TECHNICAL UNIVERSITY OF CRETE
SCHOOL OF ELECTRONIC AND COMPUTER ENGINEERING

On modeling the email traffic of the Technical University of Crete

Diploma Thesis

Spyros Boukoros

Advisor: Associate Professor Polychronis Koutsakis

Chania, September 2014

Abstract

E-mail has become a *de-facto* means of communication. Mail servers try to manage the explosive growth of e-mail usage and offer good quality of service to the users, while spam e-mails are expected to account for 90% of the e-mail traffic. The exceedingly heavy workload can lead to the replacement of existing e-mail servers due to their inability to cope with performance standards and storing capacity. In this study, we try to model the workload of the Technical University of Crete e-mail servers, for all types of traffic (spam, user and system e-mails). We collected a vast amount of e-mail logs with high variations in terms of size and volume over time. We tested some of the most popular distributions for workload characterization and used powerful statistical tests to evaluate our findings. Interestingly we come to different conclusions in comparison with previous works in the field. Our work indicates that, with the exception of some outliers, email traffic can be modeled and predicted quite well.

Acknowledgements

I would like to take this opportunity to thank wholeheartedly my thesis advisor, colleagues, family and dear friends. This thesis was made possible only through the support and guidance they offered.

Specifically, I would like to express my thanks to my advisor and mentor Polychronis Koutsakis, who has offered the guidance and support I needed as an undergraduate not only during my thesis but long before.

Thanks are also due to my teachers M. Paterakis and D. Pnevmatikatos for the interesting journeys in the fields of network modeling and computer architecture respectively and for their contribution as members of the thesis committee.

I would also like to thank P. Kontogiannis for his assistance in modeling the server's workload, by preprocessing the data, as well as the Network and Computing Infrastructure Overseeing Committee for giving me permission to use the sizes and times of arrival/departure of TUC users' emails.

Last but not least I would like to thank D. Vasileiadou for her patience and time and my friends for their support.

Special thanks to Giota Panousi for her love and emotional support all these years and to my family for teaching me not to give up.

Table of Contents

List of Figures

List of Tables

1. Introduction

E-mail has become a de-facto means of communication. It started from the infant steps of the internet and evolved with it. Nowadays almost everyone in the developed world has an email account or several, for different purposes. Its ability to attach files and multimedia content is what led to its huge success and adoption from the masses. Also, email is a very quick way of communication and nowadays it is portable using smart devices. Corporations also adopted email as one of their main communication tools. Employees can reply to their email wherever and whenever they want, send files online and most importantly notify multiple interested parties with only one email. They also tend to view emails within 6 seconds from the time they arrive [1].The above facts show why email is so much preferable against the traditional ways of communication. According to a 2010 survey [2], 64% of the participating corporations, answered that the mere adoption of email led to a significant sales increase. This happens mainly due to the larger numbers of global customers the corporation is able to reach and the multimedia content of the email, which can pass a direct message to those customers. Misuse of this powerful tool (email) is something that naturally occurs, as with every kind of technology. Irresponsible parties use its ability to carry file and/or reach to numerous customers for their own, sometimes not legal, actions. This is called junk or spam email. Spam email is defined by three characteristics: 1) it is not requested by the recipients, 2) it has commercial purpose and 3) it is always sent in bulk [3]. Spam is considered a threat for the email way of communication because it can break the trust between corporations and customers, have an effect on economy or even put into doubt whether email is indeed, a great way for communication. According to a survey, spam email is the number one time wasting technology from which corporations suffer from. It forces employees to waste time viewing numerous unimportant emails daily. This also has a direct effect on economy because it is reducing the employees' productive working time. An incident was studied [4] where Japan's Gross Domestic Product was reduced by 0.1% due to the spam traffic. Also, there are those spam emails trying to steal personal information and data by making use of social engineering techniques or other, technology based forms of attacks. Hence, corporations can lose vast amounts of money if someone gains access to their private information, bank accounts and other sensitive information. In addition, spam emails that do not have commercial value but carry malware are able to infect computers and again cause great damage to corporations. They can break trust in a corporation by forcing infected computers to spam as well and make worldwide servers block that corporation's

servers as spam servers. Then all outgoing emails are going to be lost (blocked by spam filters) and temporary isolate the corporation. Individuals can also suffer from identity theft incidents or loss of sensitive information like bank accounts codes. Spam traffic accounted for 66% of the worldwide traffic in 2013. [5]This means that Internet Service Providers (ISP), corporations and universities have to deal with millions of spam emails every day. Both spam and regular emails arrive at such great volumes that it becomes a matter of crucial importance that servers can cope with the heavy workload and do not crash. Hence, Internet Service Providers (ISPs) need to continually increase their bandwidth and storage capacity by spending tons of money.

All of the above facts, regarding regular and spam email traffic show the urgent need for accurate email traffic prediction, which will help system administrators need to take actions to optimize the way they allocate the storage space or the bandwidth that they have at their disposal. By doing so, they will be able to avoid system crashes and failures and offer users a better quality of service. Gomez *et al.* [6] found that message sizes could be represented by lognormal distributions at the body and the tail. They collected the SMTP logs of the servers they modeled for the regular traffic and the logs of the Spam-Assassin filter for the spam. Their measurement period was one week. Using the least square error (LSE) method, they determined the parameters of the lognormal distributions that fit their dataset best. Unfortunately, we do not have more clues on which other distributions they checked and which statistical tests they used to derive their results. They also modeled the arrival process and the popularity of various receivers. The Poisson arrival process was shown to fit their workload. The popularity of objects was modeled with a Zipf-like distribution i.e, some users receive or send significantly more emails than others. Bertolotti and Calzarossa [7] also collected the SMTP logs from email servers and modeled the workload. They modeled the message sizes, interarrival times and the number of recipients. Lognormal distribution was found to be the best fit for the message sizes. The interarrival times were shown to fit Weibull and Pareto distributions, in contrast with the conclusions in [6]. In [8] Shah and Noble present a large-scale study on an email server. They model various parameters various from the message sizes to the number of words emails consist of. Their measurement period lasted more than 7 months. Regarding the modeling of message sizes, which is the focus of this study, they noticed that the cumulative distribution function (CDF) is symmetric under log scale. Hence, they concluded that their data must be distributed with a lognormal distribution. Of course, this is not a methodology that can be applied in general but a more empirical approach. They modeled the main

body with a lognormal distribution while the tail was modeled with a Pareto, following the lead of [6]. In this way, they were able to model the workload with high accuracy. While Gomez *et al* [6] find that spam emails have smaller sizes than regular ones, Shah and Noble [8] claim the opposite. However, both of the above studies conclude that spam traffic is distributed with a lognormal distribution. V. Paxson modeled wide–area transport Layer Protocol (TCP) connections [9]. SMTP connections are TCP connections for transferring emails. Unlike the previously mentioned studies and in accordance to ours he used goodness of fit tests to back up his findings. Regarding the SMTP connections, he found out that the empirical distribution was bimodal and justified that from the fact that users sent either simple text mail either files. He decided to model it with two lognormal distributions, breaking the data in two populations, one below the 80^{th} percentile and the other above. The goodness of fit test he used is the Anderson – Darling test [10].

In our work, on modeling email traffic we do not use the SMTP logs like the aforementioned studies. We found that the SMTP logs from the Windows Exchange 2003 server of the Technical University of Crete could not give us detailed information about the size of the attachment each email was carrying. So we used the Message Tracking Log system (embedded in Windows Exchange systems), which proved to be a convenient alternative. Furthermore, to evaluate the accuracy of our modeling approach we used a number of goodness of fit tests.

2. Technical Background

2.1. Email Server Setup

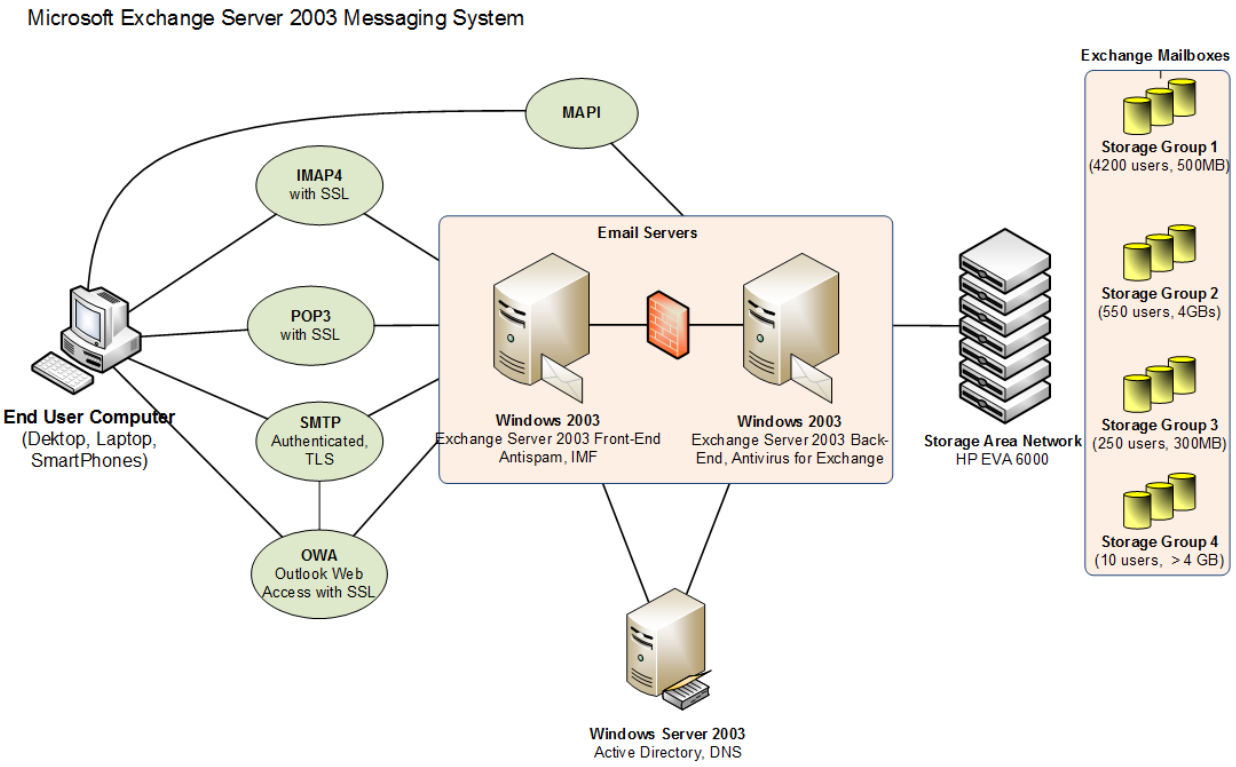

Figure 1. Email server set up

In Figure 1 the setup of the Technical University of Crete email server is presented. On the left is the end user device that connects to the server with various protocols for receiving or sending an email. Users receive emails with the MAPI, Imap4, POP3, OWA protocols which are used for client-server communication. The SMTP protocol (Simple Message Transfer Protocol) is used for server-to-server communication and with this protocol, users are able to send emails.

The e-mail server actually consists of two servers, both running Windows Server 2003 but behaving like one. They form an architecture called front-end back-end. All emails are being

served by the front server with the exception of the MAPI protocol. Then they are transferred to the back server and finally to the storage area network for storing. The front server takes care of all anti-spam measures while the back server takes care of all the antivirus measures. Below the two servers, the DNS server is shown. It contains an active directory with all the Technical university of Crete email addresses. The front server has two network adapters, one for each kind of traffic (incoming/outgoing) while the back server has one. The Exchange mailboxes have different sizes for different categories of users like faculty, staff, etc.

2.2. Anti-Spam Setup

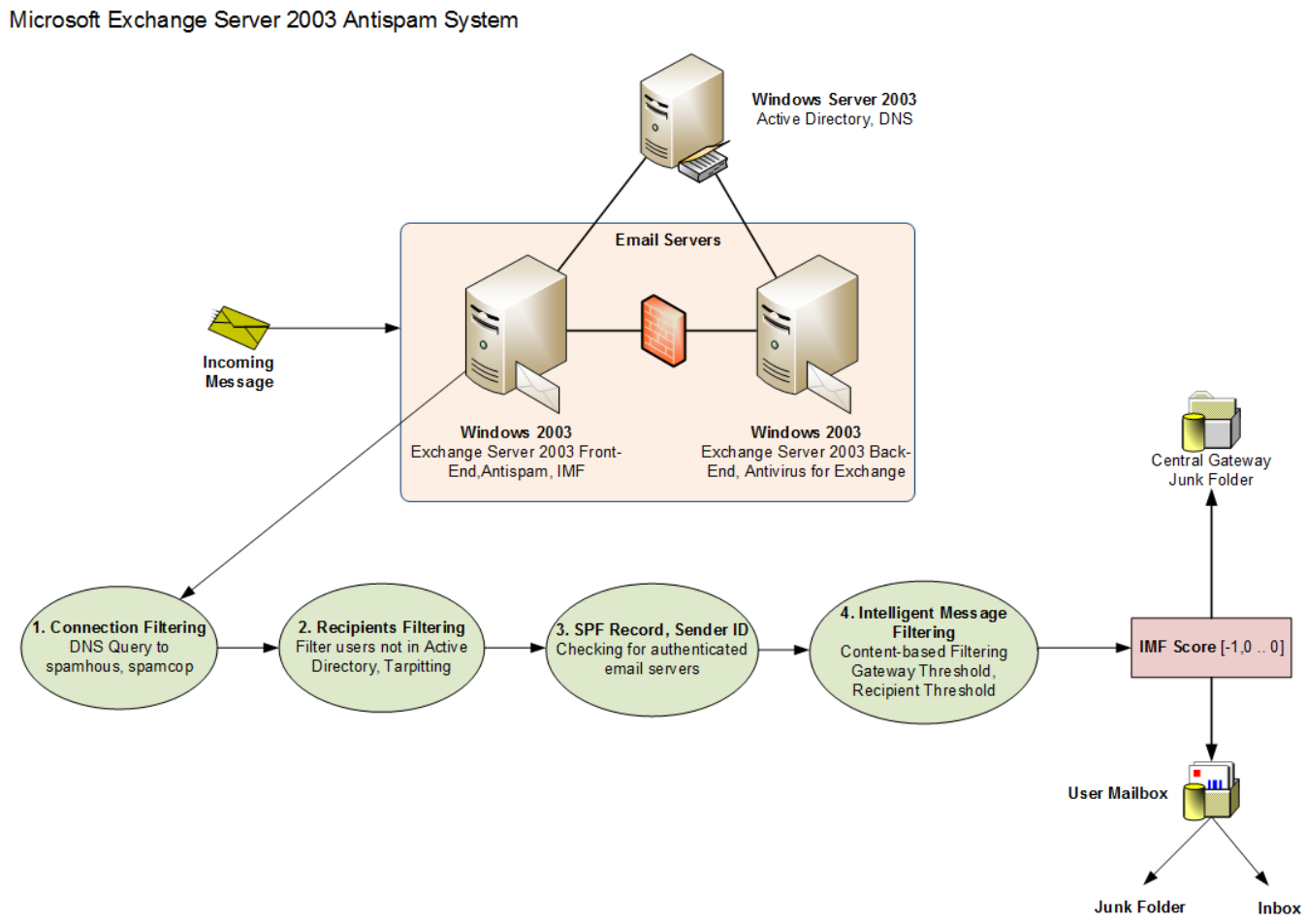

Figure 2. Anti-spam technology

In Figure 2 the anti-spam methodology used in the email server is presented. The email server is shown again, as well as an incoming message. When an email arrives, it opens an SMTP connection to the server. Then several commands follow specifying the recipient, the sender, the subject etc.

Initially the email server sends a query to two DNS servers, the “smaphous” and the “spamcop” with the email address of the sender of the message. Those DNS servers have directories of the most common spammers around the world. Therefore, if the sender of the email is a well-known

spammer, then this connection is rejected and the email never reaches the storage facility. Of course, this solves the problem of the storage capacity but not the problem of the bandwidth capacity because upon the opening of the SMTP connection the bandwidth required for the connection is lost.

The system next checks whether the recipient specified from the email is really a user in the active directory of users (DNS).In this way brute force attacks trying to find new email addresses can be avoided. In addition, connections are delayed by a small amount of time because this seems to actually “discourage “spambots and computer worms. This method is known as tarpitting. Even though emails have a small delay, email is not considered as real time communication tool so there are no side effects.

Then the server checks if the sender IP address is actually authorized to use that server name. Email servers authorize specific IP addresses to act on their behalf and communicate. If the IP address trying to send us an email is not authorized by that specific email name then most likely something is wrong and again the connection is closed.

Finally the Intelligent Message Filtering (IMF) system acts. It decides according to the email content and various other parameters if the email has a probability to be spam. It ranks the emails in a scale from -1 to 10.

-1 means that the sender and the recipient are users inside the University so the email is always sent and placed in the receiver’s inbox. If the message is ranked from 1 to 5, (although this may vary) the email is placed in the receiver’s inbox while a rank between 6 and 10 moves the email to junk.

2.3. Anti-Virus Setup

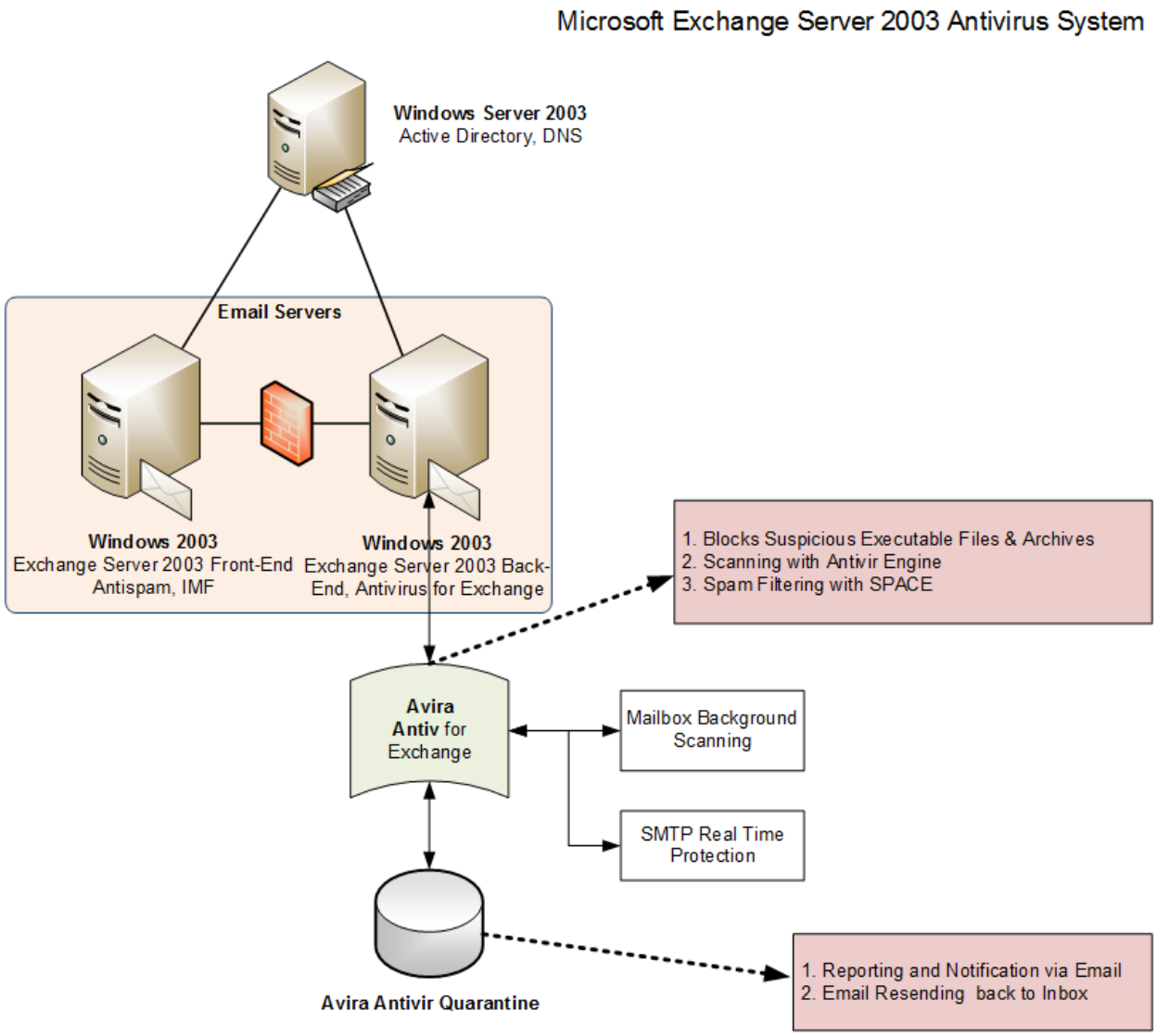

Figure 3. Anti-virus technology

The antivirus measures take place on the back end server. The email server uses the Avira antivirus. It scans the emails that arrive for harmful content. If so, it reports this with an email to the users.

3. Theoretical Background

3.1. Statistical Tests

3.1.1. Q-Q Plot

The statistical test was made with the use of Q–Q plots. The Q–Q plot is a powerful goodness-of-fit test, [11] [12] [13], which graphically compares two datasets in order to determine whether the datasets come from populations with a common distribution (if they do, the points of the plot should fall approximately along a 45 deg. reference line). More specifically, a Q–Q plot is a plot of the quantiles of the data versus the quantiles of the fitted distribution. A z-quantile of X is any value x such that $P((X \leq x) = z$. We have plotted the quantiles of the real data with the respective quantiles of the various distribution fits.

3.1.2. Kolmogorov-Smirnov Test

In order to further verify the validity of our results, we performed Kolmogorov–Smirnov test [14]. The Kolmogorov–Smirnov test (KS-test) tries to determine if two datasets differ significantly. The KS-test has the advantage of making no assumption about the distribution of data, i.e., it is non-parametric and distribution free. The KS-test uses the maximum vertical deviation between the two curves as its statistic D. As explained in [12], the use of KS-tests is a good statistical tool; however it has the drawback that KS-tests give the same weight to the difference between the actual data and the fitted distribution for all values of data, whereas many compared distributions differ primarily in their tails. It tests if the null hypothesis is accepted or rejected, usually at the 5% significance level. The null hypothesis is that the population we are testing is drawn from a

specific distribution with 5% chance of error. The Kolmogorov-Smirnov test can also be used, the way we use it in this study, as a goodness of fit test. This means that we don't actually expect to see if the test accepts or rejects a null hypothesis (even thought it would be an excellent result if the null hypothesis was accepted) but to see how "far" the actual data are from the fitted distribution.This is called Two-Sample Kolmogorov-Smirnov Test. The test measure is given by the following formula for two given cumulative distibution functions (cdf's) $F_1 \; and \; F_2$:

$$D_{n,n'} = \sup |F_{1,n}(x) - F_{2,n}(x)| \quad (1)$$

The null hypothesis is rejected at the level α significance if

$$D'_{n,n} > c(\alpha) \sqrt{\frac{(n+n')}{nn'}} \quad (2)$$

The values of $c(\alpha)$ are defined for various significance levels.

