A Framework for Quickest Detection of Traffic Anomalies in Networks

Parijat Dube [*] Ravi Mazumdar [†]

Abstract

This paper presents a stopping theoretic framework for the detection of traffic anomalies in networks. In particular we consider the case when only aggregate traffic flows can be measured. Under the assumption that the traffic changes are sudden and take place at a random time, we show that it is possible to define stopping rule which yields decisions on anomalies with minimum average detection delay in the class of all decisions based on the available information subject to a bound on the false alarm rate. We demonstrate the behavior of the detection scheme via simulations in both the case with perfect prior knowledge of the flow characteristics as well as in the robust case where only membership of classes is known. The approach is via optimal stopping theory.

1 Introduction

The most common reason for dramatic deterioration in offered network performance is due to traffic congestion. Traffic congestion can arise due to many causes, a network *hotspot*, failure of network components which can involve short term re-routing of traffic through a particular part of the network (or a particular router), and malicious users who wish to flood a particular node or site (or even a sub-net). Hence there is a need to provide network monitoring mechanisms which can detect sudden traffic overloads and react quickly to them.

Part of the difficulty is that we must be able to differentiate between statistical fluctuations, which are normal and for which the network is properly dimensioned, and genuine anomalies. Simple statistical measures such as computing mean rates are far too inefficient and too slow to provide any level of confidence. Indeed the amount of data needed (or the duration before which a decision can be made) for achieving a high level of confidence is usually too large so that the negative effects of congestion will have already set in. In the case of malicious users a server or even subnet could be brought down. Hence there is clearly a need to develop more sophisticated methods which can also be distributed across the network.

Most papers in the literature have focused on fault detection where faults could arise as a result of link failures, software failures or network errors. There are two broad based approaches in the literature, one is by *template* matching i.e. looking for signatures of the anomalies defined in terms of low, high and moderate packet rates where measurements are averages over a given window[4] and

[*] IBM T. J. Watson Research Center, Yorktown Heights, NY 10598, USA. Email: pdube@us.ibm.com

[†] Ravi Mazumdar is with Dept. Electrical and Computer Engineering, Purdue University, West Lafayette, IN, 79406. E-mail:mazum@ecn.purdue.edu

the second approach is via rule based approach for fault localization via the correlation of alarms [5, 7]. These are essentially logical procedures based on knowing exactly the fault scenarios and then correlating the alarms to locate the faults.

As mentioned above simple statistical tests of measuring averages over a given window can lead to a high false alarm rate and increasing the window size can lead to excessive delay as a result of which it may be too late to react. This has led a number of authors to seek more robust and sophisticated methods based on traffic measurements to monitor network conditions. In [6] the authors propose the use of the variance of the cumulative traffic over a fixed time interval and show that if sufficient long measurements are available then anomalies in VoIP traffic can efficiently detected. A similar traffic based statistical scheme using mean and variance estimates is presented in [9]. More recently in [10], the authors propose a Generalized Likelihood Ration (GLR) test [8] to detect traffic anomalies. This involves knowing the normal situation and the likely changes and computing the likelihood ratio based on an AR (auto regressive) characterization and choosing non-overlapping windows over which the sequences are assumed stationary using the method in [2]. The GLR is a widely known sequential test. In the Gaussian case which the authors consider, in essence the procedure checks for changes in the mean and variance. However there is no attempt at optimizing the detection delay since there is a tradeoff between choosing the window lengths and the false alarms.

One of the principal goals of our work is to investigate the suitability of traffic based methods to monitor and detect network attacks. In this context, speed of detection is the essence. In this paper we present a stopping theoretic framework for the use of traffic measurements to detect anomalies in network traffic patterns with the aim of taking preventive measures as quickly as possible. Our primary goal is to investigate the applicability and efficiency of this framework. We assume that only aggregate measures based on buffer occupancy and aggregated flows are available. We show that the theory of optimal stopping provides a very robust and elegant framework for devising quickest detection algorithms. By quickest detection it is meant that the stopping time (or alarm) has the minimum average delay amongst the class of stopping times which can be defined with respect to the same available information and subject to the same constraints on the probability of false alarm. An excellent account of the theory of optimal stopping can be found in [3, 1].

The organization of the rest of the paper is as follows: In Section 2 we formulate the problem and present the background theory which we employ to characterize the optimal stopping time. In Section 3 we present two different cases for fault (or change) detection at a node. In the first case the observed process is the arrival rate to the queue and in the second case it is the buffer occupancy at the queue at arrival epochs. Simulation results for both the cases are presented. In Section 4 we propose a robust algorithm for combined fault detection and parameter estimation for scenarios where we do not know the value of the parameter before or after the change. We provide simulation results for our proposed robust detection algorithm. We finally discuss some related issues and give some further directions in Section 5 .

2 Model: *A Markovian framework*

Let $\{X_t\}$ be the process under observation (for e.g., buffer occupancy or the aggregate arrival rate) and X_n be the observed value at the nth observation. We assume that the observed series $\{X_n\}$ forms a Markov chain. As we shall see later in Section 3, the Markovian assumption can capture the dynamics of the observed process. We next proceed to formally define our problem.

Assume that on a probability space $(\Omega, \mathcal{F}, P^{\pi,x})$ we are given random variables θ with values in

$[0,\infty)$, and random variables $\{X_n, n = 0,1,2,\ldots\}$ with values in D. With regards to our model X_n is the observed series. We assume that $P^{\pi,x}(X_0 = x) = 1$. Also define

$$\begin{aligned} P^{\pi,x}\{\theta = 0\} &= \pi, \\ P^{\pi,x}\{\theta = i\} &= (1-\pi)(1-p)^{i-1}p,\ i = 1,2,\ldots \end{aligned} \tag{1}$$

where π,p and $x(\in \mathcal{H})$ are known constants, $0 < p \leq 1$, $0 \leq \pi \leq 1$.

For each set $A = \{\omega : X_1 = x_1, \ldots, X_n = x_n\}$ we have:

$$\begin{aligned} P^{\pi,x}(A) &= \pi P^1(X_1 = x_1|X_0 = x)P^1(X_2 = x_2|X_1 = x_1)\ldots P^1(X_n = x_n|X_{n-1} = x_{n-1}) \\ &+(1-\pi)\sum_{i=1}^{n}(1-p)^{i-1}pP^0(X_1 = x_1|X_0 = x)P^0(X_2 = x_2|X_1 = x_1)\ldots \\ &P^0(X_i = x_i|X_{i-1} = x_{i-1}) \times P^1(X_{i+1} = x_{i+1}|X_i = x_i)\ldots P^1(X_n = x_n|X_{n-1} = x_{n-1}) \\ &+(1-\pi)(1-p)^n P^0(X_1 = x_1|X_0 = x)P^0(X_2 = x_2|X_1 = x_1)\ldots \\ &\ldots P^0(X_n = x_n|X_{n-1} = x_{n-1}) \end{aligned} \tag{2}$$

where P^1 and P^0 are probability measures on $(\Omega, \mathcal{F}^X)$, $\mathcal{F}^X = \sigma\{\omega : X_1, X_2, \ldots\}$, independent of π. Let for $i = 0,1$, M^i be a Markov chain on the probability space $(\Omega, \mathcal{F}^X, P^i)$.

