



A New Statistical Approach to Network Anomaly Detection

Christian Callegari¹, Sandrine Vaton², and Michele Pagano¹

¹Dept. of Information Engineering, University of Pisa, ITALY

E-mail: {christian.callegari,m.pagano}@iet.unipi.it

²Dept. of Computer Science, ENST Bretagne, FRANCE

E-mail: sandrine.vaton@enst-bretagne.fr

Abstract—In the last few years, the number and impact of security attacks over the Internet have been continuously increasing. To face this issue, the use of Intrusion Detection Systems (IDSs) has emerged as a key element in network security. In this paper we address the problem considering a novel statistical technique for detecting network anomalies. Our approach is based on the use of different families of Markovian models (namely high order and non homogeneous Markov chains) for modeling network traffic running over TCP. The performance results shown in the paper, justify the proposed method and highlight the improvements over commonly used statistical techniques.

Index Terms—Intrusion Detection System, High Order Markov Chain, Mixture Transition Model, Non-Homogeneous Markov Chain

I. INTRODUCTION

In the last few years Internet has experienced an explosive growth. Along with the wide proliferation of new services, the quantity and impact of attacks have been continuously increasing. The number of computer systems and their vulnerabilities have been rising, while the level of sophistication and knowledge required to carry out an attack have been decreasing, as much technical attack know-how is readily available on Web sites all over the world.

Recent advances in encryption, public key exchange, digital signature, and the development of related standards have set a foundation for network security. However, security on a network goes beyond these issues. Indeed it must include security of computer systems and networks, at all levels, top to bottom.

Since it seems impossible to guarantee complete protection to a system by means of prevention mechanisms (e.g. authentication techniques), the use of an Intrusion Detection System (IDS) is of primary importance to reveal intrusions in a network or in a system. IDSs are usually classified on the basis of several criteria [1].

State of the art in the field of intrusion detection is mostly represented by misuse based IDSs. Considering that most attacks are realized with known tools, available on the Internet, a signature based IDS could seem a good solution. Nevertheless hackers continuously come up with new ideas for the attacks, that a misuse based IDS is not able to block. This is the main reason why our work has focused on the development of an anomaly based IDS. In particular our goal

is to reveal intrusions carried out exploiting TCP bugs, by using Markovian models (high order and non homogeneous Markov chains) to describe the behavior of network traffic.

The use of first order homogeneous Markov chain is a well-known approach to detect two distinct kinds of “anomalies”: masqueraders (analyzing the command stream of a host) and intruders (analyzing the evolution of TCP flows in the network traffic) [2].

Vardi and Ju in [3] describe the use of high order Markov chains to detect masqueraders at the host level, and in [4][5] the authors compare performance of first order models and “generic” high order models.

After an extensive survey, to the best of our knowledge, there is no work directly related neither to the use of high order Markov chains to detect anomalies in the TCP traffic nor to the application of non-homogeneous Markov chains to anomaly detection in general.

Moreover no study at all compares the performance achievable with Markov chains of different orders and with a simple “independent” model.

The paper is structured as follows: next section provides a detailed description of the implemented system, while the subsequent section presents the experimental results. Finally section 4 concludes the paper with some final remarks..

II. SYSTEM DESIGN

In this section we provide a detailed description of the proposed anomaly based NIDS.

The aim of our work is to perform a comparison between several statistical models, which can be used to describe the behavior of TCP connections. More in detail we take into account the use of:

- first order homogeneous Markov chains
- first order non-homogeneous Markov chains
- high order homogeneous Markov chains
- stationary ECDF (Empirical Cumulative Distribution Function)
- non-stationary ECDF

Next subsections describe the training phase and the detection phase of our IDS.

A. Training Phase

To build the model which represents the “normal” behavior of the network, the system needs a training phase during which

it analyzes some network traffic, supposed to be attack free. The system analyzes raw traffic traces in libpcap format, the standard used by publicly available packet sniffer software, as Tcpdump or Ethereal. First of all the IDS performs a filtering phase so that only TCP packets are passed as input to the detection blocks.