We should bear in mind that the Kolmogorov-Smirnov two-sample test only tells us half the tale, meaning that it only tells us the maximum distance between two distributions and not which distribution our data come from.

The Kolmogorov – Smirnov test has two significant limitations.The first is that the distributions must be continuous. The second is that it tends to be more sensitive at the center of the distribution rather than the tails.

3.1.3. Anderson-Darling Test

The Anderson-Darling test [10] [15] is a modification of the Kolmogorov-Smirnov test. It gives more weight to the tails as opposed to the K-S Test. The test statistic belongs, like the Kolmogorov-Smirnov test, to the family of quadratic empirical distribution function statistics, which measure the distance between the hypothesized and the empirical CDF as

$n \int_{-inf}^{+inf} \big(F_n(x) - F(x)\big)^2 w(x) dF(x),$ (3)

with

$w(x) = \big[F(x)\big(1 - F(x)\big)\big]^{-1}$ as the weight function which favors the tails of the CDF.

Even though the Kolmogorov-Smirnov test is distribution free, there is a form of the Anderson-Darling test that is not. It makes use of the specific distribution parameters to be evaluated. The appropriate critical values need to be selected for the distribution we wish to check. This allows the test to be more sensitive but it also makes it impossible to use with a large variety of distributions. Currently, tables of critical values exist for the normal, uniform, lognormal, exponential, weibull, extreme value type I, generalized Pareto and logistic distributions.

The basic test statistic is $A^2 = -n - S$ (4)

Where

$$S = \sum_{i=1}^{n} \left(\frac{2i-1}{n} \left[\ln\big(\phi(Y_i)\big) + \ln\big(1 - \phi(Y_{n+1-i})\big) \right] \right) \quad (5)$$

A^2 is compared against the critical value of a specific distribution and, if greater, the null hypothesis is rejected.

In this study, we use the non-parametric version of the Anderson – Darling test because we are testing distributions for which no known critical values exist.

3.1.4. Kullback – Leibler Divergence Test

The Kullback - Leibler (KL) Divergence test [16] measures the information loss between two distributions.

It tells us how many extra bits we are going to need if we code samples using the Q probability distribution function instead of P. The test is non-symmetric meaning that if we reverse the P and Q (probability distributions functions) we take different results. The KL divergence for continuous probability distribution functions.

$$D_{KL}(P||Q) = \int_{-\infty}^{+\infty} \ln\left(\frac{\mathrm{P(i)}}{\mathrm{Q(i)}}\right) \mathrm{P(i)} \quad (6)$$

The 0*ln(0) appearance is interpreted as zero because

$$\lim_{x \to +\infty} xln(x) = 0$$

The result is always non-negative because of the Gibbs's inequality and zero only if P = Q everywhere

3.1.5. Relative Percentage Error

The Relative Percentage Error [17] gives a metric on how different one population is from another. By measuring the absolute difference between the two populations, we do not discriminate which one is bigger or smaller. Of course, we wish to achieve results as close to 0% as possible in order to find a modeling approach that has high accuracy.

RPE is defined as:

$$RPE = \frac{|Y-X|}{X} * 100\% \quad (7),$$

where Y is the predicted value and X the real observation.

In this study, we used the relative percentage error to quantify the results of the QQ-Plot test. i.e., we calculated the RPE of the quantiles derived from various distributions versus the quantiles of the real data.

3.1.6. Maximum Likelihood Estimation

Maximum likelihood [18] estimation is a method for finding the parameters of a statistical model based on our data. Given the model, it returns estimates for the model's parameters usually at the 95% confidence level. Every distribution has a vector, **θ**, that contains the parameters. We want to find an estimator, $\hat{\theta}$, which can be as close as possible to the true **θ**. Hence, the maximum likelihood estimation method is a way to seek the probability distribution that the data most likely obey.

This can be done by taking the joint probability function of the observations given the, unknown to us, set of parameters and assuming they are independent.

$$f(x_1, x_2, \dots, x_n|\theta) = f(x_1|\theta) * f(x_2|\theta) * \dots * f(x_n|\theta) \quad (8)$$

Then by fixing the values x, which are the observations and by considering the $\boldsymbol{\theta}$ as variable we obtain the likelihood function.

$$\mathcal{L}(\theta; x_1, x_2, \dots, x_n) = f(x_1, x_2, \dots, x_n|\theta) = \prod_{i=0}^{n} f(x_i|\theta) \quad (9)$$

Usually because it is much more useful and efficient to use the logarithmic version of the above, we have the $ln\mathcal{L}(\theta; x_1, x_2, \dots, x_n)$ which is called log-likelihood. We want to maximize the

$$\hat{\ell}(\theta|x)$$

by finding the appropriate value θ.

If a maximum exists, it can be found. For many models there is an explicit form from which the parameters can be calculated but for many others there is not, in which case optimization methods (i.e. Newton's Method) needs to be used. Sometimes more than one estimates can be found that maximize the log-likelihood function.

4. Methodology

4.1. Data Collection

The data we collected and worked on were the emails logs. We got two separate kinds of logs, for the non-spam and the spam emails. The measurement period covered one week, between February the 3rd and February 9th, 2014.

4.1.1. Non spam emails

The non-spam emails are the emails that arrived at the server and were not stopped by the filter or classified later as spam. Because our server system consists of two servers, we merged their logs. The logging system is the Message Tracking System, a tool embedded in the Microsoft Exchange Server 2003. It records all kinds of email activity at the server. Therefore, the spam traffic that is blocked from the antispam filter is not recorded because the connection is closed before the email actually arrives.

4.1.2. Spam emails

The spam emails are logged in a different way. The emails that arrive at our server but are classified as spam are saved into folders with their whole body. They are in a custom kind of file, which ends in .EML and is actually the whole message with the SMTP commands at the top.

4.2. Data Preprocessing

4.2.1. Non – Spam

The logs from the Message Tracking System were imported to a program called SawMill Flowerfire. This is a log parsing and data analytic program. It has an internal database, which is very fast and has a variety of customization methods. We were able to group the logs based on different criteria easily and export the results in .csv form. At first, we eliminated some users as

outliers. These users were 3 email addresses which were sending 117.59GB, 9.96GB and 4.81GB of data respectively.

Then we decided to break our data into 4 categories depending on whether they represent system or users' emails, and whether they are incoming or outgoing. The users' emails are those emails that refer to real people and their email addresses. The system emails are from 8 servers inside the university and consist mainly of server to server communication or diagnostic emails as well as no-reply messages sent from various users. The decision was based on the fact that the system emails are sent out in bulk, usually to the whole university to inform everyone about events. Therefore, these emails are of different nature, so we decided to consider them as a different category and model them separately, to achieve with higher accuracy.

We exported our results from Sawmill to Matlab for modeling. We created datasets for each one of the categories for emails per hour and day, senders and recipients per day and hour, total kilobytes per day and hour and volume per day and hour. We modeled the size of the emails for each one of the above categories for one week but also for every day separately.

The system servers are the

nagios@Titanas.noc.tuc.gr,

noreply@isc.tuc.gr,

noreply@tuc.gr,

postmaster@isc.tuc.gr,

webcourses@ced.tuc.gr,

eclass@isc.tuc.gr,

< >, (users with no email address)

- (users with hidden email address)

The filters we used in our database to export the major categories are the following.

For incoming traffic for the users group, we selected the entries that had a recipient address inside the domain tuc.gr. Of course, we had excluded from the database the systems' servers and the spammers.

For outgoing traffic for the users group, we selected the entries that had as senders someone with an email address inside the tuc.gr domain. Again, spammers and systems' servers were not included.

For incoming system traffic we required the sender to be one of the above 8 email addresses and the recipient to be someone inside the tuc.gr domain.

For outgoing traffic, we required just the sender to be one of the above 8 email addresses.

The number of entries collected for the outgoing and the incoming systems' emails are close because those servers rarely need to communicate with the outside world.

The program did not include any ready tool to get statistics for the senders so we had to write it ourselves. We collected all the entries from the logs and grouped them in the same way the program had grouped recipients. We were able to successfully group them into various categories and extract information for per hour and day traffic, unique message id per sender and recipients per sender.

4.2.2. Spam

Spams are incoming traffic, which, we know that targets the users and not the system. Therefore, the only thing we had to do is to write scripts to gather the spam sizes per day and per hour. Using Matlab we wrote those scripts and managed to create .csv files containing the required information like sawmill exported it. Then we were able to model the spam traffic.

4.3. Modeling

We will present statistics we gathered about ***each*** email traffic category. In addition, we will present plots with the cdf and the pdf of the real data, per hour and per day of the week.

We use the maximum likelihood estimation method to obtain the parameters of various distributions. These distributions are well known in the literature for workload characterization and modelling. The distributions used are the uniform, exponential, gamma, weibull, log logistic, lognormal and GeneralizeExtremeValue. The maximum likelihood estimation method returns a vector with the estimated parameters at the 95% significance level. After acquiring these parameters, we generate random data from each one of the distributions. We use matlab's built in functions to generate the data.

After we generate the data, it is time to apply our first test, which is the QQ-plot. We sample the various populations in a way so we have the percentiles of each population. Then using matlab's function QQ-plot with the two populations as input, we have the graphical result of the test. For each category, we have merged the QQ-plots into one plot for easier comparison. We wanted to quantify the result to make it easier to interpret the plots. Therefore, we used the Relative Percentage Error on the sampled populations. We proceeded by using the Relative Percentage Error (RPE) in order to quantify the results shown in the QQ plots. Hence, we calculated the RPE between the percentiles of the actual data and the distribution-generated values. To overcome the randomness of the latter, we generated values from each distribution 1000 times and took the mean value for each distribution.

For the Kolmogorov – Smirnov test, we used matlab's built in function (kstest2) to compare the distributions with the real data. We did the same for the Anderson – Darling however the respective built-in function in Matlab, sometimes returned an infinite result, which result means that we have many outliers. Of course, this is something we expect but we want to know the test result. Therefore, we had to rewrite the code and modify it.

At first, we created distribution objects for each one of the distributions we wanted to test. We created those using the estimates from the maximum likelihood estimation. The only exception is the uniform distribution where we need only the minimum and the maximum values.

Then we calculated the cdf on the values of the population.

The cdf when calculated on the outliers had probability 1, which means that the whole distribution is below that value. That is expected when dealing with e-mail server workload. However, because of the 1 in the cdf the result of the Anderson – Darling test is infinite due to the logs in the formula. Therefore, when the cdf of a value was 1 at we replaced it with 0.99999999999999. In this way, we do not have infinite results with the cost of a small loss in precision. In addition, where the cdf of a value was 0 we replaced it with 10^-20 for the same reason as above.

For the Kullback – Leibler test there was not a built in function, therefore we wrote the code from scratch.

5. Results

5.1. Incoming Traffic for Users

This section focuses on the incoming users' traffic. This means that we do not have system emails at all and we wanted the recipient address to match the domain tuc.gr.

First, we have some statistics:

Total Number of Emails: 134864 **avg per day:** 19266 CV: 0.28
Unique message Ids: 51889 **avg per day:** 7412 CV: 0.28
Total number of distinct senders: 19,034 **avg per day:** 2,719 CV: 0.31
Total number of distinct recipients: 2827 avg per day: 403 CV: 0.22
Total bytes: 30.39Gbytes avg **per day:** 4.34Gbytes CV: 0.74
AVG # Distinct recipients/msg: 1.4 CV: 2.9

Next, we have the traffic calculated as emails/bytes/senders/recipients per hour and per day.

The plots are normalized respectively to the maximum value.

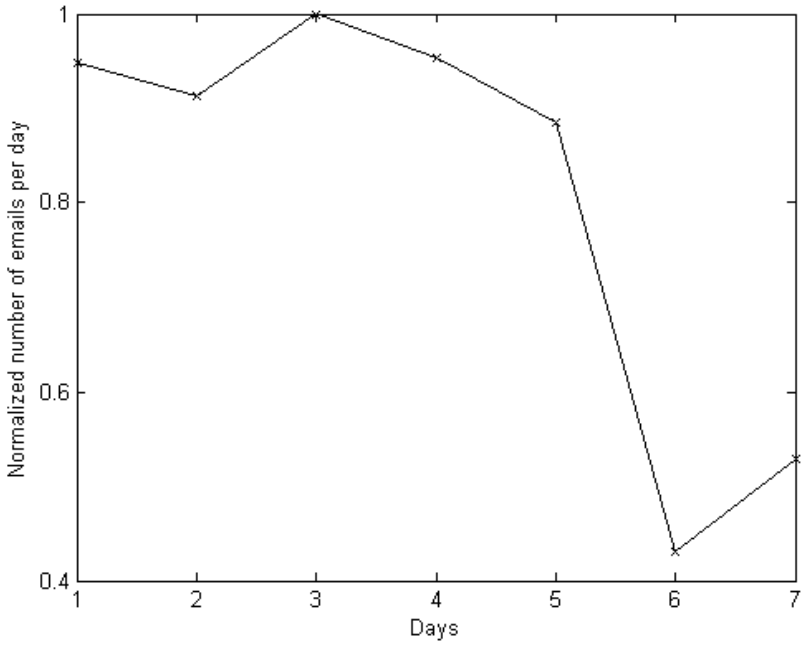

Figure 4. Incoming emails for users, emails per day

Maximum emails per day: 23281

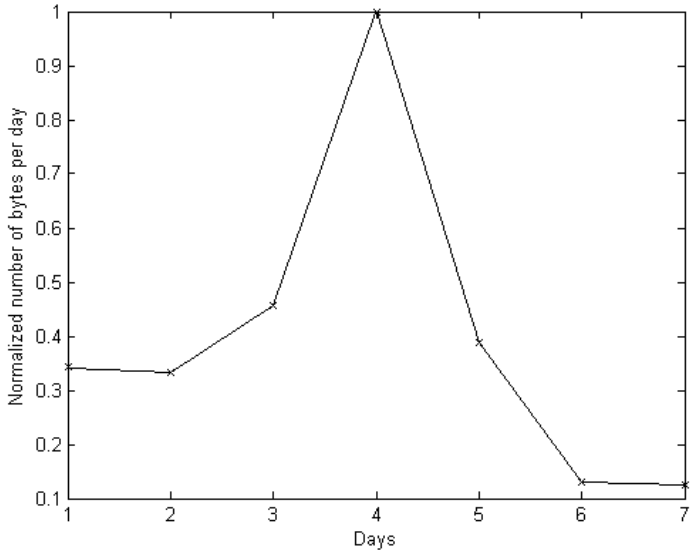

Figure 5. Incoming emails for users, bytes per day

Maximum bytes per day: 10.96GB

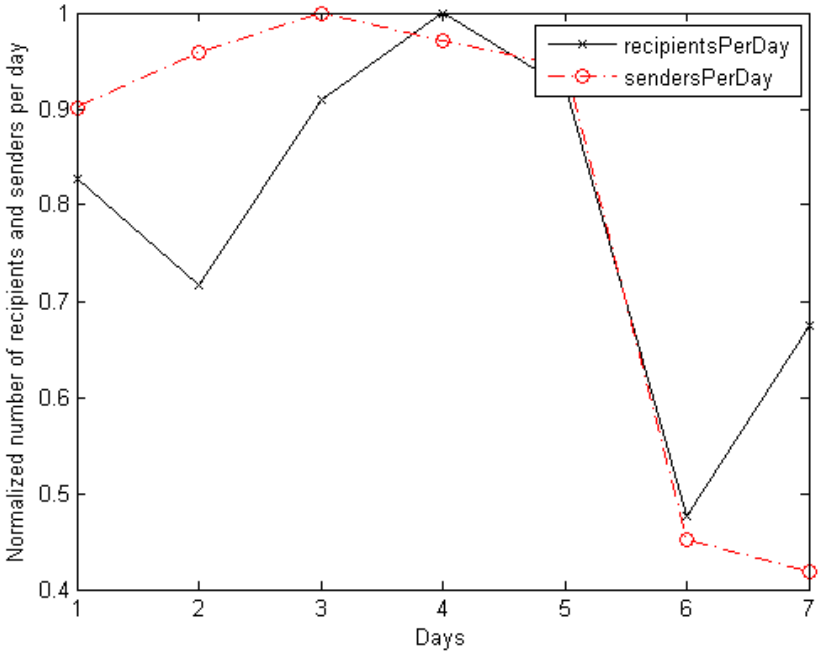

Figure 6. Incoming emails for users, recipients and senders per day

Maximum recipients per day: 1842

Maximum senders per day: 4789

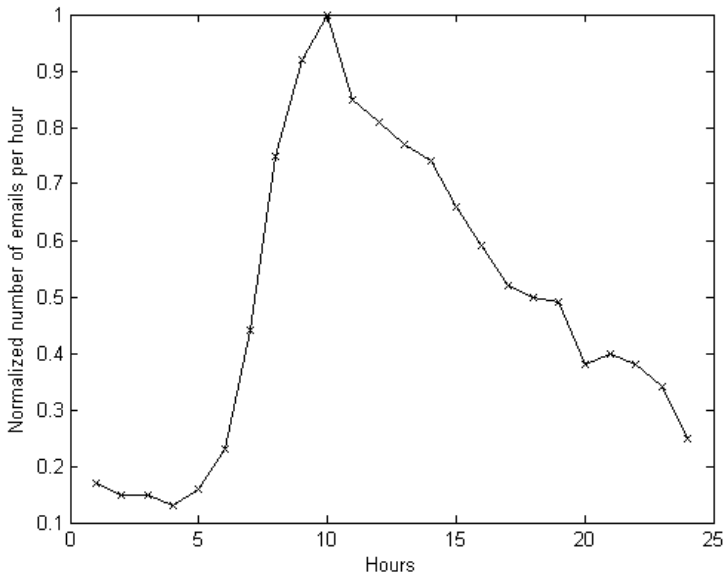

Figure 7. Incoming emails for users, emails per hour

Maximum emails per hour: 11448

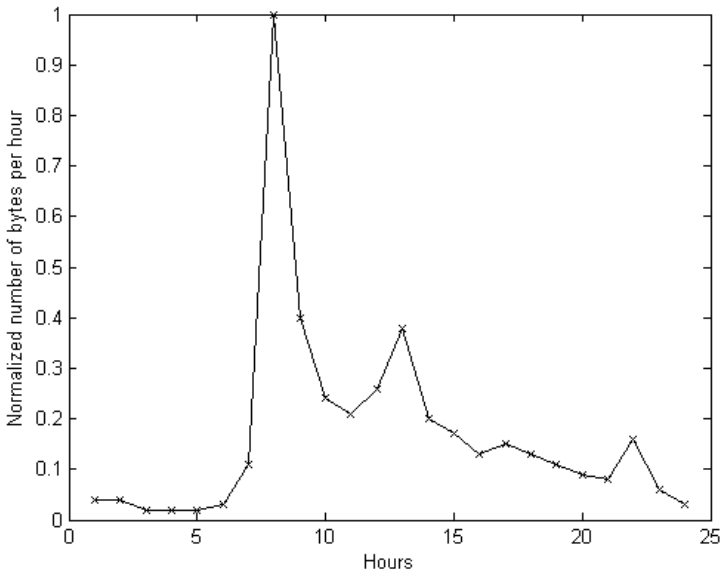

Figure 8. Incoming emails for users, bytes per hour

Maximum bytes per hour: 7.48GB

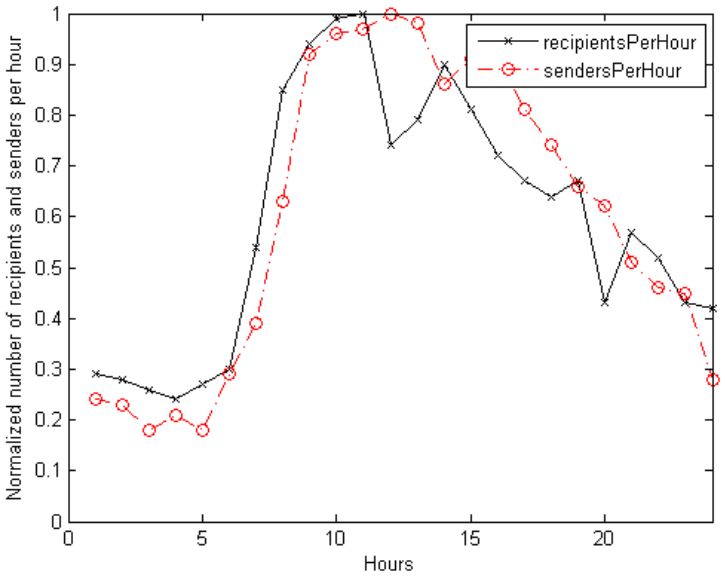

Figure 9. Incoming emails for users, recipients and senders per hour

Maximum recipients per hour: 1234

Maximum senders per hour: 2117

From the above we observe that our system deals with significantly larger traffic on the workdays than on the weekend. Wednesday and Thursday seem to be the more active days for our email server but given the measurement period of only one week this is not a conclusion that can be generalized. The large difference between workdays and weekend, however, shows that is much more beneficial to run complete antivirus scans and perform maintenance tasks on the weekends because the server is under light load, so maintenance will take less time and will be smaller the damage in the case of an unfortunate event like a system failure.

On hourly basis, we observe that the server is mostly active between 8-10 am, indicating that emails are sent first thing in the morning.

Some of our hourly statistics are shown in the tables below.

	Metric	Minimum	Maximum	Average	CV
Monday	Emails/Hour	180	1858	942	0.63
	MB/Hour	9.75	441.8	160.1	0.80
Tuesday	Emails/Hour	190	2023	906	0.61
	MB/Hour	8.77	413.8	155.1	0.80
Wednesday	Emails/Hour	205	3197	992.5	0.76
	MB/Hour	9.6	1262	212.7	1.27
Thursday	Emails/Hour	217	2062	946	0.5
	MB/Hour	8	6499	467.7	2.8
Friday	Emails/Hour	206	1969	879	0.6
	MB/Hour	15.3	735	180.7	1.04
Saturday	Emails/Hour	159	976	428	0.47
	MB/Hour	10.5	178.3	61.6	0.68
Sunday	Emails/Hour	138	1447	525.4	0.65
	MB/Hour	10.7	144.7	58.5	0.63

Table 1. Emails and traffic volume per hour

Again, as shown from the results in the table 1, the weekend is characterized by smaller amounts of traffic. The outlier on Thursday is due to an email with many recipients carrying a large PDF file.

