Bayesian analysis of a queueing system with a long-tailed arrival process

Pepa Ramirez, Rosa E. Lillo, Michael Peter Wiper

Bayesian analysis of a queueing system with a long-tailed arrival process

Journal:	*Communications in Statistics - Simulation and Computation*
Manuscript ID:	LSSP-2007-0142.R1
Manuscript Type:	Original Paper
Date Submitted by the Author:	24-Sep-2007
Complete List of Authors:	Ramirez, Pepa; Universidad Carlos III de Madrid, Estadistica Lillo, Rosa E.; Universidad Carlos III Madrid Wiper, Michael; Carlos III University Madrid, Statistics
Keywords:	Heavy tails, Bayesian inference, Queueing systems
Abstract:	Internet traffic data is characterized by some unusual statistical properties, in particular, the presence of heavy-tailed variables. In this article we deal with a mixture of two-parameter Pareto distributions as an example of heavy tailed distribution and use a Bayesian approach to fit the model. After that, we estimate some measures of interest related to the queueing system k-Par/M/1 where k-Par denotes a mixture of k Pareto distributions. Our procedure is based on recent Laplace Transform approximating results for the Pareto/M/1 system.
Note: The following files were submitted by the author for peer review, but cannot be converted to PDF. You must view these files (e.g. movies) online.	
graficos_communications.tex mixtura_pareto_afterreview3.tex	

BAYESIAN ANALYSIS OF A QUEUEING SYSTEM WITH A LONG-TAILED ARRIVAL PROCESS

Pepa Ramírez, Rosa E. Lillo and Michael P. Wiper Universidad Carlos III de Madrid Departamento de Estadística. C/Madrid, 126. 28903 Getafe, Madrid 00 34 624 9826
jrcobo@est-econ.uc3m.es

Key Words: Heavy-tailed variables, Bayesian inference, queueing systems.

ABSTRACT

Internet traffic data is characterized by some unusual statistical properties, in particular, the presence of heavy-tailed variables. A typical model for heavy tailed distributions is the Pareto distribution although this is not adequate in many cases. In this article we consider a mixture of two-parameter Pareto distributions as a model for heavy tailed data and use a Bayesian approach based on the birth-death Markov chain Monte Carlo algorithm to fit this model. We estimate some measures of interest related to the queueing system k-Par$/M/1$ where k-Par denotes a mixture of k Pareto distributions. Heavy tailed variables are difficult to model in such queueing systems because of the lack of a simple expression for the Laplace Transform (LT). We use a procedure based on recent LT approximating results for the Pareto$/M/1$ system. We illustrate our approach with both simulated and real data.

1. INTRODUCTION

Recently, the Internet has become so important in our society that it has been necessary to characterize it from a statistical point of view and to develop models able to explain and predict its behavior. It is of common interest to improve the performance of Internet and with this end, since the nineties researchers have been studying the statistical properties of internet traffic data. In particular, it has been observed that internet traffic such as packet arrivals or file sizes does not behave like a Poisson process (Paxson and Floyd 1995) and that internet traffic variables possess some unusual characteristics such as self-similarity (Willinger et al 1997, Park and Willinger 2000), long-range dependence (Beran et al 1995), burstiness and, in particular, heavy-tails (Crovella et al 1998). Many of these features are related and Paxson and Floyd (1995) suggested that self-similarity can be captured by modelling using a Pareto distribution with infinite variance.

As the number of internet users grows, it becomes all the more necessary to study models able to predict and analyze the congestion of such traffic. Queueing theory has been studying such traffic congestion problems since 1900. However, due to the features of internet data described above, classical queueing models cannot be applied in this context and new, more complex queueing models are demanded, see e.g. Fischer and Harris (1999), Greiner et al (1999), Harris et al (2000) and Fisher et al (2001). In particular, we note that if the interarrival distribution is heavy tailed, then it will often not possess a LT in closed form and therefore, the usual queueing theory techniques to calculate equilibrium distributions etc. cannot be applied.

Several approaches trying to overcome the lack of a closed expression for the LT of heavy tailed distributions can be found in the literature. One technique consists in approximating the interarrival distribution using a more tractable (but light tailed) model. For example, Feldmann and Whitt (1998) considered approximating heavy tailed distributions via mixtures of exponentials and Riska et al (2004) used a phase-type distribution.

A second approach is to try to approximate the LT. Thus, Abate and Whitt (1999) used continued fractions as an alternative to computation of LTs, via infinite-series representations. Fisher and Harris (1999) developed the *Transform Approximation method* or *TAM*, where the integral in the LT is substituted by a finite sum. Harris et al (2000) introduced an alternative way for directly finding the predictive equilibrium distributions, the *Level Crossing method*. Here we shall apply these two last methods and provide a comparison between them.

Recently, Rodríguez-Dagnino (2004, 2005) found an expression for the LT of Pareto probability distribution, in terms of the Whittaker function. In Rodríguez-Dagnino (2004) this approach is compared with the *TAM* method, obtaining very similar results.

The objective of this article is firstly to model internet arrival traffic using a mixture of Pareto distributions (denoted $k-Par$, where k is the number of components in the mixture) as a more flexible alternative to the simple Pareto distribution which should be able to better fit typical internet data samples. Secondly, we will introduce a Bayesian algorithm based on the birth death MCMC approach of Stephens (2000) in order to fit this model to a given data set. Then we shall study the k-Par$/M/1$ queueing system, by combining the *Transform Approximation method* and *Level Crossing method* techniques with the MCMC output to enable us to obtain numerical predictions of the equilibrium distributions of the system.

The paper is organized as follows. In Section 2, we introduce the Pareto mixture distribution model and then we describe how to carry out Bayesian inference for this model, illustrating the procedure with a simulated data set. In Section 3, we examine how to estimate the equilibrium queue size and waiting time distributions for the k-Par$/M/1$ system conditional on the system parameters and, predictively given the MCMC output comparing the use of *TAM* and *Level Crossing* algorithms with the simulated interarrival data of Section 3. In Section 4, we illustrate our approach with some real data sets and finally, conclusions and possible extensions to this work are considered in Section 5.

2. BAYESIAN INFERENCE FOR THE PARETO MIXTURE DISTRIBUTION.

A variable X is said to have a Pareto distribution with shape parameter $\beta > 0$ and scale parameter $b > 0$ if its density function is:

$$f_P(x|\beta, b) = \frac{\beta\, b^{\beta}}{(x+b)^{\beta+1}}, \qquad \text{for } x > 0.$$

This distribution is power tailed and only possesses moments of order less than β. In particular, the Pareto distribution has finite mean

$$E(X|\beta, b) = \frac{b}{\beta - 1}$$

if and only if $\beta > 1$.

A more general model which shall be analyzed in this article is a mixture of k Pareto densities:

$$f(x|\mathbf{w}, \boldsymbol{\beta}, \mathbf{b}) = \sum_{r=1}^{k} w_r f_P(x|\beta_r, b_r), \tag{1}$$

where $w_r > 0$ for $r = 1, \ldots, k$ and $\sum_{r=1}^{k} w_r = 1$, which can also be expressed in terms of an indicator variable:

$$f(x|\mathbf{w}, \boldsymbol{\beta}, \mathbf{b}) = \sum_{r=1}^{k} P(Z = r|\mathbf{w}) f_P(x|\beta_r, b_r), \tag{2}$$

where $P(Z = r|\mathbf{w}) = w_r, \quad \text{for } r = 1, \ldots, k.$

In the following subsection, we shall define a Bayesian procedure to fit this model to a sample of interarrival data.

2.1. BAYESIAN INFERENCE WHEN THE NUMBER OF COMPONENTS IS KNOWN

Suppose now that we have a sample $x_1, \ldots, x_n$ from the distribution in (1). Initially, we shall assume that k is known and we fix the following prior parameter distributions:

$$\begin{aligned} \mathbf{w} &\sim \text{Dirichlet}(a, a, \ldots, a), \quad \text{for constant } a \geq 0, && (3)\\ \beta_i &\propto \text{Gamma}(c, d), \quad \text{for } i = 1, \ldots, k, \quad \text{where } c > 0, d > 0, && (4)\\ b_i &\propto \text{Pareto}(\gamma, \delta), \quad \text{for } i = 1, \ldots, k, \quad \text{where } \gamma > 0, \delta > 0. && (5) \end{aligned}$$

Note that the prior distributions for (β_i, b_i) are suggested in Arnold and Press (1989) as a natural prior structure for the Pareto model. Given these prior distributions and the sample data, we are now able to calculate the conditional posterior distributions of the model parameters ($\mathbf{Z}$,$\mathbf{w}$,β,$\mathbf{b}$) where $\mathbf{Z} = (Z_1 \ldots, Z_n)$ are the indicator variables associated with $x_1, \ldots, x_n$, as defined in (2). Thus, we have

$$\begin{aligned} P(Z_i = r | x_i, \boldsymbol{\beta}, \mathbf{b}, \mathbf{w}) &= \frac{w_r f_P(x_i|\beta_r, b_r)}{\sum_{j=1}^{k} w_j f_P(x_i|\beta_j, b_j)}, \quad \text{for } i = 1, \ldots, n,\\ \mathbf{w}|\mathbf{Z} &\sim \text{Dirichlet}(a + n_1, \ldots, a + n_k), \quad \text{where } n_r = \#\{Z_i = r\},\\ \beta_r|\mathbf{x}, \mathbf{Z}, b_r &\sim \text{Gamma}\left(c + n_r, d + \sum_{i=1:Z_i=r}^{n} \log(x_i + b_r) - \log(b_r^{n_r})\right)\\ f(b_r|\mathbf{x}, \mathbf{Z}, \beta_r) &\propto b_r^{n_r}(b_r + \delta)^{-1-\gamma} \prod_{i=1:Z_i=r}^{n} (x_i + b_r)^{-1-\beta_r}. \end{aligned}$$

for $r = 1, \ldots, k$.

Thus, it is straightforward to define a Gibbs sampler algorithm to sample the joint posterior distribution, where the only complicated step is sampling the distribution of b_r. In this case, we simply use a Metropolis-Hastings step with a gamma proposal distribution.

In the following subsection, we extend the algorithm to the case where the mixture size, k, is unknown.

2.2. EXTENSION TO RANDOM k

Assume now that the mixture size, k, is unknown. Then firstly we need to define a prior distribution $p(k)$ for k, for example a truncated Poisson distribution,

$$p(k) \propto \frac{\lambda^k}{k!} \quad k = 1, \ldots, k_{max}.$$

Now, the parameter distributions introduced above can be treated as densities conditional on k. In order to sample the posterior distribution of k, there are two main approaches in the context of mixture modeling; the reversible jump sampler (Green 1995, Richardson and Green 1997) and the birth death MCMC (BDMCMC) sampler, Stephens (2000). Here we apply the BDMCMC which is somehow simpler to program and we have found to give better results in practice. We briefly outline this algorithm below.

This BDMCMC algorithm is based on a continuous time birth death process, where components are born at a fixed rate, ϱ, or die at a varying rate, in exponentially distributed time intervals with rate the sum of the birth and death rates. The birth death process is simulated during a fixed time t_0, for example, $t_0 = 1$ and the current state of the process at this time then provides the new mixture size and model parameter values.

In more detail, suppose that there are currently k components in the mixture with associated parameters $\boldsymbol{\theta} = (\mathbf{w}, \mathbf{b}, \boldsymbol{\beta})$. In order to generate a birth, parameters β_{k+1} and b_{k+1} are generated from the prior distributions. A new weight w_{k+1} is further generated from a beta distribution with parameters 1 and k and the weights of the remaining terms are normalized (by multiplying by $(1 - w_{k+1})$) so that they sum to 1.

If a death of a component occurs, then this component is removed from the mixture and the weights of the remaining components are normalised. Given the current mixture terms $k, \mathbf{w}, \boldsymbol{\beta}, \mathbf{b}$, the overall death rate is $\nu = \sum_{r=1}^{k} \nu_r$ where ν_r, the death rate for component r is given by

$$\nu_r = \varrho \frac{p(k-1)}{kp(k)} \frac{L(data|k-1, \boldsymbol{\theta}_{-r})}{L(data|k, \boldsymbol{\theta})}, \quad \text{for } r = 1, \ldots, k, \tag{6}$$

where $L(data|k, \boldsymbol{\theta})$ represents the likelihood function before the removal of component r and $L(data|k-1, \boldsymbol{\theta}_{-r})$ is the likelihood when component r is removed.

Following arguments given in Stephens (2000) it is possible to demonstrate that this algorithm provides samples from the posterior parameter distribution of k.

Below, we illustrate that the Bayesian approach can be used to well approximate Pareto mixture distributions.