The conditions given by (1) and (2) means that if $\theta = 0$, we observe a sequence of Markov random variables $X_1, X_2, \ldots$ with joint probability $P^1(X_1 = x_1|X_0)P^1(X_2 = x_2|X_1 = x_1)\ldots P^1(X_n = x_n|X_{n-1} = x_{n-1})$ (as $P^{\pi,x}(X_0 = x) = 1$). If $\theta = i$, $X_1, \ldots, X_{i-1}, X_i, \ldots$ are again Markov random variables, with joint probability $P^0(X_1 = x_1|X_0 = x)P^0(X_2 = x_2|X_1 = x_1)\ldots P^0(X_i = x_i|X_{i-1} = x_{i-1})P^1(X_{i+1} = x_{i+1}|X_i = x_i)\ldots P^1(X_n = x_n|X_{n-1} = x_{n-1})$, i.e., till observation X_i the state transitions are governed by Markov chain M^0 and from X_i onwards the state transitions are governed by M^1, where M^1 is the disturbed chain. Thus $\theta = \theta(\omega)$ is the instant of change (resulting in a faulty/abnormal behavior of the system).

Let τ be a stopping time with respect to the system of σ-algebras $F^X = \{\mathcal{F}_n^X\}, n \geq 0)$ where $\mathcal{F}_0^X = \{\Phi, \Omega\}$ and $\mathcal{F}_n^X = \sigma\{\omega : X_1, \ldots, X_n\}$. For our problem τ can be interpreted as the time at which the *alarm* is sounded to signal the change in the distribution based on an observed process which is represented by the σ-algebra. It is clearly desirable to choose τ as close as possible to the time θ (alternatively to have the least *detection delay*). Also we need to avoid *false alarms.* We note that there is a tradeoff between least detection delay and avoiding false alarms. Thus as the variable characterizing the risk associated with τ we consider ($c > 0$) a *risk function* given by:

$$\rho^{\pi,x}(\tau) = P^{\pi,x}\{\tau < \theta\} + cE^{\pi,x}max\{\tau - \theta, 0\} \tag{3}$$

where $P^{\pi,x}\{\tau < \theta\}$ can be interpreted as the probability of false alarm and $E^{\pi,x}max\{\tau - \theta, 0\}$ as the average delay of detecting the occurrence of disruption correctly (detection delay), i.e., when $\tau \geq \theta$ [1] and c is a control parameter. We next define a (π, x)-Bayes time as:

Definition 1 *For a given $\pi \in [0,1]$ and $x \in D$ we call the stopping time τ_π^* a $(\pi, x) -$ Bayes time if*

$$\rho^{\pi,x}(\tau_\pi^*) = \inf \rho^{\pi,x}(\tau),$$

where inf is taken over the class of all stopping times $\tau \in \mathcal{M}[F^X]$(with respect to the system F^X).

[1] For a particular realization $\tau(\omega)$ is the index of the observed epoch at which the fault is detected and hence an "alarm" is sounded and $\theta(\omega)$ is the true epoch at which fault occurs

We next proceed to obtain an explicit expression for evaluating τ^*_π. The problem formulation and the solution approach is inspired from [1]. The difference being that in [1] Sec. 4.3, the observation process $\{X_n\}$ was a series of mutually independent random variables whereas in our problem it is a series of Markov random variables. Thus the characterization of optimal stopping times is more technical in our case.

Theorem 1 *Let $c > 0$, $p > 0$, and let*

$$\pi_n^{\pi,x} = P^{\pi,x}\{\theta \leq n | \mathcal{F}_n^X\}$$

be the posteriori probability of disruption occurring before time n; $\pi_0^{\pi,x} = \pi$. Then the time

$$\tau^*_{\pi,x} \quad = \quad \inf\{n \geq 0 : \pi_n^{\pi,x} \geq A^*\}$$

where A^*, a constant, is a (π, x)-Bayes *time.*

Proof: The proof is along the same lines as that in [1] and is deferred to the Appendix.

For the fixed false alarm formulation, let $\pi \in [0,1), p \in (0,1]$. We shall denote by $\mathcal{M}^X(\alpha; \pi, x)$ the class of stopping times $\tau \in \mathcal{M}[F^X]$ for which

$$P^{\pi,x}\{\tau < \theta\} \leq \alpha$$

It can be shown that the optimal stopping time from the class $\mathcal{M}^X(\alpha; \pi, x)$ can be estimated by $\tilde{\tau}$:

$$\tilde{\tau} \approx \inf\{n \geq 0 : \pi_n^{\pi,x} \geq 1 - \alpha\}.$$

Remark 1 *Observe that in (1) we are taking geometrical distribution for θ which is a memoryless distribution having infinite support. This makes θ totally unpredictable apriori. Also choice of π and p will determine the performance of the algorithms but we have seen by simulations that the algorithms are robust to the values of p and π.*

3 The Observed Process

We consider two different scenarios for the observed process: (i) when the observed process is the arrival rate to a queue; (ii) when the observed process is the buffer occupancy at a queue.

3.1 Case 1: Monitoring the Aggregate Arrival Rate to a Queue

We consider a discrete time framework. Consider a simple case where the arrival rate, $\{X_n\}$ to the queue is modeled as a two state Markov chain, with states 0 and 1. When the state is 0 (1) then the arrival rate is R_1 (R_2). The disturbance occurs at time N. For $n = 1 \ldots N$, $\{X_n\}$ is governed by a transition matrix M^0 and for $n = N+1 \ldots$ transitions are governed by a different (disturbed) transition matrix M^1. We next apply the results developed in the previous section for optimal detection of the disturbance epoch in this case and provide some simulation results.

We work with $\pi = 0.1$, $\alpha = 0.001$, $x = 0$ and $p = 0.3$. Thus we are looking for optimal stopping time in that class of stopping times where the probability of false alarm is less than 0.001. We shall next consider three example with different structures of the disturbed Markov chain M^1.

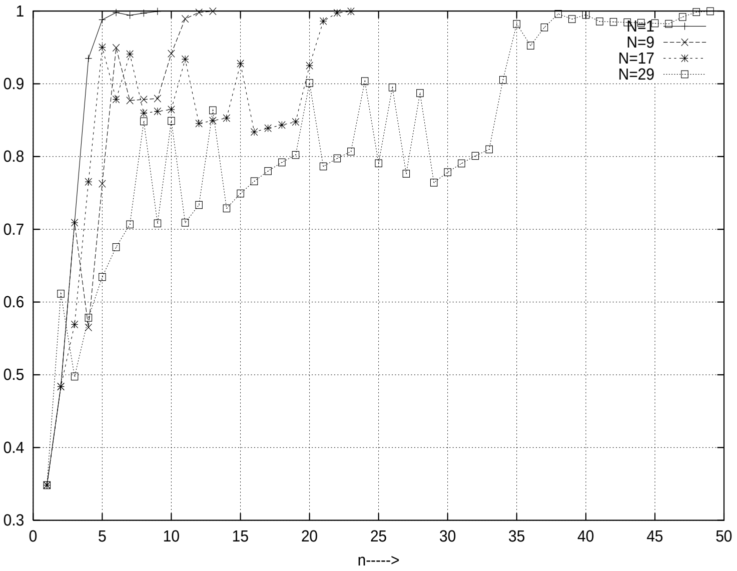

Figure 1: The posterior probability $\pi_n^{\pi,x}$ as a function of n for different N, with $\alpha = 0.001$, $p = 0.3$ and $\pi = 0.1$.