	Metric	Minimum	Maximum	Average	CV
Monday	Recipients/Hour	78	585	259	0.5
	Senders/Hour	59	480	265	0.55
Tuesday	Recipients/Hour	72	553	247	0.53
	Senders/Hour	77	548	277	0.52
Wednesday	Recipients/Hour	74	834	278	0.62
	Senders/Hour	80	523	285	0.53
Thursday	Recipients/Hour	82	745	287.6	0.58
	Senders/Hour	100	498	275.4	0.49
Friday	Recipients/Hour	77	555	260.6	0.54

	Senders/Hour	70	533	263.8	0.55
Saturday	Recipients/Hour	69	319	145.1	0.43
	Senders/Hour	70	193	131.5	0.26
Sunday	Recipients/Hour	61	488	183.4	0.61
	Senders/Hour	56	191	123.8	0.29

Table 2. Recipients and senders' numbers per hour

Figures 10 and 11 presents the CDF of the emails' size. The first figure is with no zoom and the second with zoom.

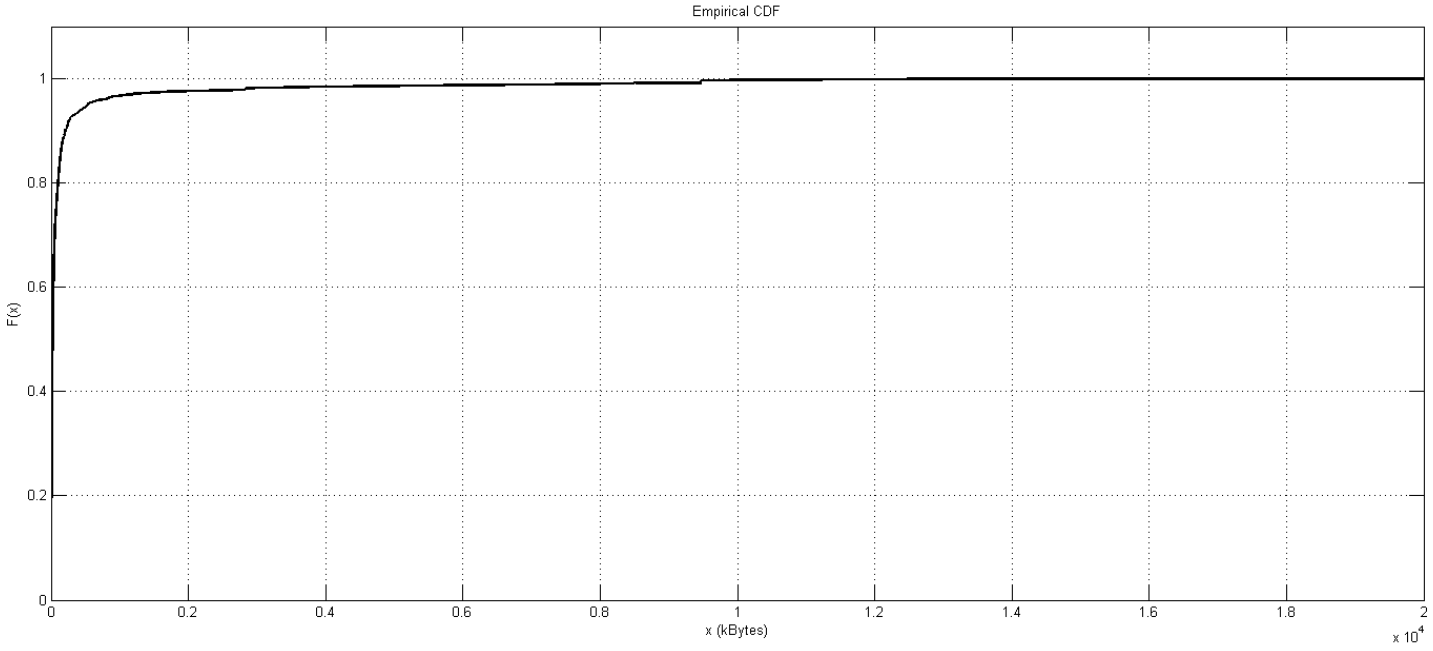

Figure 10. Incoming emails for users, CDF no zoom

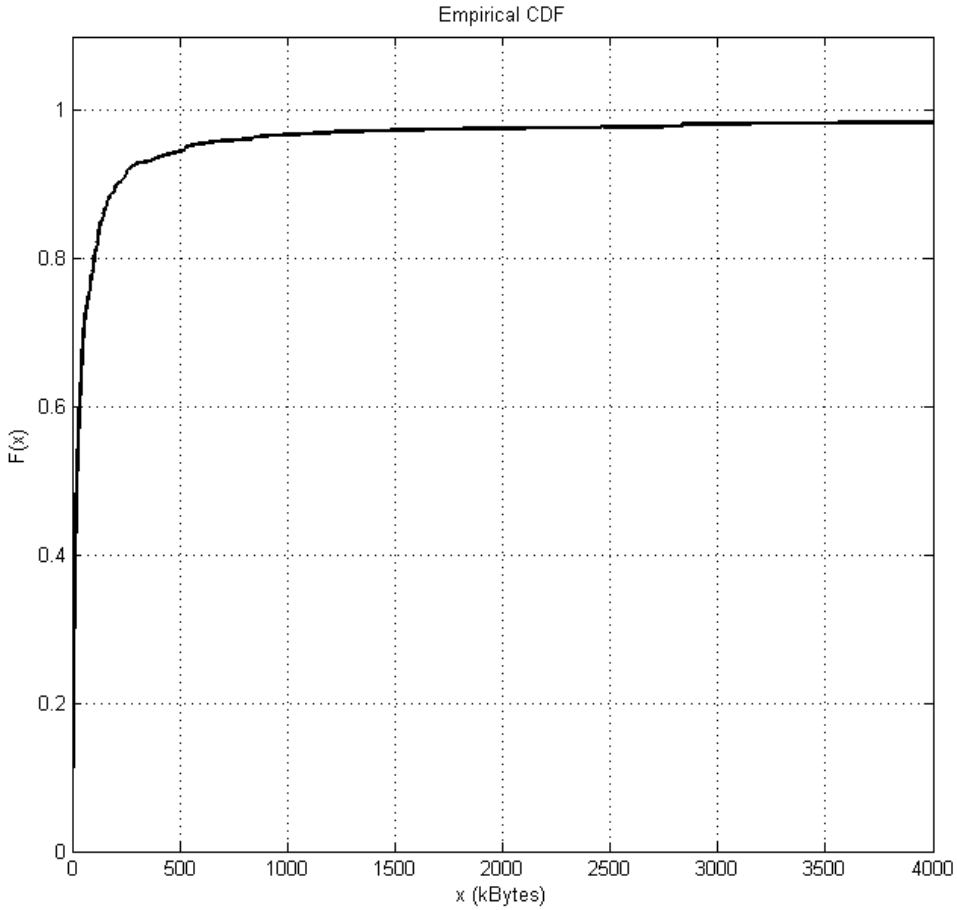

Figure 11. Incoming emails for users, zoomed

5.1.1 Statistical Tests' Results for the Overall Traffic

The results of the tests we used are shown below. The Relative Percentage Error is calculated at different percentiles, so we are able to observe how well and at which percent of the quantiles we are able to predict our traffic. We start with the QQ-plot, which is our graphical test and then we present the results for the numerical tests.

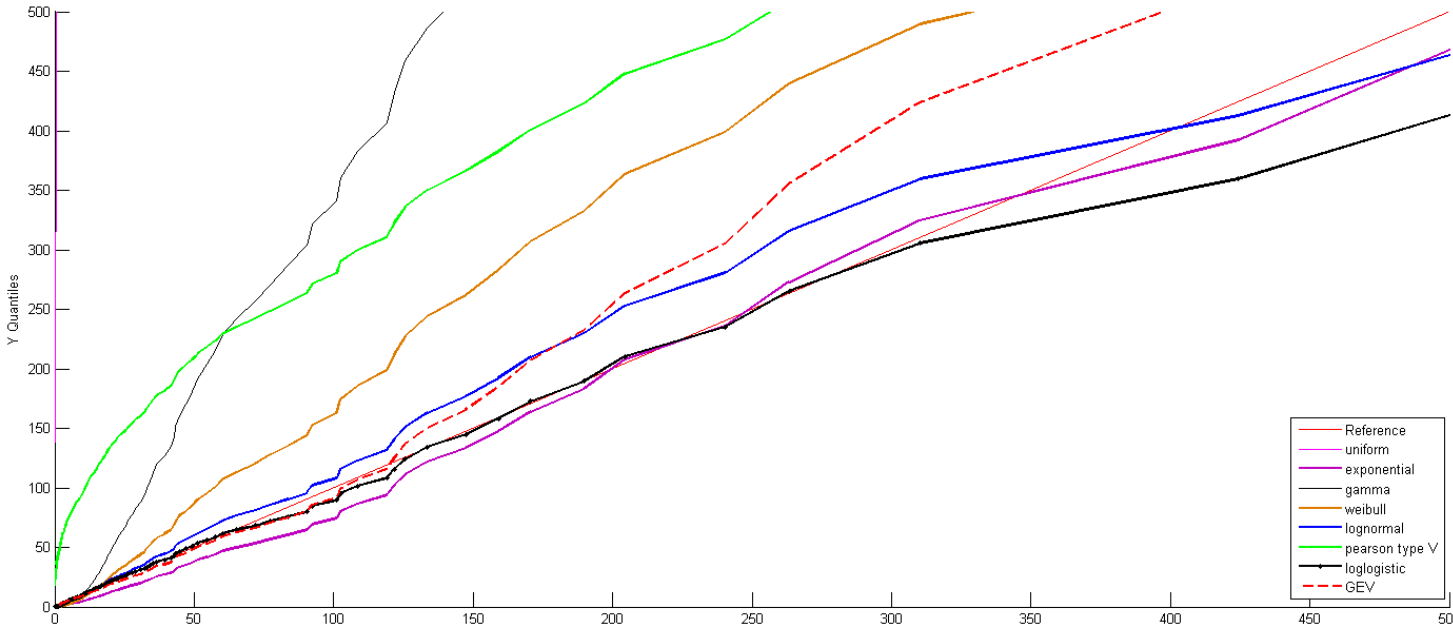

Figure 12. Incoming emails for users, QQ plot

Kolmogorov – Smirnov Test

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
0.9481	0.4957	0.2243	0.1220	0.0404	0.0276	0.0266

Anderson – Darling Test

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
754300	95700	2378900	4131500	500	200	200

Relative Percentage Error at 100%

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
7726.1	110.3	88.5	66.9	45.5	120.3	972.7

Relative Percentage Error at 98%

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
7344.9	43.97	33.543	10.026	4.0012	2.791	4.0404

Kullback – Leibler Divergence Test

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
14.3275	5.2479	1.8222	1.3598	1.0026	6.1217	10.8430

From the test results, we can see that Log Logistic distribution provides the closest fit to our data, followed by the lognormal and Generalized Extreme Value (GEV) distributions. The KS test, AD test and RPE at 98% percent of the quantiles agree. That means that the cdfs are close and we do not lose accuracy on the outliers. Since the KS test and AD test agree, Log logistic is closest to both the tails and the main body of the distribution.

The RPE results are so much different between the 98% and the 100% of the quantiles because of the outliers, which tend to have extremely large sizes, something the distributions cannot predict. Therefore, these outliers, usually the 1% of our traffic, cause bad results at the RPE at the 100%. With the exception of the outliers, we can predict with good accuracy the expected traffic, with the use of the loglogistic distribution (less than 3% RPE).

The KL test agrees with the RPE results at 100% of the quantiles, indicating in this case that the lognormal distribution provides the highest accuracy. As shown by the RPE results for the whole

traffic (100% of the quantiles), however, this "accuracy" is extremely low (45.5% error). Hence, the result of the KL test are of minor importance in our work.

In figure 13 we plot the pdf of the raw data and the pdf of the best three distributions according to the QQ –plot, AD, KS, KL tests. We can see that with the exception of some large size outliers we have a good fit for all three distributions.

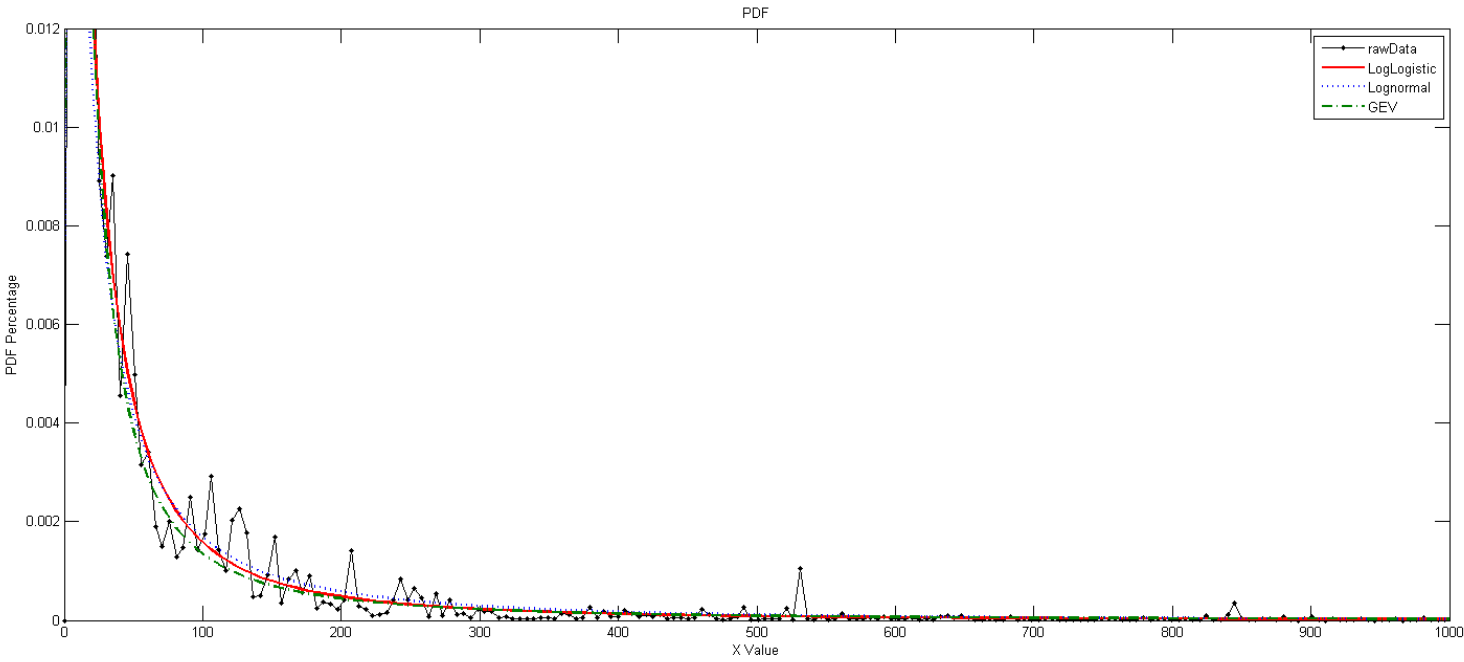

Figure 13. Incoming emails for users, PDF of real data and best fit distributions

The parameters for the above distributions are the following.

Log logistic

Mu: 3.183

Sigma: 0.9797

Lognormal

Mu: 3.2476

Sigma: 1.7843

GEV

Mu: 12.757

Sigma: 18.429

5.1.2 Statistical Test's Results for Daily Traffic

From the results derived for the daily traffic, we observe some small differences with the results for the overall traffic. Table 3 on page 42 summarizes the daily best distribution fit for the results.

Monday

KS-Test

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
0.9534	0.4587	0.1979	0.1277	0.0412	0.0362	0.0395

AD-Test

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
129950	13360	363690	686120	80	50	50

KL-Test

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
14.8874	5.7064	2.1751	1.7153	1.2348	5.7592	10.0531

RPE

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
14860	113	86	62	49	32	47

RPE at 98%

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
14040	59.387	42.711	14.296	5.7511	5.0683	7.7576

Tuesday

KS-Test

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
0.9494	0.4520	0.2122	0.1132	0.0377	0.0343	0.0466

AD-Test

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
114240	11540	362590	670350	40	30	40

KL-Test

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
14.1898	5.1843	2.1886	1.7228	1.3271	6.2814	12.1485

RPE

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
8536	108.7	87.9	64.3	52.6	32.9	61

RPE at 98%

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
8103.4	51.238	39.596	12.46	3.3756	2.4675	7.2352

Wednesday

KS-Test

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
0.9346	0.4893	0.2203	0.1246	0.0670	0.0407	0.0418

AD-Test

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
120670	14330	386950	736100	100	50	50

KL-Test

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
14.1871	5.1455	2.3124	1.9256	1.8948	8.8837	12.7816

RPE

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
1.0e+03 *						
5941.4	104.1	78.8	60.2	43.6	64	58.6

RPE at 98%

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
5653.5	47.957	33.018	13.839	7.8332	7.2985	3.9234

Thursday

KS-Test

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
0.9161	0.6196	0.2882	0.1532	0.0797	0.0437	0.0394

AD-Test

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
112190	28380	431210	698590	300	140	50

KL-Test

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
14.3368	5.7016	2.7669	2.3065	2.5366	9.3413	14.5211

RPE

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
2400.7	124.3	85.3	67.7	56.6	156.8	1731

RPE at 98%

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
2315.1	76.081	52.823	32.976	28.849	27.937	24.97

Friday

KS-Test

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
0.9293	0.4482	0.1796	0.1010	0.0330	0.0349	0.0380

AD-Test

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
94940	11860	337070	639460	40	40	30

KL-Test

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
13.6855	4.7885	2.3624	1.9872	3.2823	11.8015	15.5895

RPE

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
4462.7	98.5	74.6	50.8	35.7	15.6	117.3

RPE at 98%

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
4258.7	49.952	35.616	10.057	4.0172	5.57	10.261

Saturday

KS-Test

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
0.9560	0.4130	0.1884	0.1025	0.0518	0.0532	0.0521

AD-Test

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
54030	4320	155130	313260	30	30	30

KL-Test

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
14.7850	5.6624	2.6706	2.2129	1.7231	7.5606	11.0771

RPE

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
13714	97	75	46	35	27	34

RPE at 98%

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
12951	59.822	43.437	13.591	5.2194	6.0003	10.117

Sunday

KS-Test

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
0.9414	0.3772	0.1885	0.1110	0.0550	0.0665	0.0815

AD-Test

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
60350	4550	179630	386140	40	40	70

KL-Test

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
14.3734	5.3604	2.8900	2.4794	2.3764	8.7345	14.2054

RPE

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
4304.1	93.6	79.8	67.6	54.1	22.3	323.7

RPE at 98%

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
4119.2	27.258	18.984	7.287	3.1215	2.6302	6.9723

	Metric/best distribution	Mean	CV	mu	sigma
Monday	Email size in MB log logistic	0.0421	1.2926	3.1742	0.9063
Tuesday	Email size in MB log logistic	0.0446	1.2799	3.0445	0.9703
Wednesday	Email size in MB log logistic	0.046	1.2922	3.2761	0.9537
Thursday	Email size in MB GEV	0.0427	1.2978	12.5442 (GEV K:1.4351)	19.1569
Friday	Email size in MB lognormal	0.0441	1.3304	3.2772	1.8859
Saturday	Email size in MB lognormal	0.0421	1.2736	3.1877	1.6525
Sunday	lognormal	0.0406	1.3544	3.0538	1.6123

Table 3. Daily best distribution fit

5.2. Incoming Traffic from System

This section focuses on the incoming system traffic. This means that we only consider system emails and we want the recipient address to match the domain tuc.gr.

First, the statistics:

Total Number of Emails: 318944 **avg per day:** 45563 CV: 0,5095
Unique message Ids: 161155 **avg per day:** 23022 CV: 0.28
Total number of distinct senders: 8 **avg per day:** 1 CV: 0
Total number of distinct recipients: 3277 avg per day: 468 CV: 0.05
Total bytes: 1,82Gbytes **avg per day: 266,08**Mbytes CV: 0.4834
AVG # Distinct recipients/msg: 1.4 CV: 2.224

Next, we have the traffic calculated as emails/bytes/senders/recipients per hour and per day.

The plots are again normalized with the maximum value of each one.

