A signal processing approach to anomaly detection in networks

Kavé Salamatian
Université de Savoie
Acknowledgements
Coauthors

• Daniela Brauckhoff, Martin May, Nahur Fonseca, Mark Crovella, Nina Taft, Augustin Soule, ....
Outline

• Introduction & Methodologies
  ▫ Anomaly detection basic
  ▫ Statistical anomaly detection basis

• Parametric methods
  ▫ A refresh on Linear system theory
  ▫ Model calibration
  ▫ Filtering and processing
  ▫ Refresh on Neyman Pearson test theory

• Non Parametric Methods
  ▫ Sketches and hashes
  ▫ Random distributions and KL distance
  ▫ Sanov theorem and Stein Lemma

• Distributed anomaly detection (if we have time)
  ▫ Distributed KLT
Introduction and methodology
Network state

• Network state results from
  ▫ Traffic demand
    • Traffic matrix
    • Not observable directly
  ▫ Capacity offer
    • Routing matrix, link cap., Traffic Engineering, etc.
    • Monitored by SNMP, etc.
• Network manager goal
  ▫ To drive this equilibrium to the best beneficial point by managing capacity offer
  ▫ Traffic engineering is the art of managing offered capacity
Network monitoring

- Monitoring?
  - Being able to separate
    - What is predictable
      - Expected, normal, under control, ...
    - What is not predictable
      - Unexpected, abnormal, ...
  
- Interpretation framework
  - Only what is unpredictable have a meaning, what can be predicted does not have a information
Classes of anomaly detections

• Deterministic approaches
  ▫ Signature based
  ▫ Complexity vs. exhaustivity trade-off
• Statistical approaches
  ▫ Probabilistic
  ▫ False positive vs. detection rate trade-off
Deterministic approaches

• Each observation is assumed to result from a known causality chain
  ▫ All anomalies are characterized by causality chain that describes the signature of an anomaly
  ▫ Anomaly detection consist of back tracking the causality sequence
• Need exhaustive anomaly signature database
  ▫ Have to check all possible signature
  ▫ Not scalable
Statistical approach

• A dataset is a unique sequence of deterministic observation that is not anymore perfectly reproducible
• Statistical approach assumes that
  ▫ Observation results from a random function
  ▫ they come from a random choice in a (in)finite set of «possible» observations.
  ▫ Is this assumption sound ?
• Gives access to an arsenal of probabilistic methods
  ▫ Stationarity : Intrinsic hypothesis
    • Statistical properties hold over time
    • Bet on future
    • Possibility of radical error
• Essential False alarm vs. misdetection trade-off
Empirical modeling Challenge

• Measurement contains two components
  ▫ A structured component that reflects essential properties of the phenomenon under study
  ▫ A random component that represent fluctuations that are put aside from the model

• Modeling challenge
  ▫ Separate structure from randomness and characterize it

• Signal processing role is to do this filtering
About interpretation

• Measurements
  ▫ But what do they mean?

• Interpreting?
  ▫ Relating effects to causes
  ▫ Being able to predict the behaviours
    • At different timescales
  ▫ Being able to react
  ▫ Interpretation need *a priori*
Plato cavern allegory

_Socrate:_ « compare our nature in respect of education and its lack to such an experience as this. Picture men dwelling in a sort of cavern ... Picture further the light from fire burning higher up and at a distance behind them, and between the fire and the prisoners and above them ... men carrying past the wall implements of all kinds »

_Glaucon:_ « A strange image you speak of, and strange prisoners. »

_Socrate:_ Like to us, for, to begin with, tell me do you think that these men would have seen anything of themselves or of one another except the shadows cast from the fire on the wall of the cave that fronted them? »
Modelling approach in networking

- **Descriptive approach**
  - Much more used in measurement papers
  - Network is a black box with unknown structure
    - describe observations only through descriptive statistics
      - mean, variance, Hurst or multi-fractal parameters, etc...
  - **Top-down approach**
    - Begin with observations and derive descriptive parameters
  - **Drawbacks**
    - It does not explain why?
    - It does not answer what if?
    - It is difficult to interpret them
      - Interpretation need *a priori*
    - It does not use all the available information
      - We may have *a priori* information on the process generating the observation

- **Constructive approach**
  - **Classical approach**
  - Derivate IP performance through an explicative model of the process involved into the network
    - Network is constituted of queues and routers, ...
      - Uses simulation by ns or analytic queuing theory, network calculus, etc
    - **Down top approach**
      - Begin from input scenario and network structure and derive performance measures
  - **Drawbacks**
    - Generalization is difficult
    - Too many parameters
    - Simulation results do not describe real measurements
    - The approach is open-loop
Interpretation framework

- What is the hidden causes \((X \, \text{et} \, \theta)\) that have lead to observing \(Y\)?
  - *A priori* mode compress all of our understanding of the process involved in the generation of the observation in \(Y=M(X, \theta)\)
Interpretation

- We have to deal with two main inverse problem
  - Modelling problem
    - What are the best context parameters $\theta$ that best describe the environment
  - Interpretation problem
    - Knowing the context $\theta$ what is hidden input $X$ that will best describe observations
- A lot of measurement problem might be expressed in this framework
  - Anomaly detection
  - Flow classification and recognition
Statistical Anomaly detection Basics
Anomalies ???

