Dynamic Modeling of Internet Traffic for Intrusion Detection

E. Jonckheere, K. Shah and S. Bohacek
University of Southern California,
Los Angeles, CA- 90089-2563.
jonckhee@eudoxus.usc.edu

Abstract

Computer network traffic is analyzed via state space models and statistical techniques such as linear and nonlinear canonical correlation analyses and mutual information. As an application, the models and the statistical techniques are utilized to detect UDP flooding attacks. This work indicates that mutual information is a powerful tool for the detection of such attacks. Our approach is topology independent and our findings are tested on the so-called dumbbell and parking-lot topologies.

This research was supported by DARPA Contract N66001-00-C-8044.

1 Introduction

Traffic signal analysis has seen renewed interest over the past few years and has so far for the most part focused on modeling such phenomena as self-similarity and burstyness, building on the theory of α-stable distributions with infinite variances [1], [2]. Here, we rather focus on the dynamical aspects of the modeling, yet keeping in mind that the traffic signals inevitably contain a certain degree of randomness due to the fact that the traffic sources appear unpredictable from the observation point, typically a router. A modeling tool that fairly naturally applies in this context is the Canonical Correlation Analysis (CCA) between the past and the future of the process. The motivation for a dynamic modeling of the signals is that, if a baseline (FTP, HTTP,...) traffic model has been identified, if the model is subsequently confronted with traffic data, and if at a certain point in time the model no longer fits the data, some suspicious activity must be on going. CCA also yields as a by-product the Akaike mutual information between the past and the future, which provides a statistical signature of the signal. The latter ineluctably changes under attack and hence produces yet another intrusion detection scheme.

Several signals (link utilization, packet arrival, queue length, ...) are candidates for dynamical modeling, but here we shall focus on link utilization. No distinction between control and data packets is made at this stage. The signals are themselves generated by `ns`, the network simulator. Two different network topologies have been retained—the "dumbbell" topology and the "parking lot" topology. In both the cases, the link utilization is observed at a router. The link utilization is integrated over a sampling period ranging from 0.1 to 20 sec. Our study is somehow 4-fold: dumbbell versus parking lot topology, linear versus nonlinear, for varying sampling periods, and for varying "lags," where the lag is defined as the length of the data record utilized in the CCA.

2 Simulation Setup

We used the Network Simulator (`ns`)developed by LBNL to set up our simulation environment [3]. `Ns` is a discrete event simulator widely accepted for networking research. It provides a substantial support for simulation of TCP, routing, and multicast protocols over wired and wireless (local and satellite) networks. Moreover, `ns` generates Constant Bit Rate (CBR) traffic, TELNET, FTP, HTTP, etc. The simulator also has a small collection of mathematical functions that can be used to implement random variate generation (exponential, uniform, Pareto, etc.) We used this capability to setup the network environment that synthesized HTTP, FTP, and CBR traffic.

We performed our tests on two different topologies. The first topology under consideration was the "Dumbbell" topology (Fig. 1). We set the nodes S_i (i = 1,2,...,5) as sources and the nodes D_i (i=1,2,...,5) as destinations. Normal traffic was generated by sending a mixture of HTTP and FTP traffic from the sources (S_i) to the corresponding destinations (D_i) at random times. For HTTP traffic, the file size distribution was modeled as a general ON/OFF behavior with a combination of heavy-tailed and light tailed sojourn times, while the interpage time and the interobject per page time distributions were set to be exponential. The page

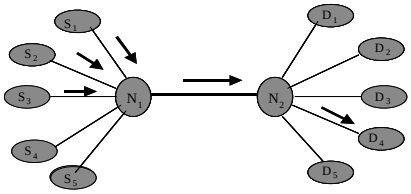

Figure 1: Dumbbell topology. Normal traffic is a mix of HTTP and FTP traffic, while UDP pakcet storm attack is simulated by sending CBR traffic from the sources S1, S2, S3 to the destination S4.