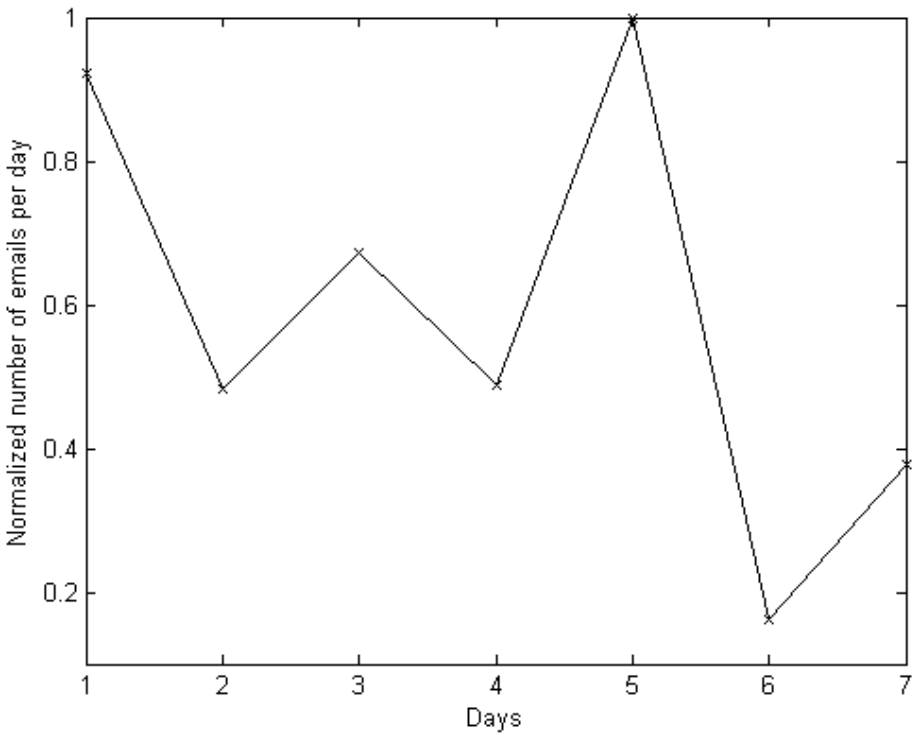

Figure 14. Incoming emails from system, emails per day

Max emails per day: 77719

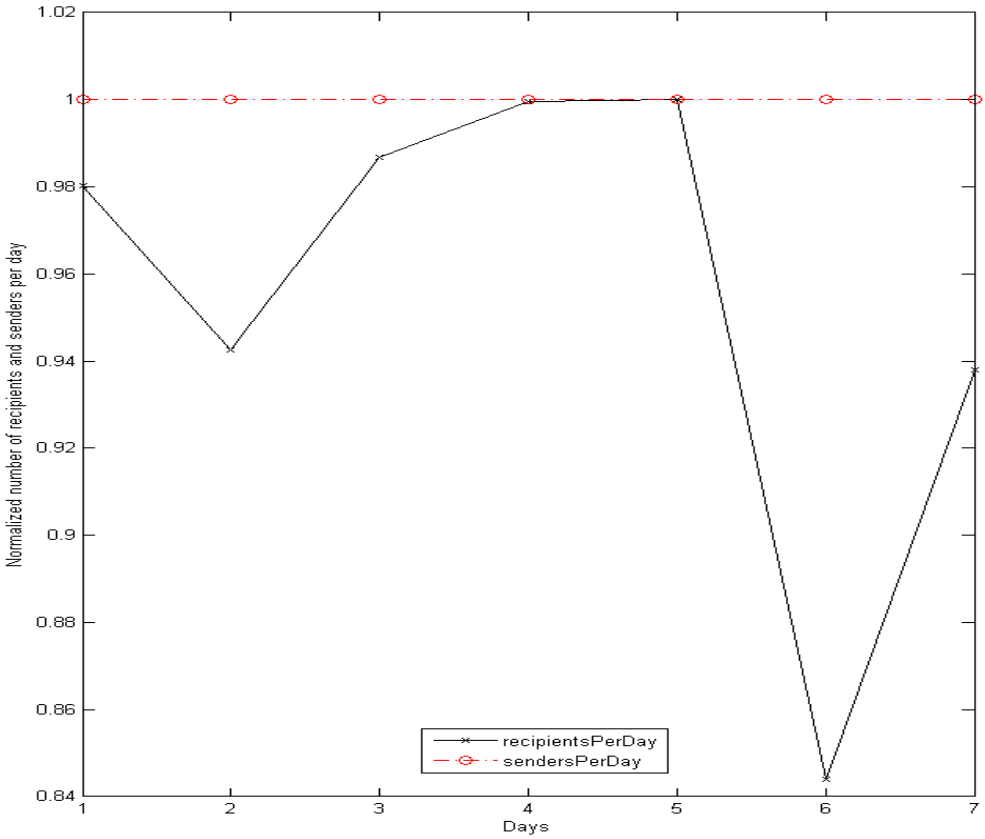

Figure 15. Incoming emails from system, recipients and senders per day

Max senders per day: 8

Max recipients per day: 2575

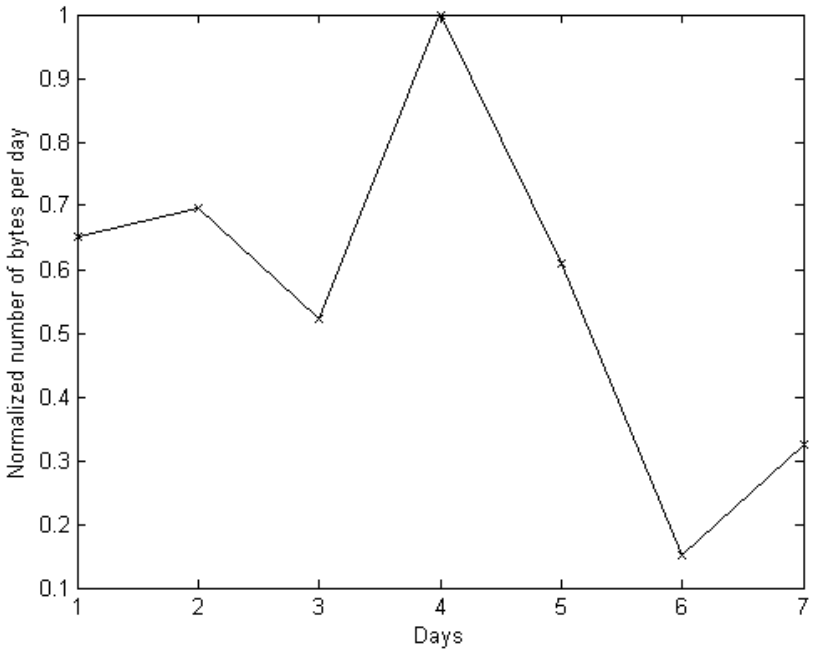

Figure 16. Incoming emails from system, bytes per day

Max bytes per day: 471MB

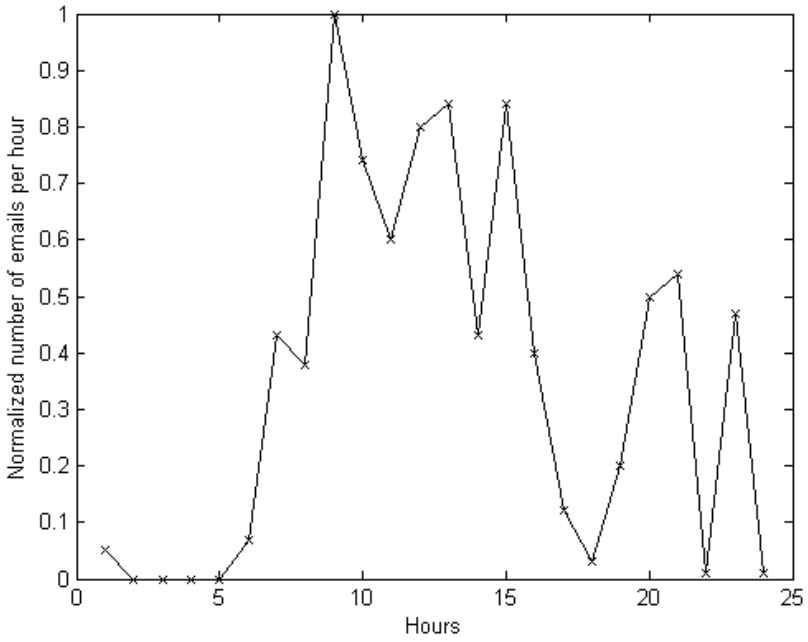

Figure 17. Incoming emails from system, emails per hour

Max emails per hour: 37766

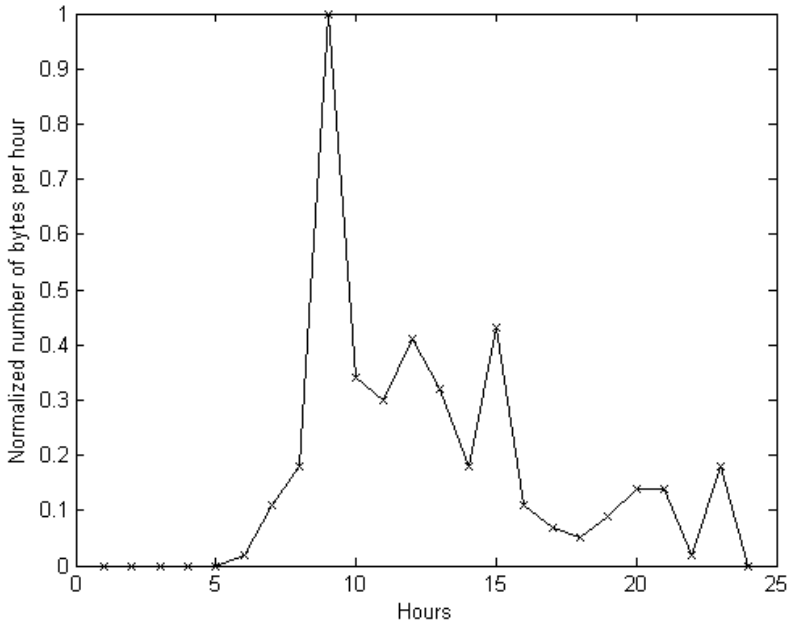

Figure 18. Incoming emails from system, bytes per hour

Max byte per hour: 446MB

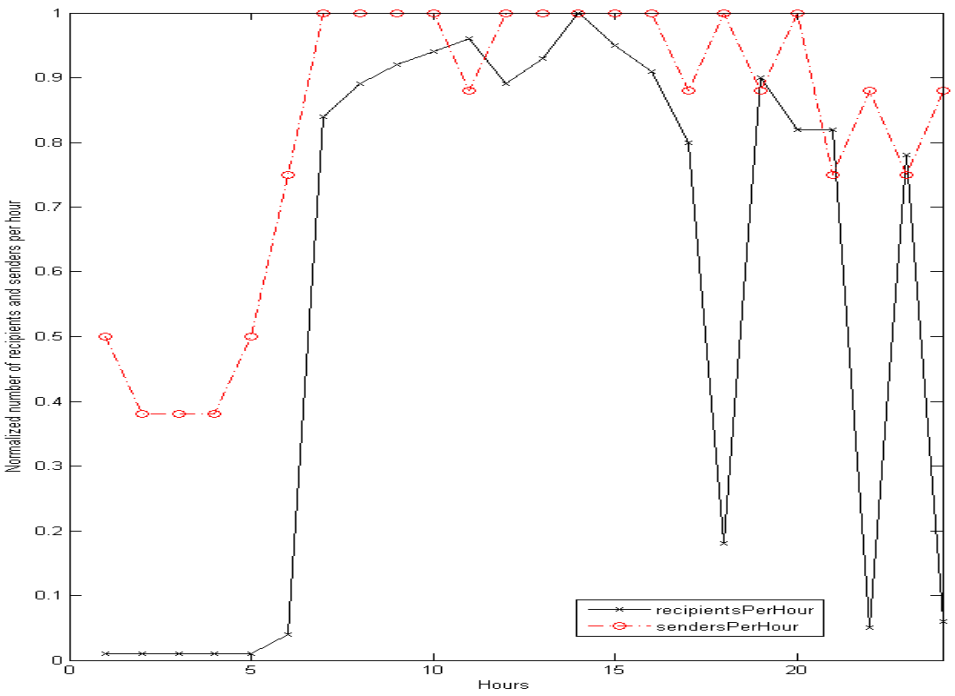

Figure 19. Incoming emails from system, recipients and senders per hour

Max senders per hour: 8

Max recipients per hour: 2473

As we can see, Thursday and Friday appear to be the most active days, i.e, the days that users receive messages from the system servers. The most active hours appear to be 8-10 am. Compared with the non-system incoming traffic we can see that the same patterns appear for the daily and the hourly traffic with the exception that the system's incoming messages are about three times larger.

Some of our hourly statistics are shown in the table 4.

	Metric	Minimum	Maximum	Average	CV
Monday	Emails/Hour	7	10464	2985	1.2299
	MB/Hour	0.1130	39.7172	12.7	1.1670
Tuesday	Emails/Hour	2	12620	1564.2	1.84
	MB/Hour	0.0127	76.227	13.504	1.50
Wednesday	Emails/Hour	2	13521	2174	1.586
	MB/Hour	0.0140	69.348	10.2571	1.5391
Thursday	Emails/Hour	3	10462	1580.8	1.688
	MB/Hour	0.0252	322.3831	19.665	3.3015
Friday	Emails/Hour	4	14581	3238.3	1.4379
	MB/Hour	0.0323	53.6914	12.1249	1.3035
Saturday	Emails/Hour	2	11388	543.60	4.3527
	MB/Hour	0.0377	35.6942	3.1105	2.6928
Sunday	Emails/Hour	1	11444	1225.9	2.083
	MB/Hour	0.0086	65.6852	6.3266	2.365

Table 4. Emails and traffic volume per hour

We observe from both the tables 4 and 5, that no significant changes take place in the number of the emails, the volume of email traffic and the number of recipients and senders throughout the week, including the weekend. In addition, the system's emails have much bigger Coefficient of Variation values, which means they are more diverse in size than the non-system incoming emails.

	Metric	Minimum	Maximum	Average	CV
Monday	Recipients/Hour	3	1987	809	1.113
	Senders/Hour	1	7	4.45	0.3912
Tuesday	Recipients/Hour	1	2167	554.2	1.5024
	Senders/Hour	1	8	4.12	0.4650
Wednesday	Recipients/Hour	2	2017	696.12	1.2216
	Senders/Hour	1	7	4.1667	0.4988
Thursday	Recipients/Hour	3	1991	576.37	1.4472
	Senders/Hour	1	8	4.208	0.4803
Friday	Recipients/Hour	2	2106	776.3750	1.1730
	Senders/Hour	1	8	4.5	0.4279
Saturday	Recipients/Hour	1	1999	108.6	3.811
	Senders/Hour	1	5	2.47	0.4847
Sunday	Recipients/Hour	1	1985	400.75	1.7937
	Senders/Hour	1	7	3,541	0,4781

Table 5. Recipients and senders' per hour

Figure 20 presents the CDF of the emails' size.

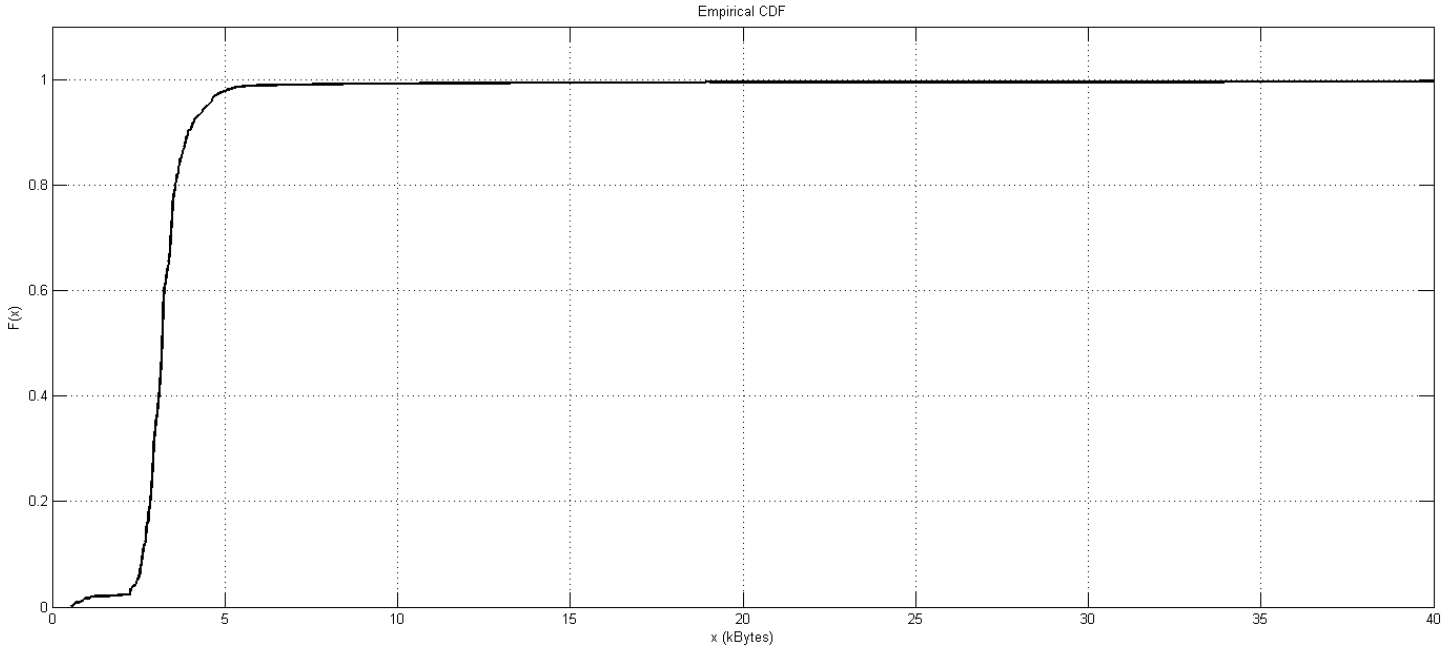

Figure 20. Incoming emails from system, CDF of real data

5.2.1 Statistical Tests' Results for the Overall Traffic

The results of the tests we used are shown below. The slight "S" shape around the x=y line in the QQ plot, which all distributions appear to have, means that we should expect some outliers at small and big kilobytes. We can see that log logistic, GEV and lognormal appear to be the best distributions.

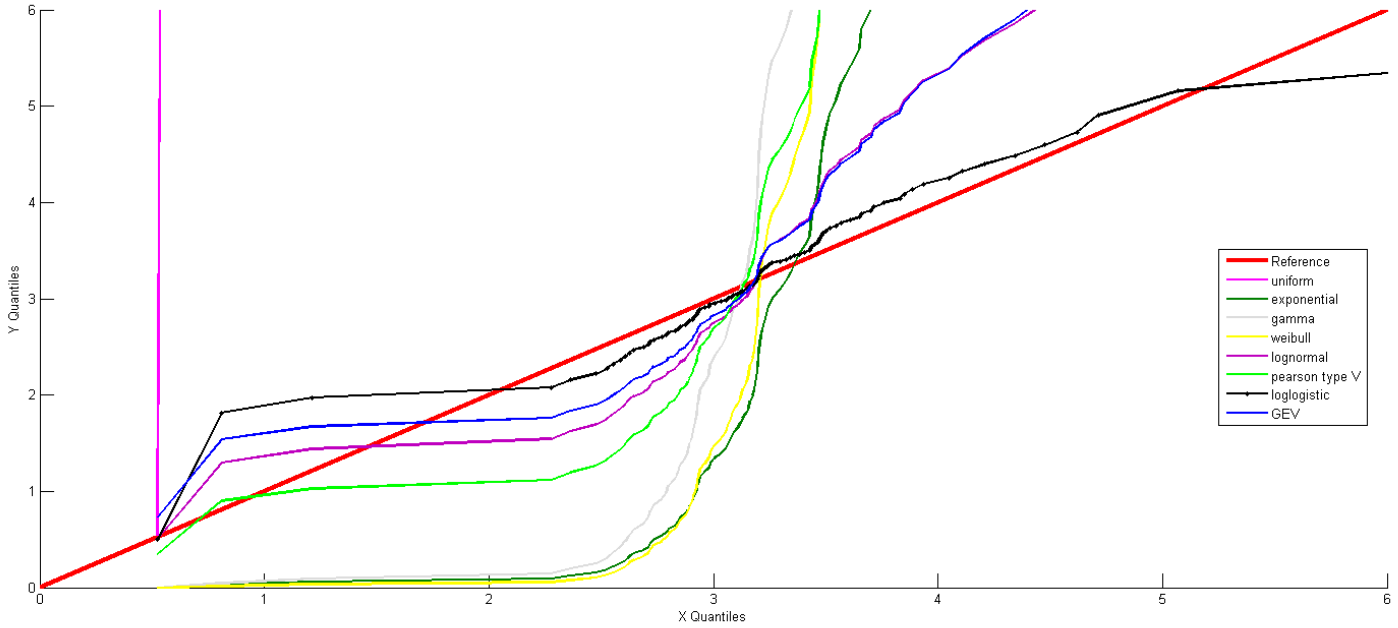

Figure 21. Incoming emails from system, QQ plot

K-S Test

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
0.9936	0.4266	0.4135	0.4332	0.2164	0.0932	0.2021

A-D Test

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
2355900	91600	334600	9557600	34800	6700	30200

K-L Test

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
18.3284	9.0661	8.4974	6.9554	28.9856	29.0730	28.9930

RPE

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
9552	99.8	99.8	98.9	96.1	95.6	96

RPE at 99.05%

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
9456.1	5.0335	5.0715	4.3081	0.80204	0.22073	0.74659

From the test results, we can see that Log Logistic seems to be closest to our real data. The lognormal and GeneralizeExtremeValue are following. The KS test, AD test and RPE at 99.05% of the quantiles agree. That means that the cdfs are close and we do not loose accuracy on the outliers. Because the KS test and the AD test agree, we know that Log logistic is closest to both the tails and the main body of the distribution.

In figure 22, we plot the pdf of the raw data and the pdf of the best three distributions according to the QQ–plot, AD and KS tests.

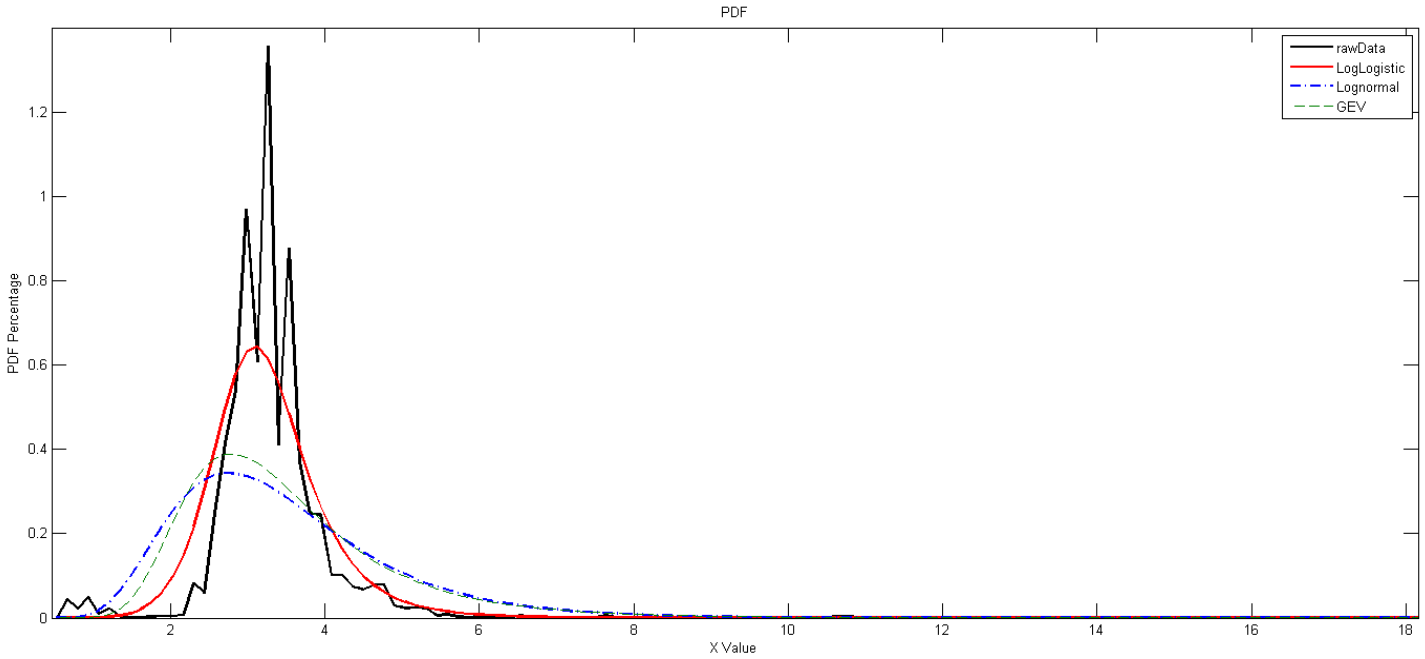

Figure 22. Incoming emails from system, PDF of real data and best fit distributions

The parameters for the above distributions are the following.

Log logistic

Mu: 1.1597

Sigma: 0.1231

GEV

Mu: 2.8793

Sigma: 0.9456

Lognormal

Mu: 1.1642

Sigma: 0.3892

We can observe that loglogistic is closer to the original shape. In addition, there are some outliers around 1 KB as QQ-Plot showed, that no distribution is able to predict.

5.2.2 Statistical Test's Results for Daily Traffic

From the results derived for the daily traffic, we observe some small differences with the results for the overall traffic. Table 6 on page 64 summarizes the daily best distribution fit for the results.