The IDS only considers some fields of the packet headers, more precisely the IP source address, the IP destination address, the source port number, the destination port number, and the TCP flags. The IP addresses and the port numbers are used to identify a connection, while the value of the flags is used to build the profile. Experimental results have shown that the stochastic models associated to the different applications strongly differ one from the other. Thus, before constructing the model, the system isolates the different services, on the basis of the server port number, and the following procedure is realized once per each application. After that the IDS reconstructs the single connections on the basis of the 5-tuple (source and destination addresses, source and destination ports, and protocol).

A value s_i is associated to each packet, according to the configuration of the TCP flags:

$$s_i = \text{syn} + 2 \cdot \text{ack} + 4 \cdot \text{psh} + 8 \cdot \text{rst} + 16 \cdot \text{urg} + 32 \cdot \text{fin} \quad (1)$$

Thus each “mono-directional” connection is represented by a sequence of symbols s_i , which are integers in $\{0, 1, \dots, 63\}$.

The training phase, as well as the detection phase, varies according to the stochastic model we are taking into account.

1) *ECDF*: In the case of the stationary ECDF the training phase simply consists of evaluating the probabilities $P(s_i)$ that the TCP flags assume the value s_i , independently of the position of the packet in the TCP connection.

For the non-stationary ECDF the system has to compute the probabilities $P_j(s_i)$ that the TCP flags of the j^{th} packet of the connection assume the value s_i . Taking into account the nature of the security attacks, for reducing the complexity of the system, we have decided to evaluate such probabilities only for the first 10 packets of a connection, i.e. $j = 1, 2, \dots, 10$.

2) *Markov Chains*: In the case of Markovian models the symbols s_i are considered as the states of a hidden discrete time finite state Markov chain.

Since not all the TCP flags configurations are observable in real traffic, the system only considers the states observed in the training phase. Moreover, to take into account the possibility that some new flags configurations could be observed during the detection phase, a rare state is added. This procedure allows us to reduce the cardinality of the state space from 64 (all the possible configurations of the six TCP flags bits) to a number K , usually smaller than ten.

Then the system estimates the transition probabilities of the Markov chain.

Since the computation of such probabilities is quite straightforward in the case of first order Markov chains (homogeneous and non-homogeneous), in the following we consider a Markov chain of order l . The main problem related to this kind of models is the “explosion” of the number of parameters,

which grows exponentially with the order, according to the rule $K^l(K-1)$. This entails the need of a parsimonious representation of the transition probabilities. The approach used in this paper is the Mixture Transition Distribution (MTD) model, first proposed in [6]. Under the MTD model, the transition probabilities of an l^{th} order Markov chain can be expressed as follows:

$$P(C_t = s_{i_0} | C_{t-1} = s_{i_1}, C_{t-2} = s_{i_2}, \dots, C_{t-l} = s_{i_l}) = \frac{1}{\sum_{j=1}^l \lambda_j r(s_{i_0} | s_{i_j})} \quad (2)$$

where C_t represents the state of the chain at step t and the quantities

$$R = \{r(s_i | s_j); i, j = 1, 2, \dots, K\} \\ \Lambda = \{\lambda_j; j = 1, 2, \dots, l\} \quad (3)$$

satisfy to the following constraints:

$$r(s_i | s_j) \geq 0; i, j = 1, 2, \dots, K \\ \sum_{s_i=1}^K r(s_i | s_j) = 1 \quad \forall j = 1, 2, \dots, K \quad (4)$$

$$\lambda_j \geq 0; j = 1, 2, \dots, l \text{ and } \sum_{j=1}^l \lambda_j = 1 \quad (5)$$

A consequence of the use of the MTD model is the reduction of the number of parameters from $K^l(K-1)$ to $K(K-1) + l - 1$. To take into account the presence of the “rare” state (labelled K), we have to fix the following quantities:

$$r(\text{rare} | s_i) = \epsilon, \quad \forall i = 1, 2, \dots, K \text{ and } \epsilon \text{ small } (\epsilon = 10^{-6}) \\ r(s_i | \text{rare}) = (1 - \epsilon) / (K - 1), \quad \forall i = 1, 2, \dots, K - 1 \quad (6)$$