Example 1:

$$M^0 = \begin{pmatrix} 0.6 & 0.4 \\ 0.8 & 0.2 \end{pmatrix} \quad M^1 = \begin{pmatrix} 0.4 & 0.6 \\ 0.2 & 0.8 \end{pmatrix}$$

Observe that this example corresponds to the interchange of columns in the modulating chain after disturbance. We plot the posterior probabilities $\pi_n^{\pi,x}$ as function of n and for different N in Figure 1.

Example 2:

$$M^0 = \begin{pmatrix} 0.6 & 0.4 \\ 0.8 & 0.2 \end{pmatrix} \quad M^1 = \begin{pmatrix} 0.001 & 0.999 \\ 0.01 & 0.99 \end{pmatrix}$$

This example captures the case when the Markov chain is in state 1 most of the times after disturbance. If $R_2 >> R_1$, then this means a sudden flood of packets enter the queue after the fault. We plot the posterior probabilities $\pi_n^{\pi,x}$ as function of n and for different N in Figure 2. The convergence is still very fast after N.

Example 3:

$$M^0 = \begin{pmatrix} 0.6 & 0.4 \\ 0.8 & 0.2 \end{pmatrix} \quad M^1 = \begin{pmatrix} 0.69 & 0.31 \\ 0.8 & 0.2 \end{pmatrix}$$

In this example we aim to detect a very little change in the transition probabilities. We plot the posterior probabilities $\pi_n^{\pi,x}$ as function of n, for different Ns in Figure 3. We see false alarms for $N = 29$, in the curve. This underlines the fact that in the absence of any actual fault or when there is

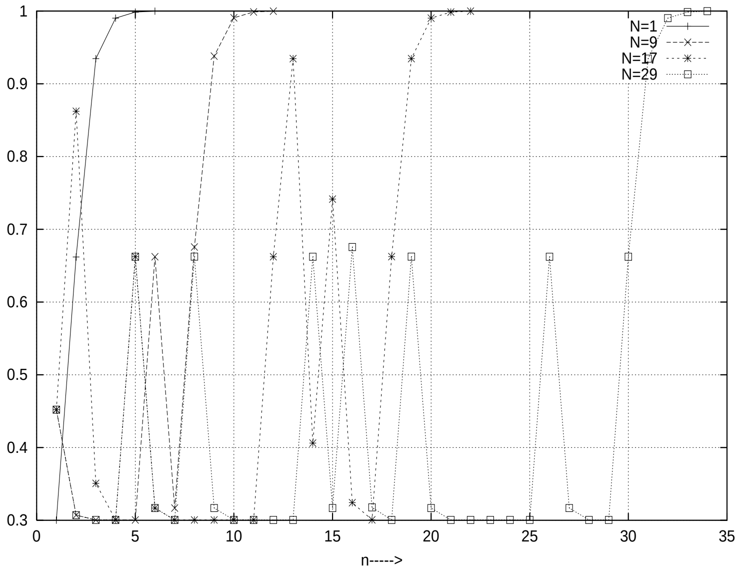

Figure 2: The posterior probability $\pi_n^{\pi,x}$ as a function of n for different N, with $\alpha = 0.001$, $p = 0.3$ and $\pi = 0.1$.

a long gap between the epoch the detection algorithm starts and the actual change occurs, a Bayesian approach like ours may give a false alarm (as for the change epoch $N = 29$, the algorithm was started at $n = 1$). This is because we assumed a geometrical distribution for priors (see (1)) which works fine if our algorithm is started at or close to the actual change epoch. However in the absence of any prior information on the change epoch one can work with a uniform distribution for priors.

3.2 Case 2: Monitoring the Buffer Occupancy in a Queue

In the previous case the observed process was the aggregate arrival rate at the queue. In this section we consider a different $\sigma - field$ of observations, namely the buffer occupancy (queue length) at arrival epochs of packets to the queue. We consider the case where we are monitoring the buffer occupancy in a $GI/GI/1$ [2]. Consider a discrete queueing system with A_{n+1} being the interarrival time between the nth and $n+1$th packet, S_n being the size of the nth packet and c_n being the server capacity during A_{n+1}. Let W_n be the buffer occupancy just before the nth arrival. We have:

$$W_{n+1} \quad = \quad (W_n + S_n - c_n A_n)^+$$

Observe that if we assume that $\{S_n\}$ and $\{A_n\}$ are i.i.d., then W_{n+1} is a Markov chain given c_n. Let before fault $c_n = c_0$ and after fault occurrence c_n changes to c_1. In this scenario we shall employ optimal stopping algorithms to detect the change in the server capacity by monitoring the buffer occupancy at each packet arrival. Let $\{A_n\}$ be i.i.d. with distribution $F_A(x)$ and $\{S_n\}$ be i.i.d. (independent from $\{A_n\}$) with distribution $F_S(x)$. We take both A_n and S_n to be discrete random variables.

[2] general and identical distribution for packet interarrival times, general and identical distribution for packet service times and a single server queue

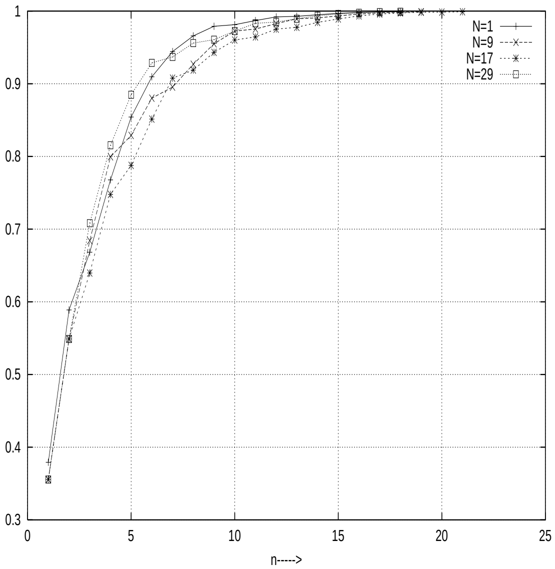

Figure 3: The posterior probability $\pi_n^{\pi,x}$ as a function of n for different N, with $\alpha = 0.001$, $p = 0.3$ and $\pi = 0.1$.

Remark 2 *Observe that the framework assumes that the changes in c_n occurs at packet arrival epochs. It is true that the server capacity may change during the interarrival times. This is a modeling assumption and it can be shown that this should not cause a significant difference in the characterization of optimal stopping times.*

Let us assume that we are given π, p and the initial value of the workload, w. Our $\sigma-$filed of observations is $\{W_n\}$. We assume that the capacity c_n remains constant for A_{n+1}. Let θ be the instant of change [3] in the value of c_n. Thus before θ, $c_n = c_0$ and after θ, $c_n = c_1$. We next write the transition probability of the Markov chain modulating $\{W_n\}$, before and after the change. We have:

$$P^{c_n}(W_{n+1} = i | W_n = j) \quad = \quad \begin{cases} P(j + S_n - c_n A_n = i) & \text{if } i > 0 \\ P(j + S_n - c_n A_n \leq 0) & \text{if } i = 0 \end{cases}$$

We can express the expression for $P^{c_n}(W_{n+1} = i | W_n = j)$ in term of the distribution functions of $\{A_n\}$ and $\{S_n\}$. Thus

$$P^{c_n}(W_{n+1} = i | W_n = j) = \begin{cases} \sum_{a=1}^{\infty} P(S_n = i - j + c_n a) F_A(a) = \sum_{a=1}^{\infty} F_S(i - j + c_n a) F_A(a) & \text{if } i > 0 \\ \sum_{a=1}^{\infty} P(S_n \leq c_n a - j) F_A(a) = \sum_{a=1}^{\infty} \sum_{s=1}^{c_n a - j} F_S(s) F_A(a) & \text{if } i \leq 0 \end{cases} \qquad (4)$$