Monday

KS-Test

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
0.9938	0.4300	0.3602	0.4787	0.2215	0.1082	0.2161

AD-Test

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
462600	19900	18500	2187700	6800	1300	6600

KL-Test

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
17.6260	8.2620	21.6161	7.4460	28.3412	28.3917	28.3461

RPE

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
52455	80	70	79	40	34	40

RPE at 99.05%

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
51450	53.427	41.545	53.239	7.8747	2.1792	8.544

Tuesday

KS-Test

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
0.9907	0.5490	0.4579	0.4379	0.3039	0.1774	0.2828

AD-Test

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
275300	15400	162200	1144100	6000	1600	4700

KL-Test

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
18.7339	9.4098	6.9181	6.5437	29.3670	29.4488	29.3320

RPE

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
61234	116	117	99	76	71	74

RPE at 99.05%

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
60083	52.245	54.626	36.175	7.7578	2.0357	5.8076

Wednesday

KS-Test

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
0.9918	0.3643	0.3793	0.4346	0.2179	0.1048	0.1821

AD-Test

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
335610	12759	16765	1561600	4003.8	624.31	3311.6

KL-Test

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
10.9240	7.2699	9.2385	6.0579	30.3506	30.5208	29.7683

RPE

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
47074	86	81	84	49	43	49

RPE at 99.05%

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
46159	48.77	42.755	47.34	8.1421	1.922	7.6808

Thursday

KS-Test

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
0.9917	0.6454	0.4829	0.4533	0.2768	0.1376	0.2542

AD-Test

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
335600	12800	16800	1561600	4000	600	3300

KL-Test

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
17.6473	8.2303	11.0989	7.0143	28.2654	2 8.3146	28.2798

RPE

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
52168	147	145	101	74	70	73

RPE at 99.05%

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
51198	86.137	89.329	39.815	6.8993	2.0807	5.19

Friday

KS-Test

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
0.9866	0.4356	0.2924	0.4142	0.2021	0.0966	0.1969

AD-Test

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
418800	20600	18800	2295400	7000	1300	6500

KL-Test

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
16.1179	7.1406	26.5316	7.6961	27.1610	27.1940	27.1094

RPE

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
1229.0	81	70	80	62	58	61

RPE at 99.05%

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
12105	28.617	15.798	27.496	6.1316	1.8377	6.1671

Saturday

KS-Test

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
0.9921	0.4690	0.4391	0.4813	0.3640	0.2668	0.3410

AD-Test

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
90410	3870	12020	369910	225 0	1130	2160

KL-Test

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
19.4165	9.9904	9.3330	7.8623	29.8042	29.8395	29.7951

RPE

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
25583	99	99	96	88	86	88

RPE at 99.05%

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
25165	16.689	16.157	14.111	3.5817	1.5278	3.4477

Sunday

KS-Test

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
0.9886	0.4263	0.4049	0.3929	0.2972	0.2053	0.2649

AD-Test

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
148150	7310	20900	786490	4140	2360	3700

KL-Test

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
17.2784	8.1166	7.7919	6.6454	27.6799	27.9326	27.9152

RPE

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
3971.5	98	97.7	95.9	84.8	82.2	84.4

RPE at 99.05%

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
3969.4	20.277	19.969	18.633	5.5171	2.6965	5.1312

	Metric	Mean	CV	mu	sigma
Monday	Email size log logistic	0.0043	13.5159	1.1456	0.0919
Tuesday	Email size log logistic	0.0086	17.4168	1.2208	0.1159
Wednesday	Email size log logistic	0.0047	11.7777	1.1402	0.1308
Thursday	Email size log logistic	0.0124	21.8120	1.1939	0.1331
Friday	Email size in MB log logistic	0.0037	4.9482	1.1541	0.1177
Saturday	Email size in MB log logistic	0.0057	23.8724	1.1380	0.1073
Sunday	log logistic	0.0052	9.4156	1.1003	0.1783

Table 6. Daily best distribution fit

5.3. Outgoing Traffic for Users

This section focuses on the outgoing user's traffic. This means that we do not have system emails at all and we wanted the senders' addresses to match the domain tuc.gr.

First, we have some statistics:

Total Number of Emails: 55645 **avg per day:** 7949 CV: 0.4638
Unique message Ids: 20146 **avg per day:** 2879.1 CV: 0.4108
Total number of distinct senders: 2594 **avg per day:** 370 CV: 0.3117
Total number of distinct recipients: 5224 avg **per day:** 746 CV: 0.3886
Total bytes: **19.83** GB avg **per day: 2.83** GB CV: 1.2480
AVG # Distinct recipients/msg: 1.27 CV: 1.16

Next, we have the traffic calculated as emails/bytes/senders/recipients per hour and per day.

The plots are normalized respectively to maximum value.

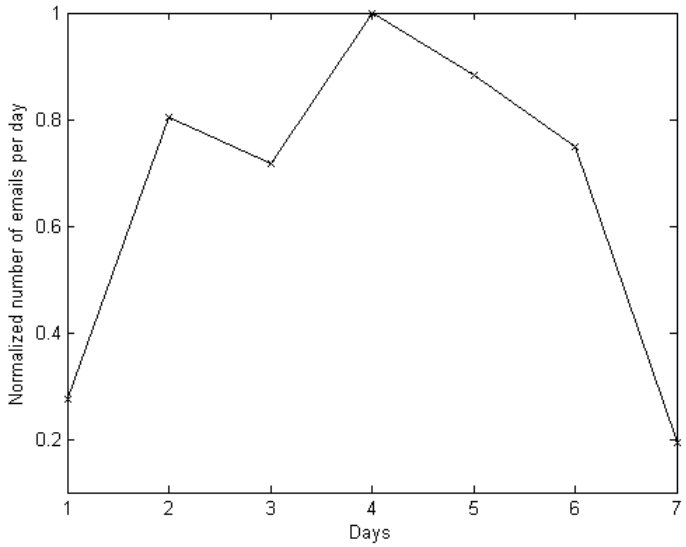

Figure 23. Outgoing emails from users, emails per day

Maximum emails per day: 12042

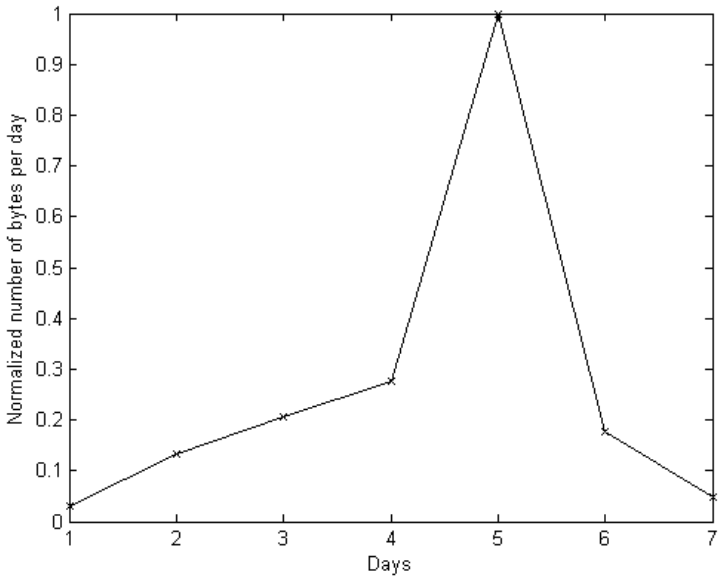

Figure 24.Outgoing emails from users, bytes per day

Maximum bytes per day: 10.58GB

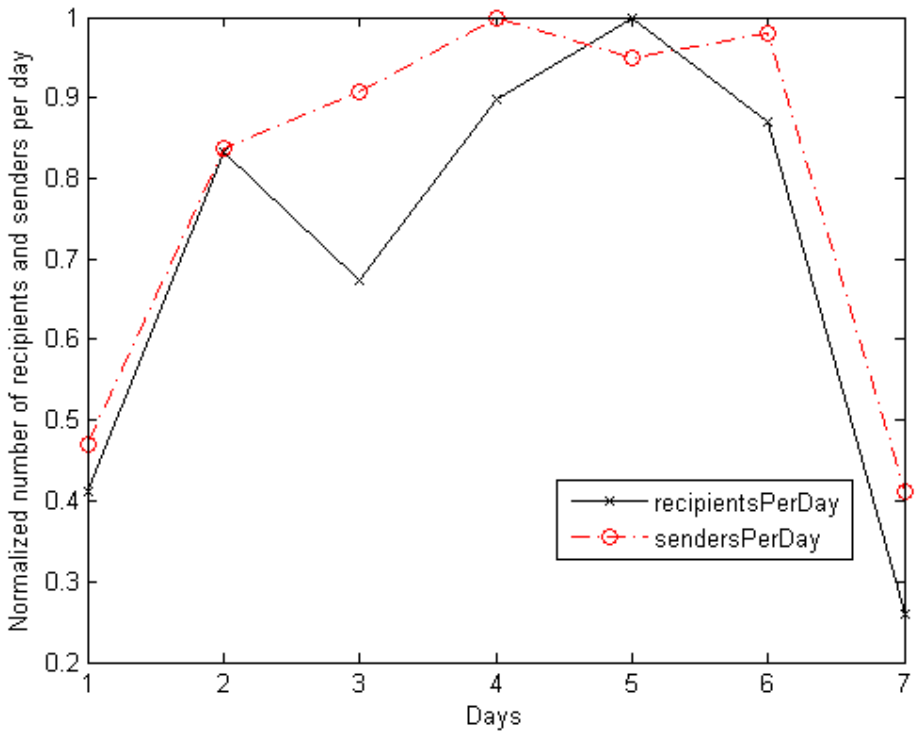

Figure 25. Outgoing emails from users, recipients and senders per day

Maximum recipients per day: 1427

Maximum senders per day: 951

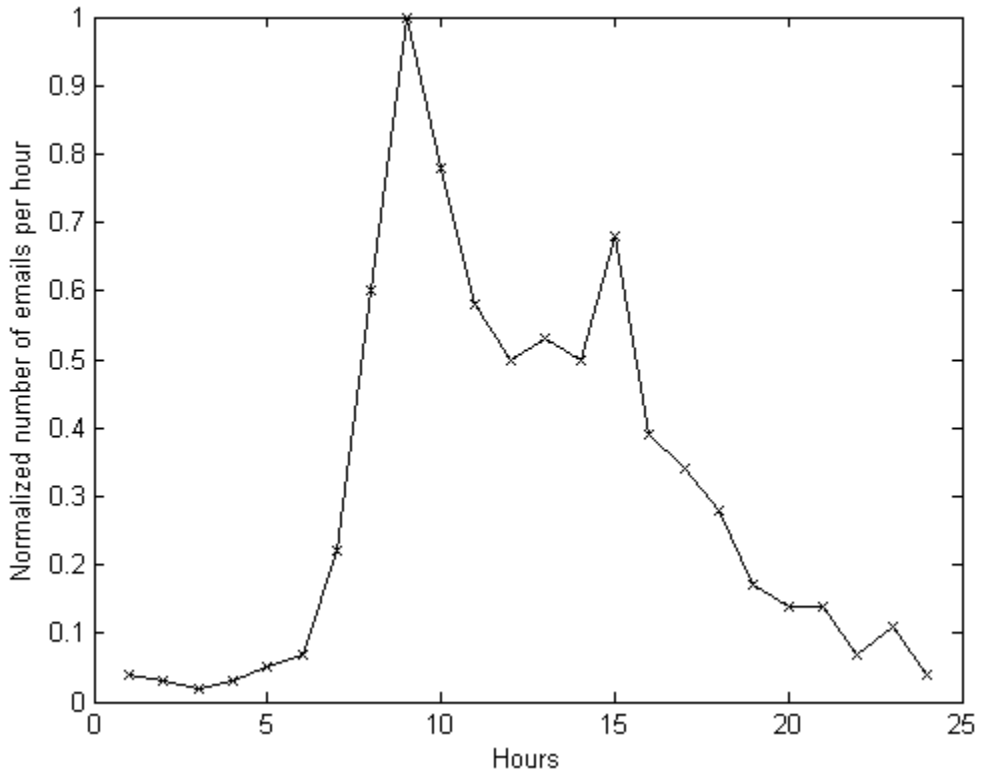

Figure 26. Outgoing emails from users, emails per hour

Maximum emails per hour: 7623

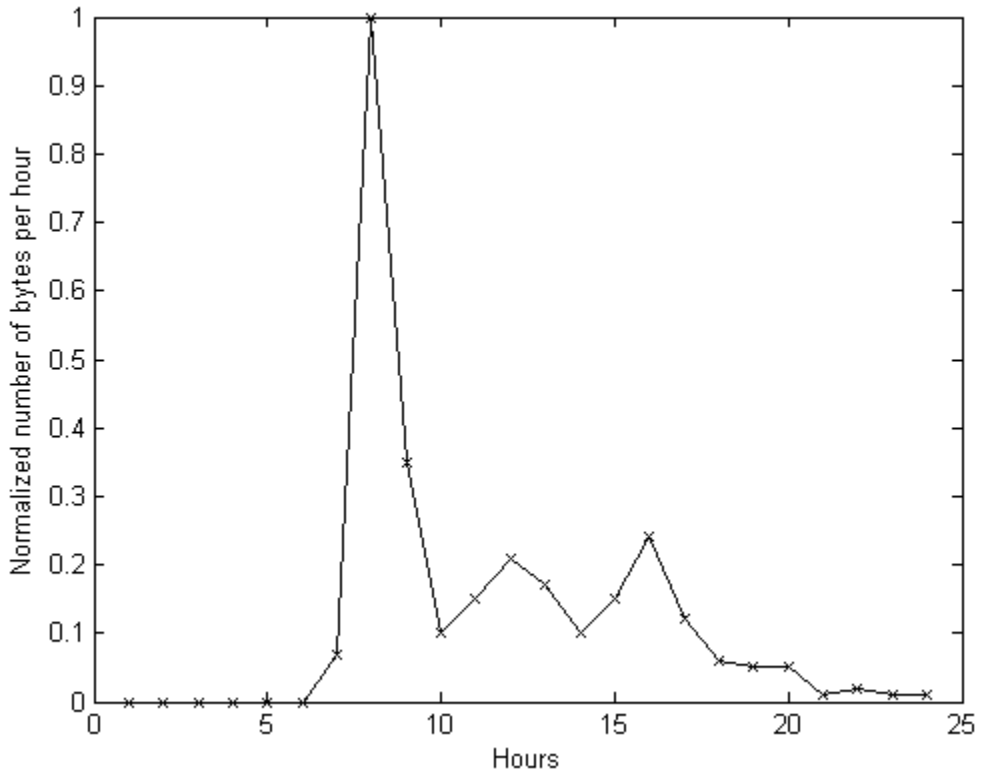

Figure 27. Outgoing emails from users, bytes per hour

Maximum bytes per hour: 6.87GB

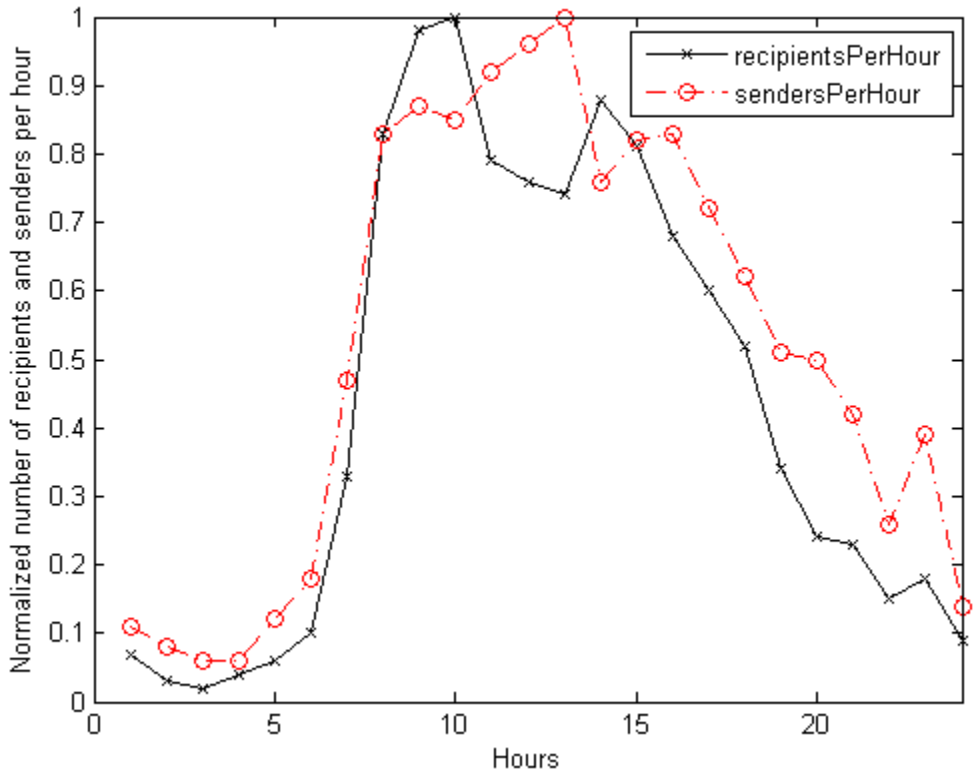

Figure 28. Outgoing emails from users, recipients and senders per hour

Maximum recipients per hour: 1427

Maximum senders per hour: 542

We observe that our users prefer to send their emails on Thursday and Friday despite the fact that they receive most emails during Tuesday and Wednesday. In addition, they prefer to send their emails early in the morning, almost the same time they receive most of the emails.

Some of our hourly statistics are shown in the tables below.

	Metric	Minimum	Maximum	Average	CV
Monday	Emails/Hour	29	1825	402.95	0.994
	MB/Hour	0.1346	306.69	60.7934	1.325
Tuesday	Emails/Hour	15	1288	359.41	1.0996
	MB/Hour	0.2558	326.4399	92.988	1.2105
Wednesday	Emails/Hour	22	2625	501.75	1.38
	MB/Hour	0.1001	469.2359	125.0944	1.1620
Thursday	Emails/Hour	12	2290	443	1.2494
	MB/Hour	0.1409	6566.1	451.489	3.021
Friday	Emails/Hour	16	1251	376.208	0.8928
	MB/Hour	0.0475	480.4257	79.9129	1.5966
Saturday	Emails/Hour	15	393	97.5417	0.9363
	MB/Hour	0.0535	168.87	22.176	1.9798
Sunday	Emails/Hour	9	861	137.6667	1.3287
	MB/Hour	0.0162	124.131	13.8329	2.0654

Table 7. Emails and traffic volume per hour

	Metric	Minimum	Maximum	Average	CV
Monday	Recipients/Hour	15	453	169.79	0.82
	Senders/Hour	8	204	93.04	0.7472
Tuesday	Recipients/Hour	8	528	141.35	0.9658
	Senders/Hour	7	248	85.583	0.8673
Wednesday	Recipients/Hour	12	815	183.5833	1.0720
	Senders/Hour	11	228	93.54	0.7972
Thursday	Recipients/Hour	7	729	172.2917	1.0450
	Senders/Hour	8	188	82.50	0.7835
Friday	Recipients/Hour	7	510	171.2083	0.8310
	Senders/Hour	6	230	90.625	0.7810
Saturday	Recipients/Hour	7	170	41.1667	0.8429
	Senders/Hour	7	102	26.5833	0.7351
Sunday	Recipients/Hour	5	351	65.667	1.2490
	Senders/Hour	4	149	37.0833	0.9130

Table 8. Recipients and senders' numbers per hour

In figure 29 we present is the CDF of the emails' size

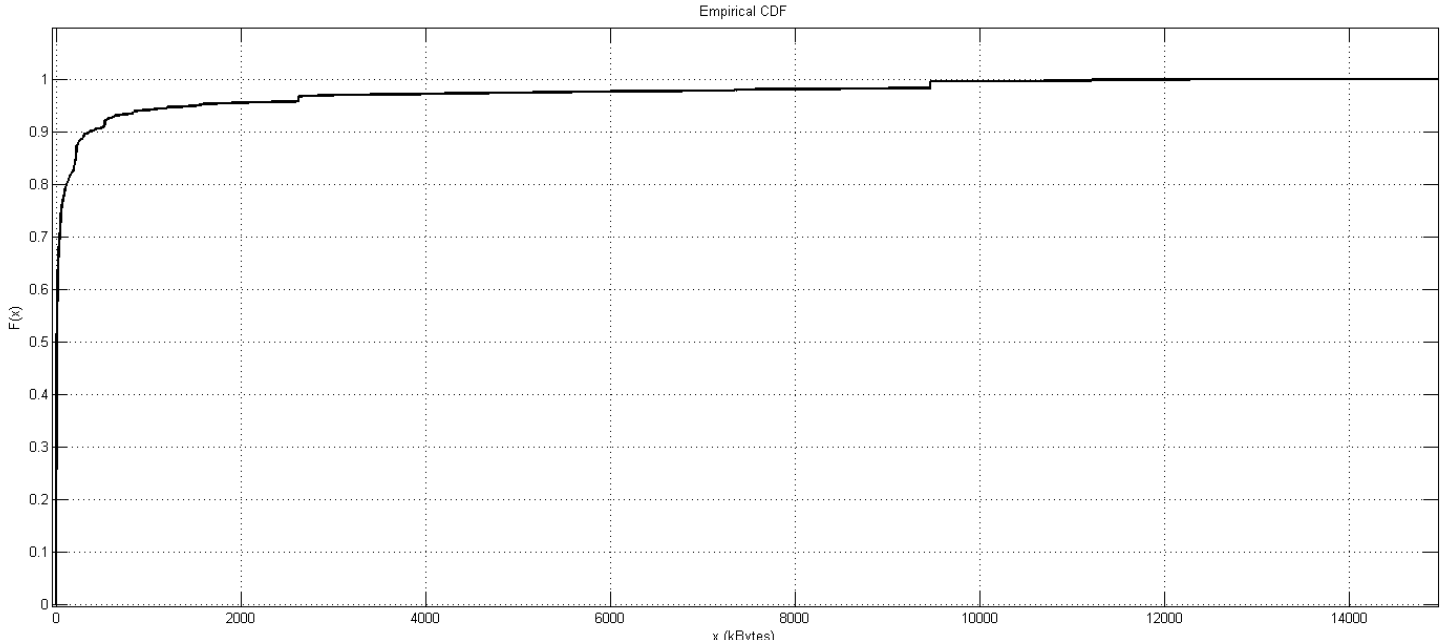

Figure 29. Outgoing emails from users, CDF of real data

We can see that with 90% probability the size of the send emails is below 500KB.