According to the MTD model the log-likelihood of a sequence (c_1, c_2, \dots, c_T) of length T is

$$LL(c_1, c_2, \dots, c_T) = \sum_{i_0=1}^K \dots \sum_{i_l=1}^K N(s_{i_0}, s_{i_1}, \dots, s_{i_l}) \log \left(\sum_{j=1}^l \lambda_j r(s_{i_0} | s_{i_j}) \right) \quad (7)$$

where $N(s_{i_0}, s_{i_1}, \dots, s_{i_l})$ represents the number of times the transition $s_{i_l} \rightarrow s_{i_{l-1}} \rightarrow \dots \rightarrow s_{i_0}$ is observed. Maximum likelihood estimation (MLE) of the chain parameters requires to maximize the right hand side of eq. (7), with respect to R and Λ , taking into account the constraints (4) and (5).

Since the original solution [7] seems to be too much computationally demanding, we have applied the procedure proposed in [3], which consists in an alternate maximization with respect to R and to Λ . This process leads to a global maximum, since LL is concave in R and Λ . For the part when R is fixed, we maximize LL with respect to Λ , and vice-versa. In the first step (estimation of Λ) we have used the sequential quadratic programming, while the second maximization step (estimation of R) is a linear inverse problem with positivity constraints (LININPOS) that we have solved applying the expectation maximization (EM) algorithm [8]. Since the first maximization step is quite trivial, in what follows we discuss

the second step, i.e. the estimation of the matrix R , with the vector Λ fixed. First of all we have re-indexed the log-likelihood in the following way:

$$\phi(s_{i_0}, s_{i_1}, \dots, s_{i_l}) = 1 + \sum_{j=0}^l (s_{i_j} - 1)K^{l-j} \rightarrow k \quad (8)$$

which takes to

$$N(s_{i_0}, s_{i_1}, \dots, s_{i_l}) \rightarrow a_k \text{ and } \sum_{j=1}^l \lambda_j r(s_{i_0} | s_{i_j}) \rightarrow b_k \quad (9)$$

Thus, at first we estimate the quantities b_k (MLE) and then we solve the linear system

$$b_k = \sum_{j=1}^l \lambda_j r(s_{i_0} | s_{i_j}) \quad (10)$$

which is a LININPOS problem.

At this point, the log-likelihood can be expressed as:

$$\sum_{k=1}^{K^{l+1}} a_k \log b_k \quad (11)$$

where [3]

$$\sum_{k=1}^{K^{l+1}} a_k = T - l \text{ and } \sum_{k=1}^{K^{l+1}} b_k = K^l \quad (12)$$

Thus a simple Lagrange method argument shows that the log-likelihood is maximized when

$$\hat{b}_k = \frac{a_k}{\sum_k a_k} \sum_k b_k = \frac{a_k}{T-l} K^l, \forall k \quad (13)$$

or, equivalently, when

$$\sum_{j=1}^l \lambda_j r(s_{i_0} | s_{i_j}) = \frac{K^l}{T-l} N(s_{i_0}, s_{i_1}, \dots, s_{i_l}), \forall (i_0, \dots, i_l) \quad (14)$$

Thus, if we consider these equations as a linear system subject to the constraints (4), we obtain a LININPOS problem, which can be solved, in the sense of the minimum Kullback-Leibler distance, using the EM algorithm. More in detail we have [9]

$$\begin{pmatrix} A \\ B \end{pmatrix} R = \begin{pmatrix} \frac{K^l}{T-l} \cdot N \\ 1 \end{pmatrix} \quad (15)$$

where

$$R^T = (r(s_1 | s_1), r(s_2 | s_1), \dots, r(s_K | s_1), \dots, r(s_K | s_K)) \\ = (r_1, r_2, \dots, r_{K^2}) \quad (16)$$