Defining π_n as in (5) we can write by Bayes formulation,

$$\pi_{n+1}^{\pi,w} \quad = \quad \frac{\pi_n^{\pi,w} P^{c_1}(W_{n+1}|W_n) + (1 - \pi_n^{\pi,w}) p P^{c_0}(W_{n+1}|W_n)}{\pi_n^{\pi,w} P^{c_1}(W_{n+1}|W_n) + (1 - \pi_n^{\pi,w}) p P^{c_0}(W_{n+1}|W_n) + (1 - \pi_n^{\pi,w})(1 - p) P^{c_0}(W_{n+1}|W_n)}$$

It can be established that $\{T_n^{\pi,x} = (W_n, \pi_n^{\pi,w})\}$ is a system of transitive statistics (by Lemma 1 in Appendix) and $\tau_{\pi,w}^* = \inf\{n \geq 0 : \pi_n^{\pi,w} \geq A^*\}$ is an optimal stopping time. We next present simulation results for this scenario.

[3] θ is the index of the arriving packet at which change in c_n occurs

We next demonstrate the working of the proposed algorithm for the case of a $Geo/Geo/1$ queueing system with $F_A(.)$ and $F_S(.)$ being geometrically distributed with parameters λ and μ respectively. Thus:

$$\begin{aligned} F_A(a) = P(A = a) = \lambda(1-\lambda)^{a-1} &\quad \text{for } a = 1, 2, 3, 4, \ldots \\ F_s(s) = P(S = s) = \mu(1-\mu)^{s-1} &\quad \text{for } s = 1, 2, 3, 4, \ldots \end{aligned}$$

Then we can write from (4)

$$P^{c_n}(W_{n+1} = i | W_n = j) \quad = \quad \begin{cases} \lambda\mu \sum_{a=\max\left(\lceil\frac{(1-i+j)}{c_n}\rceil,1\right)}^{\infty} (1-\mu)^{i-j+c_n a-1}(1-\lambda)^{a-1} & \text{if } i > 0 \\ \lambda\mu \sum_{a=\max\left(\lceil\frac{1+j}{c_n}\rceil,1\right)}^{\infty} \sum_{s=1}^{c_n a-j} (1-\mu)^{s-1}(1-\lambda)^{a-1} & \text{if } i \leq 0 \end{cases}$$

Which can be explicitly written as (with $k_1 = \max\left(\lceil\frac{(1-i+j)}{c_n}\rceil, 1\right)$ and $k_2 = \max\left(\lceil\frac{1+j}{c_n}\rceil, 1\right)$):

$$P^{c_n}(W_{n+1} = i | W_n = j) \quad = \quad \begin{cases} \frac{\lambda\mu(1-\mu)^{i-j+c_n k_1 - 1}(1-\lambda)^{k_1 - 1}}{1-(1-\mu)^{c_n}(1-\lambda)} & \text{if } i > 0 \\ (1-\lambda)^{k-1}\left(1 - \frac{\lambda(1-\mu)^{c_n k_2 - j}}{1-(1-\mu)^{c_n}(1-\lambda)}\right) & \text{if } i \leq 0 \end{cases}$$

For the simulations we took $c_0 = 2$, $c_1 = 1$, $\lambda = 0.2$, $\mu = 0.3$, $p = 0.1$, $\Pi = 0.1$, $\alpha = 0.0001$ and $w = 1000$. We plot the posteriors $\pi_n^{\pi,w}$ as a function of n for different N (the true change epoch) in Fig. 4. We observe that for $N = 100, 300$ and 500, the detection algorithm stops at $n = 225, 365$ and $n = 639$ respectively. Next keeping all other parameters constant we reduce $\alpha = 0.01$ and plotted the curve Fig 5. We observe that for $N = 100, 300$ and 500, the detection algorithm stops at $n = 142, 428$ and $n = 326$ respectively. Thus on increasing the error probability we get a *false* alarm for the case of $N = 500$ portraying the tradeoff between detection delay and probability of false alarms.

Remark 3 *In the two example scenarios studied in this section we looked at the traffic anomalies at a single queue in the network. This can be justified because attacks are known to be localized (creating* hotspots*). Also the algorithms can be independently implemented at different network nodes.*

4 Optimal Stopping Rules for Robust Fault Detection and Parameter Estimation

In the problems discussed so far we have looked at cases where we know the actual value of the changed parameter before and after the change. We now consider the case where we do not know the value of the parameter, call β (the change in the value of which is a *fault*) either before or after the change. However we do know that before the change the parameter belongs to a certain set say B_1 and after the change to a set B_2. The observed process (or the true world) is from a system for which β has a value from B_1 before change and takes a value from B_2 after change. For each set of values $\{i, j\}$, $i \in B_1, j \in B_2$ we assume that the set of modulating Markov chains (M^i, M^j) is known. The observed process is the random variable $\{X_n\}$ and is a Markov chain. We assume that for each of the pairs $(i, j : i \in B_1, j \in B_2)$ we have the prior probability $P_{i,j}$ and hence a prior probability matrix $\mathbf{P}$. We next formulate our problem and propose optimal stopping rules for combined fault detection and estimation of true values of the parameter β before and after the change.

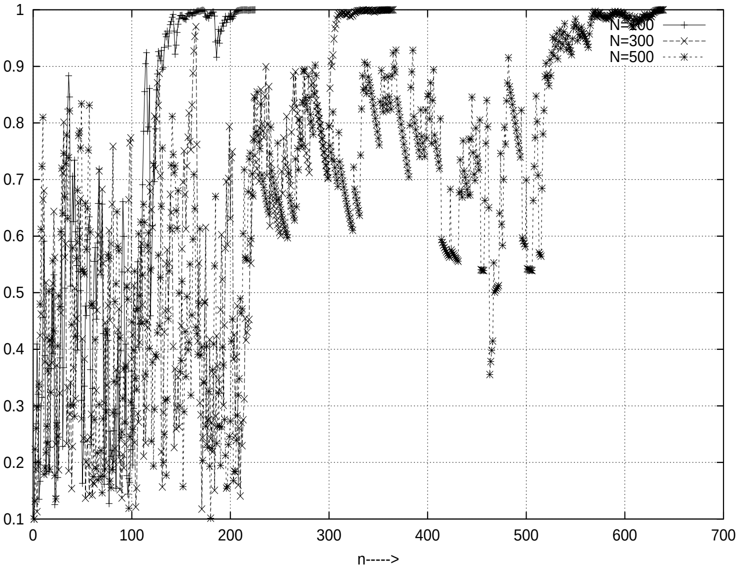

Figure 4: The posterior probability $\pi_n^{\pi,x}$ as a function of n for different N, with $\alpha = 0.0001$, $p = 0.1$ and $\pi = 0.1$, $\lambda = 0.2$, $\mu = 0.3$ and $c_0 = 2$ and $c_1 = 1$.