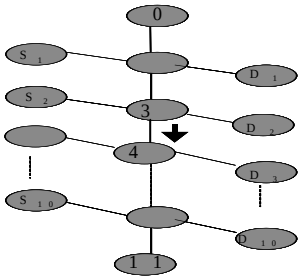

Figure 2: Parking lot topology. Baseline traffic is a mix of HTTP and FTP traffic, while UDP flooding attack is simulated by sending CBR traffic from node 3 to node 4.

size was set to be constant and the object per page size to be Pareto to replicate today's network bursty traffic [4], [2]. For FTP traffic, files of random sizes were sent at random times [5]. We monitored the traffic flowing from N_1 to N_2, the bottleneck, or "choke point," link. To simulate a UDP packet storm attack [6], a large number of small size Constant Bit Rate (CBR) packets were sent over some UDP connections from the sources S_1, S_2, S_3 to the victim destination D_4 on the top of the normal traffic. Each trial was executed for 30000 simulated seconds, logging the traffic at the 0.01 second granularity. For a particular scenario, the bottleneck link was 1.5 Mbps and the non-bottleneck links were 10 Mbps and the latency of the each link was set to 20 ms. UDP flooding attack was generated by each source having 5 UDP agents sending CBR packets of the size 200 bytes at the rate of 0.005 second/bytes to the victim.

In the more complicated "Parking Lot" topology (Fig. 2), we set the nodes S_i (i = 1,2,...,10) as sources and the nodes D_i (i = 1,2,...,10) as destinations. A dynamical model for normal TCP traffic was synthesized from the signals obtained by sending a mixture of FTP and HTTP traffic from the sources to their downstream destinations at random times. The normal traffic was monitored along the path from node 3 to node 4. In addition to this background traffic (HTTP and FTP), a large number of small size CBR packets were sent over some UDP connections from source node 3 to the victim node 4 to model the attack scenario. We monitored the link utilization along the same path, from node 3 to the node 4, and gathered the simulated attack data. Simulation results were obtained for several trials of `ns`. Each run was executed for 30000 simulated seconds, logging the traffic at the 0.01 second granularity. For a particular case, link speed was 10 Mbps and the latency of the each link was set to 20 ms. UDP packet storm was generated by 15 UDP agents sending CBR packets of a size of 200 bytes at a rate of 0.005 second/bytes to the victim.

3 Canonical Correlation Analysis

CCA is a second moment technique. In its linear version, it relies on the second moments of the process itself, and as such the analysis cannot be carried out on those self-similar traffic signal models with infinite variance [1]. One should keep in mind, however, that infinite variance processes are a convenient way of modeling *exactly* self-similar processes and that in practice self-similarity is observed only over finitely many scales. Other considerations that support the finite variance hypothesis include the small size of the network on which the traffic is simulated and the finite bandwidth of the links. These observations corroborate recent work at AT&T [7], which calls into question whether real traffic is self-similar. In the nonlinear CCA, these issues become irrelevant, because the variance analysis is applied to a nonlinear distortion of the original process, which is restricted to result in a finite variance process.

3.1 Linear state space models

Here $\{y(k) \in [-b, +b] : k = ..., -1, 0, +1, ...\}$ is the centered link utilization signal, bounded by the bandwidth, viewed as weakly stationary process with finite covariance $E(y(i)y(j)) = \Lambda_{i-j}$ defined over the probability space $(\Omega, \mathcal{A}, \mu)$. The past and the future of the process are defined, respectively, as

$$\begin{aligned} y_-(k) &= (y(k), y(k-1), ..., y(k-L+1))^T, \\ y_+(k) &= (y(k+1), ..., y(k+L))^T \end{aligned}$$

where L is the lag. The ability to devise a good model can be gauged from the Kolmogorov-Sinai, or Shannon, mutual information between the past and the future [8],[9],[10],

$$\begin{aligned} I(y_-, y_+) &= h(y_+) - h(y_+|y_-) \\ &= \int\int \log \frac{p(y_-, y_+)}{p(y_-)\, p(y_+)} p(y_-, y_+)\, dy_- dy_+ \end{aligned}$$

In the above, $h(y_+)$ is the Shannon entropy of the future and $h(y_+|y_-)$ is the conditional entropy of the fu-

ture given the past. To proceed from a numerical algebra point of view, the covariances of the past and the future are factored as

$$E(y_-(k)y_-^T(k)) = L_-L_-^T, \quad E(y_+(k)y_+^T(k)) = L_+L_+^T$$

and the canonical correlation is defined and Singular Value Decomposed (SVDed) as

$$\Gamma\left(y_-, y_+\right) = L_-^{-1}E\left(y_-(k)y_+^T(k)\right)L_+^{-T} = U^T\Sigma V$$

where U, V are orthogonal matrices and

$$\Sigma = \begin{pmatrix} \sigma_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_L \end{pmatrix}, \quad 1 \geqslant \sigma_1 \geqslant \ldots \geqslant \sigma_L \geqslant 0$$