• What is an anomaly?
  ▫ What is normal?
  ▫ What is likely?
• Assumptions
  ▫ Anomaly are anomalous
    • different from the norm
  ▫ Anomalies are rare
  ▫ Anomalous activity is malicious
  ▫ Reverse assumption: anomaly represents attacks or malicious behavior
• Are these assumptions correct ???
What is likely?

• Typical set
  ▫ The set of observed sequence $Y_{k}^{n+k}$ such that
  $$2^{-n(H(Y_{k}^{k-n})-\varepsilon)} \leq \Pr\{Y_{k}^{k+n}\} \leq 2^{-n(H(Y_{k}^{k-n})+\varepsilon)}$$
  ▫ From AEP
  $$\Pr\{A_{\varepsilon}^{(n)}\} \geq 1 - \varepsilon$$
  $$(1 - \varepsilon)2^{-n(H(Y_{k}^{k-n})+\varepsilon)} \leq |A_{\varepsilon}^{(n)}| \leq 2^{n(H(Y_{k}^{k-n})+\varepsilon)}$$
Universal anomaly detector!

- A simple anomaly detector
  - Calculate the likelihood \( \Pr\{Y_{k}^{k+n}\} \) of an observed sequence
  - Check if
  - If not, it is an anomaly
- Non parametric and universal anomaly detector
  - How to calculate the likelihood?
  - Need simplifying hypothesis
Filtering

• What if the observations are independent!

\[ \Pr\{Y_{k+n}^k\} = (\Pr\{Y_k\})^n \]

- We have to transform the observation so that it become independent!
  • Basic signal processing question

Input signal → Pre-processing → Flattened signal → Anomaly detector
Traffic dynamics

- OD 60
- OD 28
- OD 13
- OD 63
- OD 105
- OD 54
OD dynamics

• We need to capture all correlations
  ▫ Temporal Correlation
    • Short-scale, daily, weekly, etc..
  ▫ Spatial Correlation
    • Same Pop traffic
Anomaly detector structure

- two generic steps
  - Entropy reduction/Filtering
    - Remove dependencies in data
    - Compress the data: Information bottleneck
    - Remove predictable component
  - Decision/Detection
    - Apply the typicality test on the resulting hopefully independent signal
    - Statistical test
      - In the Neyman-Pearson Framework
      - Or other
        - Bayesian, etc
Taxonomy of approaches

- **Parametric**
  - Assume a parametric dependencies model
- **Non Parametric**
  - Broke the dependencies in a universal way
Parametric Approaches

- Assumes a parametric structure for dependencies
  - Normal behavior model
- Calibrate the Model
  - MLE estimation
- Filter the model prediction from observation
  - Results in an IID gaussian innovation process
- Apply a Neyman-Pearson statistical test with $H_0$: the observation is compatible with a zero mean with given variance
Non parametric approach

- Apply a de-correlating transform
  - Random projection, hashing, Sketch, etc...
- Learn the resulting distribution of normal behavior
  - Clustering, SVM, etc...
- Detect anomalies by checking cluster membership
Parametric techniques
anomaly detection steps

- Modeling
- Filtering
- Decision
How can we find a model?

- **Physical models**
  - Direct relation with physical parameters
  - High order, approximative, need complete process knowledge, physical parameters should be known
  - For example using TCP models or Bianchi csma/ca model

- **System identification**
  - Based on input/output measured data
    - Inverse statistical problem
      - Find the model that most likely has generated the observed data
  - Simple and efficient, with relatively low knowledge
  - Limited validity (operating point, type of input), sensors, measurement noise, unknown model structure
  - Frequent framework in practice
Model structure

• First step is to define a correlation/model structure
  ▫ Integrating temporal and spatial correlation
• The traffic is a dynamic entity
  ▫ Assume a state vector
    \[ X_k = \left( x^1_k, x^1_{k-1}, \ldots, x^1_{k-T}, x^2_k, \ldots, x^2_{k-T}, \ldots, x^L_k, \ldots, x^L_{k-T} \right) \]
  ▫ Generic dynamical model
    \[
    \begin{aligned}
    X_{k+1} &= f(X_k) + g(U_k) \\
    Y_k &= h(X_k) + i(U_k)
    \end{aligned}
    \]
Linear approximation