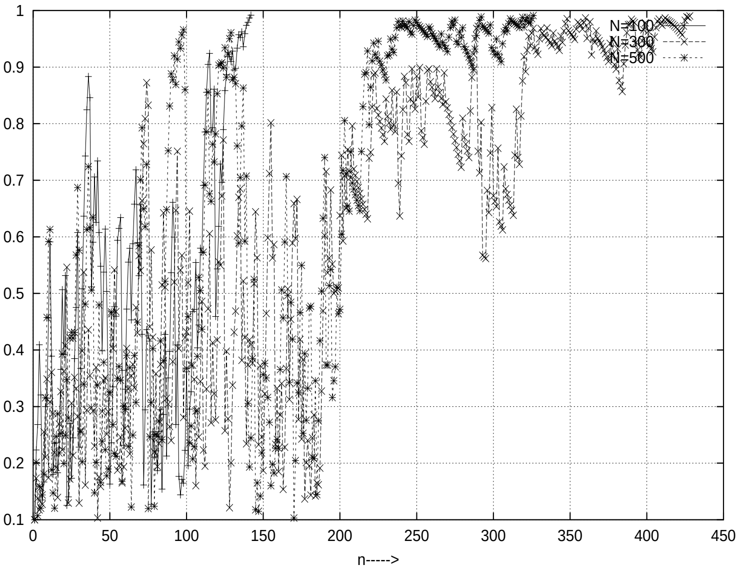

Figure 5: The posterior probability $\pi_n^{\pi,x}$ as a function of n for different N, with $\alpha = 0.01$, $p = 0.1$ and $\pi = 0.1$, $\lambda = 0.2$, $\mu = 0.3$ and $c_0 = 2$ and $c_1 = 1$.

More formally, we shall assume that on a probability space $(\Omega, \mathcal{F}, P^{\pi,x,\beta})$ we are given random variables θ with values in $[0, \infty)$, and random variables $\{X_n, n = 0, 1, 2, \ldots\}$ with values in $\mathcal{X}$. We assume that $P^{\pi,x,\beta}(X_0 = x) = 1$. Also define the joint probability

$$\begin{aligned} P^{\pi,x,\beta}\{\theta = 0, \beta_1 = i, \beta_2 = j\} &= \pi_{ij}, \\ P^{\pi,x,\beta}\{\theta = v, \beta_1 = i, \beta_2 = j\} &= (\beta_{ij} - \pi_{ij})(1 - p_{ij})^{v-1} p_{ij}, \ v = 1, 2, \ldots \end{aligned}$$

and $P^{\pi,x}\{\beta_1 = i, \beta_2 = j\} = \beta_{ij}$, where $\pi = (\pi_{ij})$, $\beta = (\beta_{ij})$, p_{ij}, $i \in B_1$, $j \in B_2$, p and $x (\in \mathcal{H} \subset \mathcal{R})$ are known constants, $0 < p_{ij} \leq 1$, $0 \leq \beta_{ij} \leq 1$, $\sum_{i \in B_1, j \in B_2} \beta_{ij} = 1$, $0 \leq \pi \leq 1$. For each set $A = \{\omega : X_1 = x_1, \ldots, X_n = x_n\}$. Also,

$$\begin{aligned} & P^{\pi,x,\beta}(A, \beta_1 = i, \beta_2 = j) = \\ & \quad \pi_{ij} P^j(X_1 = x_1 | X_0 = x) P^j(X_2 = x_2 | X_1 = x_1) \ldots P^j(X_n = x_n | X_{n-1} = x_{n-1}) \\ & \quad + (\beta_{ij} - \pi_{ij}) \sum_{k=1}^{n} (1 - p_{ij})^{k-1} p_{ij} P^i(X_1 = x_1 | X_0 = x) \ldots P^i(X_k = x_k | X_{k-1} = x_{k-1}) \\ & \quad \times P^j(X_{k+1} = x_{k+1} | X_k = x_k) \ldots P^j(X_n = x_n | X_{n-1} = x_{n-1}) \\ & \quad + (\beta_{ij} - \pi_{ij})(1 - p_{ij})^n P^i(X_1 = x_1 | X_0 = x) \ldots P^i(X_n = x_n | X_{n-1} = x_{n-1}) \end{aligned}$$

where P^i, $i \in B_1$ and P^0, $j \in B_2$ are probability measures on $(\Omega, \mathcal{F}^X)$, $\mathcal{F}^X = \sigma\{\omega : X_1, X_2, \ldots\}$, independent of π and x.
Thus, if $\theta = 0$ and $\beta_1 = i, \beta_2 = j$, we observe a sequence of Markov random variable $X_1, X_2, \ldots$ with joint probability $P^j(X_1 = x_1 | X_0) P^j(X_2 = x_2 | X_1 = x_1) \ldots P^j(X_n = x_n | X_{n-1} = x_{n-1})$(as $P^{\pi,x}(X_0 = x) = 1$). If $\theta = k$, $X_1, \ldots, X_{k-1}, X_k, \ldots$ are again Markov random variables, with joint probability $P^i(X_1 = x_1 | X_0 = x) P^i(X_2 = x_2 | X_1 = x_1) \ldots P^j(X_k = x_k | X_{k-1} = x_{k-1}) P^j(X_{k+1} = x_{k+1} | X_k = x_k) \ldots P^j(X_n = x_n | X_{n-1} = x_{n-1})$, i.e., until observation X_{k-1} the state transitions are governed by chain M^i and from X_k onwards the state transitions are governed by M^j, where M^j is the disturbed chain. We work here with scalar change, i.e., change in the value of a single parameter of the chain. The more complex case of multidimensional changes (i.e., changes in 2 or more parameters simultaneously) will be dealt in our future work. Thus θ is the instant of change.

Let for $v \geq 0$, $i \in B_1, j \in B_2$, $\pi_n^{\pi,x,\beta}(i,j) = P^{\pi,x,\beta}\{\theta \leq n, \beta_1 = i, \beta_2 = j | \mathcal{F}_n^X\}$, be the posteriori probability of disruption occurring before time n; $\pi_0^{\pi,x,\beta}(i,j) = \pi_{i,j}$ and $\beta_1 = i, \beta_2 = j$. Then by the Bayes formula, $\forall i \in B_1, j \in B_2$, we can write for $n = 1$:

$$\begin{aligned} & \pi_1^{\pi,x,\beta}(i,j) \\ & = \frac{\pi_0^{\pi,x,\beta}(i,j) P^j(X_1/X_0 = x) + (\beta_{ij} - \pi_0^{\pi,x,\beta}(i,j)) p_{ij} P^i(X_1/X_0)}{\sum_{i \in B_1, j \in B_2} \pi_0^{\pi,x,\beta}(i,j) P^j(X_1/X_0) + \sum_{i \in B_1, j \in B_2} (\beta_{ij} - \pi_0^{\pi,x,\beta}(i,j)) P^i(X_1/X_0)} \end{aligned}$$