2.3. ILLUSTRATION WITH A SIMULATED DATA SET

In order to test the Bayesian algorithm, we simulated 1000 data from the density

$$f(x|\mathbf{w},\boldsymbol{\beta},\mathbf{b}) = 0.55 f_P(x|\beta_1 = 0.5, b_1 = 1) + 0.45 f_P(x|\beta_2 = 10, b_2 = 4). \quad (7)$$

We assumed a truncated Poisson prior distribution with mean 2 for k and relatively uninformative prior distributions for the remaining parameters setting $a = 1$, $(c, d) = (0.05, 0.05)$, $\gamma > 1$, (for example $\gamma = 10.01$), and $\delta = 1$ in (3), (4) and (5) respectively, and ran the BDMCMC algorithm for 20000 iterations with initial values set to $k^{(0)} = 1$ and $(\beta^{(0)}, b^{(0)}) = (\hat{\beta}, \hat{b})$, the maximum likelihood estimates for these values.

Figure 1 illustrates the mixing properties of the algorithm in terms of the evolution of the mixture size k.

FIGURE 1 ABOUT HERE
Graph of mixture components k versus iterations.

We can see that the chain mixes quite well and in fact the acceptance rate of the BDMCMC algorithm for k was around 15%.

Figure 2 shows the posterior distribution of k and it can be seen that the algorithm correctly predicts that the data come from a two component mixture.

FIGURE 2 ABOUT HERE
Posterior distribution of k.

Note also that, conditional on $k = 2$, the mean posterior parameter estimates (after imposing the order condition $w_1 > w_2$ to make the model identifiable) were close to the true data generating values. Thus, for example, we found $E[\mathbf{w}|k = 2, \text{data}] \approx (0.57, 0.43)$.

Figure 3 illustrates the empirical, predictive (dotted lines) and theoretical (pointed dotted lines) cumulative distribution function estimated for the first 900 data (top left), for the 900th-980th data (top right) and for the 980th-990th data (bottom) respectively.

FIGURES 3.1, 3.2 AND 3.3 ABOUT HERE

Empirical (solid line), predictive (pointed dotted line) and theoretical (dotted line) cumulative distribution functions for the first 900 data (top left), the 900th-980th data (top right) and the 980th-990th data (bottom).

It is easy to see that the fitted results are close to the empirical and theoretical cumulative distribution functions.

Finally, in order to check prior sensitivity, we ran the algorithm with various different prior distributions. In particular, we found that there was little sensitivity (in the posterior parameter estimation) to changes in the prior distributions for $\mathbf{w}$, $\mathbf{b}$ and $\boldsymbol{\beta}$ although, as would be expected, there was slightly more sensitivity to changes in the prior distribution for k. Increasing the prior variance leads to more variation in the posterior parameter estimates. However, the predictive distribution functions in each case were indistinguishable.

Similar results were also obtained with other simulated data sets, illustrating that the BDMCMC algorithm well approximates Pareto mixture data.

3. THE k-Par/M/1 QUEUEING SYSTEM

In this section, we shall consider the previously outlined Pareto mixture distribution as a model for the arrival process in a single-server queueing system with independent, exponential service times. This queueing system, which we denote as k-Par/M/1 is an example of a $G/M/1$ queueing system. The main properties of such systems are outlined below.

Firstly, if A and S denote the interarrival and service time distributions and $E(A) = 1/\lambda$ and $E(S) = 1/\mu$ are the expected interarrival and service times, then the traffic intensity is defined by $\rho = \lambda/\mu$. When $\rho < 1$ the system is stable and then the steady-state probability for the number of customers Q in system just before an arrival is, for all $n \in \mathbb{N}$:

$$P(Q = n) = (1 - r_0)r_0^n,$$

where $r_0 \in (0, 1)$ is the unique real root of the equation

$$r_0 = f^* \left(\mu(1 - r_0)\right), \tag{8}$$

and $f^*(\cdot)$ is the Laplace-Stieltjes transform of the interarrival-time density function $f(\cdot)$ defined as

$$f^*(s) = \int_0^\infty e^{-sx} f(x)dx, \qquad \text{for } Re(s) > 0.$$

Moreover, let W_q represent the stationary time spent queueing for service and W the time spent in the system. Then

$$F_{W_q}(x) = P(W_q \leq x) = 1 - r_0 e^{-\mu(1-r_0)x},$$

and

$$F_W(x) = P(W \leq x) = 1 - e^{-\mu(1-r_0)x}.$$

For the k-Par/M/1 system with parameters $\boldsymbol{\theta} = (\mathbf{w}, \boldsymbol{\beta}, \mathbf{b})$ as given in (1), then the mean interarrival time does not exist if any element of $\boldsymbol{\beta}$ is less than or equal to one. In this case, the queueing system is automatically stable whatever the service rate μ. Otherwise, the traffic intensity is given by

$$\rho = \frac{1}{\mu \sum_{r=1}^{k} w_r \frac{b_r}{\beta_r - 1}}. \tag{9}$$

However, the evaluation of the LT of the Pareto distribution is difficult, requiring numerical techniques. Thus, the standard techniques for finding the root of (8) cannot be easily applied and an alternative approach to obtaining the steady state distributions is needed. One method is suggested in Rodríguez-Dagnino (2004, 2005) where the LT is expressed in terms of the Whittaker function for which efficient estimation routines are available. Here we consider two alternative algorithms: the *Transform Approximation* method and the *Level Crossing* method.

3.1. THE TRANSFORM APPROXIMATION METHOD

The *Transform Approximation Method* was proposed in Fischer and Harris (1999) for the case of the single parameter Pareto distribution and is based on approximating the LT of the interarrival time distribution. Here we describe the approach of Fisher and Harris in the more general case of the mixture of two-parameter Pareto distributions.

The basic idea of the *TAM* approach is to select values, say $x_1, \ldots, x_M$ from the support of the arrival distribution when we can define an $M-$point approximation of the interarrival distribution LT as

$$f_M^*(s) = \frac{1}{M} \sum_{i=1}^{M} e^{-sx_i}$$

where the values x_i (called the *TAM* samples), are chosen to cover the support of the original interarrival random variable. In particular, we approximate the LT of a two-parameter Pareto (β, b) distribution, by selecting values x_i given by

$$x_i = b\left[\left(1 - \frac{i}{M+1}\right)^{-1/\beta} - 1\right],$$

when it can be proved that, as M increases, the *TAM* approximation converges to the true LT.

Now, for the Pareto mixture model with density given by (1), we can approximate the LT as

$$\sum_{r=1}^{k} w_r f_M^*(s|\beta_r, b_r)$$

where $f_M^*(s|\beta_r, b_r)$ is the *TAM* approximation to $f^*(s|\beta, b)$, the LT of the two-parameter Pareto distribution with parameters (β, b).

Now, assuming that the queueing system is stable, the equilibrium distributions can be found by approximating the LT in (8) to give

$$\sum_{i=1}^{M}\sum_{g=1}^{k} w_g e^{-\mu(1-r_0)b_g\left[\left(1-\frac{i}{M+1}\right)^{-1/\beta_g}-1\right]} = r_0 M. \tag{10}$$

which can be easily solved numerically.

Shortle et al (2004) propose a generalization of the *TAM* method called *geometric TAM* or *GTAM*. Their idea is to pick quantiles further out in the tail of the distribution. This could be important for heavy-tailed distributions, since in that way, the tail behavior could be better captured. The M-point approximation of the interarrival distribution Laplace Transform is given in this case by

$$f_M^*(s) = \sum_{i=1}^{M} p_i e^{-sx_i}$$

where the *TAM* samples are the quantiles corresponding to some geometric probabilites $F(x_i) = y_i = 1 - q^i$ for some value of $q \in (0, 1)$ and $i = 1, \ldots, M$. For the Pareto distribution defined in Section 2,

$$x_i = b\left(q^{-i/\beta} - 1\right).$$

The values of the weights $p(i)$ are defined as

$$\begin{aligned} p_1 &= \frac{y_1 + y_2}{2} \\ p_i &= \frac{y_{i+1} + y_{i-1}}{2} \qquad \text{for } i = 2, \ldots, M-1 \\ p_M &= 1 - \frac{y_{M-1} + y_M}{2} \end{aligned}$$

(10) becomes for this method,

$$\sum_{i=1}^{M}\sum_{g=1}^{k} w_g p_i e^{-\mu(1-r_0)b_g\left(q^{-i/\beta_r}-1\right)} = r_0. \tag{11}$$

We also tried this extension in our simulations; see Subsection 3.4.

3.2. LEVEL CROSSING METHOD

Harris et al (2000) introduce an alternative algorithm for directly finding the root r_0 of (8), the *Level Crossing Method*. They demonstrate that for any stable $G/M/1$ queueing system that if $F(x)$ is the cumulative distribution function of the interrarival time, then the associated survival function, $\bar{F}(x) = 1 - F(x)$ verifies:

$$\int_0^\infty \bar{F}(u) e^{-\gamma} du = \frac{1}{\mu}, \tag{12}$$

where $\gamma = \mu(1 - r_0)$.

In the case where the interarrival distribution is the Pareto mixture, (12) becomes:

$$\int_0^\infty \sum_{r=1}^{k} w_r \frac{b_r^{\beta_r}}{(u + b_r)^{\beta_r}}\ e^{-\gamma} du = \frac{1}{\mu}. \tag{13}$$

This equation can be solved numerically to estimate γ and therefore r_0

Thus, conditional on the system parameters and assuming that the queuing system is stable, either the *TAM* or the *Level Crossing* algorithm can be used to estimate r_0 and facilitate the estimation of the predictive equilibrium distributions of queue size etc.

It is not clear which of the two root estimation methods is more efficient. We consider this aspect in Section 3.4 where we analyze the real and simulated queueing data.

3.3. INFERENCE FOR THE k-Par/M/1 QUEUEING SYSTEM

Given a sample of interarrival data, we have seen that the BDMCMC algorithm can be used to produce a sample of values $\left\{k^{(i)}, \mathbf{w}^{(i)}, \boldsymbol{\beta}^{(i)}, \mathbf{b}^{(i)}\right\}$ for $i = 1, \ldots, N$ from the posterior distribution of the interarrival parameters.

Supposing that the service rate μ is known, then it is straightforward to estimate the probability that the system is stable,

$$P(\rho < 1|\text{data}) \simeq \frac{1}{N}\sum_{i=1}^{N} I\left(\rho^{(i)} < 1\right)$$

where $\rho^{(i)}$ is the value of ρ calculated from (9) using the i'th set of interarrival parameters and μ and $I(\cdot)$ is an indicator function. Given that this probability is high, then for each set $\left\{\mu, k^{(i)}, \mathbf{w}^{(i)}, \boldsymbol{\beta}^{(i)}, \mathbf{b}^{(i)}\right\}$ of generated parameters such that $\rho^{(i)} < 1$, the root $r_0^{(i)}$ can be generated and the posterior predictive distributions of queue size etc. can be estimated by Rao Blackwellization.

One point to note however is that, as commented in Wiper (1997), it can be shown that the predictive means of the equilibrium queue size or waiting time distributions do not exist. This is a typical feature for Bayesian inference in $G/M/\cdot$ or $M/G/\cdot$ queueing systems. Thus, if posterior summaries of these distributions are required, it is preferable to use the median and quantiles.

Note that in the case that μ is unknown, we can consider the experiment of observing a number of service times as, for example, in Armero and Bayarri (1994). Then, a sample of size N could also be generated from the posterior distribution of μ and, combined with the interarrival parameter sample, can be used to estimate the traffic intensity and predictive distributions as above.

3.4. COMPARISON OF THE *TAM* and *Level Crossing* METHODS WITH SIMULATED DATA

Here we consider the simulated arrival data analyzed in Section 2.3 and we shall assume a single exponential server with known service time $E[S] = 1$. In this case, the queueing system is stable, as the expected interarrival time is infinite. Given the true interarrival time distribution, then the value of r_0 can be computed to be equal to 0.5873 using either the *TAM* or *Level Crossing* method. We also tried the *GTAM* method. The problem we found with this generalized *TAM* is the fact that the method performs well only for some "appropriate" values of q. If q is very close to 1 (0.999) the tail is not captured, but if it is close to the 0, the body of the distribution is not

taken into account. In Shortle et al (2004), the authors propose choosing q so that the TAM mean matches the mean of the original distribution. Our experiments showed that in that way, the obtained value for r_0 (0.5872) is similar to that found with the original TAM but given the extra effort needed to find the optimal q, the computational time increases.