are the unknowns, $r(s_i | s_j) = r_{i+K(j-1)}$, and

$$N = \begin{pmatrix} N(s_1, s_1, \dots, s_1) \\ N(s_1, s_1, \dots, s_2) \\ \vdots \\ N(s_1, s_1, \dots, s_K) \\ \vdots \\ N(s_K, s_K, \dots, s_1) \\ \vdots \\ N(s_K, s_K, \dots, s_K) \end{pmatrix} = \begin{pmatrix} N_1 \\ N_2 \\ \vdots \\ N_K \\ \vdots \\ N_{1-K+K^{l+1}} \\ \vdots \\ N_{K^{l+1}} \end{pmatrix} \quad (17)$$

where $N(s_0, s_1, \dots, s_l) = N_i$, $i = \phi(s_{i_0}, s_{i_2}, \dots, s_{i_l})$

$$A = \{a_{ij}\}_{K^{l+1} \times K^2} \\ \text{where } a_{ij} = \sum_{k=0}^l \lambda_k I[j = i_0 + K(i_k - 1)], \\ (i_0, \dots, i_l) = \phi^{-1}(i) \quad (18)$$

The matrix B looks like:

$$B = \begin{pmatrix} 1, \dots, 1 & 0, \dots, 0 & 0, \dots, 0 & 0, \dots, 0 \\ 0, \dots, 0 & 1, \dots, 1 & 0, \dots, 0 & 0, \dots, 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0, \dots, 0 & 0, \dots, 0 & 0, \dots, 0 & 1, \dots, 1 \end{pmatrix}_{K \times K^2} \quad (19)$$

At this point the EM iteration step is the following:

$$r_j \leftarrow \frac{a_{.j}}{a_{.j} + b_{.j}} \hat{r}_j(A, \frac{K^l}{T-l} \cdot N, R) + \frac{b_{.j}}{a_{.j} + b_{.j}} \hat{r}_j(B, 1, R) \quad (20)$$

where

$$\hat{r}_j(W, u, v) \equiv \frac{v_j}{w_{.j}} \sum_i \frac{w_{ij} u_i}{\sum_k w_{ik} v_k}, \quad j = 1, 2, \dots, K^2 \quad (21)$$

for matrix $W = \{w_{ij}\}$ and vectors $u = \{u_i\}$, $v = \{v_i\}$, and

$$a_{.j} = \sum_i a_{ij} = \sum_{(i_0, \dots, i_l) = \phi^{-1}(i)} \sum_{k=0}^l \lambda_k I[j = i_0 + K(i_k - 1)] \\ = \sum_{k=0}^l \lambda_k \sum_{i_0} \dots \sum_{i_l} I[j = i_0 + K(i_k - 1)] \\ = \sum_{k=0}^l \lambda_k K^{l-1} = K^{l-1} \quad (22)$$

and $b_{.j} = \sum_i b_{ij} = 1$.

The choice of the initial values for R and Λ is a key point. Experimental tests have shown that good results are obtained choosing $\lambda_i = 1/l$, $i = 1, 2, \dots, l$ and setting R to the first order transition probabilities.