• Linearizing the generic dynamic model
  ▫ Using Taylor series expansion
  \[ f(X_k) = f(0) + X_k \frac{\partial f}{\partial X_k} \bigg|_0 + X_k^2 \frac{\partial^2 f}{\partial X_k^2} \bigg|_0 + \cdots \]
  • Can extend the state vector to integrate powers of state
  \[ X_k^2, \cdots, X_k^N \]

• Any non-linear dynamical model can be represented by a linear dynamical model with large enough state vector
Linear Time Invariant systems

- After linearization

\[
\begin{align*}
X_{k+1} &= AX_k + BU_k \\
Y_k &= CX_k + DU_k
\end{align*}
\]

- A, B, C and D can change with time (Time variant system)
  - The system might be without input
  - Or have a random input!
Linear system representations

- **time-domain representation**
  - input/output
  - state-space

- **frequency-domain representation**
  - Frequency response
  - Pole and zero

- All four are equivalent!
  - Sometimes one of these representation is more useful
Time domain representations (1)

\[ y(k) = x(k) * h(k) = \sum_{l=0}^{k} x(k - l)u(l) \]

Input/output: \[ y(k) = H[x(k)] \]

If one knows the impulse response of a system, he can derive the answer of the system to any input.

\[ h(k) = H[\delta(k)] \]
Time representation(2)

- State space representation

\[
\begin{align*}
X_{k+1} &= AX_k + BU_k \\
Y_k &= CX_k + DU_k
\end{align*}
\]

- For random input \( \varepsilon_k \) iid
  - Autoregressive model when \( B=I \)
  - ARMA model in general

\[
\begin{align*}
X_{k+1} &= AX_k + B\varepsilon_k \\
Y_k &= CX_k + D\varepsilon_k
\end{align*}
\]
Frequency representation (1)

• Frequency response

\[ \mathcal{F}(h(k)) = H(\omega) \]
\[ Y(\omega) = H(\omega)X(\omega) \]

- For random signals

\[ S_X(\omega) = \mathcal{F}(R_{XX}(\omega)) \]
\[ S_Y(\omega) = |H(\omega)|^2 S_X(\omega) \]
Frequency representation (2)

- Z transform (or Laplace transform)
  
  \[ X(z) = \mathcal{Z}(x(k)) = \sum_{k=0}^{\infty} x(k)z^k \]

  \[ Y(z) = H(z)X(z) \]

  - Poles and Zeros values of \( z \) for which the value of the transfer function \( H(z) \) becomes infinity or zero respectively.
  - For stochastic processes the mean of \( z \)-transform gives the characteristic function
    - The moment of the distribution are derivative of the characteristic function.
Example

\[ x(k) = 0.91x(k - 1) + 0.068x(k - 2) + 0.04x(k - 3) + 0.084x(k - 4) \]
Model calibration

- Several approaches
  - Linear regression
    \[ x(k) = a_1 x(k-1) + a_2 x(k-2) + a_3 x(k-3) + a_4 x(k-4) \]
    - Issue with correlation between values
  - Correlation fitting
    - Equivalent to Maximum Likelihood for Gaussian data
    - Find the linear coefficient that will fit the correlation
      - Different definition of the correlation
        - Burg, Yule-Walker, etc...
    - Will use them in the lab
  - Approximation with Minimal Squared Error
    - PCA/Karhunen Loeve expansion
    - Similar to linear regression but take care of the correlation

- Spectral estimation = model calibration
Biased/unbiased correlation

Example: Autocorrelation function estimate of a zero mean discrete white noise $e(k)$ with $N = 500$.

Biased estimate

$$R_{ee}(h) \approx \frac{1}{N} \sum_{k=0}^{N-h-1} e(k)e(k+h)$$

Unbiased estimate

$$R_{ee}(h) \approx \frac{1}{N-h} \sum_{k=0}^{N-h-1} e(k)e(k+h)$$

![Graphs showing biased and unbiased correlation](a) and (b)
Principal Component Analysis

• Let’s suppose we have \((X_1, \ldots, X_K)\) correlated centered gaussian RV
  ▫ PCA finds an orthonormal basis \(P = (\phi_1, \ldots, \phi_K)\), where such that the variance along the principal components is maximized
  ▫ After applying PCA, one can write \(X\) in the new coordinate system as \(X = \sum_{j=1}^{K} Y_j \phi_j\)
  ▫ How to derive the basis? Using the covariance matrix \(\Sigma = E\{XX^T\}\) \(\Sigma \phi_i = \lambda_i \phi_i\) \(\text{var}\{Y_i\} = \lambda_i\)
PCA as an approximation

- Transform matrix
  \[ U = [\phi_1 | \phi_2 | \cdots | \phi_K] \]
  \[ Y = UX \]