Based on the sampled interarrival data, the posterior probability that the queueing system was stable was estimated to be 0.9999. Thus, in Table 1 we are able to compare the true distribution of the equilibrium queue size and the posterior predictive distributions calculated using the sampled interarrival data via the TAM and *Level Crossing* methods respectively.

Table 1: True and predicted equilibrium queue size distributions

n	*True* $P(Q=n)$	*TAM* $P(Q=n\|\text{data})$	*Level Crossing* $P(Q=n\|\text{data})$
0	0.4127	0.3940	0.3929
1	0.2424	0.2385	0.2382
2	0.1423	0.1414	0.1416
3	0.0836	0.0876	0.0878
4	0.0409	0.0531	0.0534
5	0.0288	0.0323	0.0325
6	0.0169	0.0196	0.0198
7	0.0099	0.0119	0.0120
8	0.0058	0.0073	0.0073
9	0.0034	0.0044	0.0045

The results are almost identical but from the computational point of view the TAM method is up to four times faster than the *Level Crossing* algorithm. Although this is unimportant when the root has only to be evaluated once, within the MCMC algorithm where many evaluations are made this time difference can be considerable. Thus, we would suggest that the TAM method be preferred in this context.

4. RESULTS FOR REAL DATA SETS

In this section, we examine two real data sets taken from the `xtremes` site:

`http://www.xtremes.de/xtremes/xtremes/download/download.htm`.

The first of these (t2a) consists of 50000 interarrival times in seconds of a trace of 1 million ethernet packets, and the second one (t4a) is composed of 226386 world wide web transfer packet sizes in bytes.

Figure 4 shows the empirical and fitted distribution functions, based on 20000 MCMC iterations and given the same prior distributions as outlined in Section 2.3, for both data sets fitted on two scales in each case so that the tail behaviour can be observed.

FIGURES 4.1, 4.2, 4.3 AND 4.4 ABOUT HERE

Empirical (solid line) and Predictive (dotted line) distributions for the t2a data (top) and the t4a data (bottom).

We can see that the t4a data in particular are very long tailed. The t4a data appear very well fitted although the fit of the t2a data is not quite so good, although the tail behaviour is well estimated.

Table 2 gives the posterior distributions of the number of mixture components for both data sets. There is a high posterior probability of 2 mixture components for the t2a data set and more uncertainty for the t4a data although the posterior mode is also 2 components.

Table 2: Posterior probabilities of numbers of mixture components

	Data set	
	t2a	t4a
k	$P(k\|\text{data})$	$P(k\|\text{data})$
1	0.03	0.375
2	0.763	0.4096
3	0.198	0.2113
4	0.009	0.0042

Now we shall consider the queueing aspects. Firstly we consider the t2a data. Figure 5 gives the posterior probability that the system is stable for a various different values of the service rate μ. From the table, it is clear that there is a high probability that the system is stable for values of μ greater than 385.

FIGURE 5 ABOUT HERE

Posterior probabilities of stability for various values of μ.

In the following figures we illustrate the predictive queueing and system

time distributions and the distribution of the number of clients in the system in equilibrium for values of μ greater than 385.

FIGURES 6.1 AND 6.2 ABOUT HERE
Predictive system waiting time and queue waiting time distributions for `t2a` data set.

FIGURE 7 ABOUT HERE
Predictive system size distribution just before an arrival for `t2a` data set.

We can see that as the service rate increases, then the median waiting and system times and the numbers of clients in the system decrease, as would be expected.

We now consider the `t4a` data set. There exists a relationship between the file size and the transfer time: if a file has a large size, its transmission (or download) time is longer, so these data can be also thought as interarrival times in a k-Par/M/1 queueing system for some time measure depending on the transfer system features.

As the expected interarrival time (predictive size mean) is infinite, this queue is stable for all values of μ and no congestion traffic problems happen. Figure 8 shows the predictive queueing and system time distributions when $\mu = 0.01$.

FIGURES 8.1 AND 8.2 ABOUT HERE
Predictive system waiting time and queue waiting time distributions for `t4a` data set if $\mu = 0.01$.

Finally, Table 3 gives the distribution of the equilibrium queue size when $\mu = 0.01$.

Table 3: Predictive system size distribution just before an arrival for `t4a` data set when $\mu = 0.01$.

n	0	1	2	3	4
$P(Q = n)$	0.8263	0.1419	0.0255	0.0049	0.001

5. CONCLUSIONS AND EXTENSIONS

In this paper, we have illustrated how to carry out Bayesian inference for a mixture of Pareto distributions and then combined this approach with techniques from the queueing literature in order to estimate predictive equilibrium distributions for the k-Par/M/1 system.

An advantage of our mixture model is the flexibility it provides in fitting data sets where the single Pareto model is not adequate. From the queueing point of view, we may also conclude that the *TAM* algorithm should be preferred to the *Level Crossing* approach for root estimation in this context as, although both methods produce similar results, the computational expense in implementation of the *TAM* algorithm is somewhat lower.

A number of extensions are possible. Firstly, following Harris et al (2000), we could extend our results to the case of a multiple number of servers, that is, to study the behaviour of the queueing system k-Par/M/c. A different interesting extension would be to compare both *TAM* and *Level Crossing* methods with the approach proposed in Rodríguez-Dagnino (2005).

Another important objective in the queueing context is optimal control of the system, that is, when to open and close the system or optimizing the number of servers. Methods of Bayesian decision theory can be combined with Bayesian inference to do this. See for example, Ausín et al (2004).

In this article, we have considered just exponentially distributed service times. One possible extension is to consider the case of more general service time distributions, in particular the so called phase type service type distributions.

Finally, we would like to develop Bayesian inference methods and queueing results for another heavy-tailed distribution variables or internet-related variables, such us the double Pareto or double Pareto lognormal distribution, recently suggested to be very suitable in the Internet context. See for example, Mitzenmacher (2003) or Reed and Jorgensen (2004).

BIBLIOGRAPHY

Abate, J. and Whitt, W. (1999). Computing Laplace transforms for numerical inversion via continued fractions. *INFORMS Journal on Computing*, **11**, 394–405.

Armero, C. and Bayarri, S. (1994). Bayesian prediction in *M/M/1* queues. *Queueing Systems*, **15**, 401–418.

Arnold, B. and Press, S.J. (1989). Bayesian estimation and prediction for Pareto data. *Journal of the American Statistical Association*, **84**, 1079–1084.

Ausín, M.C., Lillo, R.E. and Wiper, M. (2004). Bayesian control of the number of servers in a *GI/M/c* queueing system. *Working paper 04-69*, Statistics and Econometrics Series 17, Universidad Carlos III de Madrid.

Beran, J., Sherman, R., Taqqu, M.S. and Willinger, W. (1995). Long-range dependence in variable-bit-rate video traffic. *IEEE Transactions on Communications*, **43**, 1566–1579.

Crovella, M., Taqqu, M. and Bestavros, A. (1998). Heavy-tailed probability distributions in the World Wide Web. In R.J. Adler, R.E. Feldman and M.S. Taqqu (eds), *A Practical Guide to heavy tails*, New York: Chapman and Hall, 3–26.

Feldmann, A. and Whitt, W. (1998). Fitting mixtures of exponentials to long-tail distributions to analyze network performance models. *Performance Evaluation*, **31**, 245–258.

Fischer, M. and Harris, C. (1999). A method for Analyzing Congestion in Pareto and Related Queues. *The telecommunications review*, Chap 2, 15–27.

Fischer, M. and Gross, D. and M. Bevilacqua, D. and Shortle, J. (2001). Analyzing the waiting time process in Internet queueing systems with the transform approximation method. *The telecommunications review*, **12**, 21–32.

Green, Peter J. (1995). Reversible jump Markov Chain Monte Carlo computation and Bayesian model determination. *Biometrika*, **82**, 711–732.

Greiner, M., Jobmann, M. and Lipsky, L. (1999). The Importance of Power-Tail Distributions for Modeling Queueing Systems. *Operations Research*, **47**,

313–326.

Harris, C., Brill, P. and Fisher, M. (2000). Internet-type queues with power-tailed interarrival times and computational methods for their analysis. *INFORMS Journal on Computing*, **12**, 261–271.

Mitzenmacher, M. (2003). Dynamic models for file sizes and double pareto distributions. *Internet Math.*, **1**, 305–333.

Park, K. and Willinger, W (2000). *Self-Similar Network Traffic and Performance Evaluation*. John Wiley & Sons. Inc. New York, NY, USA.

Paxson, V. and Floyd, S. (1995). Wide-Area Traffic: The Failure of Poisson Modeling. *IEEE Transactions in Networking*, **3**, 226–244.

Reed, W.J. and Jorgensen, M.A. (2004). The double Pareto-lognormal distribution - A new parametric model for size distributions. *Communications in Statistics: Theory and Methods*, **33**, 1733–1753.

Richardson, S. and Green, P. (1997). On bayesian analysis of Mixtures with an unknown number of components. *Royal Statistical Society*, **59**, 731–792.

Riska, A., Diev, V. and Smirni, E. (2004). Efficient fitting of long-tailed data sets into phase-type distributions. *Performance Evaluation Journal*, **55**, 147–164.

Rodríguez-Dagnino, R.M. (2004). On the $Pareto/M/c$ and $Pareto/M/1/K$ queues. *Proc. SPIE, ITCom 2004, Performance, Quality of Service, and Control of Next-Generation Communication Networks II*, **5598**, 183–193.

Rodríguez-Dagnino, R.M. (2005). Some remarks regarding asymptotic packet loss in the $Pareto/M/1/K$ queueing system. *IEEE Communications Letters*, **9:10**, 927–929.

Shortle, J.F., Brill, P.H., Fischer, M.J., Gross, D. and Masi, D.M.B. (2004). An algorithm to compute the waiting time distribution for the $M/G/1$ queue. *INFORMS Journal on Computing*, **16**, 152–161.

Stephens, M. (2000). Bayesian Analysis of mixtures with an unknown number of components -An alternative to reversible jump methods. *Annals of Statistics*, **28**, 40–74.

Wiper, M.P. (1997). Bayesian analysis of Er/M/1 and Er/M/c queues. *The Journal of Statistical Planning and Inference*, **69**, 65–79.

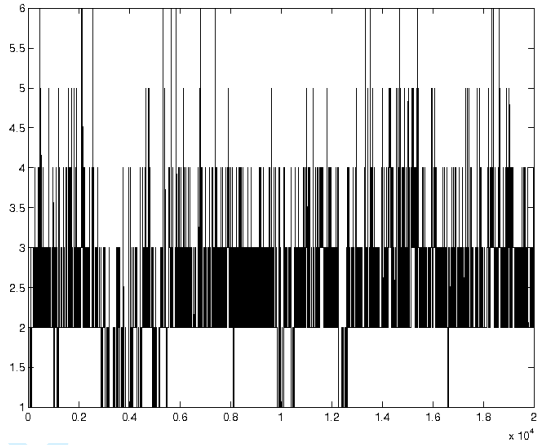


Figure 1: Graph of mixture components k versus iterations.

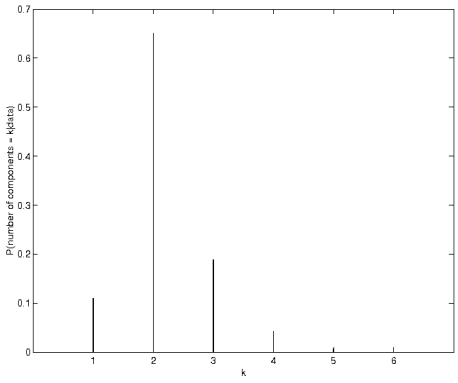


Figure 2: Posterior distribution of k.