The σ's are called canonical correlation coefficients (CCC's). If the process is Gaussian, it is well known that

$$\begin{aligned} \Delta(y_-, y_+) &= I(y_-, y_+) \ \textit{where}, \\ \Delta(y_-, y_+) &: = -\frac{1}{2}\log\det\left(I - \Gamma^T\left(y_-, y_+\right)\Gamma\left(y_-, y_+\right)\right) \end{aligned}$$

At this stage, it is customary to assume that there are only a restricted number $D \leq L$ of significant CCC's, which we group in Σ_1, and we further partition Σ and the orthogonal matrices conformably as

$$U = \begin{pmatrix} U_1 \\ U_2 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \Sigma_1 & 0 \\ 0 & 0 \end{pmatrix}, \quad V = \begin{pmatrix} V_1 \\ V_2 \end{pmatrix}$$

The canonical past and the canonical future [11] are defined as

$$\overline{y_-(k)} = U_1L_-^{-1}y_-(k), \quad \overline{y_+(k)} = V_1L_+^{-1}y_+(k)$$

The state is defined as the minimum collection of past-measurable random variables necessary to predict the future, that is, $E(y_+\left(k\right)|y_-\left(k\right))$. A basis of such collection of random variables is given by

$$x(k) = E\left(\overline{y_+(k)}|\overline{y_-(k)}\right) = \Sigma_1\overline{y_-(k)}$$

The state transition matrix A is defined as the least squares fit regression matrix of $x(k+1)$ on $x(k)$, viz.,

$$\begin{aligned} A &= E\left(x(k+1)x^T(k)\right)\left(Ex(k)x^T(k)\right)^{-1} \\ &= \Sigma_1U_1L_-^{-1}\overrightarrow{\Lambda}L_-^{-T}U_1^T\Sigma_1^{-1} \end{aligned}$$

where $\overrightarrow{\Lambda}$ denotes Λ shifted to the right by one position, that is,

$$\overrightarrow{\Lambda} = \begin{pmatrix} \Lambda_1 & \Lambda_2 & \cdots & \Lambda_L \\ \Lambda_0 & \Lambda_1 & \cdots & \Lambda_{L-1} \\ \vdots & \vdots & \ddots & \vdots \\ \Lambda_{L-2} & \Lambda_{L-3} & \cdots & \Lambda_1 \end{pmatrix}$$

It is a bit tedious to show (although it is implicitly contained in Akaike [11]) that the residual noise $w(k)$ is white and furthermore

$$Q = E(w(k)w^T(k)) = \Sigma_1^2 - A\Sigma_1^2A^T$$

Next, a regression of $y(k+1)$ on $x(k)$ is done, yielding the matrix C as

$$\begin{aligned} C &= E\left(y(k+1)x^T(k)\right)\left(E(x(k)x^T(k)\right)^{-1} \\ &= \left(\Lambda_1, \Lambda_2, ..., \Lambda_L\right)L_-^{-T}U_1^T\Sigma_1^{-1} \end{aligned}$$

Again, the residual error $v(k)$ can be shown to be white and

$$R = E(v(k)v(k)) = \Lambda_0 - C\Sigma_1^2C^T$$

Finally, it is also readily found that

$$S = E(w(k)v(k)) = \Sigma_1U_1L_-^{-1}\Lambda^{(2)} - A\Sigma_1^2C^T$$

Where, $\Lambda^{(2)}$ is the 2^{nd} row of $\overrightarrow{\Lambda}$. Hence, we have a state space model of the form [12]

$$\begin{aligned} x(k+1) &= Ax(k) + w(k) \\ y(k+1) &= Cx(k) + v(k) \end{aligned}$$

In order to confront the data with the model, we need to know the state $x(k)$, which could be computed as $\Sigma_1\overline{y_-(k)}$. It is, however, more efficient to get an estimate of the state provided by the Kalman filter

$$\begin{aligned} \widehat{x}(k+1|k+1) &= A\widehat{x}(k|k) + K(y(k+1) - C\widehat{x}(k|k)) \\ y(k+1) &= C\widehat{x}(k|k) + (y(k+1) - C\widehat{x}(k|k)) \end{aligned}$$