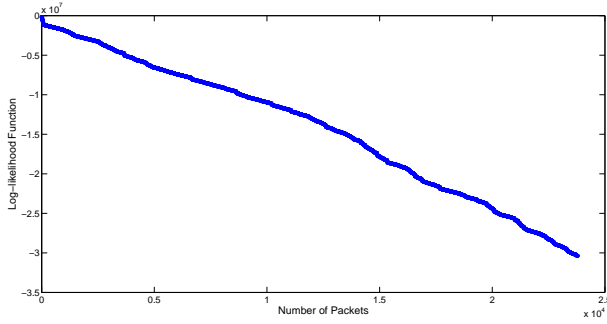


Fig. 1. Log-likelihood function of a “normal” connection

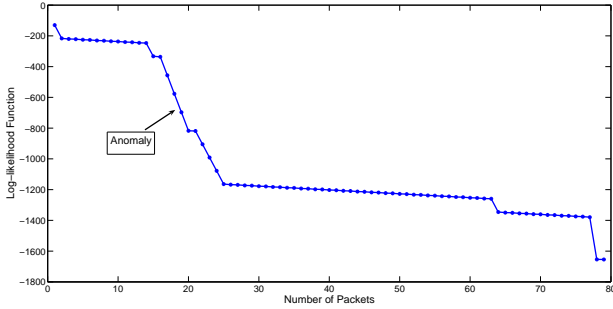


Fig. 2. Log-likelihood function of an “anomalous” connection

B. Detection Phase

Once the training phase has been performed, the IDS has a model of the “normal” behavior of the network, represented by the computed *profile*.

As for the training phase, the input is given by raw traffic traces in libpcap format, which are processed so as to extract sequences of TCP flags configurations.

Thus, given an observed sequence (c_1, c_2, \dots, c_T) , the system has to decide between the two hypotheses:

$$\begin{aligned} H_0 &: \{(c_1, c_2, \dots, c_T) \sim \text{computed model}\} \\ H_1 &: \{\text{anomaly}\} \end{aligned} \quad (23)$$

The problem is to choose between a single hypothesis H_0 , which is associated to the estimated stochastic model, and the composite hypothesis H_1 , which represents all the other possibilities. No optimal result is presented in the literature about this decision theory problem, thus the best solution is represented by the use of the Generalized Likelihood Ratio (GLR) test [10].

Since the problem is quite straightforward for ECDF, in the following we only consider the case of Markovian models, for which the GLR test is defined as follows:

$$H(X) = \begin{cases} H_0 & \text{if } X < \xi \\ H_1 & \text{if } X > \xi \end{cases} \quad (24)$$

where the threshold ξ is set by means of MonteCarlo simulations and the quantity X is given by:

$$X = \left(\frac{\text{Max}_{v \neq u} L(c_1, c_2, \dots, c_T | \Lambda_v, R_v)}{L(c_1, c_2, \dots, c_T | \Lambda_u, R_u)} \right)^{\frac{1}{T}} \quad (25)$$

where the vector (Λ_u, R_u) represents the parameters corresponding to the model computed during the training phase (hypothesis H_0) and the component $\frac{1}{T}$ is introduced to take into account that each observed sequence is characterized by a different length T . It is worth noticing that this test is equivalent to decide on the basis of the Kullback-Leibler divergence between the model associated to H_0 and the one computed for the observed sequence.

III. EXPERIMENTAL RESULTS

In this section we compare the performance of the different statistical models over the 1999 DARPA evaluation project [11] data set.

For sake of brevity, in the following we only present the results related to the Telnet traffic, since they appear to be representative of the overall performance.

To test the correctness of the computed models we have calculated the log-likelihood function of some sequences. Figure 1 corresponds to a “normal” connection. As expected from the theory, the function decreases almost linearly with the number of packets; its slope is equal to the entropy of the model, which, for first order Markov chain, is defined as:

$$H(\text{MC}) = \sum_i \sum_j \pi(i) P(s_j | s_i) \log P(s_j | s_i)$$

where $\pi(i)$ is the stationary distribution of the Markov chain. The given definition can be easily extended for higher order Markov chains.

On the other hand the effect of an anomaly is an abrupt jump in the log-likelihood function, as highlighted by figure 2. Both these figures refer to a first order model, but the behavior of the log-likelihood function does not significantly vary with the order of the Markov chain.

To evaluate the performance we have used a Receiver Operating Characteristic (ROC) curve, which plots detection rate vs. false positive rate, obtained varying the value of the threshold ξ .

Figure 3 shows the ROC curves for Markov chains of different orders. We have considered Markov chains of order up to 4, since higher orders imply a heavy processing time, not suitable for on-line detection. Since the results obtained using a model based on a Markov chain of order 1 are already very good for these traffic traces, it is not easy to realize that we achieve some improvements with high order models. To be noted that the ROC curves are almost ideal, since we have a detection rate close to 100% with a negligible false alarm rate. Nevertheless the zoomed area inside the figure shows that with the model of order 4 we are able to achieve the best results, obtaining a detection rate of 53% with a false alarm rate which is about one half of that related to the Markov chain of order 1.