- An L dimensional approximation of X can be derived if we choose L<K column of U
  \[ U_L = [\phi_1 | \phi_2 | \cdots | \phi_L] \]
  \[ \hat{X} = U_L^T Y_L = U_L^T U_L X \]
  \[ Y_L = U_L X \]
  \[ \hat{X} = \sum_{j=1}^{L} Y_j \phi_j \]
  \[ E \left\{ (X - \hat{X})^2 \right\} = \sum_{j=L+1}^{K} \lambda_j \]

- This gives the MMSE approximation of dimension L
Extension to Stochastic Processes

• Karhunen-Loeve expansion theorem
  ▫ one can rewrite a zero-mean stationary stochastic process $X(t)$ as an orthogonal series expansion
  \[ X(t) = \sum_{i=1}^{\infty} Y_i \Phi_i(t) \]
  ▫ Where $Y_i$ are pair-wise independent RV and $\Phi_i(t)$ are pair-wise orthogonal deterministic functions
  ▫ The KLE theorem for discrete processes is given by
  \[ X[k] = \sum_{i=1}^{\infty} Y_i \Phi_i[k] \]
How to derive KLE basis

- For time continuous process
  \[ \sum_{i=1}^{K} \int_{0}^{\infty} \sigma_{i,l}(s) \Phi_{i,j}(s-t) ds = \lambda_{l,j} \Phi_{l,j}(t), \quad j > 0 \]

- Galerkin approximation
  - Apply PCA to the following matrix

\[
\begin{bmatrix}
X_1[k] \\
\vdots \\
X_K[k]
\end{bmatrix} = \sum_{i=1}^{K} \sum_{j=1}^{N} Y_{i,j} \Phi_{i,j}[k]
\]

\[
X = \begin{bmatrix}
x_1(1) & \cdots & x_1(n-N) \\
x_1(2) & \cdots & x_1(n-N+1) \\
\vdots & \ddots & \vdots \\
x_1(N) & \cdots & x_1(n) \\
\vdots & \ddots & \vdots \\
x_K(1) & \cdots & x_K(n-N) \\
x_K(2) & \cdots & x_K(n-N+1) \\
\vdots & \ddots & \vdots \\
x_K(N) & \cdots & x_K(n)
\end{bmatrix}
\]
KLE for modeling

- Let’s assume a signal with model
  \[ X_{k+1} = AX_k + \varepsilon_k \]
  \[ Y_{k+1} = U^T AUY_k + U^T \varepsilon_k \]

- In PCA coordinates

- Using KLE approximation

\[
\begin{align*}
\mathcal{Y}_R[k+1] &= A_M \mathcal{Y}_R[k] + B_M \varepsilon[k] \\
\hat{\mathbf{x}}[k] &= U_{ML} \mathcal{Y}_R[k] + D_M \varepsilon[k] \\
A_M &= U_{LM}^T A U_{LM}; B_M = U_{LM}^T [(A S + I) \mid - S] \\
S &= -A^{-1} + U_{LM} A_M^{-1} U_{LM}^T; D_M = S.
\end{align*}
\]

- Effects on poles and zeros

\[
\begin{align*}
\frac{\hat{\mathbf{x}}(z)}{\varepsilon(z)} &= (z I - A)^{-1} \\
\frac{\hat{\mathbf{x}}(z)}{\varepsilon(z)} &= U_{ML} (z I - A_M)^{-1} B_M + D_M
\end{align*}
\]
Normal behaviour?

- What is the normal behavior characterized by the model
  - Time and frequency representation are equivalent
PCA model
Second step : Filtering

- Filter separate the signal space into two subspace
  - Signals that pass through the filter
  - Signals that are rejected by the filter
- Anomaly detection filter
  - Pass everything is compatible with the normal behavior
  - Block everything is not compatible with the normal behavior
Simplistic approach

• Deduce the signal as predicted by the normal behavior model from the observation.

\[ e(t) = X(t) - \hat{X}(t) \]

• What to do if we have measurement error
  ▫ How to ensure that the error is not propagating in prediction
Kalman Filter

• Filtering what is compatible with the model
  ▫ Two steps
    • Prediction
      \[
      \hat{X}^{-}(t+1) = A \hat{X}(t)
      \]
    • Correction
      \[
      \hat{X}(t+1) = \hat{X}^{-}(t+1) + K(t+1) \left( Y(t+1) - C \hat{X}^{-}(t+1) \right)
      \]
  ▫ Innovation:
    \[
    \eta(t+1) = Y(t+1) - C \hat{X}^{-}(t+1)
    \]
Kalman filter effect
Kalman filter results
Innovation

Recalibration needed
Recalibration!
Innovation after recalibration
Remarks about Kalman Filter

- When the signal and the noise are gaussian the Kalman filter is the MMSE conditional filter
  - Different from KLE that was the non-conditional MMSE
  - The innovation process is a centered iid gaussian process with known variance
    - Variance derived by the Kalman filter
- Because of feedback Kalman filter is somewhat robust toward deviation from hypothesis
  - The innovation might have some residual dependencies and the distribution would not be gaussian
    - The resulting ROC curve would have slightly lower performance
- For more general settings particle filter seems to be the solution
Decision step

- For gaussian processes the innovation is a centered iid gaussian process with known variance
  - If the observations follow the normal behavior model innovation should be compatible with the given gaussian parameters.
- For PCA we have asymptotic theorems that state that the filter output converge to gaussian
  - Based on Central Limit Theorem
    - More generally convergence of measure
Hypothesis Testing

\[ H_0 : \theta \in \Theta_0 \]
\[ H_a : \theta \in \Theta_a \]

• Define a decision function \( T = r(x) \) and a decision region \( R = \{ x : T > c \} \) for some constant threshold \( c \).

