inter-departure Packets for Exponential

Critical Value

Critical Value Table =level of significance, α =5%

Decision

Since test statistic is less the critical value i.e. 0.001< 0.05. Hence we reject H_0 and accept H_1. This decision confirms the information from the graph in figure 4 since the empirical sample points cut cross the theoretical limits. Moreover, the P-value is a probability that measures the evidence against the null hypothesis. Smaller P-values provide stronger evidence against the null hypothesis.

Conclusion of second Hypothesis

From Figure 4, theoretically larger values of the Anderson-Darling statistic indicate that the data do not follow an exponential distribution. Hence Anderson-Darling statistic of 9.998 is an indication that the data do not follow exponential distribution. Moreover, since the Empirical data points does not lies along the straight line, the proposed Empirical data point does not fit Exponential distribution. Since the inter-departure packets are not exponentially distributed.

Justification why Poisson model failed

Poisson distribution have two key characteristics: the interarrival and inter-departure times are both exponentially distributed, and independent. Since from first and second hypothesis, the interarrival and inter-departure are not exponentially distributed, we therefore conclude that the Population from which inter arrival and inter-departure rate were drawn from is not Poisson distributed. Hence Poisson has failed.

Third Hypothesis

Hypothesis:

H_0 : Population from which the interarrival sample was drawn is Normally Distributed.

H_1 : Population from which the interarrival sample is drawn is not Normally Distributed

at α = 5% level of significance

Test Statistic:

Figure 5 below is the probability plot of interarrival packets which gives information about the P-value

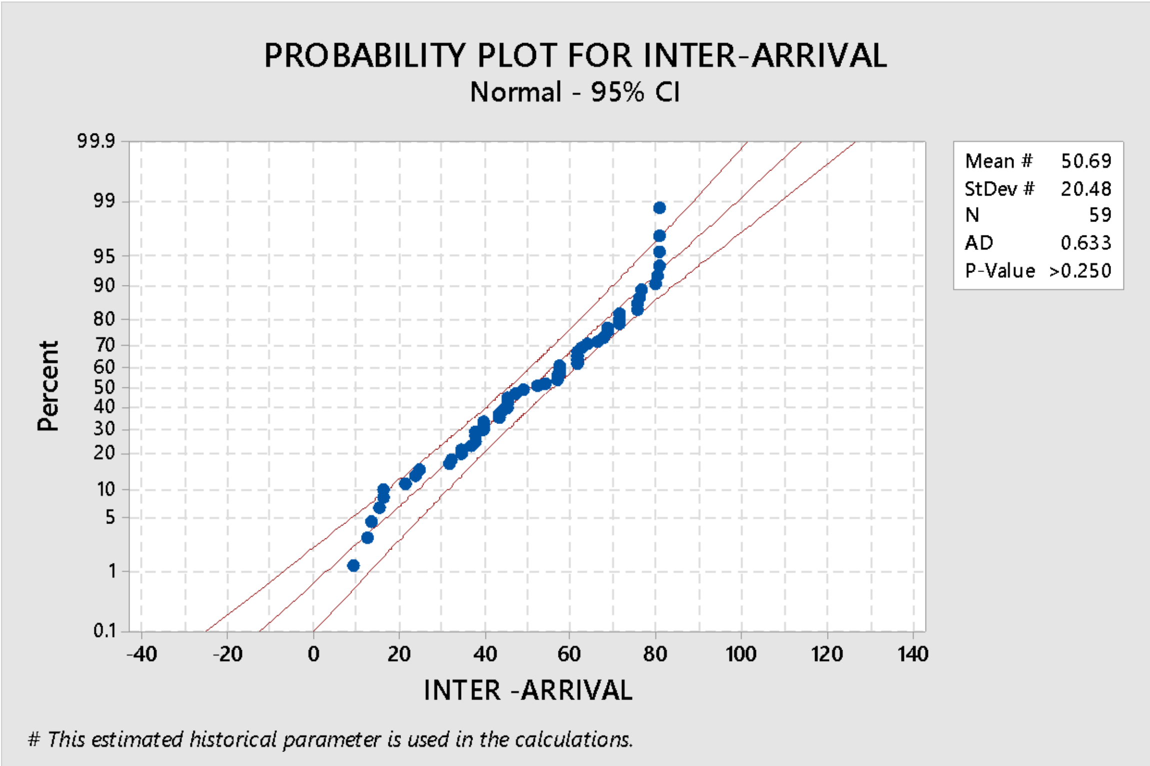

Figure 5: Probability Plot of Inter- arrival Packets for Normal

Critical Value

For Minitab, Critical Value =level of significance, α =5%

Decision

Since test statistic is greater than the critical value i.e. 0.250 > 0.05. Hence, we fail reject H_0 .