5.3.1 Statistical Tests' Results for the Overall Traffic

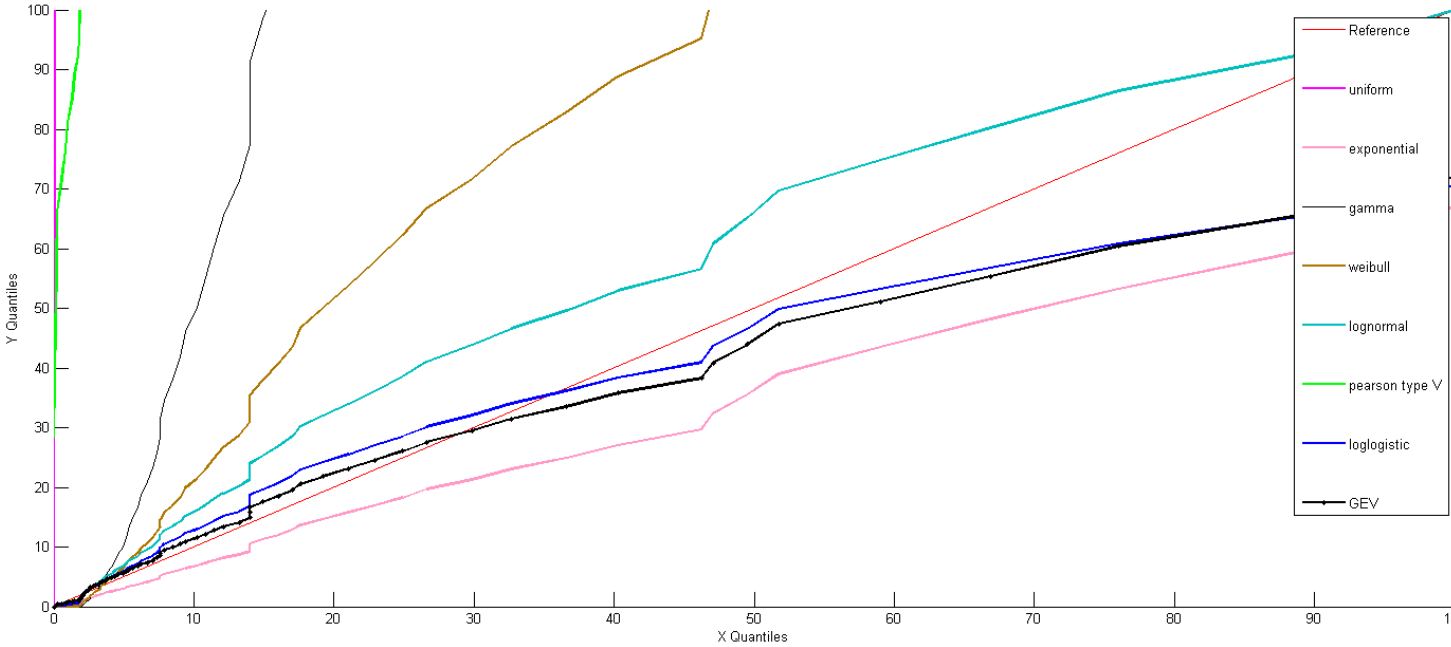

Figure 30. Outgoing emails from users, QQ plot

K-S Test

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
0.9487	0.1791	0.2406	0.1552	0.0947	0.0901	0.0608

A-D Test

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
379600	75200	990500	649700	900	800	300

K-L Test

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
16.4917	7.3711	2.0439	1.6460	1.7064	7.2543	11.4637

RPE

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
16845	118	74	51	64	308	3077

RPE at 99%

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
15913	62.96	37.144	14.298	14.517	16.846	8.464

RPE at 99% of the quantiles has the same results with A-D, QQ-plot and KS test while KL test agrees with RPE at 100% of the quantiles.

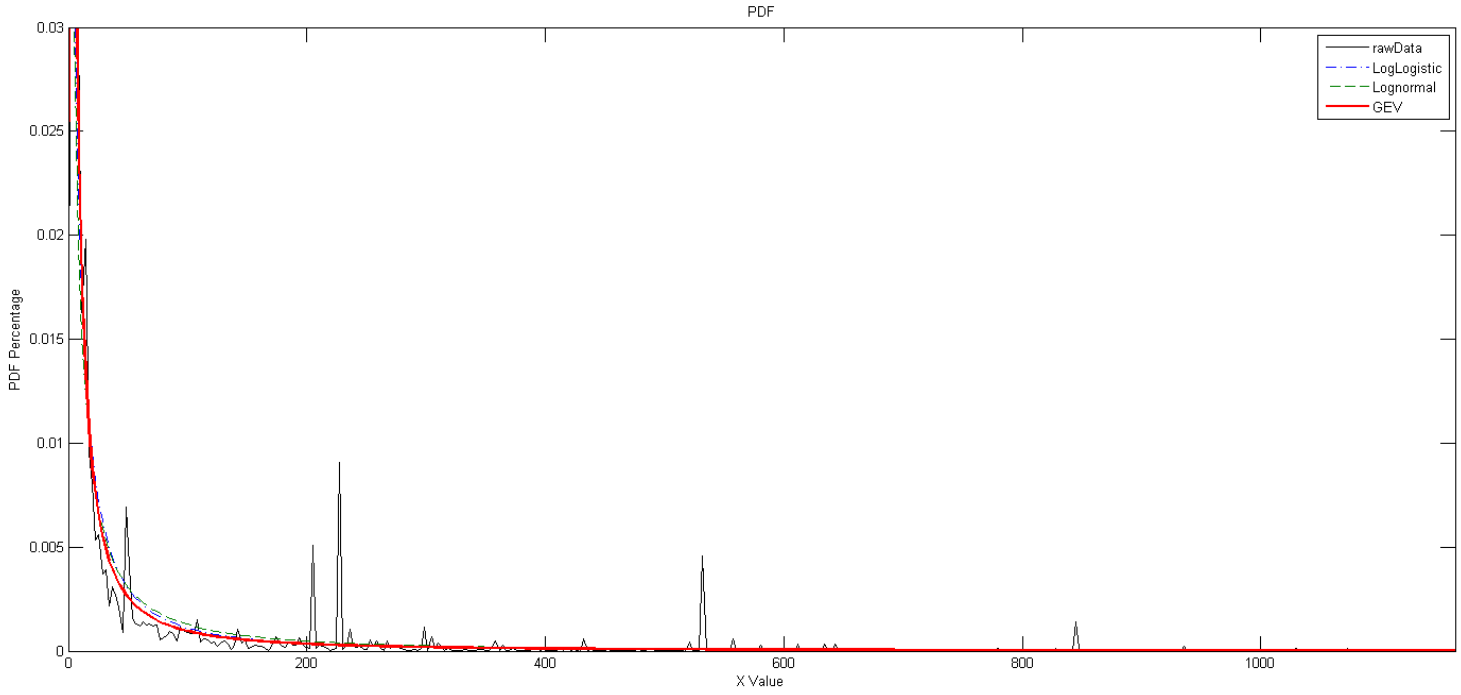

Figure 31. Outgoing emails from users, PDF of real data and best fit distributions

In figure 31 we plot the Pdf's of the best fit distributions as well as the one of the raw data. As expected, we have outliers who affect directly our accuracy.

GEV seems to be the best distribution for this kind of traffic. This is the only exception, since all the other kinds of traffic are modeled with a log logistic distribution.

The parameters for the above distributions are the following.

Log logistic **Lognormal**

Mu: 2.569 Mu: 2.7704

Sigma: 1.2689 Sigma: 2.2781

GEV

Mu: 6.059

Sigma: 10.72

5.3.2 Statistical Test's Results for Daily Traffic

We can observe that the daily results follow the same distribution as the whole week, with rare exceptions mostly happening at the weekends. Table 9 summarizes the daily best distribution fit.

Monday

KS-Test

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
0.9547	0.2227	0.2789	0.1994	0.1210	0.1014	0.0827

AD-Test

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
65340	12790	182700	281000	190	110	60

KL-Test

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
16.1191	7.2674	2.3998	1.9098	1.4398	4.1847	7.3140

RPE

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
19587	151	111	78	75	76	61

RPE at 99%

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
18526	85.244	59.613	12.836	6.2404	6.8784	4.5055

Tuesday

KS-Test

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
0.8913	0.1674	0.2024	0.1430	0.0808	0.0769	0.0732

AD-Test

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
38450	8600	138240	258040	100	100	40

KL-Test

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
14.5161	5.8076	3.2374	3.2477	5.3330	14.0198	17.7904

RPE

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
2295.1	108.8	710	53	43.5	70.3	762.5

RPE at 99%

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
2221	55.394	29.725	11.005	11.381	11.273	8.22

Wednesday

KS-Test

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
0.9667	0.1433	0.1841	0.1123	0.1244	0.1254	0.1122

AD-Test

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
77150	7730	156780	366410	150	180	150

KL-Test

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
16.7766	7.7407	3.1450	2.6559	2.2938	9.0617	14.1881

RPE

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
26698	98	72	51	33	91	1098

RPE at 99%

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
25163	47.355	31.842	10.404	9.7095	14.902	41.92

Thursday

KS-Test

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
0.8938	0.1600	0.2901	0.2183	0.2002	0.1514	0.1287

AD-Test

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
61220	19560	170120	324830	540	450	220

KL-Test

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
16.8440	7.9041	3.9134	4.5072	5.7837	11.7547	15.0268

RPE

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
4513.6	102	49.3	64.5	86.5	156.6	842.6

RPE at 99%

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
4272.8	87.458	43.264	48.631	55.452	59.739	50.671

Friday

KS-Test

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
0.9324	0.1878	0.1972	0.1468	0.0979	0.0959	0.0777

AD-Test

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
44040	8680	140720	261900	140	140	70

KL-Test

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
14.9695	6.1986	3.0836	2.9131	5.2740	13.0749	16.7743

RPE

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
3521.9	97.6	72	55.3	41.2	68.9	497.5

RPE at 99%

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
3374.3	41.707	24.898	10.283	11.76	13.842	19.546

Saturday

KS-Test

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
0.9365	0.1534	0.2596	0.1408	0.0608	0.0642	0.0544

AD-Test

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
12347	3186	41060	63823	14	15	9

KL-Test

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
15.2440	6.8192	3.0978	2.7148	3.5571	10.9084	15.9571

RPE

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
4034.9	148.5	108.4	72.5	72	61.6	65.2

RPE at 99%

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
3862.3	82.063	50.712	7.8928	4.3875	4.065	7.8841

Sunday

KS-Test

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
0.9211	0.2778	0.2910	0.2142	0.2111	0.1794	0.1736

AD-Test

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
19088	4508	56263	95517	194	163	116

KL-Test

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
16.1112	7.2764	3.5294	3.2812	3.1735	7.0327	10.9248

RPE

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
4817.3	118.7	80.9	68.3	73.1	63.6	157.9

RPE at 99%

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
4594.8	64.649	39.427	23.828	25.123	26.709	23.931

In table 9, we observe that the mean size of an email is roughly the same every day with the exception of Thursday where we have some outliers. The mean size is high and the CV is low which means that someone sent an email to many recipients carrying a large payload.

	Metric	Mean	CV	mu	sigma
Monday	Email size GEV	0.15	6.5869	3.9751 K=1.2060	5.5421
Tuesday	Email size GEV	0.2587	4.1858	6.2714 K=1.73	11.2009
Wednesday	Email size GEV	0.2493	6.5077	9.6553 K=1.7557	17.5032
Thursday	Email size GEV	1.0192	2.7097	7.9624 K=1.9317	15.5912
Friday	Email size in MB GEV	0.2124	4.7393	5.8214 K=1.7230	10.5480
Saturday	Email size in MB GEV	0.2273	5.4924	3.847 K=1.703	6.789
Sunday	Email size in MB GEV	0.1005	5.1468	2.5689 K=1.2373	3.4123

Table 9. Daily best distribution fit

5.4. Outgoing Traffic for System emails

This section focuses on the outgoing system's traffic. This means that we have system emails that are leaving our server.

First, we have some statistics:

Total Number of Emails: 320973 **avg per day:** 45853 **CV**: 0.5101
Unique message Ids: 161954 **avg per day:** 23137 **CV**: 0.5126
Total number of distinct senders: 8 **avg per day:** 1 **CV**: 0
Total number of distinct recipients: 3635 **avg per day:** 519 **CV**: 0.066
Total bytes: **1.90** GB **avg per day:** 278.01 Mbytes **CV**: 0.4711
AVG # Distinct recipients/msg: 29.75 **CV**: 1.3164

The total number of outgoing system emails is quite close to the incoming ones. This happens because these servers are used mainly for communication inside the university.

Next, we have the traffic calculated as emails/bytes/senders/recipients per hour and per day.

The plots, as always, are normalized with the maximum value of each one.

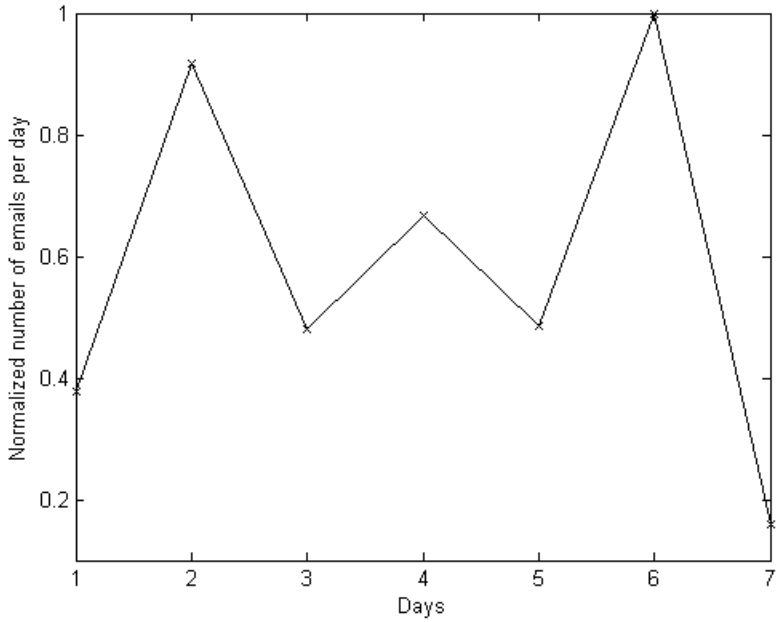

Figure 32. Outgoing emails from system, emails per day

Maximum emails per day: 78495

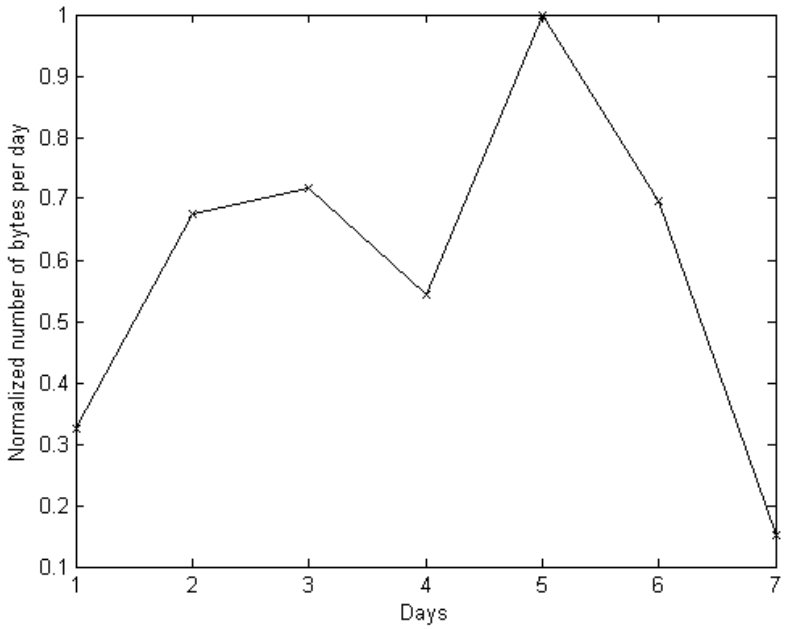

Figure 33. Outgoing emails from system, bytes per day

Maximum bytes per day: 0.4626 GB

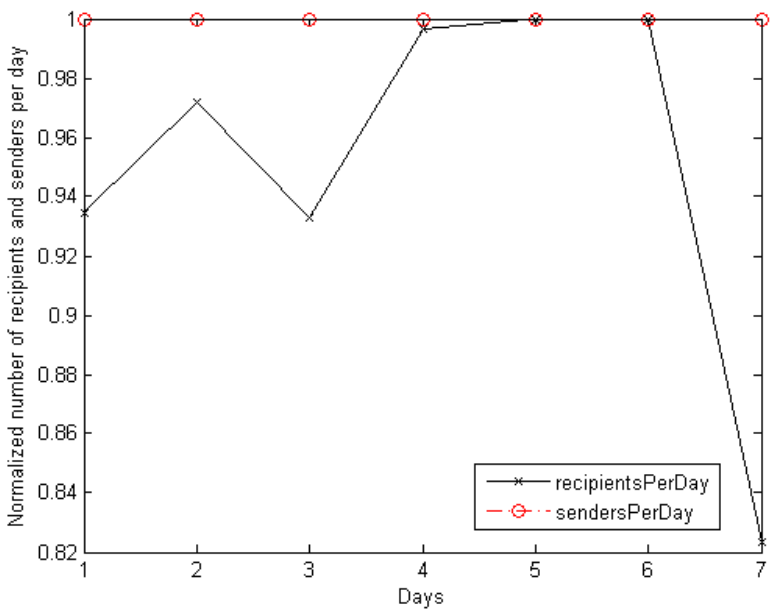

Figure 34. Outgoing emails from system, recipients and senders per day

Maximum recipients per day: 2675

Maximum senders per day: 8

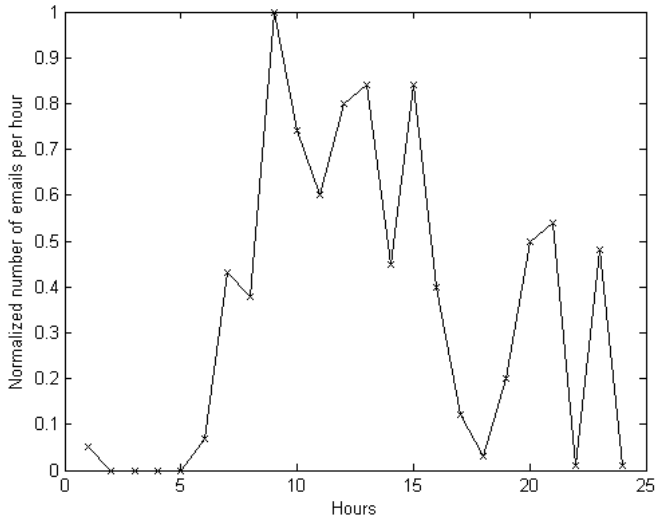

Figure 35. Outgoing emails from system, emails per hour

Maximum emails per hour: 37870

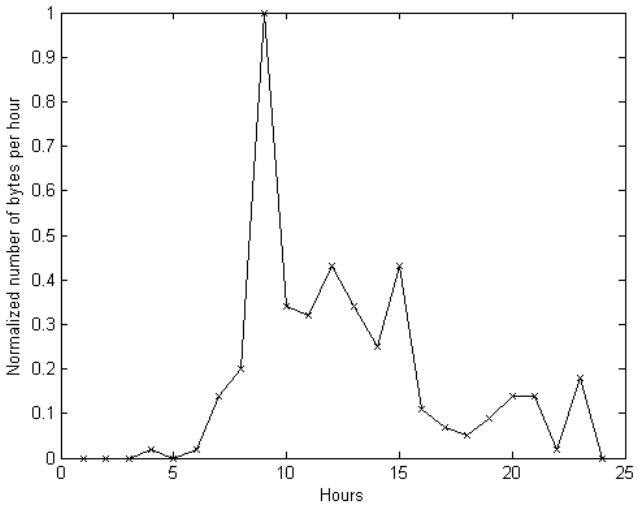

Figure 36. Outgoing emails from system, bytes per hour

Maximum bytes per hour: 0.4406 GB

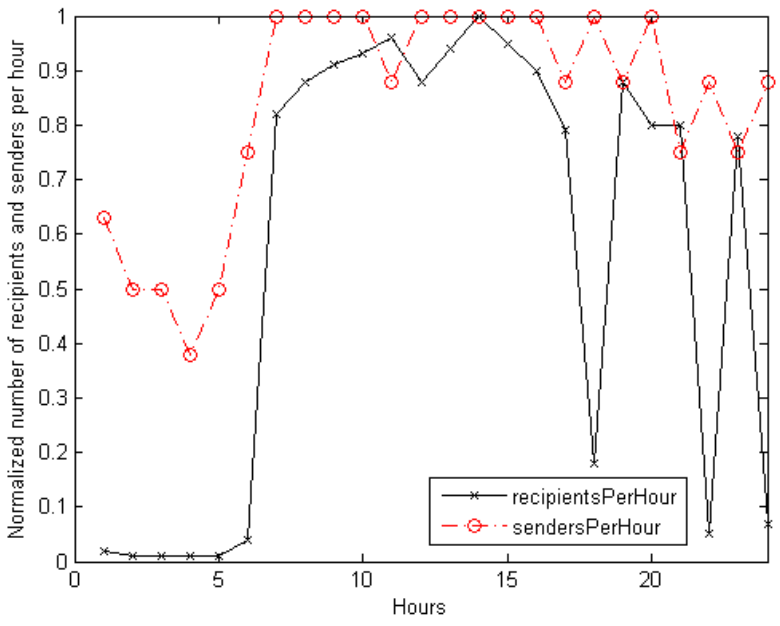

Figure 37. Outgoing emails from system, recipients and senders per hour

Maximum recipients per hour: 2552

Maximum senders per hour: 8

We noticed increased traffic from 9:00 until 13:00 but we observe that the day with the most traffic is Saturday. The address noreply@isc.tuc.gr is the one responsible for this increase by sending 10.000 emails on Saturday, which was the 86% of the traffic that day.

Some of our hourly statistics are shown in the tables 10 and 11.

	Metric	Minimum	Maximum	Average	CV
Monday	Emails/Hour	7	10475	2996	1,2284
	MB/Hour	0,1287	39.7779	13.1584	1.1573
Tuesday	Emails/Hour	2	12636	1573.9	1.8363
	MB/Hour	0.0127	76.36	14.19	1.4571
Wednesday	Emails/Hour	3	13544	2184	1,5653
	MB/Hour	0,2176	76.7906	10.81	1.5251
Thursday	Emails/Hour	4	10486	1591.5	1.6752
	MB/Hour	0.1852	322.6726	19.80	3.2800
Friday	Emails/Hour	5	14601	3270.6	1.4216
	MB/Hour	0.1463	53.8406	13.50	1.1923
Saturday	Emails/Hour	3	11407	524.87	4,4201
	MB/Hour	0.08	35.94	3.1236	2.643
Sunday	Emails/Hour	4	11462	1232.4	2.076
	MB/Hour	0.0086	65.7677	6.4896	2.304

Table 10. Emails and traffic volume per hour

	Metric	Minimum	Maximum	Average	CV
Monday	Recipients/Hour	4	1998	814.70	1.1109
	Senders/Hour	1	7	4.6667	0.3666
Tuesday	Recipients/Hour	1	2175	558.6250	1.4945
	Senders/Hour	1	8	4,50	0.3986
Wednesday	Recipients/Hour	2	2025	703.4167	1.2179
	Senders/Hour	2	7	4.50	0,4040
Thursday	Recipients/Hour	4	2038	584.3333	1.4423
	Senders/Hour	2	8	4.7083	0.3783
Friday	Recipients/Hour	4	2113	783.1667	1.1647
	Senders/Hour	2	8	4.9583	0.3337
Saturday	Recipients/Hour	2	2008	106.5	3.819
	Senders/Hour	1	5	3.125	0,3443
Sunday	Recipients/Hour	3	1992	405.79	1.7859
	Senders/Hour	1	7	3.91	0.3682

Table 11. Recipients and sender's numbers per hour

In figure 38 we present the CDF of the emails' size.