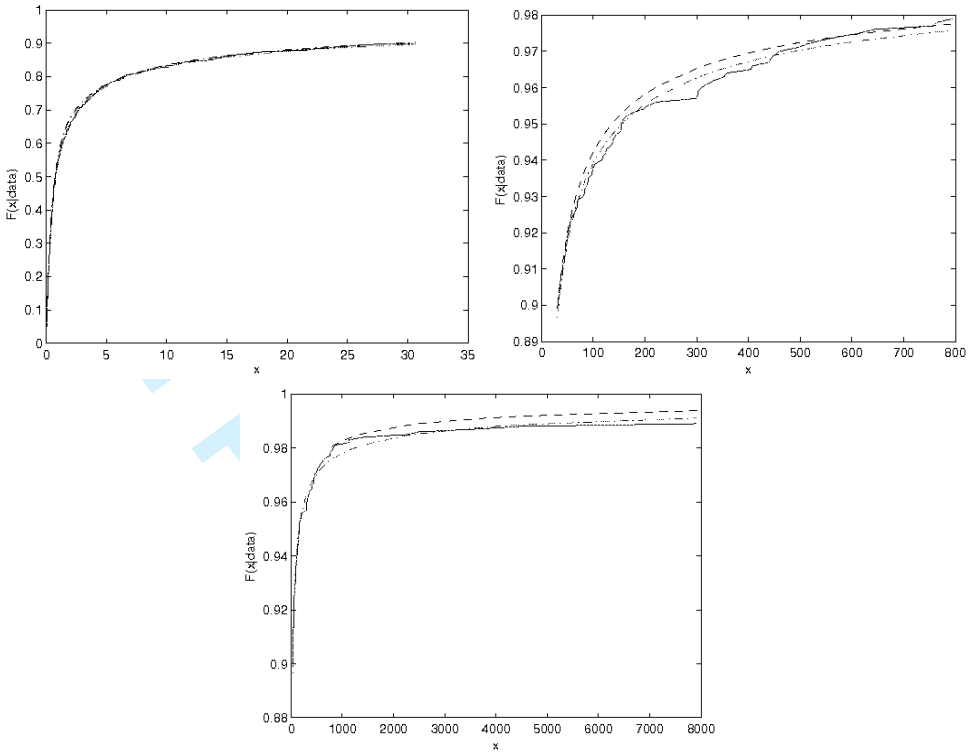


Figure 3: Empirical (solid line), predictive (pointed dotted line) and theoretical (dotted line) cumulative distribution functions for the first 900 data (top left), the 900th-980th data (top right) and the 980th-990th data (bottom).

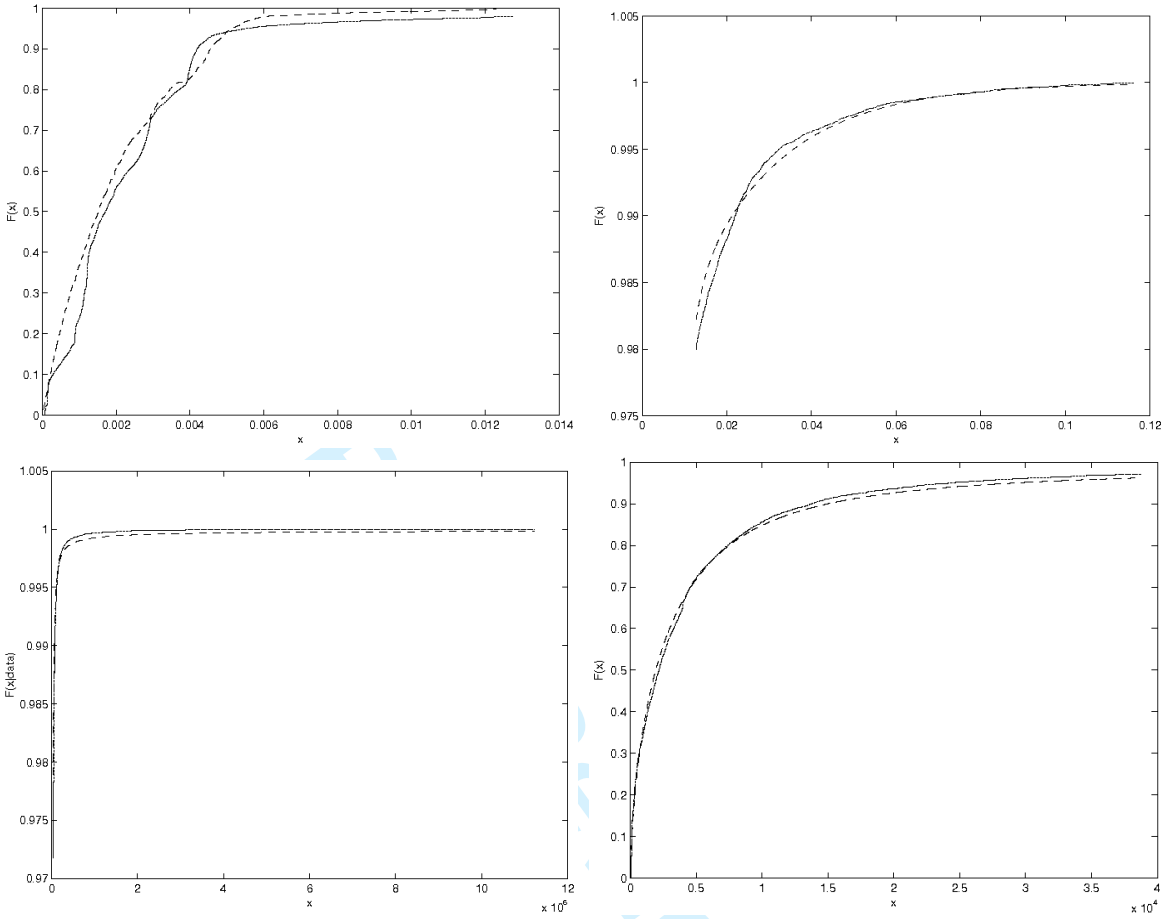


Figure 4: Empirical (solid line) and Predictive (dotted line) distributions for the `t2a` data (top) and the `t4a` data (bottom).

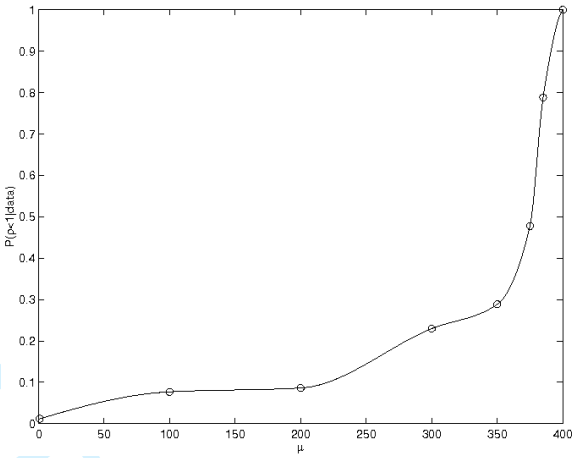


Figure 5: Posterior probabilities of stability for various values of μ.

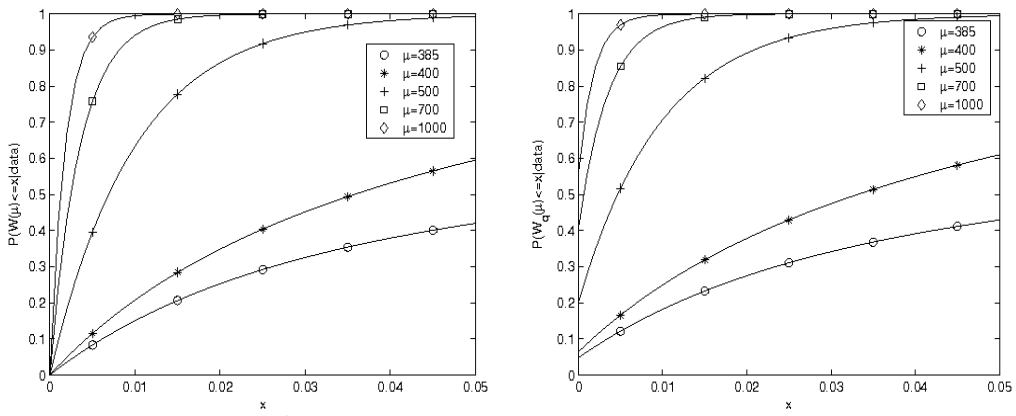


Figure 6: Predictive system waiting time and queue waiting time distributions for `t2a` data set.

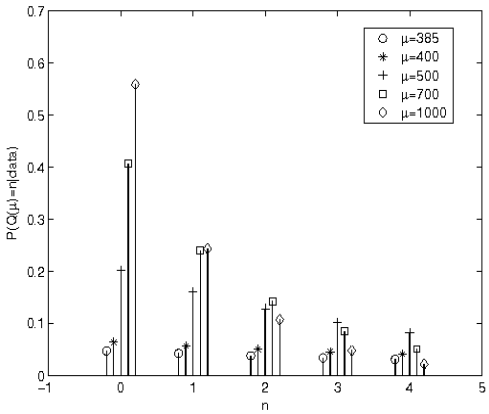


Figure 7: Predictive system size distribution just before an arrival for `t2a` data set.