Since $y(k+1) - C\widehat{x}(k|k)$ is well known to be a white noise, called *innovation*, the Kalman filter provides yet another state space model, referred to as *innovation representation* [13]. The Kalman gain is given by

$$K = -\left(R + CPC^T\right)^{-1}\left(B^TPA^T + S\right)$$

and $P = E(x(k) - \widehat{x}(k|k))(x(k) - \widehat{x}(k|k))^T$ is the stabilizing solution to the discrete-time algebraic Riccati equation

$$\begin{aligned} P &= APA^T + Q \\ &\quad -(APB + S^T)\left(R + CPC^T\right)^{-1}\left(B^TPA^T + S\right) \end{aligned}$$

A few numerical remarks: It is customary to define $L_\pm$ to be lower triangular (Cholesky factorization), although $L_\pm$ could be defined upper triangular ("anti-Cholesky" factorization), in which case Γ is near-Hankel and in fact will be Hankel for $L = \infty$. The particular factorization does not affect the CCC's. $E(y_\pm(k)y_\pm^T(k))$ might be marginally positive definite, resulting in problems in the Cholesky factorization; there is thus a need to monitor the condition number of $E(y_\pm(k)y_\pm^T(k))$.

3.2 Nonlinear state space models

Here, we allow the zero-mean process $\{y(k) \in \Re : k = ..., -1, 0, +1, ...\}$ to be of infinite variance (for example, an α-stable, H-self-similar process [1]). The nonlinear CCA [8],[9] is an attempt to reach the mutual information, in the nongaussian setup, as

$$\sup_{f,g} \left(\Delta(f(y_-), g(y_+))\right) \leq I(y_-, y_+)$$

where $f, g : \Re^L \longrightarrow \Re^L$ are measurable, bijective functions such that $E(f) = E(g) = 0$, $E\left(ff^T\right) < \infty I$, $E\left(gg^T\right) < \infty I$. Equality is achieved iff $f\left(y_-\right), g\left(y_+\right)$ can be made jointly Gaussian (Cramer-Wold theorem; see [8],[9]), in which case the joint past/future process is called *diagonally equivalent to Gaussian*. Since the canonical correlation is unaffected by scaling of f, g, it is convenient to choose $E_- f^T f = 1, E_+ g^T g = 1$, where $E_\pm$ denotes the mathematical expectation relative to the probability space $\left(\Omega, \mathcal{A}_\pm, \mu_\pm\right)$ of future/past random variables. Here, instead of the approach taken in [8], [9], we propose a more computationally viable one based on the fact that the components of $f\left(y_-\right), g\left(y_+\right)$ can be expressed as linear combinations of polynomials $p_j(y_-), q_j(y_+); j = 1, 2, ...$ such that $E_- p_j = E_+ q_j = 0$, $E_-\left(pp^T\right) < \infty I$, $E_+\left(qq^T\right) < \infty I$, and forming bases of the Lebesgue spaces of zero-mean measurable functions such that $E_- f^T f < \infty, E_+ g^T g < \infty$, respectively. The problem clearly reduces to $\sup_{\phi,\gamma}(\Delta(\phi p\left(y_-\right), \gamma q\left(y_+\right)))$. If $L = \infty$, the expression between parentheses is in fact independent of ϕ, γ provided they are bounded along with their inverses. This yields $\Delta\left(p\left(y_-\right), q\left(y_+\right)\right)$ as the absolute upper bound that can be reached by this analysis. If $L < \infty$, the above supremum is nontrivial and is easily accomplished via linear CCA of $p\left(y_-\right), q\left(y_+\right)$, that is, via SVD of $\Gamma\left(p\left(y_-\right), q\left(y_+\right)\right)$. Specifically, do the factorizations $E(p\left(y_-\right)p\left(y_-\right)^T) = L_- L_-^T$, $E\left(q(y_+)q(y_+)^T\right) = L_+ L_+^T$ along with the SVD

$$\Gamma\left(p\left(y_-\right), q\left(y_+\right)\right) = \begin{pmatrix} U_1 \\ U_2 \end{pmatrix}^T \begin{pmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{pmatrix} \begin{pmatrix} V_1 \\ V_2 \end{pmatrix}$$

Here, we take Σ_1 to be $L \times L$ and we retain only those L CCC's. The motivation is to allow for easy comparison with the full-dimensional linear case and therefore gauge how much increase in the CCC's is gained by going to the nonlinear analysis. The coefficients of the optimal distortion functions are given by $\phi = U_1 L_-^{-1}, \quad \gamma = V_1 L_+^{-1}$.