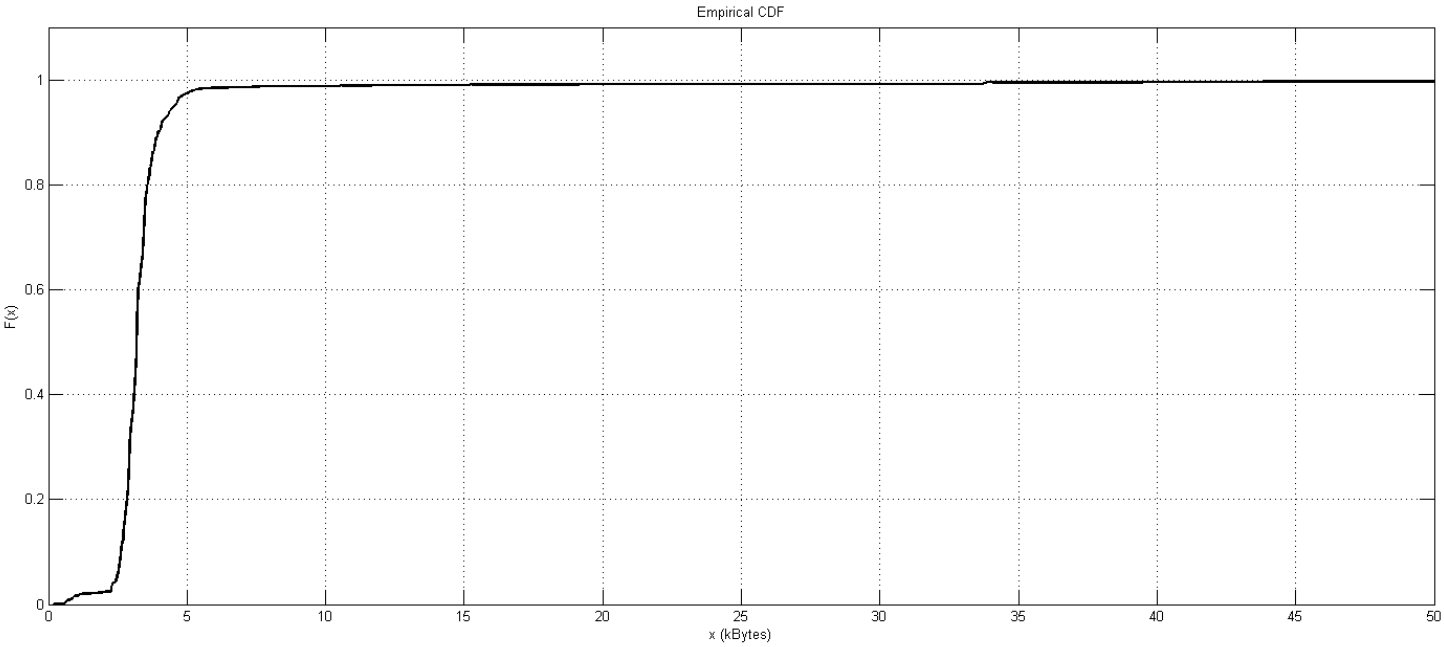

Figure 38. Outgoing emails from system, CDF of real data

We can observe that with probability almost 99% these senders will send an email around 6 kilobytes. This means that these servers rarely send attachments. Instead, they send short plain messages. The emails we send are more correlated than those we receive and larger.

5.4.1 Statistical Tests' Results for the Overall Traffic

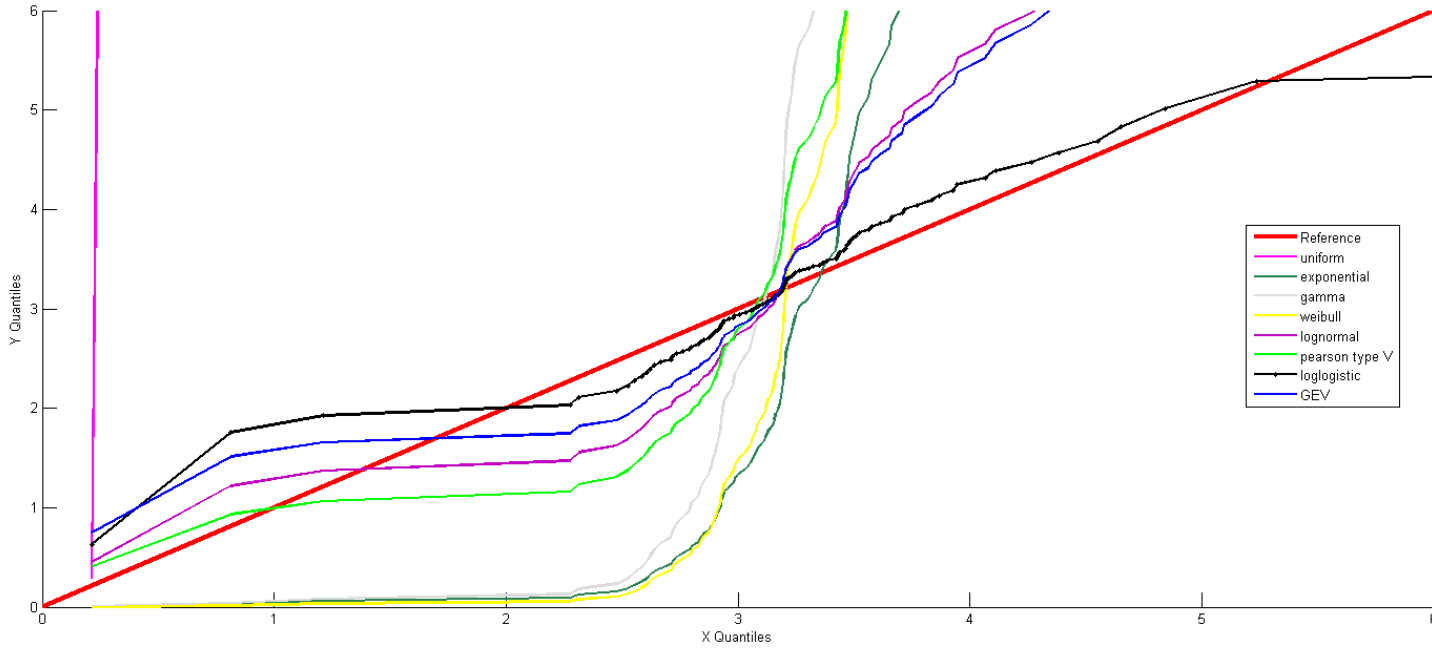

Figure 39. Outgoing emails from system, QQ plot

K-S Test

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
0.9931	0.4906	0.4170	0.4275	0.2232	0.1015	0.2072

A-D Test

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
2333500	93400	397300	9618100	37800	8200	31200

K-L Test

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
18.3359	9.0154	8.1902	6.8373	29.0327	29.0914	29.0197

RPE

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
21728	100	100	97	91	90	91

RPE at 98%

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
20539	10.633	10.785	8.7861	1.9582	0.51177	1.6864

Plotting the pdf in figure 40 we can observe that the K-S, A-D, QQ-plot and RPE at 99% of the quantiles tells us. That log logistic seems to be the best distribution to model our data. We have

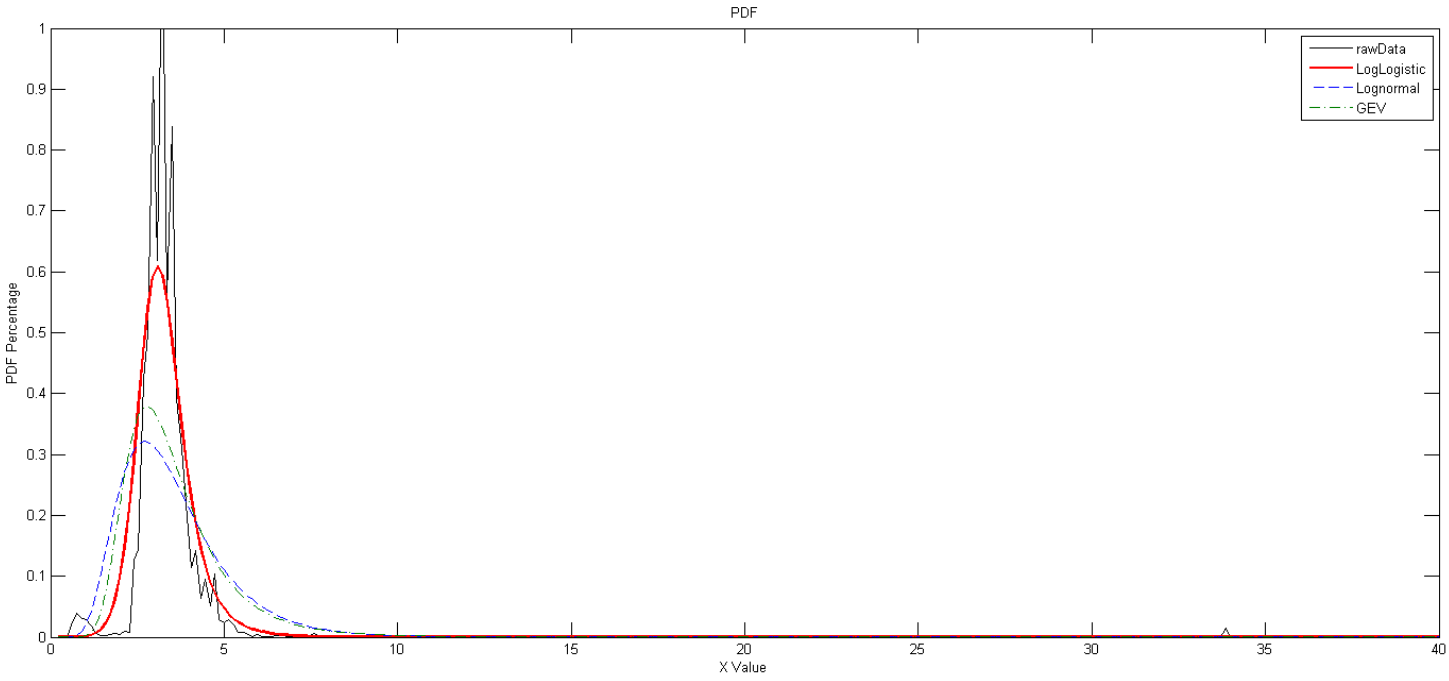

Figure 40. Outgoing emails from system, PDF or real data and best fit distributions

some outliers around 2 kilobytes which no distribution is able to predict. In addition, we have some outliers at 35 kilobytes and some larger than 40 kilobytes.

The parameters for the best three distribution are shown below.

Log logistic

Mu: 1.1606

Sigma: 0.13013

GEV

Mu: 2.8808

Sigma: 0.9745

K: 0.1151

Lognormal

Mu: 1.1723

Sigma: 0.4180

5.4.2 Statistical Test's Results for Daily Traffic

We can observe that the daily results follow the same distribution as the whole week, with rare exceptions mostly happening at the weekends. Table 12 summarizes the daily best distribution fit.

Monday

KS-Test

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
0.9919	0.5405	0.3660	0.4741	0.2337	0.1131	0.2177

AD-Test

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
463600	19800	19700	2194800	7700	1500	6900

KL-Test

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
17.6114	8.2971	18.9318	7.3584	28.3460	28.3717	28.3031

RPE

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
10445	96	94	96	88	87	88

RPE at 98%

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
9910.7	9.8795	8.0433	9.7435	1.7306	0.44235	1.6653

Tuesday

KS-Test

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
0.9901	0.5049	0.4600	0.4346	0.3100	0.1832	0.2807

AD-Test

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
276400	16000	188100	1151700	6200	1700	4700

KL-Test

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
11.8715	8.2184	5.5590	5.2994	30.6736	31.2196	30.3348

RPE

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
62639	119	119	100	76	70	73

RPE at 98%

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
59023	50.86	51.784	32.916	8.0055	1.748	5.7067

Wednesday

KS-Test

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
0.9905	0.4982	0.3891	0.4320	0.2231	0.1088	0.1861

AD-Test

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
331100	13000	21700	1567700	4300	700	3400

KL-Test

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
10.9813	7.3267	8.5214	6.0142	30.2405	30.5778	29.8886

RPE

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
14167	96	95	95	85	83	84

RPE at 98%

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
13412	13.943	12.784	13.118	2.5041	0.6112	2.1919

Thursday

KS-Test

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
0.9913	0.4963	0.4817	0.4485	0.2787	0.1410	0.2496

AD-Test

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
271700	22800	549600	1160300	4700	1100	3400

KL-Test

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
11.9234	8.3319	4.8992	5.0209	30.3686	31.4289	30.6574

RPE

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
86015	178	176	103	58	51	55

RPE at 98%

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
81015	127.64	125.84	56.146	11.126	2.6886	8.16

Friday

KS-Test

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
0.9842	0.4973	0.3327	0.4053	0.2248	0.1185	0.2087

AD-Test

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
462100	19400	16300	2373700	9000	2200	7300

KL-Test

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
10.3568	6.7582	15.8062	6.7630	29.1195	29.9555	29.1808

RPE

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
35339	67	56	67	38	30	35

RPE at 99%

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
33282	47.365	33.514	47.123	13.439	3.5629	11.44

Saturday

KS-Test

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
0.9910	0.5451	0.4406	0.4723	0.3603	0.2793	0.3422

AD-Test

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
90710	3910	14180	372830	2360	1260	2190

KL-Test

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
12.4103	8.7462	7.9476	6.6931	31.0214	31.6374	31.2124

RPE

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
153020	100	100	90	30	20	30

RPE at 98%

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
144000	93.623	90.776	76.868	21.515	7.0693	19.076

Sunday

KS-Test

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
0.9868	0.4600	0.4073	0.3888	0.2933	0.2068	0.2673

AD-Test

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
148680	7410	23170	791220	4140	2370	3700

KL-Test

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
9.7176	6.1392	5.7655	4.7270	25.7111	27.9896	27.2853

RPE

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
14716	95	94	87	46	36	44

RPE at 98%

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
13879	66.637	65.655	59.334	18.599	9.1243	17.001

In table 12, we can see the results for the daily best distributions. We can see that always, the best daily distribution is the same with the weekly one having close parameters. In addition, the mean size of an email is roughly the same every day.

	Metric	Mean	CV	mu	sigma
Monday	Email size in MB log logistic	0.0044	13.1895	1.1462	0.0958
Tuesday	Email size in MB log logistic	0.0090	17.3208	1.2217	0.1220
Wednesday	Email size in MB log logistic	0.0050	11.8603	1.1402	0.1341
Thursday	Email size in MB log logistic	0.0124	21.7321	1.1928	0.1367
Friday	Email size in MB log logistic	0.0041	6.0191	1.1574	0.1326
Saturday	Email size in MB log logistic	0.0060	22.8783	1.1398	0.1193
Sunday	Email size in MB log logistic	0.0053	9.293	1.0997	0.1822

Table 12. Daily best fit distribution fit

5.5. Spam Traffic

Total Number of Emails: 1577 **avg per day:** 225 CV: 0.147

Total bytes: 28,98MB **avg per day:** 4.1MB CV: 0.601

CV Emails per Hour: 0.379

CV Bytes per Hour: 0.968

We observe that our total spam emails are very few. The reason, as explained in Section4, is that the TUC spam filter is very effective.

Next, we have the traffic calculated as emails/bytes/senders/recipients per hour and per day.

The plots, as always, are normalized with the maximum value of each one.

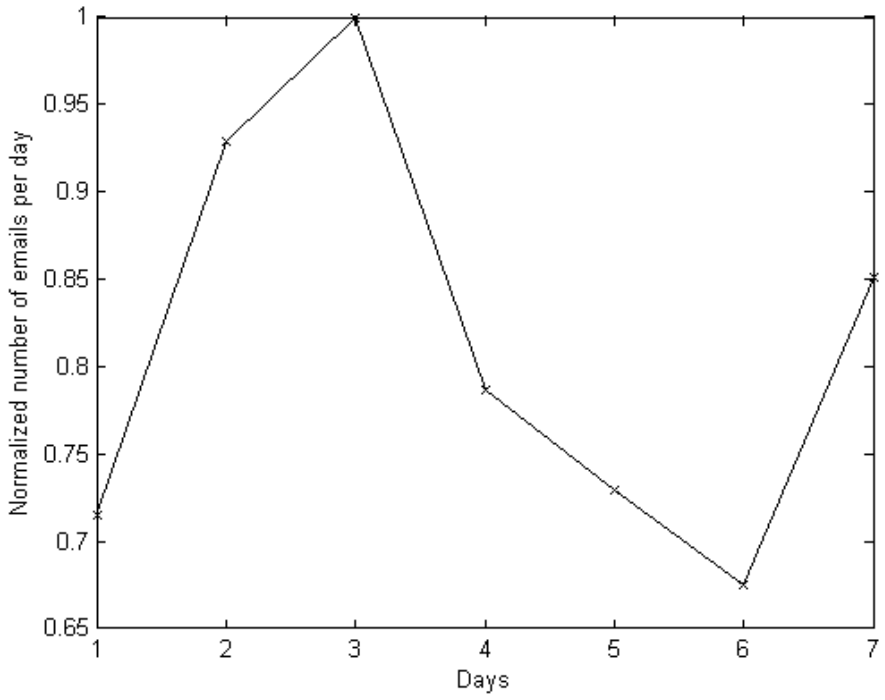

Figure 41. Spam emails, emails per day

Maximum emails per day: 295

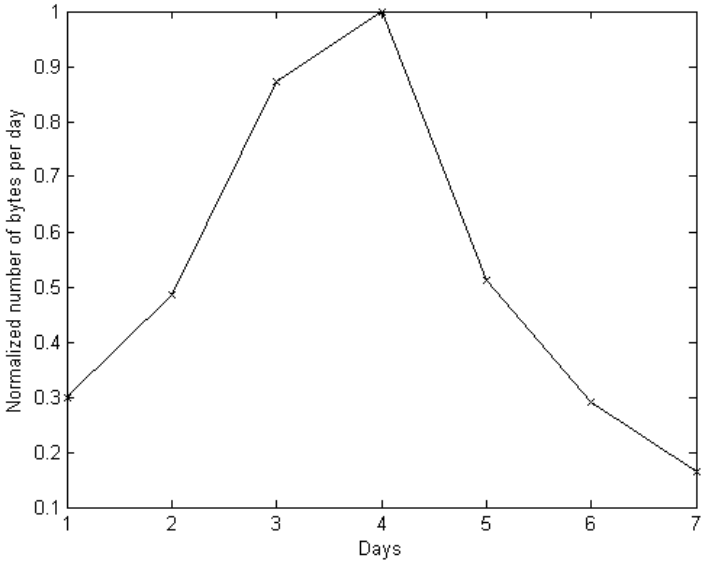

Figure 42. Spam emails, bytes per day

Maximum bytes per day: 8MB

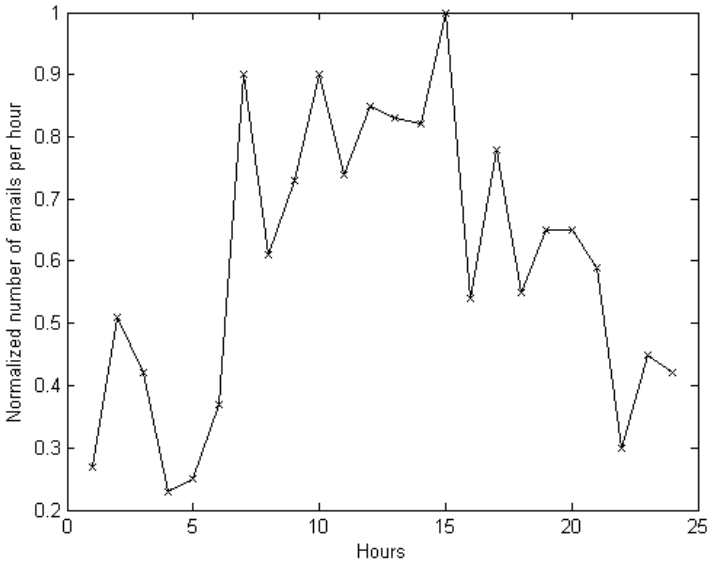

Figure 43. Spam emails, emails per hour

Maximum emails per hour: 110

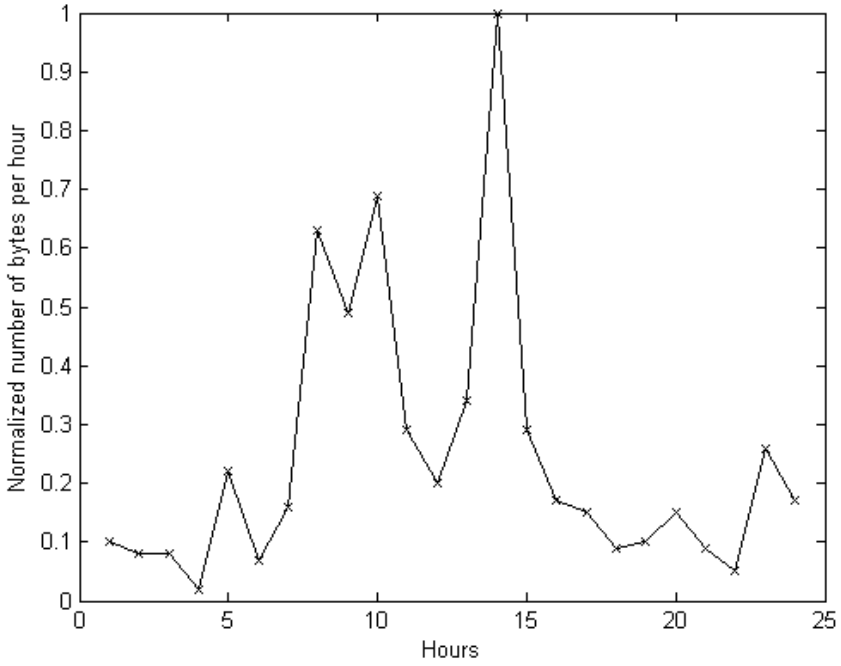

Figure 44.Spam emails, bytes per hour

Maximum bytes per hour: 4.94MB

We can see that we have minor variations during the days and the hours. This is in agreement with previous works done on spams. We show in figure 45 that indeed spam traffic is almost the same every day. We expect the same to happen if we had received the whole spam traffic and not only the portion not blocked by the antispam filter. In table 13, we have summarized the characteristics of the spam traffic.

.