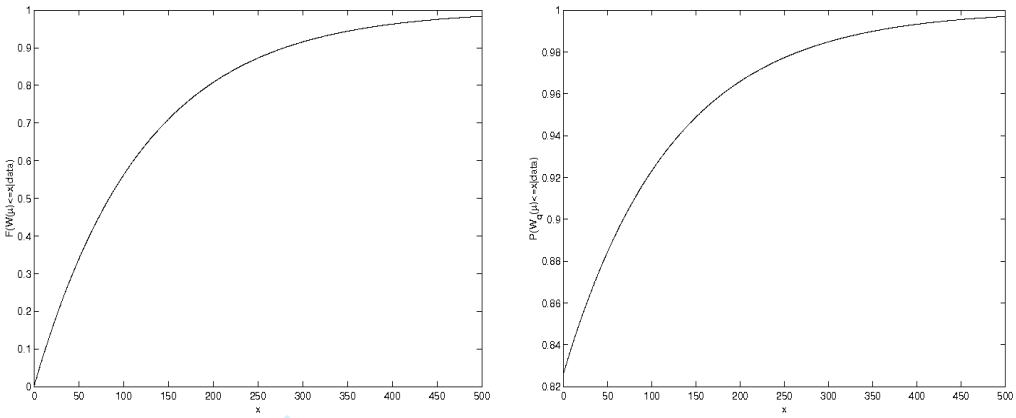
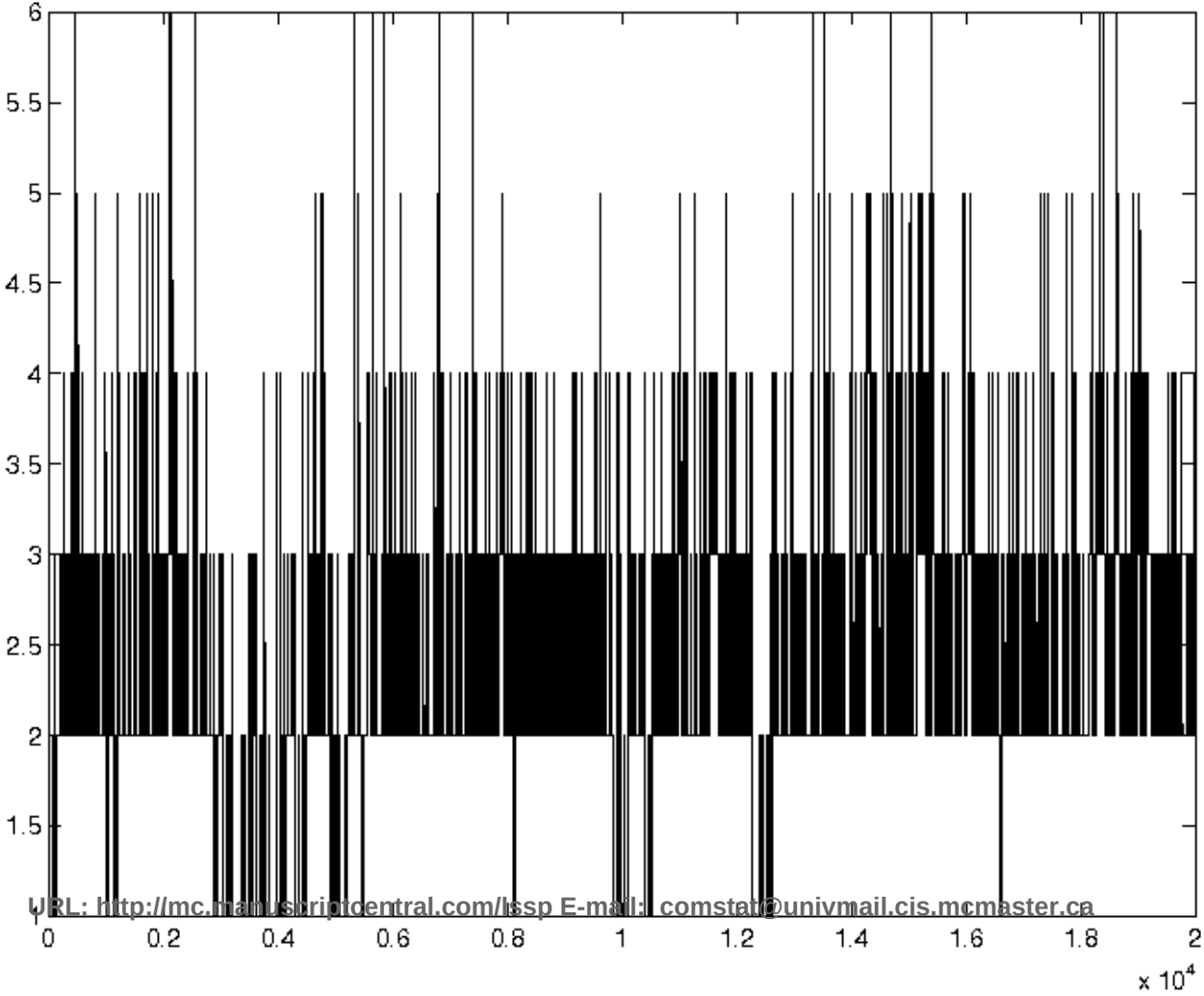
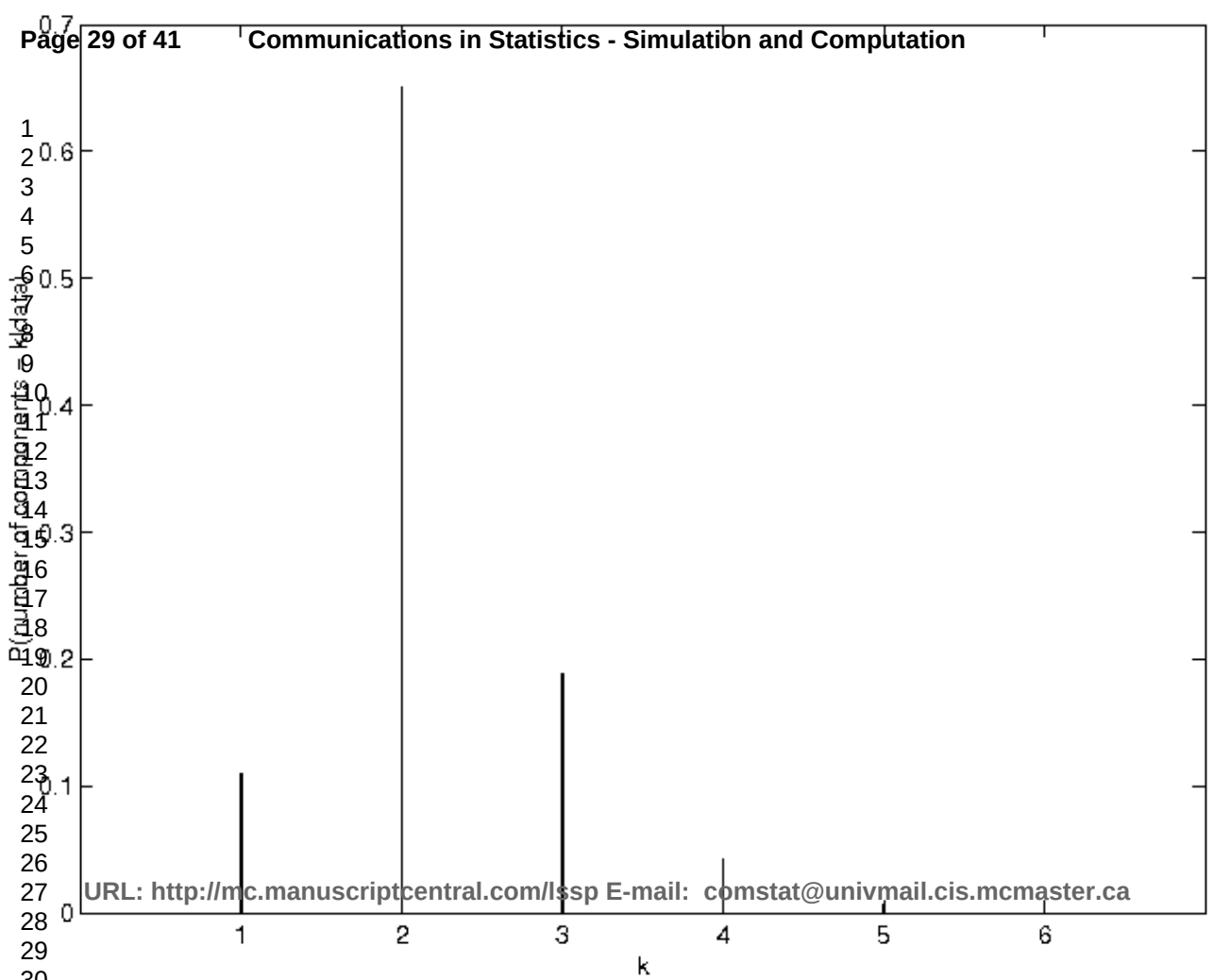
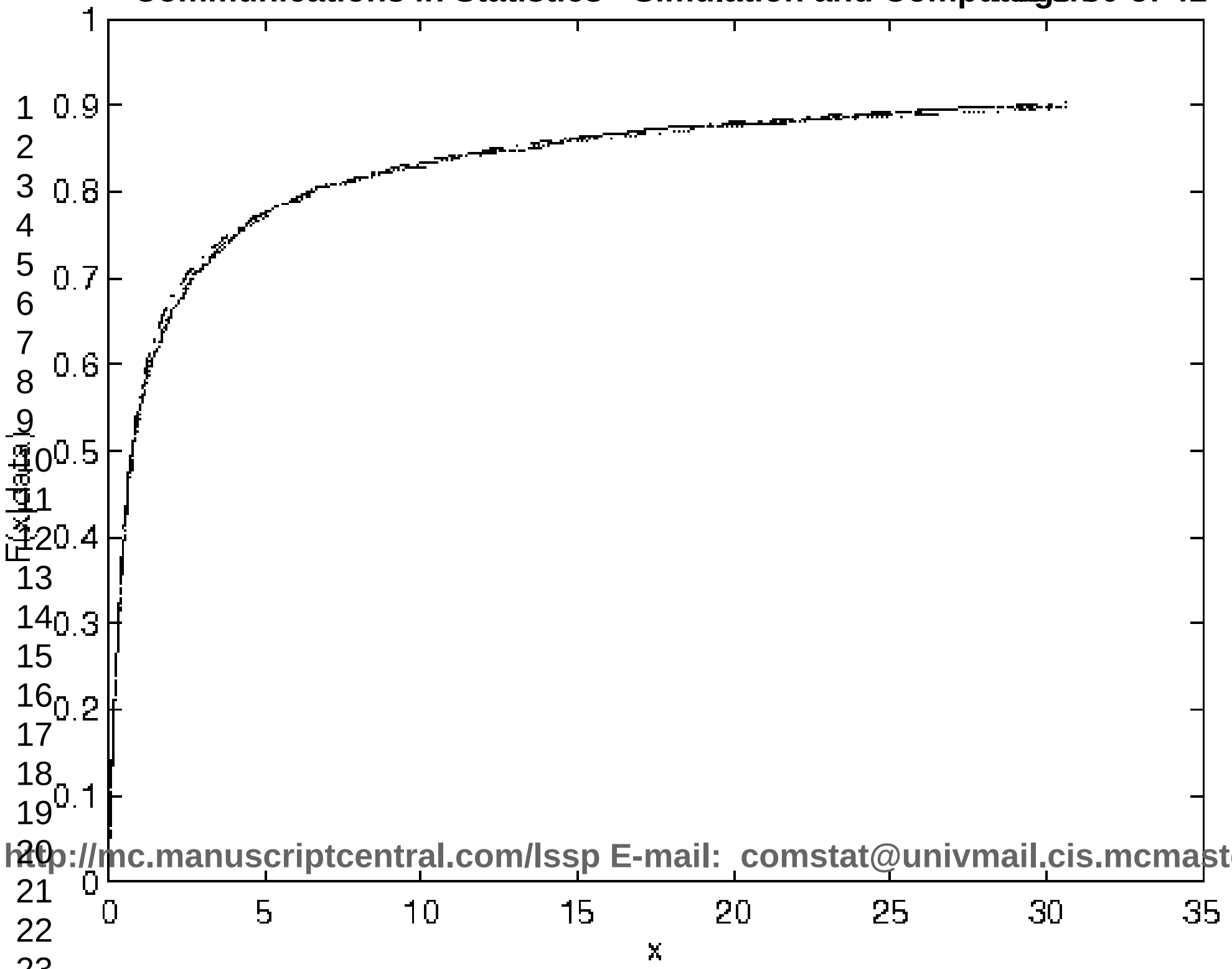
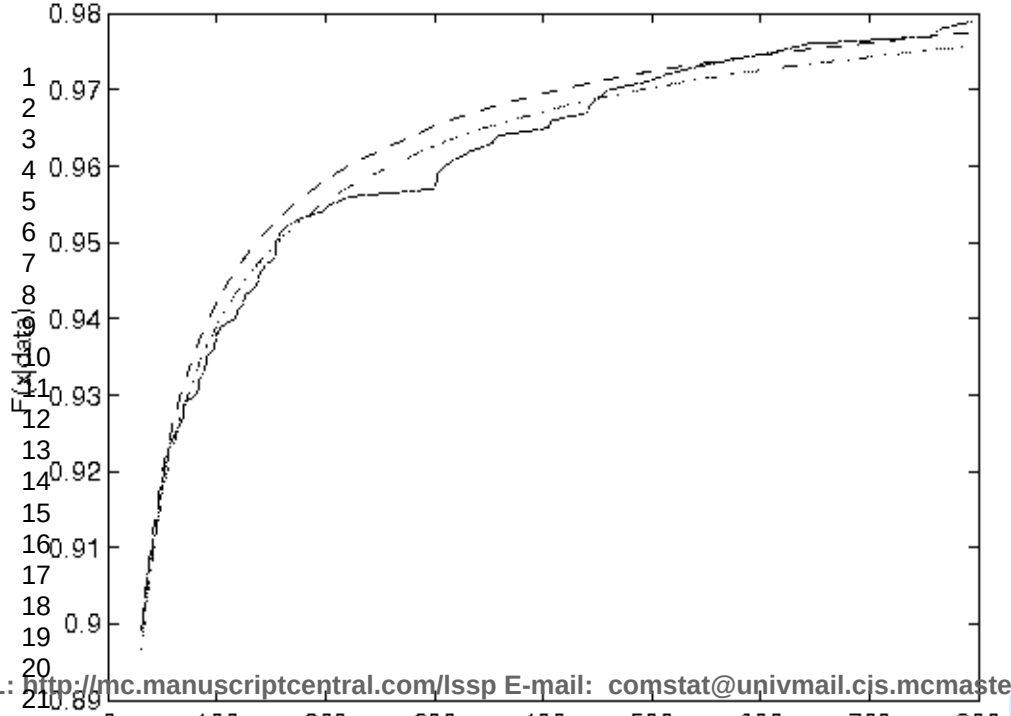


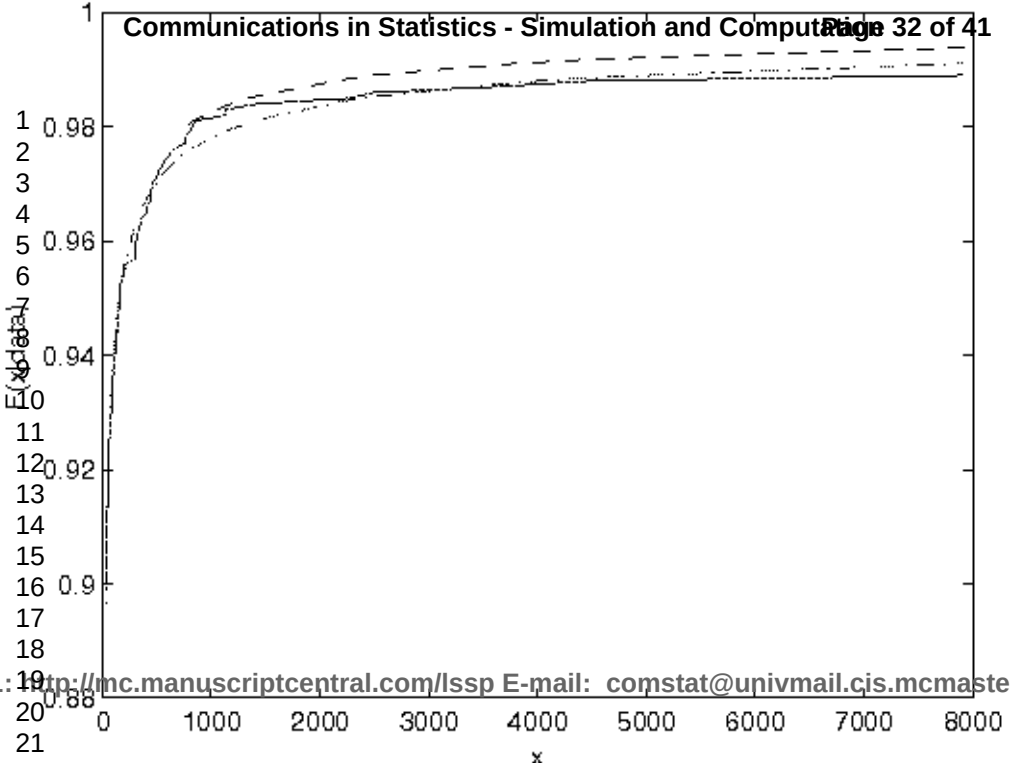
Figure 8: Predictive system waiting time and queue waiting time distributions for `t4a` data set if $\mu = 0.01$.