The following figure shows the performance of the ECDF, while figure 5 presents a comparison between the first order

Markov chains and the time dependent models described in the paper.

Since a detection performed analyzing only the first 10 packets of each connection is obviously worse than the one based on the entire connections, also the time independent model has been computed only considering the first ten packets of each connection. It is easy to conclude that the homogeneous Markov chain achieves a detection rate almost 10% bigger than the other two models.

This apparent paradox can be justified by the fact that the non-homogeneous models have been computed with a relatively short, and so “incomplete”, training phase. Indeed, on one side the whole training data set has been used to compute only one homogeneous model, while on the other side, the same quantity of data is partitioned into ten subsets corresponding to the first ten steps in the time evolution of each connection. In particular this can lead to almost deterministic probabilities for the first steps of the non homogeneous models, thus a single flag configuration at step i , present in the training data set only at steps $j \neq i$ (and hence captured by the time independent model), may generate a false alarm.

Finally, we have taken into account that an intrusion should be detected as soon as the anomaly appears. Thus, in figure 6, we show the performance of the homogeneous Markov chain model as a function of the number of analyzed packets for each connection (both for building the model and for the detection phase). The results highlight that good performance are achieved with a small number of packets, demonstrating that such statistical models are suitable for on line anomaly detection.

IV. CONCLUSIONS

In this paper we have presented an anomaly based network intrusion detection system, which detects anomalies using statistical characterizations of the TCP traffic. We have compared several stochastic models, such as first order homogeneous and non-homogeneous Markov chains, high order homogeneous Markov chains, and stationary and non-stationary ECDF. We have detailed the estimation of the parameters of the models and we have shown the results obtained with the DARPA 1999 data set. The performance analysis has highlighted that the best results are obtained with the use of homogeneous Markov chains and that some improvements can be achieved using high order Markovian models: for instance, 4th order Markov chains lead to the same detection rate of first order models, with almost one half of false alarms.

Moreover, we have shown that, since only a small quantity of packets is sufficient to reveal intrusions in the TCP traffic, this kind of approach is suitable for on line detection.

V. ACKNOWLEDGMENTS

This work was partially supported by the Euro-NGI Network of Excellence funded by the European Commission and partly by the RECIPE project funded by MIUR.

REFERENCES

- [1] Kemmerer, R.A., Vigna, G., *Intrusion Detection: a Brief History and Overview*, IEEE Security and Privacy (supplement to Computer, vol. 35, no. 4) pp 27-30, April 2002
- [2] Ye, N., Yebin Zhang, Y., and Borrer, C.M., *Robustness of the Markov-Chain for Cyber-Attack Detection*, IEEE Transactions on Reliability, Vol. 53, no. 1, pp. 116-123, March 2004
- [3] Ju, W-H and Vardi, Y., *A Hybrid High-order Markov Chain Model for Computer Intrusion Detection*, NISS, Technical Report Number 92, February 1999.
- [4] Schonlau, M., et al., *Computer Intrusion: Detecting Masquerades*, NISS, Technical Report Number 95, March 1999.
- [5] Ye, N., Ehiabor, T., and Zhanget, Y., *First-order Versus High-order Stochastic Models for Computer Intrusion Detection*, Quality and Reliability Engineering International, 18:243-250, 2002.
- [6] Raftery, A.E., *A model for high-order Markov chains*, Journal of the Royal Statistical Society, series B, 47, 528-539, 1985.
- [7] Raftery, A.E. and Tavare, S. *Estimation and modelling repeated patterns in high-order Markov chains with the mixture transition distribution (MTD) model*, Journal of the Royal Statistical Society, series C - Applied Statistics, 43, 179-200, 1994.
- [8] Vardi, Y. and Lee, D., *From Image deblurring to Optimal investments: Maximum Likelihood Solutions for Positive Linear Inverse Problem*, Journal of the Royal Statistical Society, series B, 55, 569-612, 1993.
- [9] Iusem, A.N. and Svaiter, B.F., *A New Smoothing-Regularization Approach for a Maximum-Likelihood Estimation Problem*, Applied Mathematics and Optimization, 29:225-241, 1994.
- [10] Mood, A.M., Graybill, F.A., and D. C. Boes, D.C., *Introduction to the Theory of Statistics* 3rd ed. Tokyo, Japan: McGraw-Hill, 1974.
- [11] Lippmann, R., et al., *The 1999 DARPA Off-Line Intrusion Detection Evaluation*, Computer Networks Volume 34, Issue 4 , October 2000, Pages 579-595.