• Possible error
  ▫ Type I Error:
    • Rejecting \( H_0 \) when \( H_0 \) is true
    • False alarm
  ▫ Type II Error:
    • Accepting \( H_0 \) when \( H_0 \) is false
    • Misdetection
Test Result

- without anomaly
- With anomaly

False Positives
True negatives

Test Result
Test Result

False negatives
Neyman-Pearson Theorem

- when performing a hypothesis test between two point hypotheses $H_0: \theta = \theta_0$ and $H_1: \theta = \theta_1$, then the likelihood-ratio test which rejects $H_0$ in favour of $H_1$ when

$$r(x) = \frac{l(\theta_0|x)}{l(\theta_1|x)} \leq \eta,$$  

where $\Pr\{r(x) \leq \eta|H_0\} = \alpha$

is the most powerful test of size $\alpha$ for a threshold $\eta$.

- There is a fundamental trade-off between false alarm and misdetection probability
  - ROC curve present this trade-off
Moving the Threshold: right
Moving the Threshold: left
True Positive Rate (sensitivity)

False Positive Rate (1-specificity)

ROC curve
ROC curve comparison

A good test:

A poor test:
Best Test:

Worst test:

The distributions don’t overlap at all

The distributions overlap completely
Area under ROC curve (AUC)

- *Overall measure* of test performance
- *Comparisons* between two tests based on differences between (estimated) AUC
- For continuous data, AUC equivalent to *Mann-Whitney U-statistic* (nonparametric test of difference in location between two populations)
AUC for ROC curves

AUC = 100%

AUC = 50%

AUC = 90%

AUC = 65%
Interpretation of AUC

- AUC can be interpreted as the probability that the test result from a randomly chosen anomaly is more indicative of anomaly than that from a randomly chosen non-anomalous sample: $P(X_i \geq X_j \mid D_i = 1, D_j = 0)$

- So can think of this as a nonparametric distance between anomalous/normal test results
Evaluation

• Dataset
  ▫ Two weeks of Netflow data from one border router of a medium-size ISP that connects Swiss universities and research labs

• Ground truth
  ▫ Manually labeling of dataset by visual time series inspection and top-n queries on the data itself

• Metrics
  ▫ Total of 14 different metrics (volume and entropy) computed at 15 minute intervals
  ▫ Data comes from single link
ROC curves for PCA (varying $k$)

ROC Curves:
- Plot False Positive Rate (FPR) vs. True Positive Rate (TPR) for different thresholds
- One curve per $k$ setting
- FPR is logarithmic

FPR: 1%
max(TPR): 30%
Explanation for bad PCA Performance

- Problems with PCA identified by Ringberg et al.
  - Dimension $k$ of projection space is hard to set
  - Projection space is polluted with large anomalies
  - But... no solution provided
- Additionally we find that...
  - Convergence to a stable distribution does not happen unless the image is uncorrelated
  - Traffic data are spatially and temporally correlated
  - Just removing spatial correlation is not enough!
    - Temporal correlation subsists ➔ No convergence happens
    - The optimal projection subspace rotates with time
MSE Distribution (PCA, $k=2$)

PDF of Decision Variable

Normal Probability Plot

heavy-tail and bias
Solution: Remove Temporal and Spatial Correlation

- Replace PCA by Multidimensional Karhunen-Loeve Expansion
- Classical PCA:
  - Diagonalize $KxK$ matrix where $K$ is the number of correlated RV
- Discrete Multidimensional KLE:
  - Diagonalize a $NKxNK$ matrix where $N$ is the correlation range
  - KLE with $N=1$ corresponds to classical PCA
Distribution of the Decision Variable
(KLE, $L=2/M=5$)

less bias, no heavy-tail
ROC Curves for PCA/KLE ($L/k = 2$)

- FPR: 1%
- TPR(PCA): 30%
- TPR(KLE): 43%

$M = \text{Temporal correlation range}$
ROC Curves for PCA/KLE ($L/k = 3$)

- FPR: 1%
- TPR(KLE): 58%
- FPR: 1%
- TPR(PCA): 17%

$M = \text{Temporal correlation range}$
ROC Curves for PCA/KLE ($L/k = 4$)