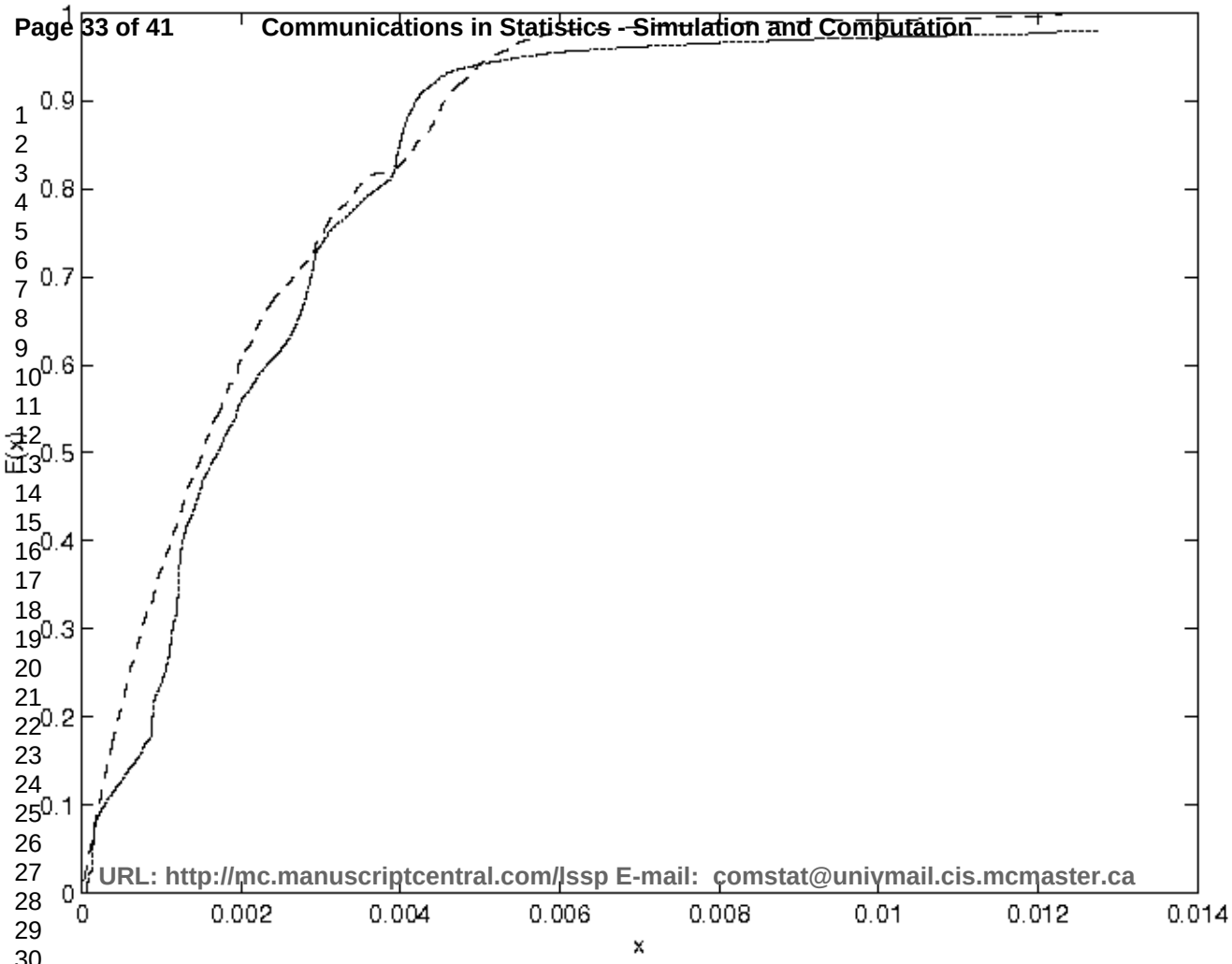

6
5.5
5
4.5
4
3.5
3
2.5
2
1.5
1
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
2
x 10^4

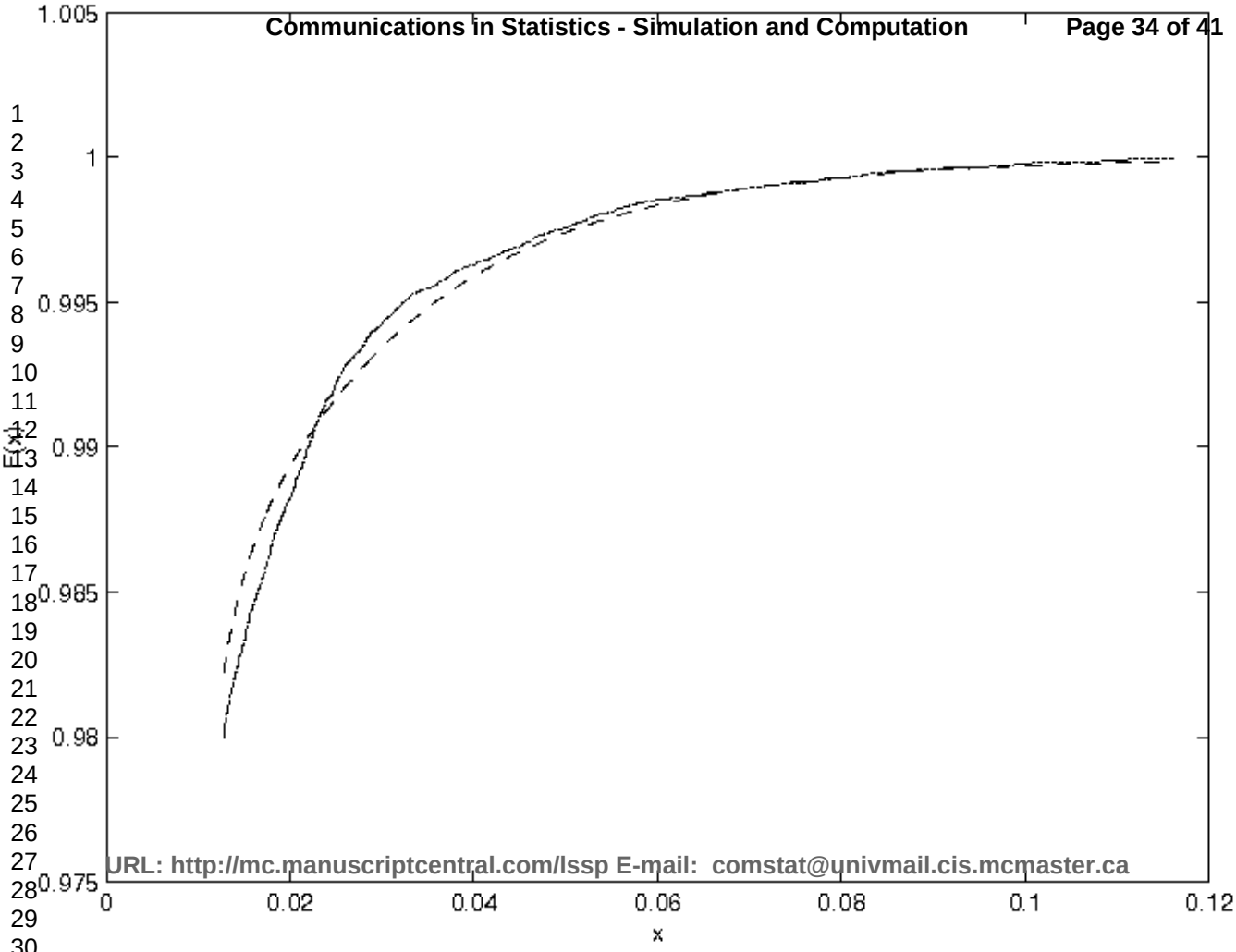

P(number of components = k|data)
k
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
1
2
3
4
5
6

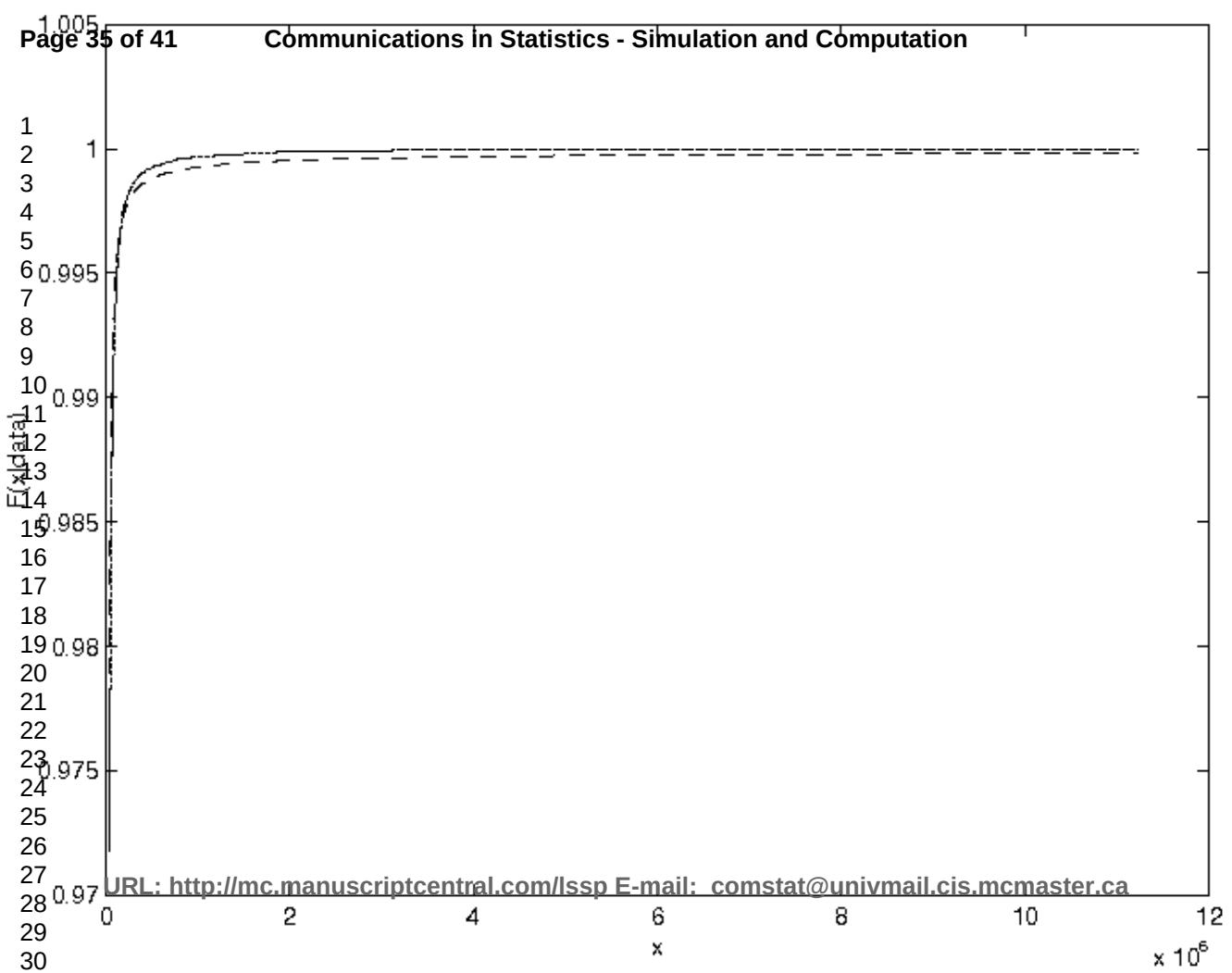

F(x|data)
x
0
5
10
15
20
25
30
35
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