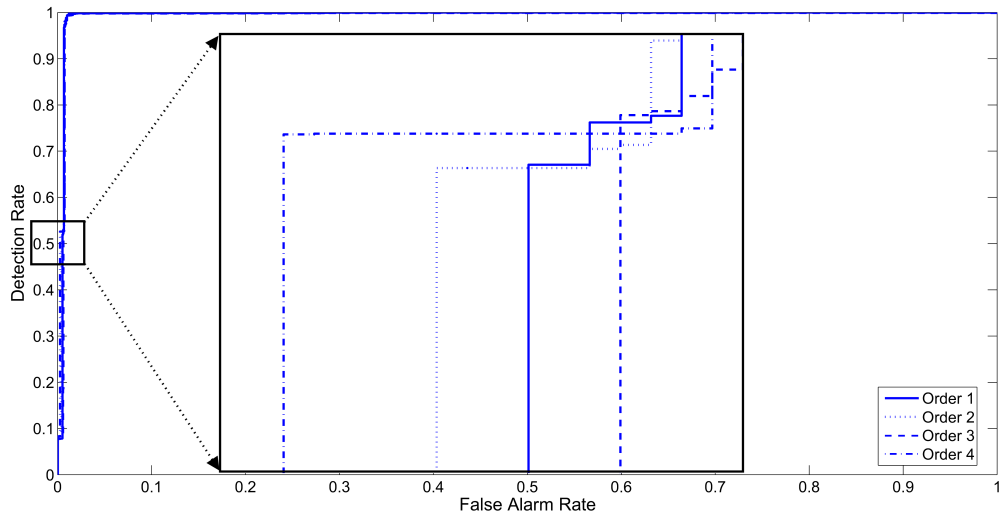


Fig. 3. Performance of Markovian models of different orders

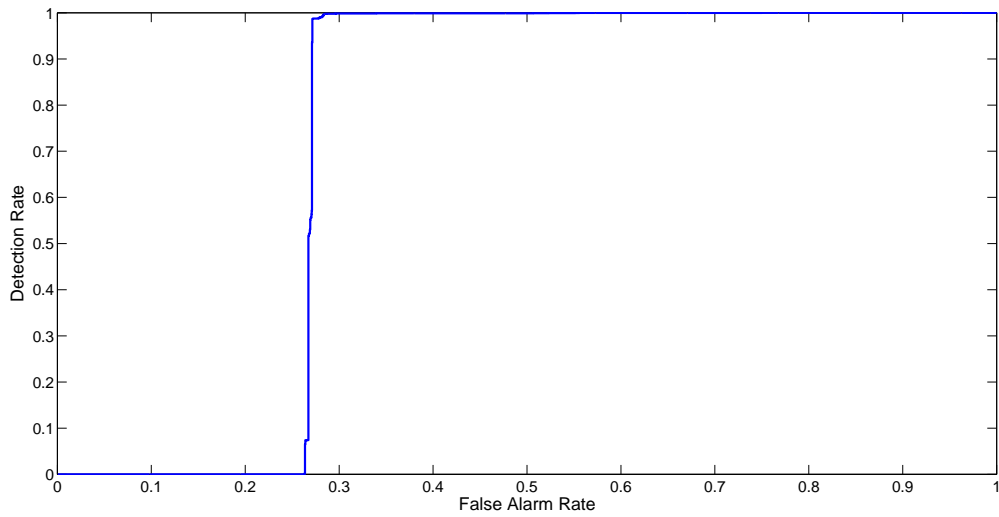


Fig. 4. Performance of ECDF model

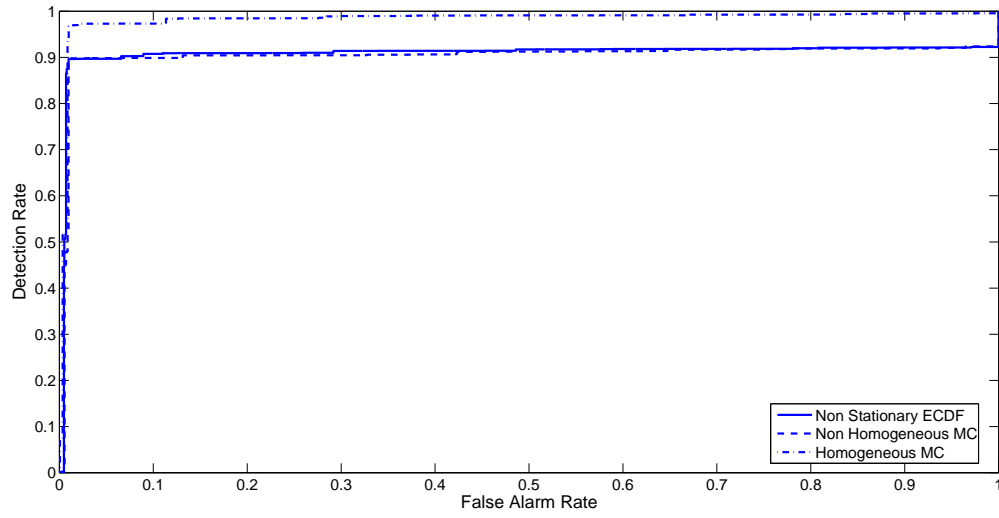