- FPR: 1%
- TPR(PCA): 15%
- TPR(KLE): 49%

$M = \text{Temporal correlation range}$
Anomaly recognition

- Traffic model
- Tracking system
- SNMP measurements
- Traffic matrices estimates
- Innovation process
- Matched filter bank
- Decision
- Attack detection
- OD flow
  Measure
- Traffic modelling
- Multiscale analysis
- Change detection
PACKET SAMPLING AND ANOMALY DETECTION
Motivations

• Packet sampling is mandatory
  ▫ Router CPU time is expensive
  ▫ In practice packet sampling rate up to 1:1000!
• Literature report that packet sampling hinder highly anomaly detection
  ▫ How can we do anomaly detection in presence of packet sampling
  ▫ Can we relate AD performance to sampling rate
    • Similar to the theory of detection in noise.
Motivation

Data Preprocessing

Packet Sampling \hspace{1cm} Aggregation \hspace{1cm} Time Series

Anomaly Detection

Entropy Reduction \hspace{1cm} Anomaly Decision

Impact?
Data Preprocessing

- (Random) Packet Sampling
  - Why?
    - To reduce amount of packet capture by the capturing device
    - To cope with link speed, CPU load, memory speed
  - How?
    - Selects randomly packets with probability $p$

- Aggregation
  - Why?
    - Translate to granularity of interest
    - Filtering fluctuations
    - Data compression
  - How?
    - Summing up all packets within a window of size $w$ and derive temporal mean
      - Granularity of interest in the order of several $w$
      - The larger the window, the smaller the fluctuations
Generic anomaly detection structure

- Anomaly detection occurs in two steps
  - Entropy reduction
    - Remove the irrelevant information from the data
  - Anomaly Decision
    - Use a statistical test for decision
      - $H_0$: is the observation compatible with what we will see without anomalies
Entropy reduction

• Parametric approaches
  ▫ Through filtering a parametric model of expected normal behavior from observation
    • Parametric model obtained through
      • MMSE : PCA
      • ML : Linear Dynamical System Estimation
    • Use of Kalman/particle filter or simple substractor to remove predictable part and generate uncorrelated signal

• Non parametric approaches
  ▫ Applies a generic entropic reduction
    • Hash function, sketches, etc...
  ▫ Generate almost uncorrelated signal
Detection step

• Uses signal after entropy reduction
  ▫ Parametric case
    • The resulting signal is an uncorrelated signal with known distribution
      • Frequently gaussian signal
        ▫ For PCA, Kalman Filter
    • The test is a simple mean statistical test
  ▫ Non parametric case
    • The resulting signal is an almost uncorrelated signal with unknown distribution
      • We have first to learn the distribution
    • Frequently becomes a distribution fitness test
      • Stein Lemma framework
• ROC curve captures trade-off between false positives and false positives
  ▫ Issue with ground truth
Assumptions

• We analyse the parametric framework
  ▫ Framework of PCA, Kalman filter, etc...
• Normal behavior model relies on estimation of the correlation structure
  ▫ Spatial and temporal
  ▫ Anomaly detection needs correct correlation structure estimation
• But: packet sampling and aggregation modify temporal correlation of packet signal
  ▫ How to quantify modification and determine its impact?
Spectral estimation

- If the spectrum of a signal is not well estimated, its temporal evolution model needed for AD cannot be well estimated.
- Analysing effect of sampling on signal spectra gives the effect of sampling on all representation using the correlation structure

\[ S_X(\Omega) = \mathcal{F}(R_X(\tau)) = \int_{-\infty}^{+\infty} X(t)e^{j\Omega t} dt. \]
Spectra of internet traffic

- From a layer 3 perspective Internet traffic is
  \[ X(t) = \sum_i L_i \delta(t - T_i) \]
  - If packet arrival forms a renewal process
    \[ S_X(\Omega) = \lambda \Phi(\Omega) - \lambda^2 \mathbb{E}\{L\}^2 \delta(\Omega) \]
  - This is not case in Internet
  - In reality we have short pulses
    - In a 1 Ghz link with minimal packet size of 60 Bytes the bandwidth is in the order of several Mhz.
Effect of sampling on spectra

- Regular sampling
  - Spectra is reproduced at fixed interval
    - Risk of aliasing if the bandwidth is more than \(2\pi f\) ≤ \(\Omega_S\)
    - Perfect reconstruction by using a low pass filter with a bandwidth \(2\pi B ≤ 2\pi f ≤ \Omega_S\)

- Irregular sampling
  - Packet thinning: random choice with probability \(p\)
    - \(S_X(\Omega) = E\{Z\}^2 S_X(\Omega) + \lambda \text{Var}\{Z\}\)
    - No aliasing anymore but white noise added
    - Proportional to \(p/(1-p)\)
Verification on Synthetic Trace (Packet Sampling)
Validation real trace