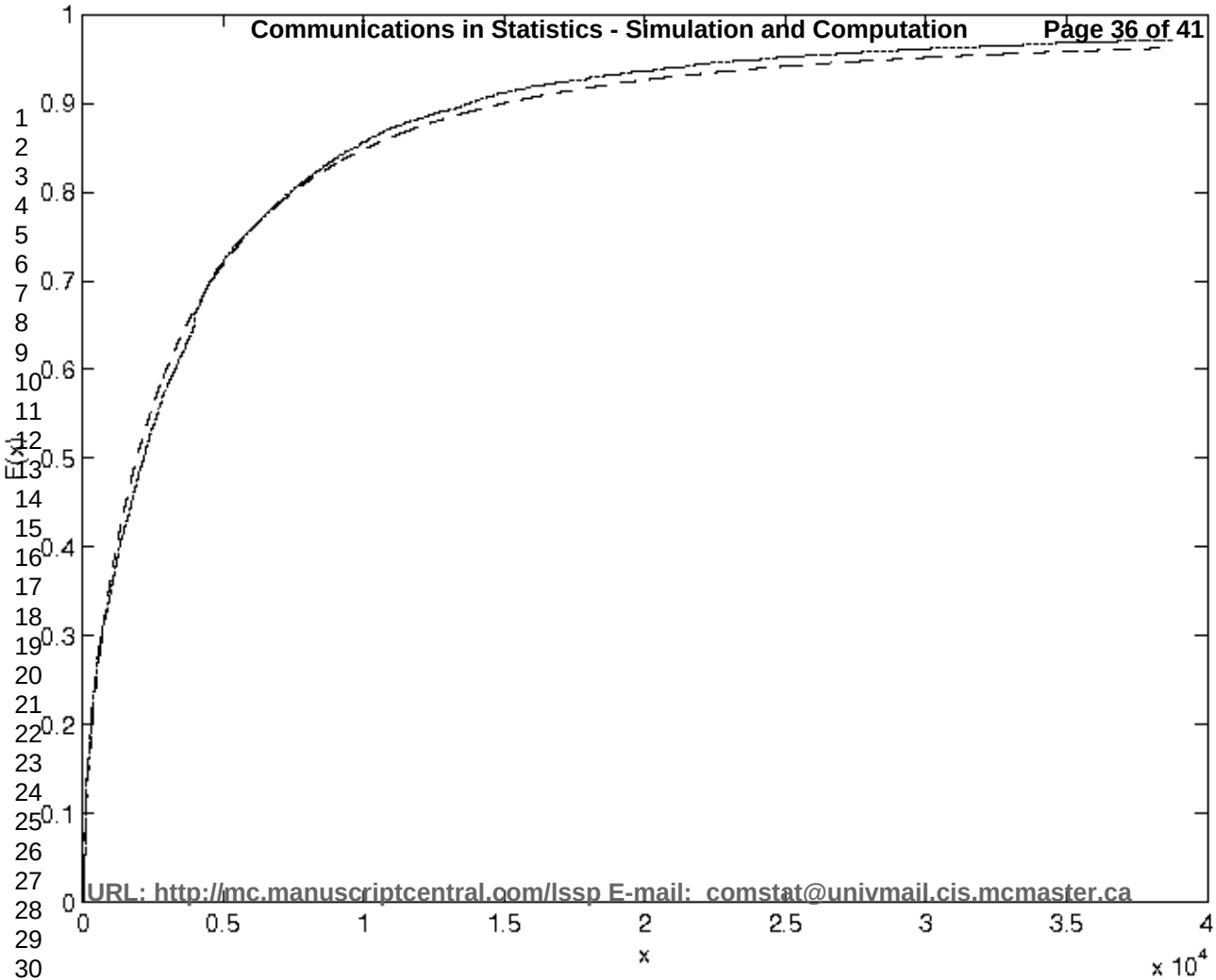

0.98
0.97
0.96
0.95
0.94
0.93
0.92
0.91
0.9
0.89
F(x|data)
0
100
200
300
400
500
600
700
800
x

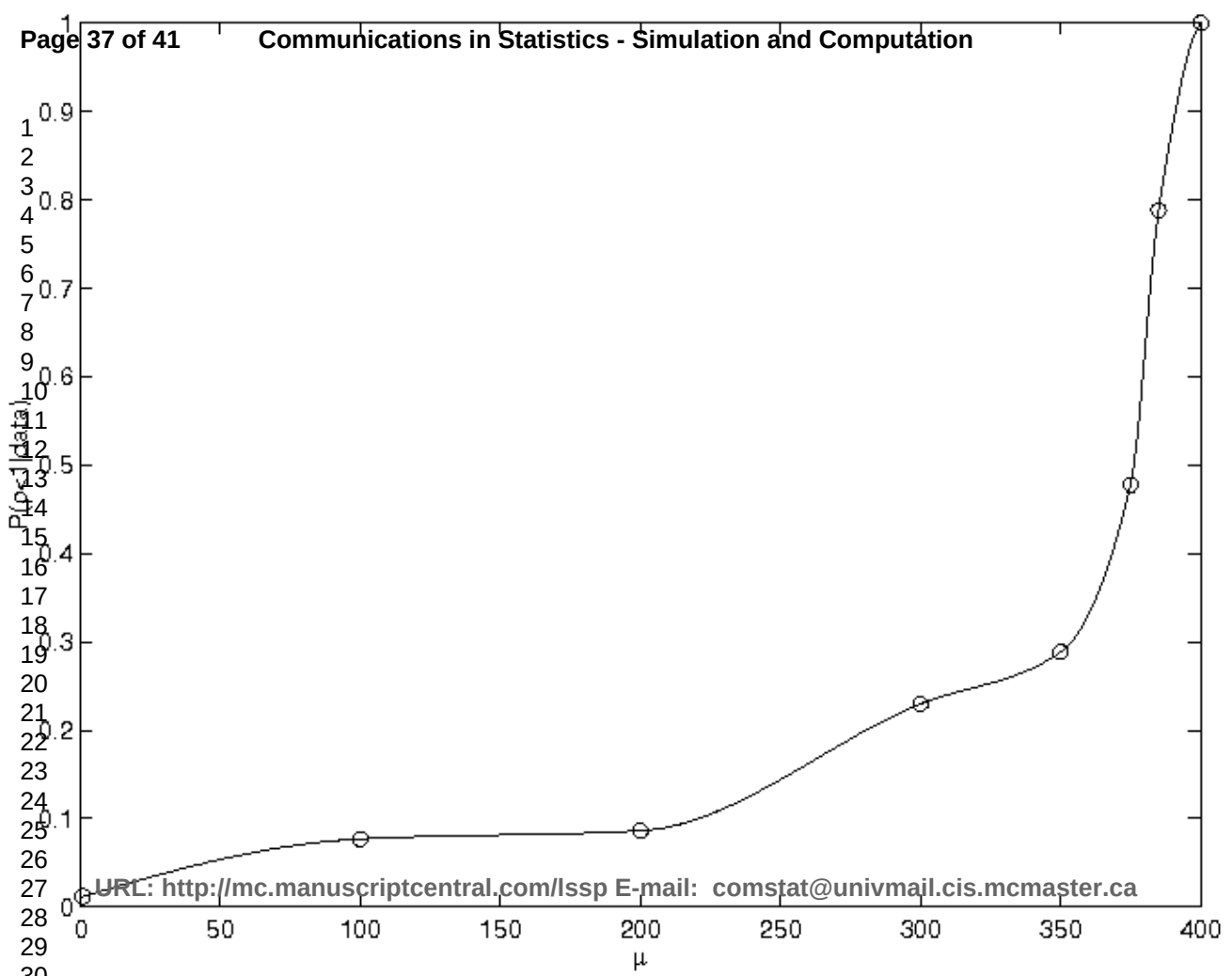

1
0.98
0.96
0.94
0.92
0.9
0.88
E(xbar)
0
1000
2000
3000
4000
5000
6000
7000
8000
x

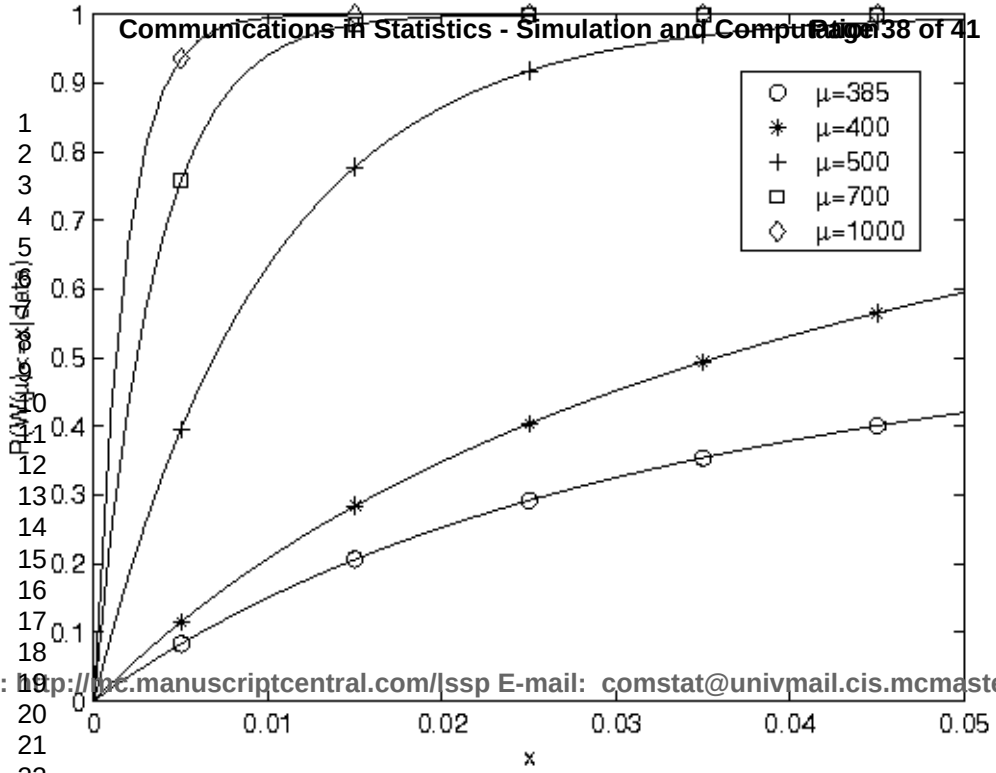

F(x)
x
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
0.002
0.004
0.006
0.008
0.01
0.012
0.014

