REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
AUGUST 2010	Conference Paper (POSTPRINT)	February 2008 – August 2010

4. TITLE AND SUBTITLE

ENTROPY-BASED HEAVY TAILED DISTRIBUTION TRANSFORMATION FOR NETWORK TRAFFIC ANALYSIS

5a. CONTRACT NUMBER
In House

5b. GRANT NUMBER
N/A

5c. PROGRAM ELEMENT NUMBER
61102F

6. AUTHOR(S)

Keesook J. Han

5d. PROJECT NUMBER
231G

5e. TASK NUMBER
IH

5f. WORK UNIT NUMBER
02

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

AFRL/RIGG
525 Brooks Road
Rome, NY 13441-4505

8. PERFORMING ORGANIZATION REPORT NUMBER
N/A

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

AFRL/RIGG
525 Brooks Road
Rome NY 13441-4505

10. SPONSOR/MONITOR'S ACRONYM(S)
N/A

11. SPONSORING/MONITORING AGENCY REPORT NUMBER
AFRL-RI-RS-TP-2010-31

12. DISTRIBUTION AVAILABILITY STATEMENT
APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED. PA #: 88ABW-2010-1908
Date Cleared: April 7, 2010

13. SUPPLEMENTARY NOTES
 This Article was published in the Proceedings of the ASME 2010 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference IDETC/CIE 2010. August 15-18, 2010, Montreal, Canada.

14. ABSTRACT
In general, network traffic data has a heavy-tailed probability distribution. The Entropy-Based Heavy Tailed Distribution Transformation (EHTDT) has been developed to convert the heavy tailed network traffic data distribution into a transformed probability distribution. In practice, the entropy distribution of the transformed probability distribution exhibits a type of linearity that gives rise to an eigenstructure that allows the characterization of network traffic data to effectively lossily compress network traffic data via the Rate Controlled Eigen-Based Coding. The aforementioned eigenstructure is motivated by singular value decomposition theory. A very high compression ratio can be achieved by the proposed method. Results of applying the methods to real network traffic data network traffic data are presented.

15. SUBJECT TERMS
Entropy-based Heavy Tailed Distribution Transformation, Network Traffic Analysis, Anomaly Detection, Principal Component Analysis, Singular Value Decomposition, Compression

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Keesook J. Han
U	U	U	UU	7	19b. TELEPHONE NUMBER (*Include area code*) N/A

POSTPRINT

Proceedings of the 2010 ASME 2010 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference
IDETC/CIE 2010
August 15-18, 2010, Montreal, Quebec, Canada

DETC2010-28522

ENTROPY-BASED HEAVY TAILED DISTRIBUTION TRANSFORMATION FOR NETWORK TRAFFIC ANALYSIS

Keesook J. Han
Air Force Research Laboratory
Rome, NY 13441, USA
Email: Keesook.Han@rl.af.mil

ABSTRACT

In general, network traffic data has a heavy-tailed probability distribution. The Entropy-Based Heavy Tailed Distribution Transformation (EHTDT) has been developed to convert the heavy tailed network traffic data distribution into a transformed probability distribution. In practice, the entropy distribution of the transformed probability distribution exhibits a type of linearity that gives rise to an eigenstructure that allows the characterization of network traffic data to effectively lossily compress network traffic data via the Rate Controlled Eigen-Based Coding. The aforementioned eigenstructure is motivated by singular value decomposition theory. A very high compression ratio can be achieved by the proposed method. Results of applying the methods to real network traffic data network traffic data are presented.

INTRODUCTION

Intrusion Detection Systems (IDS's) must be capable of detecting unknown attacks. The problem with building an anomaly detection model is that observed activities deviate significantly from established normal usage profiles. Reliable anomaly detection modeling requires training huge datasets regularly in order to learn legitimate behaviors. There is an enormous cost in collecting, storing, and analyzing intrusion datasets. A difficult problem in handling intrusion detection data is that one is not able to store and manage efficiently a huge amount of intrusion detection data with the current data mining and data management technologies.

In general, anomaly detection usually involves computation on massive datasets. There has been an increased interest in data mining based approaches for intrusion detection. The major difficulty of data mining is that it is computationally expensive to find correlations between attributes in massive intrusion detection datasets. It is desirable to perform statistical processing on reduced datasets instead of the original full datasets. The reduced data sets must of course contain enough information for effective segmentation and classification. To efficiently measure similarity in appearance within object classes, one must first determine which features are most effective at describing anomalies of objects. A standard linear method for data feature extraction is that of principal component analysis (PCA). This reduction is achieved by selecting the first few principal components. These components capture the most relevant features use to classify a group of objects to be recognized. However, intrusion detection technologies based on PCA are still immature because of dynamic behaviors and heavy tailed distributions in network traffic.