To further motivate this optimization, consider a linear regression of $g\left(y_+\right)$ on $f\left(y_-\right)$. It is easily found that

$$\begin{aligned} & \min_A E \left\| g\left(y_+\right) - A f\left(y_-\right) \right\|^2_{(L_+ L_+^T)^{-1}} \\ = \quad & L - Trace\left(\Gamma^T \Gamma\left(f\left(y_-\right), g(y_+)\right)\right) \end{aligned}$$

for $A = \Sigma_1$. Clearly, the best choice of f, g is the one that maximizes $Trace\left(\Gamma^T \Gamma\left(f\left(y_-\right), g(y_+)\right)\right)$ and it is readily seen that this is achieved for the same distortion functions.

The canonical past and future, that is, orthonormal bases of the past/future such that $E\overline{y_-}(k)\overline{y_+}(k)^T = \Sigma_1$, are given by $\overline{y_-}(k) = U_1 L_-^{-1} p\left(y_-(k)\right), \quad \overline{y_+}(k) = V_1 L_+^{-1} q(y_+(k))$. The state, that is, a convenient basis for the set of random variables $E(y_+\left(k\right)|y_-\left(k\right))$, is defined as

$$x(k) = E(g(y_+(k))|f(y_-(k))) = E(\overline{y_+}(k)|\overline{y_-}(k))$$

If the past/future process is diagonally equivalent to Gaussian, we have $x(k) = \Sigma_1 \overline{y_-}(k)$. If not, the ACE algorithm would yield the correct nonlinear relationship between $x(k)$ and $\overline{y_-}(k)$ as $\theta_i(x(k)) = \sum_j \phi_{ij}((\overline{y_-}(k))_j)$. To obtain the state space equation, we have to do a regression of $x(k+1)$ on $x(k)$. If the past/future process is diagonally equivalent to Gaussian, this yields $x(k+1) = \Sigma_1 U_1 L_-^{-1} \overrightarrow{\Lambda} L_-^{-T} U_1^T \Sigma_1^{-1} x(k) + w(k)$. However, in general, $x(k), x(k+1)$ will fail to be jointly Gaussian and the regression is most easily accomplished by running the Alternating Conditional Expectation (ACE) algorithm [14], which produces a relationship of the form

$$\theta_i(x_i(k+1)) = \sum_j \phi_{ij}(x_j(k)) + w(k), \quad \left\|\theta_i\right\|_{L^2} = 1$$

For the output equation, we again use the ACE algorithm, which yields

$$\theta_y(y(k+1)) = \sum_j \phi_{yj}(x_j(k)) + v(k), \quad \left\|\theta_y\right\|_{L^2} = 1$$

Because of the θ function emanating from the ACE algorithm, we obtain a descriptor, generalized state space system. However, simulation results have shown that θ is linear in a neighborhood of 0 and then saturates, so that the generalized nonlinear state space system does not exhibit much singularity.

NUMERICAL REMARK: Practically, p, q are chosen as simple monomials in the components of the past, future. It is important to scale the large power appearing in $p(y_-), q(y_+)$, for otherwise the high power terms become dominant over the low power terms. In such a nonparametric procedure as ACE, the distortion functions θ, ϕ need to be interpolated from clusters of data points, with inevitable inaccuracies. Thus, contrary to the linear case where A^k is fairly reliable, it is not quite so in the nonlinear case, where the k-fold composition of the ϕ's yields inaccurate k-step predictions beyond $k = 5$.