Fig. 5. Performance comparison of the analyzed time dependent models (10 packets only)

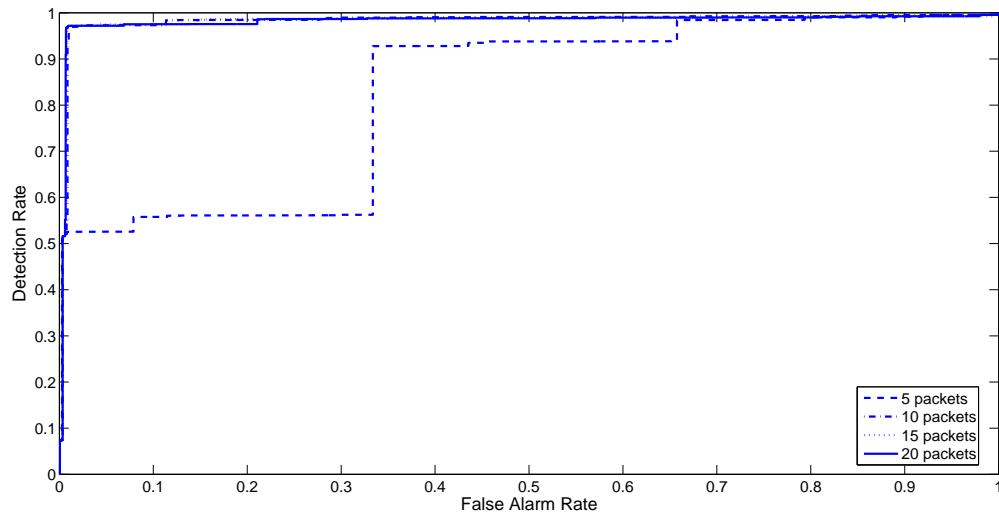


Fig. 6. Performance of the homogeneous Markov chain model, as a function of the number of processed packets