Also define the posteriors, $\Gamma_n^{\pi,x,\beta}(i,j) = P^{\pi,x,\beta}(\beta_1 = i, \beta_2 = j | \mathcal{F}_n^X)$ with $\Gamma_0^{\pi,x,\beta}(i,j) = \beta_{ij}$. Observe that $\Gamma_n^{\pi,x,\beta}(i,j) = \pi_n^{\pi,x,\beta}(i,j) + \bar{\pi}_n^{\pi,x,\beta}(i,j)$, where $\bar{\pi}_n^{\pi,x,\beta}(i,j) = P^{\pi,x,\beta}(\theta > n, \beta_1 = i, \beta_2 = j | \mathcal{F}_n^X)$ We have

$$\bar{\pi}_1^{\pi,a,\beta} = \frac{(\beta_{ij} - \pi_0^{\pi,x,\beta}(i,j))(1 - p_{ij}) P^i(X_1/X_0)}{\sum_{i \in B_1, j \in B_2} \pi_0^{\pi,x,\beta}(i,j) P^j(X_1/X_0) + \sum_{i \in B_1, j \in B_2} (\beta_{ij} - \pi_0^{\pi,x,\beta}(i,j)) P^i(X_1/X_0)}$$

and thus

$$\Gamma_1^{\pi,x,\beta}(i,j) = \frac{\pi_0^{\pi,x,\beta}(i,j) P^j(X_1/X_0 = x) + (\beta_{ij} - \pi_0^{\pi,x,\beta}(i,j)) P^i(X_1/X_0)}{\sum_{i \in B_1, j \in B_2} \pi_0^{\pi,x,\beta}(i,j) P^j(X_1/X_0) + \sum_{i \in B_1, j \in B_2} (\beta_{ij} - \pi_0^{\pi,x,\beta}(i,j)) P^i(X_1/X_0)}$$

Below we enumerate the steps for sequential updation of posteriors $\pi_n^{\pi,x,\beta}(i,j)$ and $\Gamma_n^{\pi,x,\beta}(i,j)$.

4.1 Steps for Sequential Updating of Posteriors

1. Take $\pi_0^{\pi,x,\beta}(i,j) = \pi_{ij}$ and $\Gamma_0^{\pi,x,\beta}(i,j) = \beta_{ij}$, $\forall i \in B_1, j \in B_2$.
2. At the $n+1$th observation, X_{n+1}, update the posteriors $\pi_{n+1}^{\pi,x,\beta}(i,j)$ and $\Gamma_{n+1}^{\pi,x,\beta}(i,j)$ as follows

$$\begin{aligned}
\pi_{n+1}^{\pi,x,\beta}(i,j) &= \frac{\pi_n^{\pi,x,\beta}(i,j)P^j(X_{n+1}/X_n) + (\Gamma_n^{\pi,x,\beta}(i,j) - \pi_n^{\pi,x,\beta}(i,j))p_{ij}P^i(X_{n+1}/X_n)}{\sum_{i\in B_1, j\in B_2}\pi_n^{\pi,x,\beta}(i,j)P^j(X_1/X_0) + \sum_{i\in B_1, j\in B_2}(\Gamma_n^{\pi,x,\beta}(i,j) - \pi_n^{\pi,x,\beta}(i,j))P^i(X_1/X_0)} \\
\Gamma_{n+1}^{\pi,x,\beta}(i,j) &= \frac{\pi_n^{\pi,x,\beta}(i,j)P^j(X_{n+1}/X_n) + (\Gamma_n^{\pi,x,\beta}(i,j) - \pi_n^{\pi,x,\beta}(i,j))P^i(X_{n+1}/X_n)}{\sum_{i\in B_1, j\in B_2}\pi_n^{\pi,x,\beta}(i,j)P^j(X_1/X_0) + \sum_{i\in B_1, j\in B_2}(\Gamma_n^{\pi,x,\beta}(i,j) - \pi_n^{\pi,x,\beta}(i,j))P^i(X_1/X_0)}
\end{aligned}$$

4.2 Robust Algorithm

Since we do not know the true value of parameters before and after the change we propose the following optimal stopping algorithm. Let (i_n^*, j_n^*) be the estimated value of the parameters (β_1, β_2) (the true world) and π_n^* be the value of the posterior $P^{\pi,x,\beta}(\theta \leq n, \beta_1 = i^*, \beta_2 = j^* | \mathcal{F}_n^X)$ at the nth observation.
Algorithm

1. At $n = 0$, let i_0^*, j_0^* be $(i_0^*, j_0^*) = \text{Argmax}_{i\in B_1, j\in B_2}\pi_{ij}$ and $\pi_0^* = \pi_{i_0^*, j_0^*}$.
2. At the $n+1$th observation update $\pi_{n+1}^{\pi,x,\beta}(i,j)$ and $\Gamma_{n+1}^{\pi,x,\beta}(i,j)$ by the procedure listed in Section (4.1). Also update (i_{n+1}^*, j_{n+1}^*) as $(i_{n+1}^*, j_{n+1}^*) = \text{Argmax}_{i\in B_1, j\in B_2}\pi_{n+1}^{\pi,x,\beta}(i,j)$ and $\pi_{n+1}^* = \pi_{n+1}^{\pi,x,\beta}(i_{n+1}^*, j_{n+1}^*)$.
3. If $\pi_{n+1}^* \geq 1 - \alpha$, stop.

Thus at the $n+1$th observation, the estimated values of parameters β_1 and β_2 is that set that maximize the combined probability $P^{\pi,x,\beta}(\theta \leq n+1, \beta_1 = i, \beta_2 = j | \mathcal{F}_{n+1})$ and the value of π_{n+1}^* is the value of $\pi_{n+1}^{\pi,x,\beta}$ corresponding to this set.

Remark 4 *A formal proof of the optimality of the proposed robust algorithm shall be done based on the theory of optimal stopping in our immediate future work.*

4.3 Simulation Results

We next study the efficiency of our proposed algorithm for combined fault detection and parameter estimation through the first simulation scenario that we studied in Sec. (3). For Case I in Sec. (3) where we observe the arrival rate to a queue we consider a case where $B_1 = (\beta_{a1}, \beta_{a2})$ and $B_2 = (\beta_{b1}, \beta_{b2})$. The arrival rate process $\{X_n\}$ is a two state (states 1 and 0) Markov chain parameterized by some $\beta_1 \in B_1$ before fault and some $\beta_2 \in B_2$ after fault. We assume that we know the Markov chains modulating the arrival process for all the pair of parameter values $\{i, j; i \in B_1, j \in B_2\}$. For the simulations we take

$$M^{\beta_{a1}} = \begin{pmatrix} 0.2 & 0.8 \\ 0.9 & 0.1 \end{pmatrix}, \quad M^{\beta_{a2}} = \begin{pmatrix} 0.10 & 0.90 \\ 0.01 & 0.99 \end{pmatrix},$$