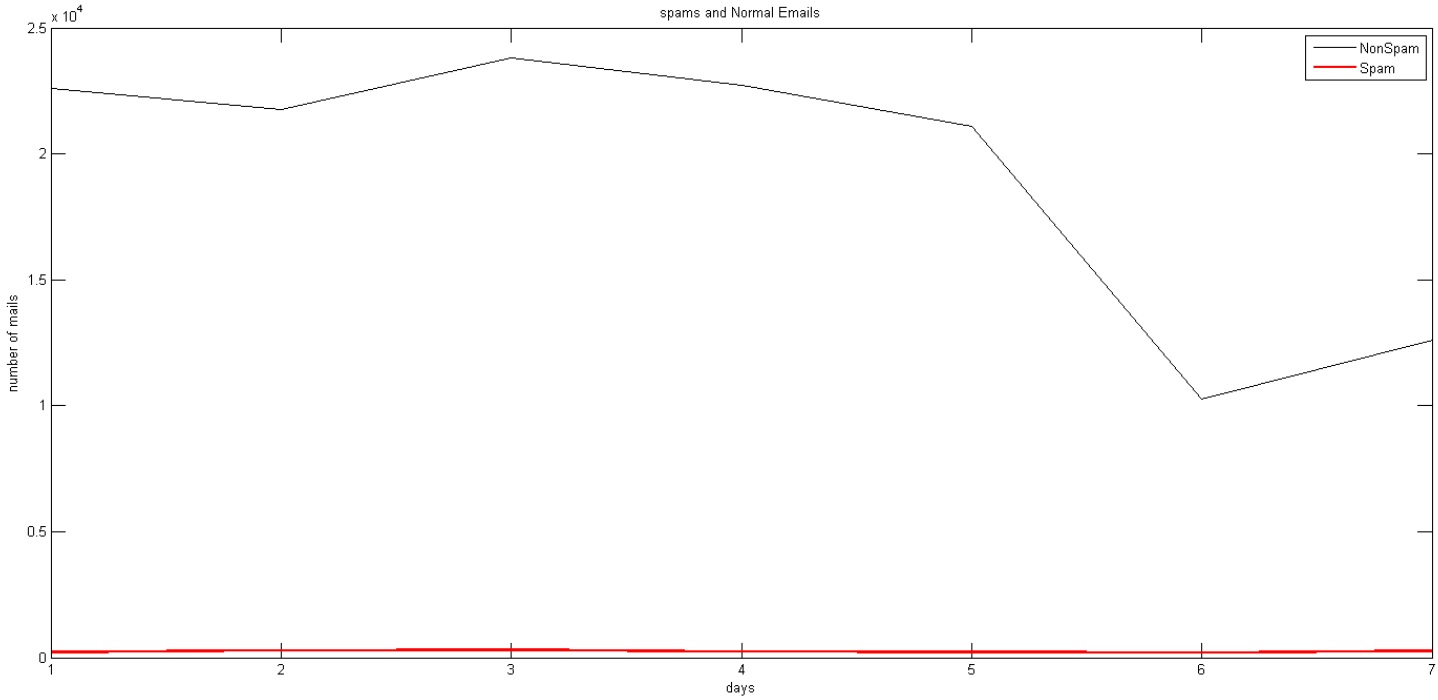

Figure 45.Spam traffic vs regular incoming traffic

	Metric	Minimum	Maximum	Average	CV
Monday	Emails/Hour	0	23	8.7	0.69
	MB/Hour	0	0.5	0.1	1.121
Tuesday	Emails/Hour	1	36	11.41	0.743
	MB/Hour	0.0014	0.656	0.1619	0992
Wednesday	Emails/Hour	2	25	12.29	0.5810
	MB/Hour	0.0064	2.466	0.2902	2.05
Thursday	Emails/Hour	2	25	9.667	0.5865
	MB/Hour	0.0086	1.98	0.3333	1.6568
Friday	Emails/Hour	2	25	8.9583	0.5610
	MB/Hour	0.01	1.46	0.17	1.75
Saturday	Emails/Hour	2	26	8.29	0.6882
	MB/Hour	0.0059	0.4776	0.0966	1.12
Sunday	Emails/Hour	1	15	6.29	0.5787
	MB/Hour	0.0088	0.3396	0.0548	1.26

Table 13. Spam emails and traffic volume per hour

The CDF of our data in figure 46.

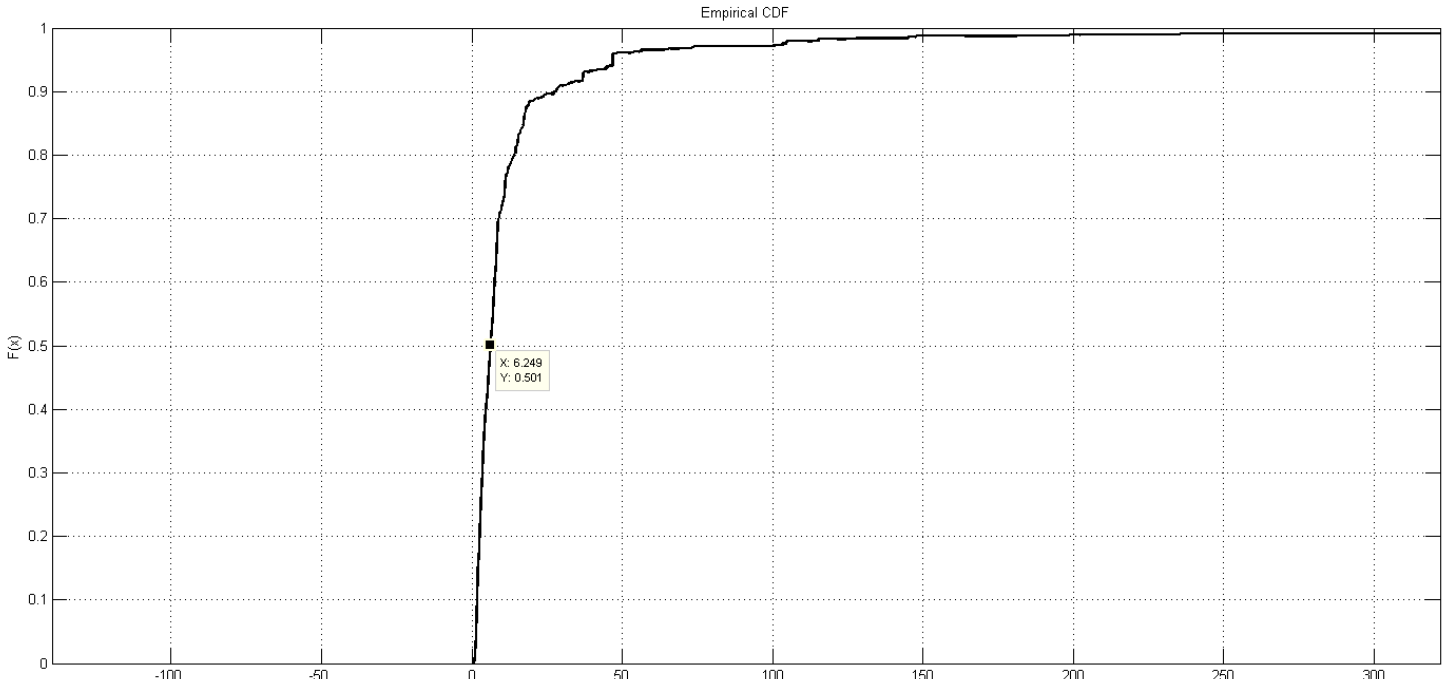

Figure 46. Spam emails, CDF of real data

5.5.1 Statistical Tests' Results for the Overall Traffic

K-S Test

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
0.9424	0.2234	0.1881	0.1876	0.0716	0.0525	0.0607

A-D Test

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
6948	277	10702	49111	11	10	5

K-L Test

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
13.8849	4.7542	3.1854	2.7502	3.1975	3.5674	10.1334

RPE

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
4451.3	81.9	70.9	63.4	58.6	48.2	42.3

RPE at 99.05%

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
4417	35.164	25.195	19.177	13.185	11.696	5.4596

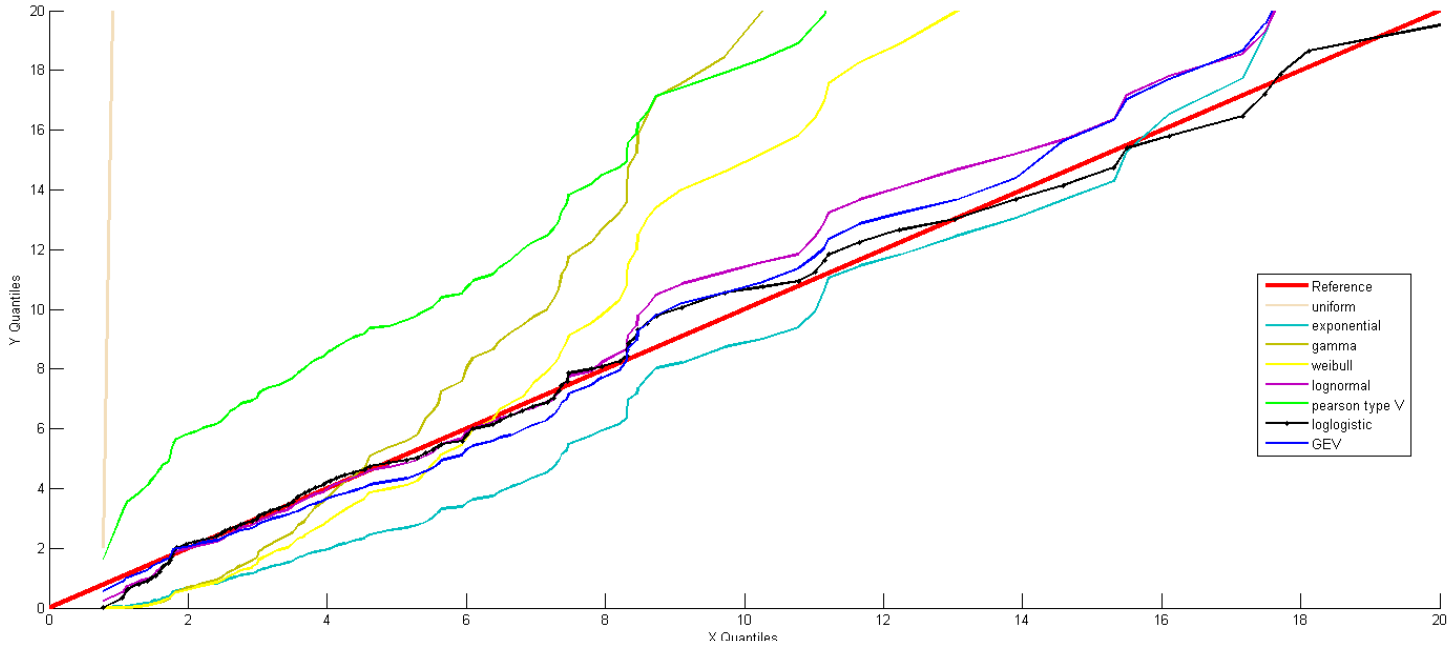

Figure 47. Spam emails, QQ plot

Once again log logistic, GEV and lognormal seem to provide the best fits.

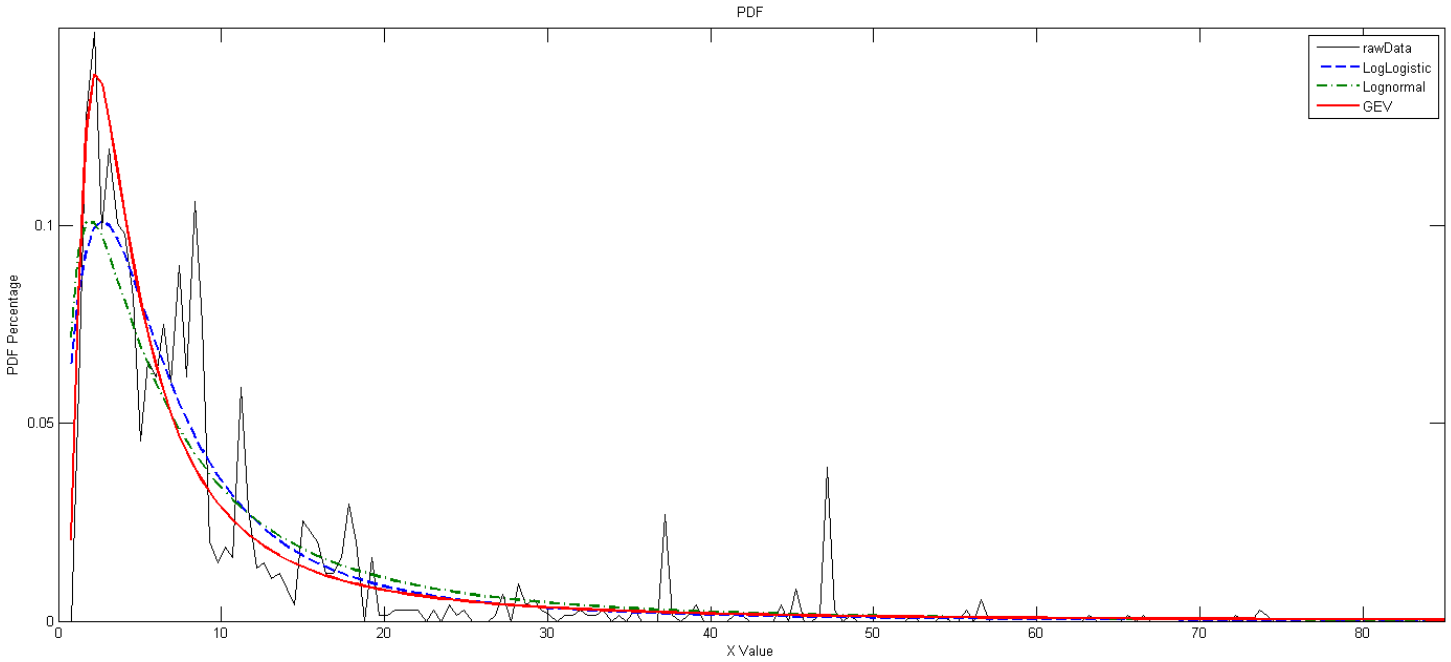

Figure 48. Spam emails, PDF of real data and various distributions

The parameters for the above PDFs are shown below.

Log logistic

Mu: 1.753

Sigma: 0.57063

GEV

Mu: 1.8267

Sigma: 3.3088

K: 0.7866

Lognormal

Mu: 1.8267

Sigma: 1.0655

GEV provides the best distribution fit according to PDFs and the test above. Of course, we have outliers but appear with small probability.

5.5.2 Statistical Test's Results for Daily Traffic

Monday

KS-Test

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
0.8542	0.2415	0.1830	0.1734	0.1354	0.1098	0.1201

AD-Test

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
475.6	10.6	139.7	6311.8	3	4.3	1.6

KL-Test

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
15.0467	6.9225	6.7498	6.6459	7.2679	15.8971	18.5980

RPE

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
848.9443	46.8111	40.4552	39.9683	35.1480	29.3861	39.9555

RPE at 99.05%

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
850.13	34.379	26.267	28.36	22.881	18.068	17.865

Tuesday

KS-Test

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
0.8339	0.2694	0.1915	0.1917	0.1443	0.1333	0.1380

AD-Test

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
520.5	22.1	23.9	8508.5	5.6	5.7	3.5

KL-Test

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
14.6956	7.0947	7.0597	7.1042	7.4954	16.9420	20.3333

RPE

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
621.2970	52.7618	43.2136	45.2687	35.8751	32.2913	28.6879

RPE at 99.05%

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
386.53	15.556	10.121	12.488	4.4444	3.7145	4.2309

Wednesday

KS-Test

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
0.9500	0.2574	0.2708	0.2724	0.0979	0.0892	0.1047

AD-Test

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
1336.2	105.1	2902	9163.5	3.3	4.6	1.9

KL-Test

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
15.4747	6.5827	4.5346	4.0906	5.8757	4.1426	9.2512

RPE

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
3951.6	107.8	85.8	76.3	71.2	66.0	66.9

RPE at 99.05%

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
3918.7	47.153	25.061	16.971	7.7726	6.8965	6.3184

Thursday

KS-Test

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
0.9107	0.2134	0.2285	0.2132	0.1352	0.1206	0.0993

AD-Test

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
768.8	61.2	2328.2	7247	4.7	6.8	1.6

KL-Test

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
15.3226	6.8019	5.5934	5.3843	5.9176	14.3282	21.5397

RPE

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
2048.9	80.2	48.7	43.4	48.3	40.2	190.2

RPE at 99.05%

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
2005.2	59.179	28.137	26.508	27.562	24.068	36.209

Friday

KS-Test

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
0.9407	0.2840	0.2877	0.2915	0.1003	0.0897	0.0929

AD-Test

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
896.4	60.7	1311.7	6716.0	2.8	4.3	1.4

KL-Test

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
15.2418	6.3208	4.6479	4.2449	6.6210	4.7582	13.2796

RPE

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
2737.1	100.6	79.1	74.3	69.4	65.7	81.7

RPE at 99.05%

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
2691.9	43.292	22.823	19.302	9.1861	8.7419	8.981

Saturday

KS-Test

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
0.7178	0.2474	0.2004	0.1807	0.1195	0.1075	0.0994

AD-Test

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
313.3	9.4	316	6106	4.8	6.4	1.6

KL-Test

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
14.9395	8.5701	8.7696	9.1780	11.1910	20.8147	24.6647

RPE

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
438.4346	30.3388	33.3126	21.7382	30.4721	35.4291	249.2557

RPE at 99.05%

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
436.46	25.23	26.853	17.477	25.52	21.212	50.922

Sunday

KS-Test

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
0.9077	0.2514	0.2339	0.2251	0.1277	0.0959	0.1098

AD-Test

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
451.1	8.3	72.0	3918.5	2.9	4.4	1.7

KL-Test

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
15.2510	6.5310	5.9386	5.6570	7.5922	11.0223	16.0555

RPE

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
1743.5	54.1	43.1	44.0	38.8	31.7	48.9

RPE at 99.05%

Uniform	exponential	gamma	weibull	lognormal	log logistic	GEV
1733.1	31.095	18.46	22.351	13.8	11.653	17.082

In table 14, we can see the results for the daily best distributions. We can see that usually, the best daily distribution is the same with the weekly one having close parameters.

	Metric	Mean	CV	mu	sigma
Monday	Email size in MB GEV	0.0114	1.9930	4.2433 K=0.6486	3.5381
Tuesday	Email size in MB Log logistic	0.0142	1.9439	1.8887	0.4445
Wednesday	Email size in MB GEV	0.0236	6.6335	3.8327 K=0.7988	3.1672
Thursday	Email size in MB GEV	0.0345	3.3378	4.55 K:1.2991	5.01
Friday	Email size in MB Log logistic	0.0191	5.3370	1.5484	0.4781
Saturday	Email size in MB GEV	0.0116	1.5150	3.43 (GEV K:1.01)	3.15
Sunday	Log logistic	0.0087	3.1036	1.3338	0.4711

Table 14. Daily best distribution fit

6. Conclusions

In this work, we tried to model the workload of the email servers of the Technical University of Crete. We evaluated various well-known distributions from relevant literature on workload characterization, in terms of their fitting accuracy to our data. We have shown that, with the exception of few outliers, we can predict the workload with very high accuracy. For 98% of the traffic (i.e. excluding the outliers) we never achieve lower than 92% accuracy. In contrast with previous work in the field, we found that the lognormal distribution does not provide the best fit for any of the categories that we divided our traffic. Instead, the best fit is provided by the log logistic distribution, followed by the Generalized Extreme Value distribution. We believe that

these results offer a solid basis for future work on email traffic modeling which will acquire data from a much larger pool of servers (i.e., not just from the Technical University of Crete) and for a larger measurement period.

References

[1] T. Jackson, R. Dawson, and D. Wilson, "The Cost of Email Interruption," *Journal of Systems & Information Technology, vol 5,* pp. 81-92, 2001.

[2] Get Response, "Email Marketing Trends Survey," 2010. [Online]. Available: https://www.getresponse.com/documents/core/reports/2010_Email_Marketing_Trends_Survey.pdf.

[3] Y.R. Bujang and H. Hussin, "Should We Be Concerned with Spam Emails? A Look at Its Impacts and Implications," in *Proceedings on Information and Communication Technology for the Muslim World*, 2013.

[4] T. Takemura, and H. Ebara, "Spam Mail Reduces Economic Effects," in *IEEE , Second International Conference on the Digital Society*, 2008.

[5] A. Kashyap, A. Horbury, A. Catacutan et al., "Internet Security Threat Report 2014 :: Volume 19," 2014. [Online]. Available: http://www.symantec.com/content/en/us/enterprise/other_resources/b-istr_main_report_v19_21291018.en-us.pdf.

[6] L.H. Gomez, C. Cazita , J.M. Almeida et al, "Workload models of spam and legitimate e-mails," *Performance Evaluation ,* vol. 64, no. 7-8, pp. 690-714, 2007.

[7] L.Bertolotti and M.C. Calzarossa, "Workload Characterization of Email Servers," in *Proceedings of SPECTS*, 2000.

[8] S. Shah and B.D. Noble , "A study of e-mail patterns," *Software - Practice and Experience,* vol. 37, no. 14, pp. 1515 - 1538, 2007.

[9] V. Paxson, "Empirically-Derived Analytic Models of wide-area TCP connections," *IEEE/ACM Transactions on Networking,* vol. 2, no. 4, pp. 316-336, 1994.

[10] "Anderson - Darling test," [Online]. Available: http://www.mathworks.com/help/stats/adtest.html.

[11] D.P. Heyman, A. Tabatabai, T.V. Lakshman, ", Statistical analysis and simulation study of video teleconference traffic in ATM networks," *IEEE,* vol. 2 (1), pp. 49-59, 1992.

[12] A.M. Law, W.D. Kelton, Simulation Modeling & Analysis, 2nd ed., McGraw-Hill, 1991.

[13] A. Lazaris, P. Koutsakis and M. Paterakis, "A new model for video traffic originating from multiplexed MPEG-4 videoconference streams," *Performance Evaluation,* vol. 65, pp. 51-70, 2008.

[14] F. j. Massey, "The Kolmogorov-Smirnov Test for Goodness of Fit," *Journal of the American Statistical Association, Vol. 46, No. 253,* pp. 68-78, 1951.

[15] "Engineering Statistics Handbook," [Online]. Available: http://www.itl.nist.gov/div898/handbook/eda/section3/eda35e.htm.

[16] S.Kullback and R.A. Leibler, "On Information and Sufficiency," *The Annals of Mathematical Statistics,* vol. 22, no. 1, pp. 79-86, 1951.

[17] L. I. Lanfranchi and B.K. Bing, "MPEG-4 Bandwidth Prediction for Broadband Cable Networks," *IEEE TRANSACTIONS ON BROADCASTING,* pp. 741-751, 2008.

[18] I. J. Myung, "Tutorial on maximum likelihood estimation," *Journal of Mathematical Psychology 47,* pp. 90-100, 2003.