Conclusion

We therefore conclude that the population from which the interarrival sample was drawn is Normally Distributed. This evidence is also clear form the probability plot in Fig.5 since the empirical sample point lies with the theoretical range.

Fourth Hypothesis:

H_0 : Population from which the departure sample is drawn is Normally Distributed.

H_1 : Population from which the departure sample is drawn is not Normally Distributed

at α = 5% level of significance

Test Statistic:

Figure 6 below is the probability plot of inter-departure packets which gives information about the P-value

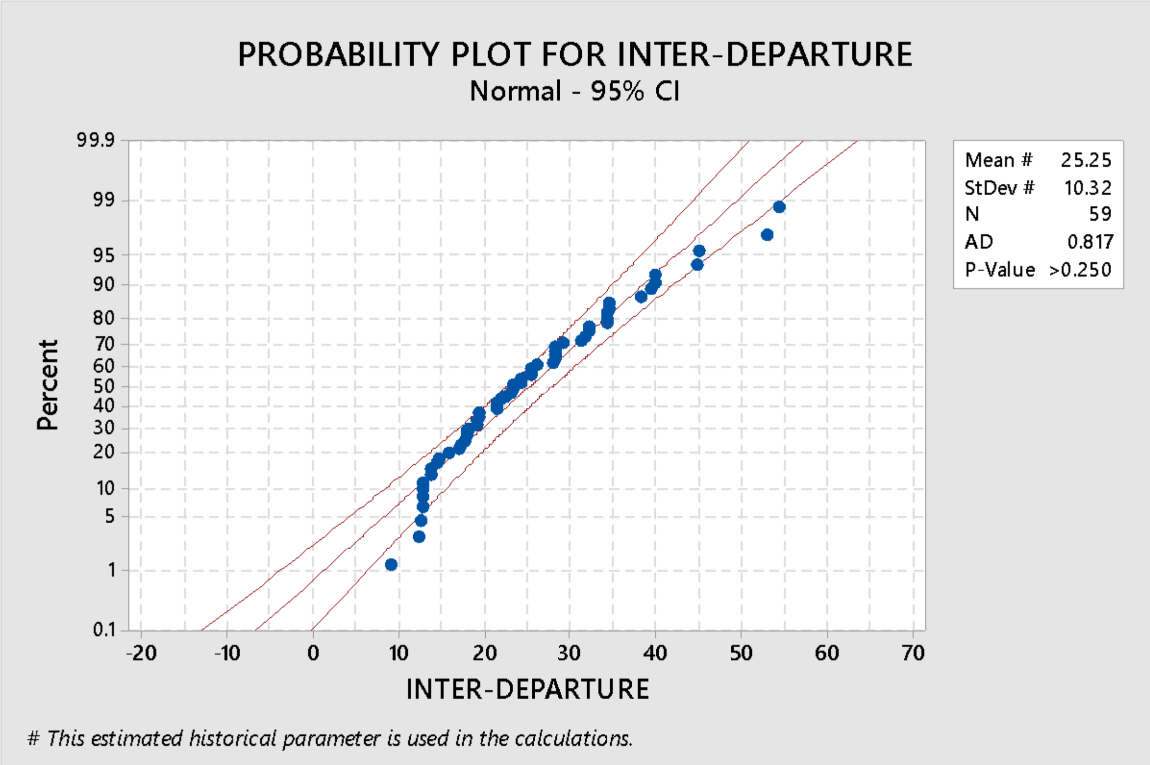

Figure 6: Probability Plot of Inter- departure Packets for Normal

Critical Value

For Minitab, Critical Value =level of significance, α =5%

Decision

Since test statistic is greater than the critical value i.e. 0.250 > 0.05. Hence, H_0 failed to be rejected.

Conclusion

We therefore conclude that the population from which the inter-departure sample was drawn is Normally Distributed. This evidence is also clear form the probability plot in Fig.5 since the empirical sample point lies with the theoretical range or limits.

Determination of Burstiness of Internet Traffic in Finite Session

Apply N-Flow model, from equation 5 we have

$$b = 1 - \frac{\kappa}{\lambda_p}$$

$\kappa = 3.85 Mb/s$ $\lambda_p = 80.57 Mb/s$ for arrival packets,

$$b = 1 - (3.85/80.57) = 0.95220$$

$b = 0.9522 \rightarrow 1, then\ \overline{\Delta N} = 0$, which means that all packets in a burst are transmitted simultaneously. This also implies there is self –similar process with a degree of either long- or short -range dependence. This dependency account for the reason why Poisson failed. How long or short the dependency is depends on the value of the Hurst parameter. For short range dependency the Hurst parameter is between zero and half [i. e. 0< Hurst parameter<0.5] and that of long range dependency 0.5< Hurst parameter<1.