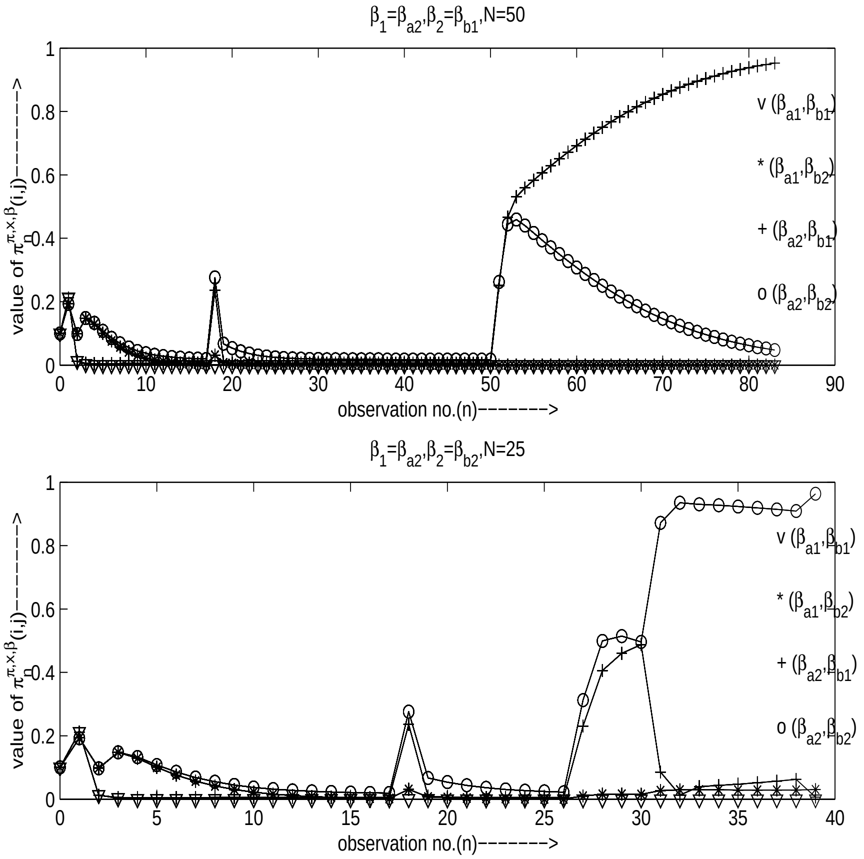

Figure 6: The posterior probability $\pi_n^{\pi,x,\beta}(i,j)$ as a function of n for different i,j, $i \in B_1, j \in B_2$ with (i)$N = 50$, $\beta_a = \beta_{a2}, \beta_b = \beta_{b1}$, (ii) $N = 25$, $\beta_a = \beta_{a2}, \beta_b = \beta_{b2}$.

$$M^{\beta_{b1}} = \begin{pmatrix} 0.99 & 0.01 \\ 0.90 & 0.10 \end{pmatrix}, \quad M^{\beta_{b2}} = \begin{pmatrix} 0.9 & 0.1 \\ 0.3 & 0.7 \end{pmatrix},$$

$\pi_{ij} = 0.05, \beta_{ij} = 0.25, \pi_{ij} = 0.01 \forall i,j$, $\alpha = 0.001$ and $x = 0$. Let N be the epoch at which fault occurs. We next present the results using our proposed algorithm for optimal fault detection and parameter estimation for two different examples.

- **Case 1**: $\beta_1 = \beta_{a2}$, $\beta_2 = \beta_{b1}$, $N = 50$.
  Fig. (6). The algorithm stops at iterate 84 with estimated values of β_1 and β_2 as β_{a2} and β_{b1}.
- **Case 2**: $\beta_1 = \beta_{a2}$, $\beta_2 = \beta_{b2}$, $N = 25$.
  Fig. (6). The algorithm stops at iterate 40 with estimated values of β_1 and β_2 as β_{a2} and β_{b2} respectively.

5 Perspectives and Future Work

In real situations, our proposed detection/detection-estimation algorithms can be implemented at network nodes. Since the posteriors are updated using Bayes formula if we run these algorithms for sufficiently long time then there will be an alarm even when there is no actual change, thus giving a false alarm. This is because as $n \to \infty$, $\pi_n^{\pi,x} \to 1$ which is easy to see. Thus, in practice the algorithms should be run for some fixed (or variable) time, say W and then restarted with the value of $\pi_n^{\pi,x}$ corresponding to the initial priors. Also the efficiency of Bayesian method depends on the value of priors. However, in the absence of any information about priors one can even take uniform distribution for θ on W which corresponds to the maximum ignorance case. It is however known that the Bayes procedures are not very sensitive to priors if there are drastic changes. It is in this context that we are naturally interested as for example during denial of service attacks. The efficiency of the detection algorithms depends on how drastic the changes are. This situation can be taken into account within the robust framework presented earlier. The prior and post change sets need to be chosen accordingly. Such algorithms can also be distributed across the network where only partial observations are available (i.e. we cannot measure the occupancies directly but only through secondary effects such as increase in delays in receiving acknowledgements). The key is to be able to compute the change statistic with respect to the observed σ-field or history.

We emphasize that the goal of this paper is to explore the potentials of stopping theory based framework for formulating algorithms for optimal detection of traffic anomalies in communication networks. Our preliminary simulations of example scenarios highlight the potentials of this optimal stopping theoretic framework. Our immediate challenge now is to demonstarte the efficiency of our proposed framework by doing experiments with live data with actual traffic overloads by comparing the performance of our approach with traditional statistical measures involving mean and variance changes (e.g., the Cumulative SUM (CUSUM) tests). We would like to remark that the threshold in our tests depend on apriori given distribution. But in CUSUM the threshold depends on observed statistics . If the observations are not Gaussian (as ours) then the calculated statistics is not mean and variance but is $\Pi^{\pi,x}$. Of course, have the observations been Gaussian, $\Pi^{\pi,x}$ can be related to mean and variance.

In future work we will address how to use such schemes with other information not necessarily directly connected with traffic but other indicators such as acknowledgments and delays and to thus come up with an end-to-end anomaly detection framework. Also, in our model and simulations we assumed perfect knowledge of the observed process. In reality the observed $\sigma - field$, for e.g., the arrival rate, the buffer occupancy etc. will be obtained by measurements and thus will be corrupted with noise. We need to extend this framework to scenarios where noisy estimates of observed process is available and to study the efficiency of the proposed algorithms.

Appendix

We will first show that the problem of determining the $(\pi, x)-$Bayes stopping time can be reduced to an optimal stopping problem for a Markov sequence.
By the Bayes formula, ($P^{\pi,x}$ − almost surely($a.s.$)) we write,

$$\pi_{n+1}^{\pi,x} \quad = \quad \frac{\pi_n^{\pi,x} P^1(X_{n+1}|X_n) + (1-\pi_n^{\pi,x})pP^0(X_{n+1}|X_n)}{\pi_n^{\pi,x} P^1(X_{n+1}|X_n) + (1-\pi_n^{\pi,x})pP^0(X_{n+1}|X_n) + (1-\pi_n^{\pi,x})(1-p)P^0(X_{n+1}|X_n)}$$

We thus observe that $\pi_{n+1}^{\pi,x}$ for each n can be written as a function of $\pi_n^{\pi,x}$, X_{n+1} and X_n. For each n we define a vector, $T_n^{\pi,x} = (X_n, \pi_n^{\pi,x})$. We further look at some standard definitions in the optimal stopping theory [4].