3.3 Nonlinear auto-regressive models

Here, we develop a simplified approach that relies on

$$\sup_{f}(\Delta(f\left(y_{-}\right),y_{+})$$

The primary motivation is that this method leads to simple nonlinear Auto-Regressive (AR) models. The simplified nonlinear CCA procedure goes as follows: As before, let $f=\phi p$. Define $E\left(p\left(y_{-}\right)p\left(y_{-}\right)^{T}\right)=L_{-}L_{-}^{T}$, $E\left(y_{+}y_{+}^{T}\right)=L_{+}L_{+}^{T}$ along with the SVD $\Gamma\left(p\left(y_{-}\right),y_{+}\right)=U^{T}\Sigma V$. There are L canonical correlation coefficients and to allow for comparison with the previous case, we take all of them into consideration. Under these circumstances, the supremum is trivial, that is, the supremum is achieved for all f's; however, it is convenient to choose the optimal distortion as $\phi=U_{1}L_{-}^{-1}$. The canonical past and future are defined as $\overline{p\left(y_{-}\right)}=UL_{-}^{-1}p\left(y_{-}\right)$, $\overline{y_{+}}=VL_{+}^{-1}y_{+}$. Now, we do the linear regression of y_{+} on $p\left(y_{-}\right)$. It is easily seen that

$$\begin{aligned}&\min_{A}E\left\|y_{+}-Ap\left(y_{-}\right)\right\|_{(L_{+}L_{+}^{T})^{-1}}^{2}\\ =\;& L-Trace\left(\Gamma^{T}\left(p\left(y_{-}\right),y_{+}\right)\Gamma\left(p\left(y_{-}\right),y_{+}\right)\right)\end{aligned}$$

for $A=L_{+}\Gamma^{T}L_{-}^{-1}$ (in canonical coordinates $\overline{A}=\Sigma$). Observe that $E(y_{+}|p\left(y_{-}\right))\neq L_{+}\Gamma^{T}L_{-}^{-1}p\left(y_{-}\right)$ (in canonical coordinates $E(\overline{y_{+}}|\overline{p\left(y_{-}\right)})\neq\Sigma\overline{p\left(y_{-}\right)}$) unless the processes $y_{+},p\left(y_{-}\right)$ $(\overline{y_{+}},\overline{p\left(y_{-}\right)})$ are jointly Gaussian, which is unlikely to occur without nonlinear processing of the past. Define $\widetilde{y}_{+}=\overline{L_{+}\Gamma^{T}L_{-}^{-1}p\left(y_{-}\right)}$ (in canonical coordinates $\widetilde{\overline{y_{+}}}=\Sigma\overline{p\left(y_{-}\right)}$). With this notation, we get a model of the form $\widetilde{y}_{+}(k)=L_{+}\Gamma^{T}L_{-}^{-1}p(y_{-}(k))$. Taking the first row of the above yields the AR model.

4 Results and Interpretation

Figures 3 and 4 show that, in the case of the dumbbell topology, an attack can easily be detected by observing the link utilization. However, in the case of the parking-lot topology, Figures 7 and 8 show no significant difference between the attack and nonattack link utilization, calling for more sophisticated techniques to detect the attack.

The mutual information plots for the dumbbell topology are shown in Fig. 5 and Fig. 6, while those of the parking lot topology are shown in Fig. 9 and Fig. 10. The first observation is that the nonlinear CCC's are consistently higher than the linear CCC's, as expected, confirming the existence of nonlinearities in the signals. Also fairly consistent is the increase of the mutual information with both the sampling period and the lag. The increase of the mutual information with the sampling period can be justified as follows: As the sampling period increases, the signal is more integrated and hence

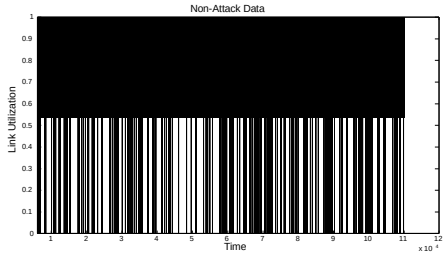

Figure 3: Link utilization time-series for non-attack data at sampling period 0.01 sec. for dumbbell topology.

smoothed over and hence looks more deterministic. As the lag increases, more random variables in both the past and the future are included, resulting in increased mutual information. However, the change in mutual information resulting from an attack can go either way: In the dumbbell topology, the mutual information increases under attack while in the parking lot topology, it decreases. The explanation for the increase under attack is as follows: CBR traffic is a deterministic signal, and if CBR occupies most of the link utilization, the sequence is more predictable and hence the information increases. For the parking lot topology, CBR occupies a small part of the link utilization, under attack, the signal is more mixed and hence less predictable, resulting in a decrease of mutual information. (A similar fact– that the Kolmogorov complexity could go both ways under attack–has also been observed in [15] .)