Moreover, burstiness, b approaching one also suggest that the intervals of number of Poisson arrivals was overlapped which makes the data not independent. Giambene (2014) said “the number of Poisson arrivals in disjoint intervals are statistically independent where as in a situation where arrivals are in overlapped intervals, data are not independent”.

Theoretical Analysis

Poisson models rely on the Central Limit Theorem for measurement of parameters. This theorem loosely states that as we group more and more data from a distribution or distributions, the sample distribution will approach a normal distribution, and the mean will "smooth" toward a constant number. This allows easy measurement of intrinsically and mostly non-normal distributions. This close guess is good enough as long as the sample size, $n \geq 30$. The close guess improves as the sample size increases and the average is taken over more random variables. This can be observe in the graphs below.

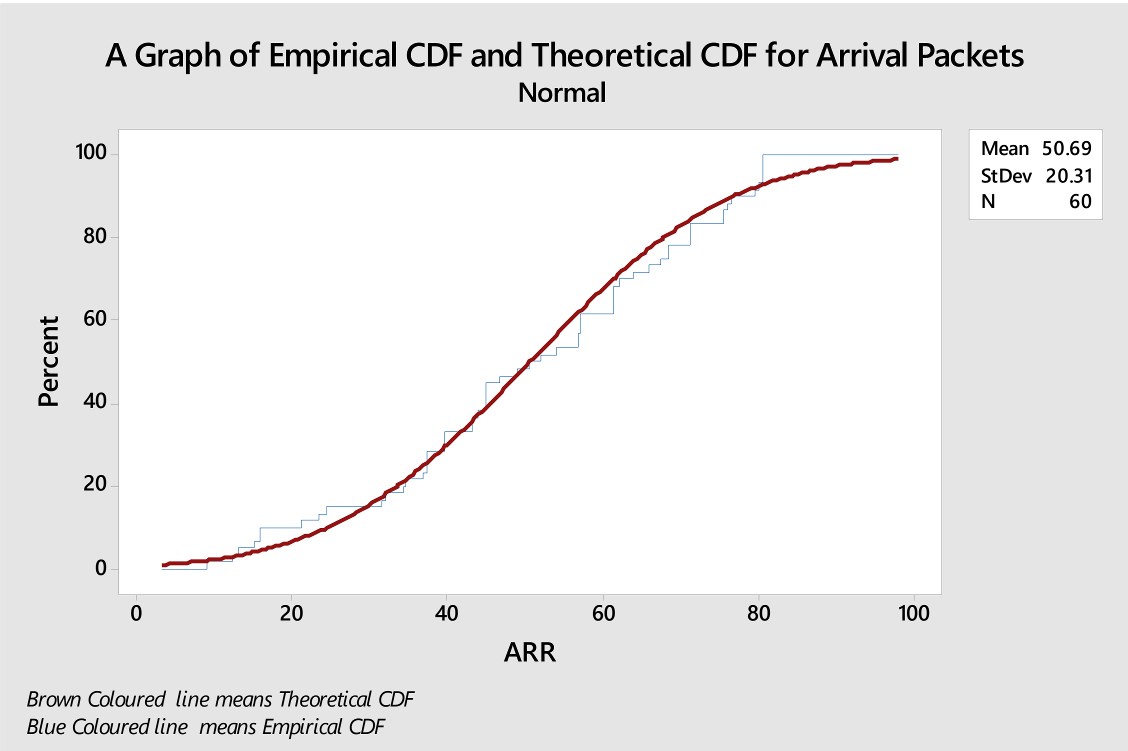

Figure 7: Empirical and Theoretical CDF’s for Arrival packets

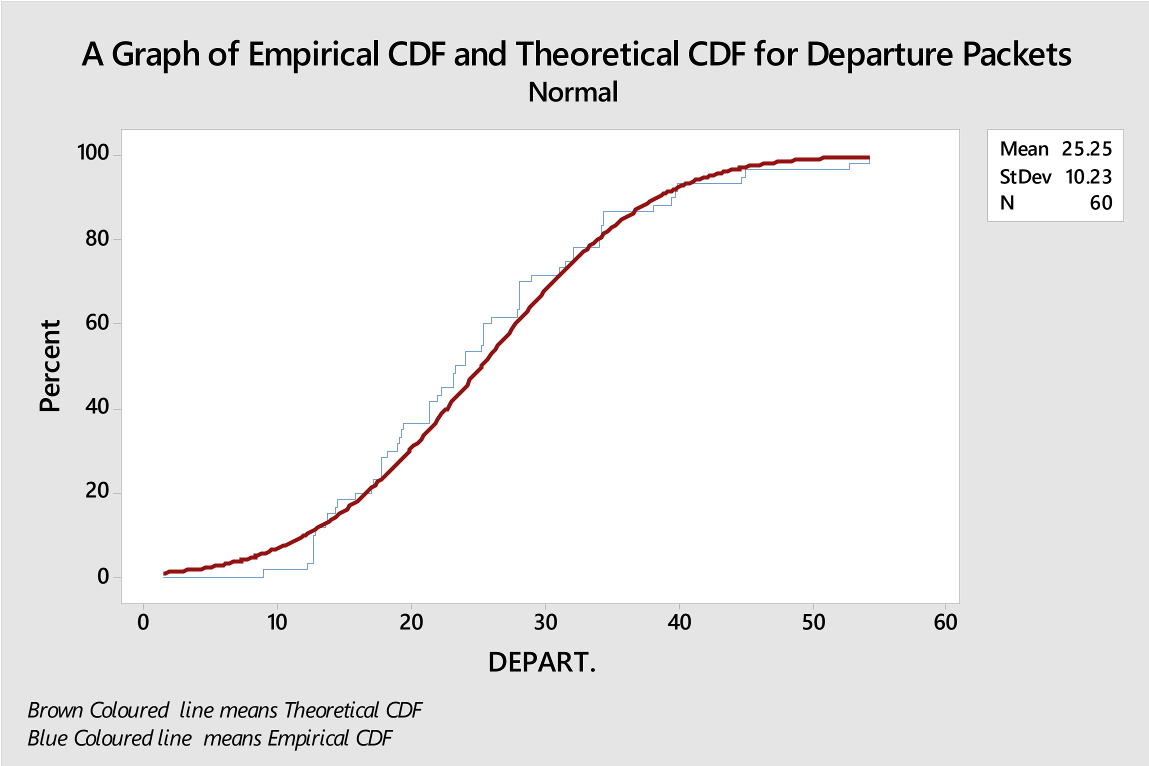

Figure 8: Empirical and theoretical CDF's for Departure packets

The Empirical CDF's from Figures 7 and 8 illustrate real time traffic distribution for arrival and departure packets respectively. Both graphs exhibit similar characteristics. The discrepancies are seen along the path but reduces at the tails. Both graphs reveals that the distribution has normalized. This confirms the central limit theory.

Moreover, as the packets increases in number the standard deviation shrinks. This explains why at the upper tail Empirical and Theoretical CDF's becomes the same and moves parallel to the arrival and departure axes. For arrival $n \geq 100$ and for the departure $n \geq 55$