Definition 2 *Let $F^X = \{\mathcal{F}_n^X\}$, $n = 0, 1, \ldots$, where $\mathcal{F}_n^X = \sigma\{\omega : X_0, \ldots, X_n\}$. The system of random elements $\eta = (\eta_0, \eta_1, \ldots)$ with values in $(Y, \mathcal{Y})$ is referred to as a system of* transitive statistics(with respect to F^X)*, if:*

1. *η_n is $\mathcal{F}_n^X/\mathcal{Y}$-measurable, $n = 0, 1, \ldots$;*
2. *For each $n = 1, 2, \ldots$ there exists a $\mathcal{Y} \times \mathcal{X}/\mathcal{Y}$-measurable function $\phi_n = \phi_n(y, x)$ such that with probability 1 $\eta_n(\omega) = \phi_n(\eta_{n-1}(\omega), X_n(\omega))$.*

Definition 3 *Let $\eta = (\eta_0, \eta_1, \ldots)$ be the system of transitive statistics(with respect to F^X) with values in $(Y, \mathcal{Y})$. If, for each $n = 0, 1, \ldots$ with probability 1*

$$P\{X_{n+1} \in B | \mathcal{F}_n^X\} \quad = \quad P\{X_{n+1} \in B | \eta_n\}, \; B \in \mathcal{X}, \tag{5}$$

then the elements $(\eta_n, \mathcal{F}_n^X, P)$, $n = 0, 1, \ldots$, form a Markov random function:

$$P\{\eta_{n+1} \in A | \mathcal{F}_n^X\} \quad = \quad P\{\eta_{n+1} \in A | \eta_n\}, \; A \in \mathcal{Y}. \tag{6}$$

Thus we have the following Lemma:

Lemma 1 *$T = (T_0^{\pi,x}, T_1^{\pi,x}, \ldots)$ with values in $(R \times D, \mathcal{B} \times \mathcal{X})$ [5] is a system of transitive statistics and the elements $\mathcal{T}^{\pi,x} = (T_n^{\pi,x}, \mathcal{F}_n^X, P^{\pi,x}), n \geq 0$ forms a Markov random function.*

Proof: $T_{n+1}^{\pi,x}$ is a function of $T_n^{\pi,x}$ and X_{n+1}; $T_n^{\pi,x}$ is $\mathcal{F}_n^X/(\mathcal{Y} \times \mathcal{X}_\sigma)$ measurable. We will now show that indeed the system $\mathcal{T}^{\pi,x} = (T_n^{\pi,x}, \mathcal{F}_n^X, P^{\pi,x}), n \geq 0$ forms a Markov random function (for a given π and x). Using Definition 3 we need only to verify (6) (with η_n replaced by $T_n^{\pi,x}$, $n \geq 0$), whose validity is evident from the following chain of equations:

$$\begin{aligned} P^{\pi,x}\{X_{n+1} \in A | \mathcal{F}_n^X\} &= P^{\pi,x}\{X_{n+1} \in A | \mathcal{F}_n^X, \theta \leq n\}\pi_n^{\pi,x} + P^{\pi,x}\{X_{n+1} \in A | \mathcal{F}_n^X, \theta > n\}(1 - \pi_n^{\pi,x}) \\ &= P^{\pi,x}\{X_{n+1} \in A | X_n, \theta \leq n\}\pi_n^{\pi,x} + P^{\pi,x}\{X_{n+1} \in A | X_n, \theta > n\}(1 - \pi_n^{\pi,x}) \\ &= P^{\pi,x}\{X_{n+1} \in A | X_n, \pi_n^{\pi,x}\} = P^{\pi,x}\{X_{n+1} \in A | T_n^{\pi,x}\} \end{aligned}$$

Thus, the family of Markov random functions $\{\mathcal{T}^{\pi,x}, 0 \leq \pi \leq 1, x \in \mathcal{X}\}$ can be associated with a two-dimensional Markov process with discrete time, $T = (T_n, \mathcal{F}_n, P_{\pi,x}), n \geq 0$, having the same transition probabilities as each Markov random function $\mathcal{T}^{\pi,x}, \pi \in [0,1], x \in \mathcal{X}$. From [1] we can write the risk function $p^{\pi,x}(\tau^*)$ (with E as expectation) as $p^{\pi,x}(\tau) = \inf_{\tau \in \mathcal{M}[F^X]} \rho^{\pi,x}(\tau)$ where,

$$\rho^{\pi,x} = \inf_{\tau \in \mathcal{M}[F^X]} E^{\pi,x} \left\{ (1 - \pi_\tau^{\pi,x}) + c \sum_{k=0}^{\tau-1} \pi_k^{\pi,x} \right\}$$

Thus [see Sec. 2.15 in [1]] to find the $(\pi, x)-$ Bayes time τ_π^* we need only to find the optimal stopping time in the problem

$$\rho(\pi, x) \quad = \quad \inf E_{(\pi,x)} \left[(1 - \pi_\tau) + c \sum_{k=0}^{\tau-1} \pi_k \right], \tag{7}$$

[4] See also [1].

[5] R is the set of real numbers and $\mathcal{B}$ is the $\sigma - algebra$ of Borel subsets of R

where inf is taken over the class of stopping times

$$\mathcal{M}^1(F) = \left\{ \tau \in \mathcal{M}(F) : E_{(\pi,x)} \sum_{k=0}^{\tau-1} \pi_k < \infty,\ \pi \in [0,1] \right\}.$$

Let $g(\pi, x) = (1 - \pi)$ [6], and let $Qg(\pi, x) = \min\{g(\pi, x), c\pi + Kg(\pi, x)\}$, where K is the operator defined as $Kg(\pi, x) = E_{\pi,x} g(T_1)$, where $T_1 = (\pi_1, x_1)$. By Theorem 2.23 [1], it can be shown that the time,$\tau_0 = \inf\{n \geq 0 : \rho(T_n) = g(T_n) = 1 - \pi_n\}$ is an optimal stopping time.

Observe that $g(\pi, x)$ and $Qg(\pi, x)$ are concave and nonincreasing in π. Let $A^* = \max_{x \in D} A^*(x)$ where $A^*(x)$ is defined as

$$A^*(x) = \inf\{a \in [0,1] : c\pi + K\rho(\pi, x) \geq g(\pi, x) \text{for} \pi \geq a\}$$

Then the time τ_0, such that $\tau_0 = \inf\{n \geq 0 : \pi_n \geq A^*\}$ is an optimal stopping time in the problem posed by (7) and the time $\tau^*_{\pi,x} = \inf\{n \geq 0 : \pi_n^{\pi,x} \geq A^*\}$ is (π, x)-Bayes for any $\pi \in [0,1]$ and $x \in \mathcal{H}$ (in this case the threshold is independent of (π, x)) which establishes Theorem 1.

References

[1] A.N.Shiryayev. *Optimal Stopping Rules.* Springer-Verlag, 1978.

[2] U. Appel and A. V. Brandt. Adaptive sequential segmentation of piecewise stationary time series. *Information Sciences*, 29:27–56, 1983.

[3] Y. S. Chow, H. Robbins, and D. Siegmund. *Great Expectations: The Theory of Optimal Stopping.* Houghton Mifflin Company, Boston, 1971.

[4] F. Feather and R. Maxion. Fault detection in an ethernet network using anomaly signature matching. In *ACM Sigcomm*, volume 23, 1993.

[5] I. Katzela and M. Schwarz. Schemes for fault detection in communication networks. *IEEE/ACM Trans. on Networking*, 3:753–764, 1995.

[6] M. Mandjes, I. Saniee, and A. Stolyar. Load characterization and anomaly detection for voice over IP traffic. In *Proceedings of the ACM Sigmetrics 2001*, pages 326–327, 2001.

[7] I. Rouvellou and G. Hart. Automatic alarm correlation for fault identification. In *Proceedings of the IEEE INFOCOM'95*, pages 553–561, 1995.

[8] D. Siegmund. *Sequential Analysis: Tests and confidence intervals.* Springer-Verlag, 1985.

[9] R. R. Talpade, G. Kim, and S. Khurana. NOMAD: Traffic-based network monitoring framework for anomaly detection. In *Proceedings of the 4th. IEEE Symposium on Computers and Communication*, 1998.

[10] M. Thottan and C. Ji. Adaptive thresholding for proactive network problem detection. *IEEE Network: Special Issue on Network Management*, Oct. 1998.

[6] Notice that g depends on $T(= (\pi, x))$ only through its component π