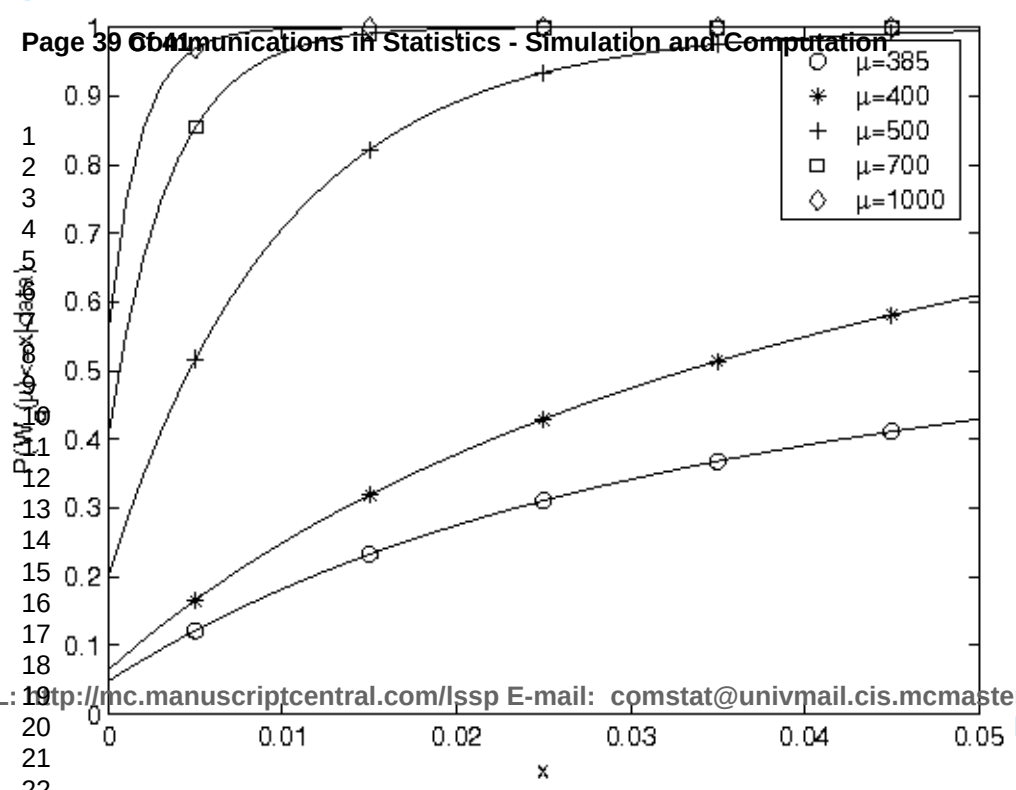

1.005
1
0.995
0.99
0.985
0.98
0.975
E(x)
0
0.02
0.04
0.06
0.08
0.1
0.12
x

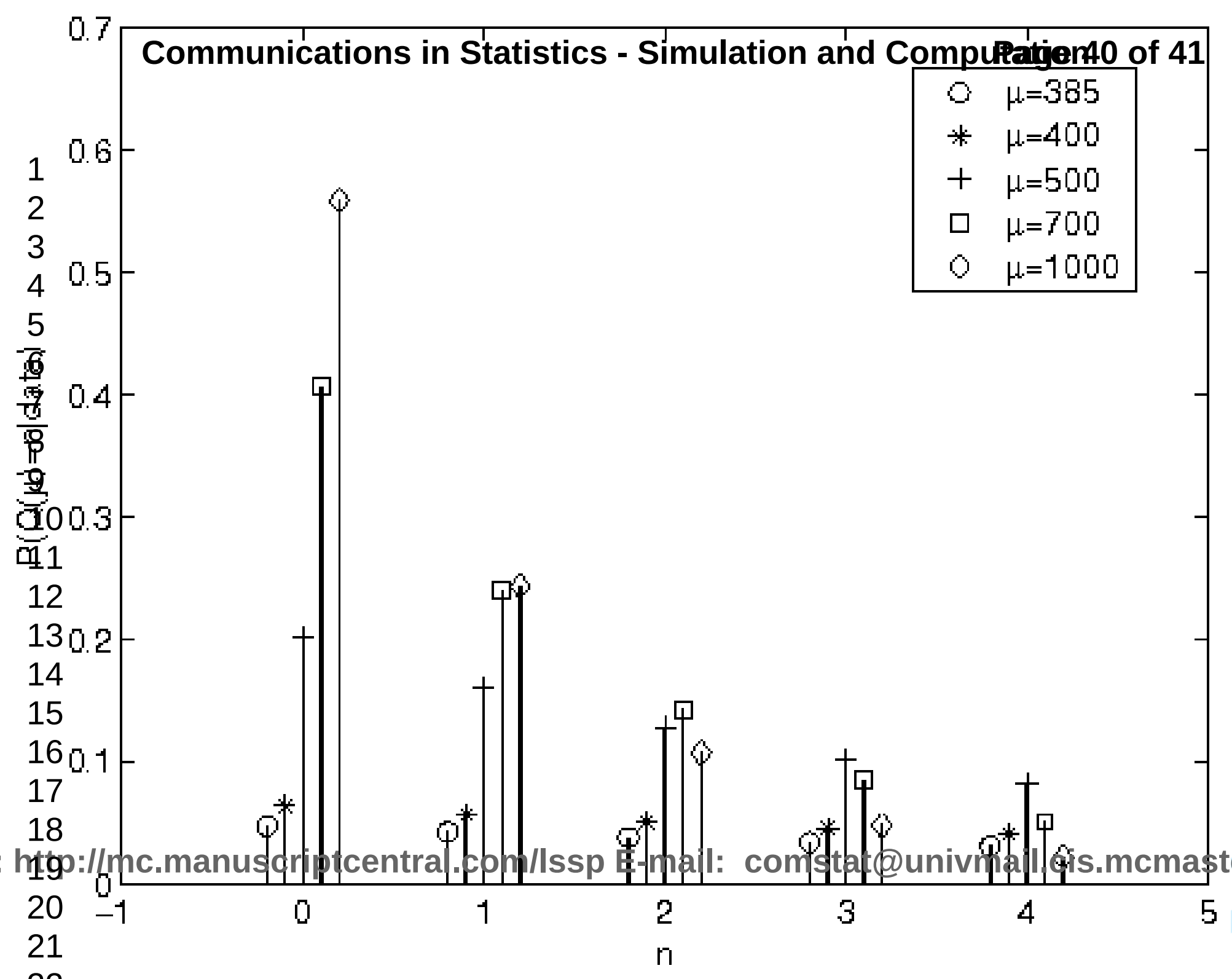

1.005
1
0.995
0.99
0.985
0.98
0.975
0.97
F(x|data)
0
2
4
6
8
10
12
x
x 10^6

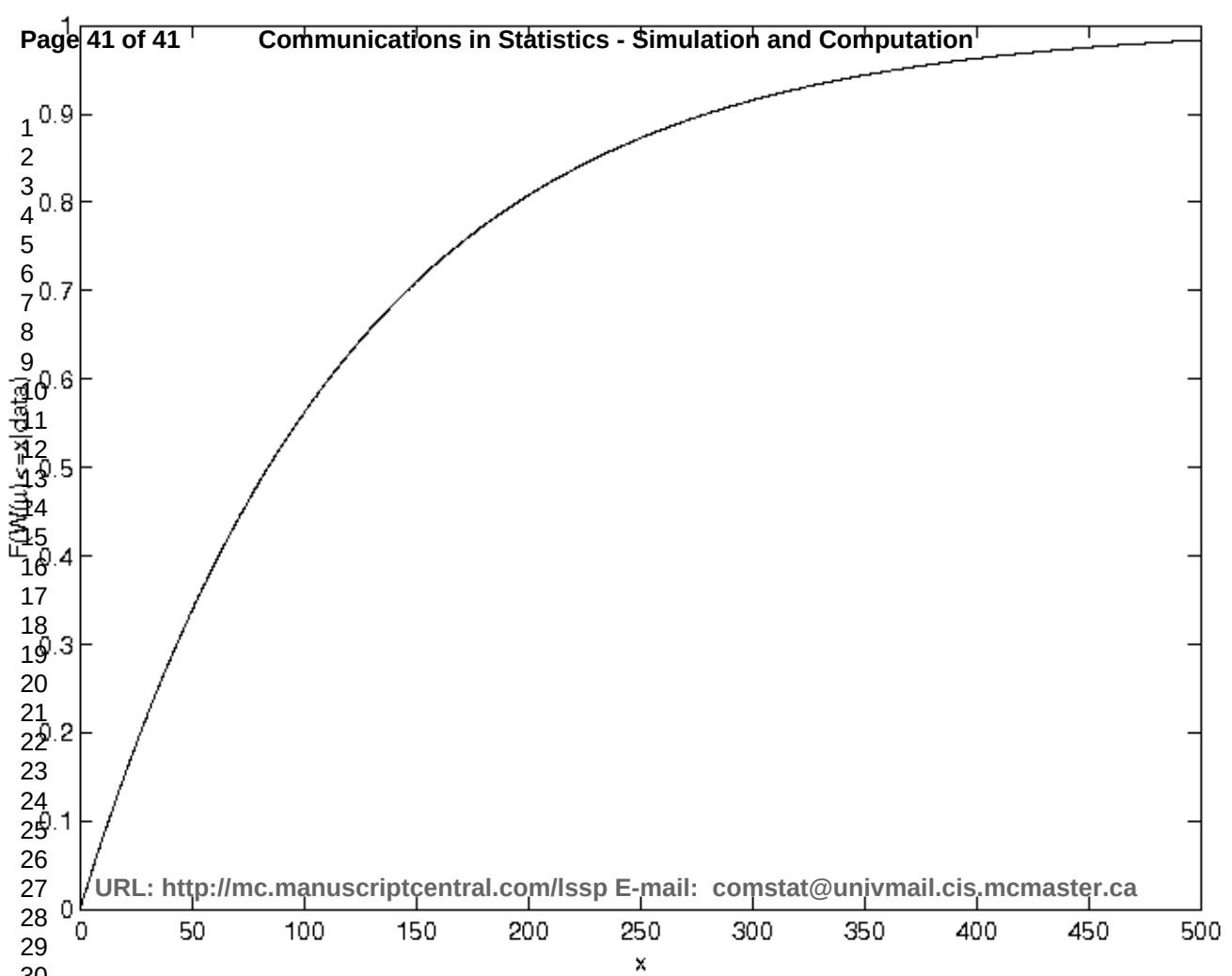

F(x)
x
x 10^4
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
0.5
1
1.5
2
2.5
3
3.5
4

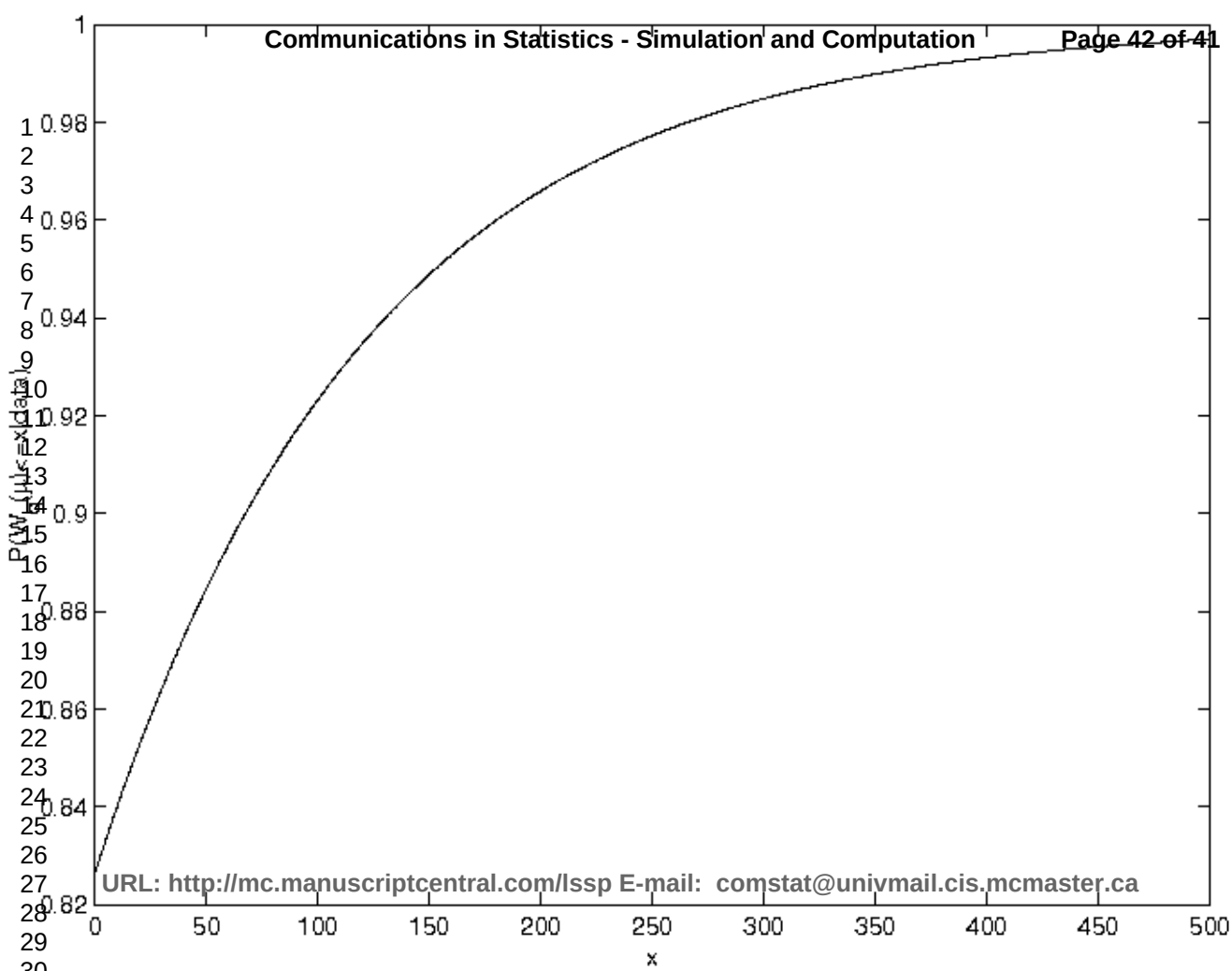

P(p<1 | data)
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
50
100
150
200
250
300
350
400
μ


μ=385
μ=400
μ=500
μ=700
μ=1000
x
0
0.01
0.02
0.03
0.04
0.05
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1


μ=385
μ=400
μ=500
μ=700
μ=1000
P(W_(μ,σ)<x|data)
x
0
0.01
0.02
0.03
0.04
0.05
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1


μ=385
μ=400
μ=500
μ=700
μ=1000
n


F(W|μ,s=x|data)
x
0
50
100
150
200
250
300
350
400
450
500
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1


1
0.98
0.96
0.94
0.92
0.9
0.88
0.86
0.84
0.82
0
50
100
150
200
250
300
350
400
450
500
x