Deep-down opinion or analysis hints that if the period of time (or space) of grouping is increased well enough, the degree of data relationships will eventually become unimportant by scaling. Unfortunately, the central limit theorem only applies if data is independent and has limited variance. Moreover, when burstiness approaches one (i.e. $b \rightarrow 1$), central limit true theorem do not apply and the traffic becomes self-similar (i.e. self -almost the same). Self-almost the same traffic shows either SRD or LRD qualities which cannot be eliminated by grouping. This SRD or LRD makes worthless or meaningless all results from queuing theory got using Poisson process. This further explains and gives reasons why Poisson failed.

CONCLUSION

Based on the results achieved, it can be concluded that Poisson has failed in wide –area traffic in Ghana Technology University College. Moreover, from experimental and theoretical analysis, it is clear that the internet traffic behaviour in Ghana Technology University College is normally distributed.

REFERENCE

Montgomery, D.C. & Runger, G.C.(2014) Applied statistics and probability for engineers (6th edition).USA. John Wiley & Sons, Inc.

Gupta, B.C. & Guttman, I. (2013) Statistics and probability with applications for engineers and scientists. Canada. John Wiley & Sons, Inc.

Stephens, M.A. (1974) Empirical distribution function test for goodness of fit and some comparisons, *J.Am.Start.Assoc*.,Vol 69,pp730-737.

Stephens, M.A. (1977) Goodness of fit for the extreme value distribution, *Biometrika*, vol 64, pp585-588.

Siriwong, K. Lipsky, L. & Ammar, R. (2007). Study of bursty internet traffic. *Sixth IEEE International Symposium on Network Computing and Applications* (NCA 2007), pp.53-60,

Giambene, G.(2014). Queuing theory and telecommunications networks and applications (2nd edition). USA. Springer. Also available online at http://extras.springer.com with ISBN (Online)978-1-4614-4084-0