The prediction error plots for the parking lot topology are shown in Fig. 11,12 and 13. The main conclusion is that the normal/attack gap increases as we go from simple linear prediction, to nonlinear AR prediction, and eventually to nonlinear statespace prediction.

5 Concluding Remarks

These early investigations have demonstrated that some specific flooding attack scenarios, while not visible to the naked eye, create dynamical shift substantial enough for the mutual information to be affected and for the corrupted data to depart from the prediction of the baseline models. It appears that the most reliable way to detect the attack is by analysis of the link utilization along a bottleneck link. Other attacks, like SYN, which disrupts the normal sequencing of control and data packets, require a distinction between control and data packets, and will be reported elsewhere. Here the signal was treated as stationary although the autocorrelation test shows that the incremental signal is more stationary although modeling the incremental signal, did not appear to improve results.

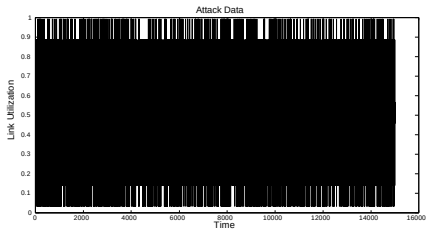

Figure 4: Link utilization time-series for attack data at sampling period 0.01 sec. for dumbbell topology.

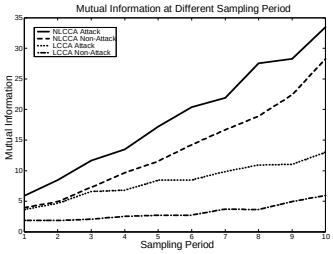

Figure 5: Mutual information versus sampling period for dumbell topology. Note that the mutual information for the NLCCA is higher than that of the LCCA, indicating presence of nonlinearity in the signal.

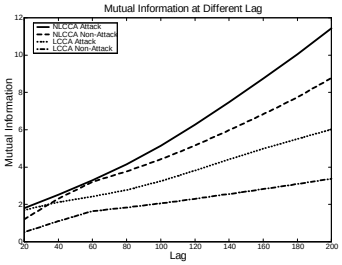

Figure 6: Mutual information versus lag for the dumbbell topology.

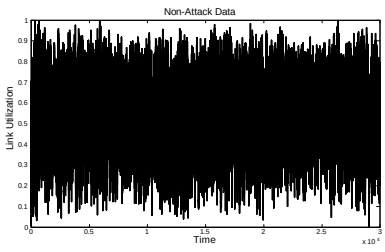

Figure 7: Link utilization time-series for non-attack data at sampling period 0.01 sec. for parking lot topology.

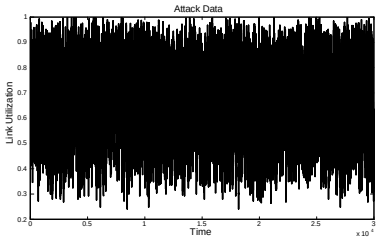

Figure 8: Link utilization time-series for attack data at sampling period 0.01 sec. for parking lot topology.

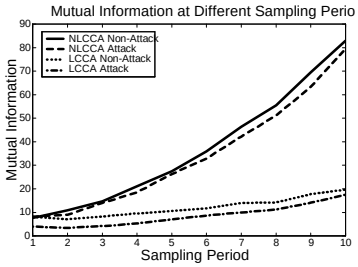

Figure 9: Mutual information versus sampling period for parking lot topology. Clearly, there is a substantial increase in mutual information in the NLCCA case as compared with the LCCA case.

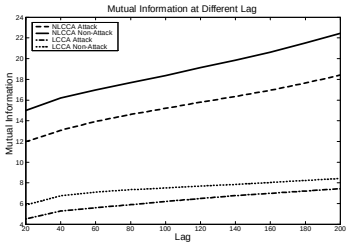

Figure 10: Mutual information versus lag for parking lot topology. Note that the difference between the mutual informations of the non-attack and attack cases is higher in the NLCCA case than in the LCCA case.