Estimated PSD for different random sampling probability

- No sampling
- 1:2 sampling
- 1:5 sampling
- 1:10 sampling
- 1:100 sampling

Frequency in MHz

PSD (dB/Hz)
Aggregation effect

- **Aggregation**
  
  \[ X_{\Xi}(t) = \frac{1}{\Xi} \int_{t-\Xi}^{t} X(s)ds. \]

  - A rectangular window + an aggregator + a regular sampler
    
    \[ S_{X_{\Xi}}(\Omega) = \frac{1}{\Xi} \sum_{k=-\infty}^{k=+\infty} \text{sinc}^2 \left( \frac{(\Omega - k\Omega_{\Xi})\Xi}{2} \right) S_X(\Omega - k\Omega_{\Xi}) \]
    
    - Linear distortion by the sinc term
    - Equivalent to a non-sharp low-pass filter
    - Reproduction of the spectra
    - Beware of aliasing with side lobes !!!!!!!
Signal processing refreshment

- **ALWAYS USE A PRE-ALIASING FILTER BEFORE DOING ANY TEMPORAL SAMPLING**
  - Pre-aliasing filter ensure that aliasing does not occur and the signal is not corrupted
- **Realistic scenario**
  - Packets captured over a 1 Gbps link with minimal packet size of 40 bytes
    - 6 Mhz bandwidth
  - Aggregation of 1 sec
    - Sampling rate of 1 hz for a signal of 6 Mhz Bandwidth !!!!
      - Aliasing will surely occurs
Verification on Synthetic Trace
Our proposition

• In place of an aggregation use a sharp low pass filter as our signal processing profs learn us to do!
  ▫ The bandwidth of the low pass filter should be chosen in the order of $1/(2T)$
    • $T$ is the timescale of interest
Validation (MAWI trace)
Structure

- Packets on the link
  - Packet Sampling
    - Sampling rate $p$
  - Aggregator
  - Low Pass Filter
  - Temporal Sampling rate $T$
    - Discrete Traffic Signal
  - Bandwidth $B$
  - Anomalies
    - Anomaly Decision
      - Innovation Process
        - Decision
          - Threshold $D$
            - AR Model
              - Model Calibration
                - Model order $k$
Road Map

- Show that main problem is aliasing introduced by aggregation (packet sampling is inevitable)
- Introduce our solution to use low-pass filter instead of aggregation
- Derive fundamental trade-off between sampling rate and detectable anomaly scale
- Evaluate proposal on synthetic and real traffic traces
SNR Verification on MAWI Trace

- Theoretical SNR

\[ SNR = \frac{\mathbb{E}\{Z\}^2}{\lambda B \text{Var}\{Z\}} \int_{-B/2}^{B/2} S_X(\Omega) d\Omega \]

- Measured SNR

Each point corresponds to different sampling rate.
Evaluation Outline

- Show impact on entropy reduction step for Kalman filter and MAWI trace
- Theoretically derive upper bounds for false positive and false negative probabilities
- Show impact on anomaly decision step for MAWI trace with synthetic anomalies
- Show impact on anomaly decision step for MetroSec trace with real anomalies
Effect of sampling on modeling

- We use an AR model for normal behavior of traffic
  \[ y[k] = \sum_{i=1}^{p} \alpha_i y[k - i] + \epsilon_i \]
  - We use a model of order 5 to 7 chosen by a MDL criterion
  - Estimation process give an estimation error variance value
Effect of sampling on the model
Modelling noise
AR Model and Kalman Filter Frequency Response (MAWI trace)
Effect on the detection step

- Input to Kalman filter $a[k] + n[k]$
- Anomaly effect $e_a[k] = a[k] * w[k]$  
  $M(a) = \max |e_a[k]|$
- Misdetection probability $P^M_D = 2 * Q\left(\frac{M(a) - D}{\sqrt{V_i(p)}}\right)$
  - The noise variance can be estimated by theoretical formula proportional to $p/(1-p)$
- False alarm probability $P^{FA}(p, D) < 2 * Q\left(\sqrt{p\frac{D}{\sqrt{C}}}\right)$
False Negative Rate for MAWI Trace
True Positive Rate for MetroSec Trace

Bandwidth of filter reduced by same factor as sampling rate
Readjustement

• Trade-off between sampling rate and bandwidth of interest
  ▫ When the sampling rate decrease, the noise in the system increase
  ▫ By reducing the bandwidth of the LPF we can control the amount of noise in the system and fix an SNR
How to implement the low pass filter

- Generate the packet arrival signal
  \[ X(t) = \sum_i L_i \delta(t - T_i) \]

- Digital approach
  - A cascade of decimation (division of sampling rate by 2) and applying an FIR LPF
  - Have a non negligible complexity as at least 6 level of cascade are used