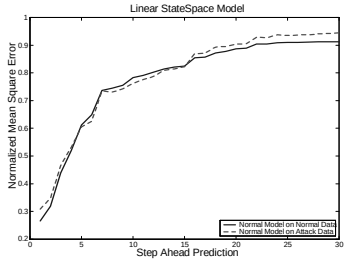

Figure 11: The normalized mean square linear state space prediction error under normal and attack conditions. Observe that the normal/attack gap is small.

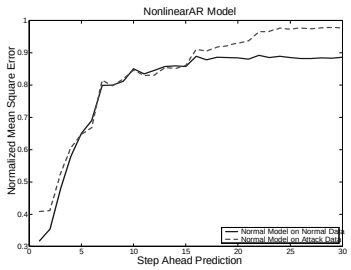

Figure 12: The normalized mean square nonlinear AR prediction error in the normal and attack cases versus the number of steps ahead. Observe the degradation of the prediction under attack for a large number of steps ahead.

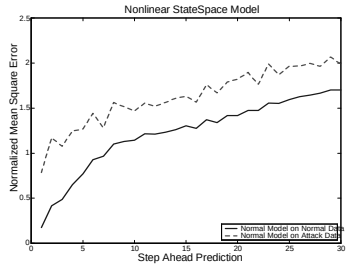

Figure 13: The normalized mean square nonlinear state space prediction error in the normal and attack cases versus the number of steps ahead. Observe the substantial increase in the normal/attack gap compared to the linear case. Note, however, that the plot is valid only for a small number of steps ahead because of the numerical unreliability of compounding nonlinear functions.

References

[1] G. Samorodnitsky and M. S. Taqqu, *Stable Non-Gaussian Random Processes, Stochastic Models with Infinite Variance*. Chapman Hall, 1994.

[2] W. Leland, M. Taqqu, W. Willinger, and D. Wilson, "On the self-similar nature of ethernet traffic(extended version)," pp. 1–15, IEEE/ACM Transactions on Networking, 2, 1994.

[3] S. Floyd, "Simulation tests. available at ftp://ftp.ee.lbl.Gov/Papers/Simtests.ps.z . ns is available at http://www-nrg.ee.lbl.Gov/Nrg.," July 1995.

[4] P. Pruthi and A. Erramilli, "Heavy-tailed ON/OFF source behavior and self-similar traffic," pp. 445–450, IEEE, 1995.

[5] "RFC959: File transfer protocol,"

[6] "CERT advisory CA-96.01: UDP port denial-of-service attack." CERT, 2/8/1996. updated 9/24/1997.found at ftp://info.cert.org/pub/cert_advisories/ca-96.01.udp_service_denial,"

[7] J. Cao, W. S. Cleveland, D. Lin, and D. X. Sun, "On the nonstationrity of Internet traffic," *Proceeding ACM SIGMETRICS '01*, pp. 102–112, 2001.

[8] E. Jonckheere and B.-F. Wu, "Mutual kolmogorov-sinai entropy approach to nonlinear estimation," *IEEE Conference on Decision and Control*, pp. 2226–2232, Tucson, Arizona, Dec 1992.

[9] B.-F. Wu, "Identification and control of chaotic processes-TheKolmogorov-sinai entropy approach, ph.d. dissertation, dept. of electrical engineering–systems, university of southern california," *Ph.D. dissertation, Dept. of ElectricalEngineering-Systems, University of Southern California*, 1992.

[10] S. Kullback, *Information Theory and Statistics*. Dover, 1968.

[11] H. Akaike, "Markovian representation of stochastic processes by canonicalvariables," *SIAM J. Control*, pp. Vol.13, pp. 162–173, Jan. 1975.

[12] M. Aoki, *State Space Modeling of Time Series*. Springer-Verlag Berlin Heidelberg New York London Paris Tokyo, 1987.

[13] E. Jonckheere and J. Helton, "Power spectrum reduction by optimal hankel norm approximation of the phaseof the outer spectral factor," *IEEE Transactions on Automatic Control*, vol. vol. AC-30, No. 12, pp. 1192–1201, December 1985.

[14] L. Breiman and J. H. Friedman, "Estimating optimal transformations for multiple regression and correlation," *Journal of the American Statistical Association*, vol. 80, pp. 580–619, 1985.

[15] S. Evans, S. F. Bush, and J. Hershey, "Information assurance through kolmogorov complexity," *Accepted publication at the DARPA Information Survivability Conference and Exposition 2*, June 2001.