- Analog approach
  - Use an D/A to generate an analogic packet arrival signal
  - Apply an analogic lowpass filter
    - With capacitance and coils 😊
  - No complexity at all, just real electrical engineering
Non parametric techniques
Symbolic view of traffic

- **Common view:**
  - Timeseries of volume
  - Counts of bytes/
  - Packets

- **Symbolic perspective**
  - Each packet results in a symbol
    - derived from packet content
- **Symbol sequence** $X = (x_1, x_2, \ldots, x_n)$
  - Histogram or Type
    - $|\text{types}| = O(n^k)$ but $|\text{sequences}| = O(k^n)$
  - largest type has almost all sequences
Issues with symbolic view

• The space is huge
• Types do not converge to something stable
• Solution
  ▫ To compress the space
    • Using coarser view of the space by symbol aggregation
      • Equivalent to choosing rectangular bins
    • Using hashes
      • Equivalent to using random binning
Anomaly detection

- The normal behaviour of a network should be assessed as a global joint distribution
  - An anomaly will have effects at different place of the network
- Basic underlying question how to infer the joint distribution in a distributed way?
  - In an efficient way
Classical distribution estimation

- Simplest density estimator: multidimensional histogram.

- Rectangular binning of the data.
Rectangular grid issues

• Quality of the binning depends of
  ▫ bin size
  ▫ origin
  ▫ grid orientation
    • following principal directions
    • A single grid could either be quite good or quite bad, depending how much it is oriented like the data.
Mosaic-based histogram

- A rectangular binning can be viewed as a set of regularly-spaced hyperplanes on each dimension, each bin being an hyper-rectangle delimited by hyperplanes.
Random Mosaic

• Throw randomly the same number $M$ of hyperplanes to obtain a mosaic covering the data space
Mosaic-based binning

- The binning $Y_i$ of the data sample $X_i$ on each local site is made very simply:
  $$Y_i = \text{sgn}(A \cdot X_i - b)$$

where $A$ and $b$ are random matrices.

- Only need to transmit number of hyperplanes and random seed to each local site!
Mosaic-based binning

- We can associate a binary numerotation to each of the regions (convex polytopes) in a natural way:
Density estimation

Let the system of cells be $\{C_j\}$ Then the estimate at a point $t \in \mathbb{R}^d$ is

$$\hat{f}(t) = \frac{1(X \in C_j)}{N_X \cdot Vol(C_j)}, \quad t \in C_j$$
Spatial distribution

• Distribution of the quartet (@SRCIP,#SRCPRT,@DSTIP,#DSTPRT)
  ▫ Hard to obtain
    • Too many bins
    • Not really stable

• Sketch based approach
  ▫ Regroup header through random hash functions
    • Results in small number of bins
      • For example 16 hash function
    • Stable behavior
      • Because of aggregation
  ▫ Because of random hash function it is robust to evasion
Comparison of regular vs. Mosaic.

- **Regular**: Mean $1.3 \times 10^{-5}$
- **Mosaic-based**: Mean $3.9 \times 10^{-6}$
Back to anomaly detection

- This performs better than a fixed grid when the local distributions differ.

- Anomaly detection:
  - fix a mosaic.
  - periodically update the density estimate
  - if the current density estimate $f_t(\cdot)$ is “too far” from the density of the “normal” state $f_0(\cdot)$
    => Anomaly
Distribution distance

- Kullback-Leibler distance is easier to evaluate

\[
D(f_0 || f_t) = \sum_{\text{cells } C_j} f_0(C_j) \cdot \log \left( \frac{f_0(C_j)}{f_t(C_j)} \right)
\]

since we can take it as a discrete distribution over the bins.
Distance test

• Sanov theorem

\[ \Pr \left\{ \hat{P}_n \text{ is close to } P \right\} \sim \exp\left( -nD\left( \hat{P}_n \parallel P \right) \right) \]

  ▫ Valid for iid processes

• Extension to Process with memory

\[ \Pr \left\{ \hat{P}_n \text{ is close to } P \right\} \sim \exp\left( -nI\left( \hat{P}_n \parallel P \right) \right) \]

  ▫ \( I(\hat{P}_n \parallel P) \) is the large deviation rate function for the memory structure
  ▫ Could be obtained for HMM structure
Anomaly detection methodology

- **Steps**
  - **Reference window analysis**
    - Map header values using sketches to a small number of values
    - Calibrate an HMM over a reference window of size $L$
    - Use this HMM as the reference memory structure
  - **Running window analysis**
    - Map header values using sketches to a small number of bins
    - Obtain the histogram for a window size $L$
    - Compare the histogram to the HMM obtained over a reference period
  - **Make a decision test**
    - $H_0$: the observed histogram can come from the reference HMM
    - Use the Extension of Sanov theorem to detect the anomaly
Comparison of KL distance
Application to anomaly detection
Anomalies show up the tail
References

- Combining Filtering and Statistical Methods for Anomaly Detection.