The study of heavy tailed distributions in network traffic has been an important research topic in various network applications [1][2][4][5][6][7][8][10]. Characterization of heavy tailed network traffic plays a critical role to improve the Quality of Services. More efficient intrusion detection data modeling and management methods are required to characterize heavy tailed network traffic data with greater reliability and faster retrieval rates.

This paper provides combined network traffic characterization and the PCA approaches that are applied to minimize model complexities and maintenance problems in IDS design. The proposed Entropy-Based Heavy Tailed Distribution Transformation and the Rate Controlled Eigen-Based Coding method are effective methods to extract meaningful features from heavy tailed datasets. These feature

extraction functions are useful for traffic analyzers and intrusion detection tools.

STATISTICAL ANOMALY DETECTION MODELING

There has been recently a big increase in the number of studies related to the statistical analysis to characterize traffic traces. One of the open problems in understanding the dynamic nature of network traffic is when the statistical distributions of traffic traces are non-Gaussian and heavy tailed [3][9]. Heavy tails refer to the power decrease of the marginal distributions. It is evident that many important problems with heavy tailed anomalies are poorly described by standard statistical models.

This research aims to develop a new statistical model to represent a heavy tailed distribution in a compact form with great generality and several feature extraction properties. A wide range of shapes of the distribution can be investigated by choosing the parameters. This approach is novel because these estimators are extracted to take advantage of the anomaly detection.

Future research will focus on temporal granularity and statistical characteristics, how to detect and measure these quantities and identify other potential characteristics, especially within apparent heavy tailed regions. In this approach, principal component based statistical characteristics are extracted from the heavy tailed distribution data, and stored in a database that is updated regularly and automatically to determine dynamic thresholds for discriminant functions.

ENTROPY-BASED HEAVY TAILED DISTRIBUTION TRANSFORM (EHTDT)

Network traffic characterization has been studied extensively, but an accurate characterization of network traffic still remains elusive due to difficulty of parameter estimations. This section describes the transformation procedures to characterize network traffic data. The simple transformation process has the ability to predict the behavior of large-scale network traffic. This section describes the transformation procedures with real network traffic data.

The plot of real network traffic connection is shown in Figure 1. Frequency and ordering properties of network traffic datasets are important features of anomaly detection models. The most common way to detect anomalies is to use statistical distributions represented by a discrete distribution with a specified number of bins and the relative frequency of a value appearing in that bin. Real network traffic exhibits heavy tailed distributions in Figure 2.

One of the most challenging characteristics of heavy tailed distribution is to parameter estimation. Known statistical procedures can be used to estimate parameters, but it is infeasible due to computational complexity for real-time network traffic characterization. This research addresses a new method for estimating the parameters of heavy tailed distributions using the EHTDT.

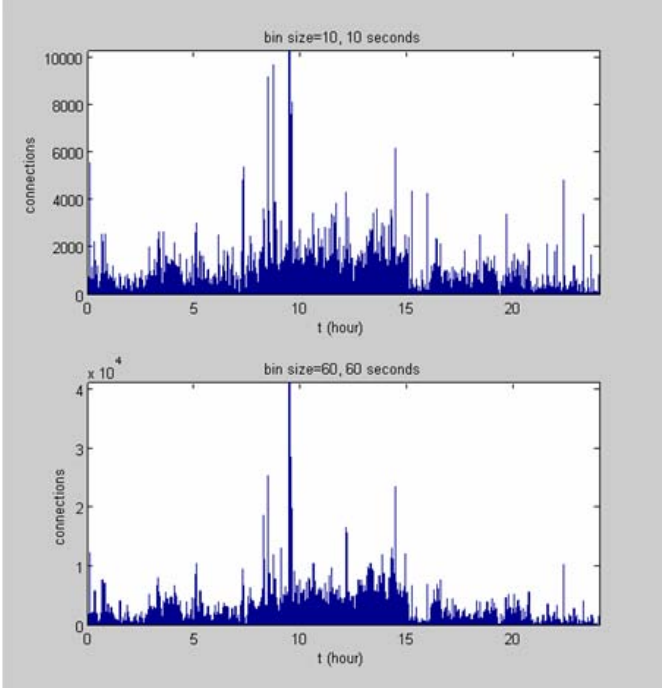

Figure 1. PLOTS OF DAILY CONNECTIONS.

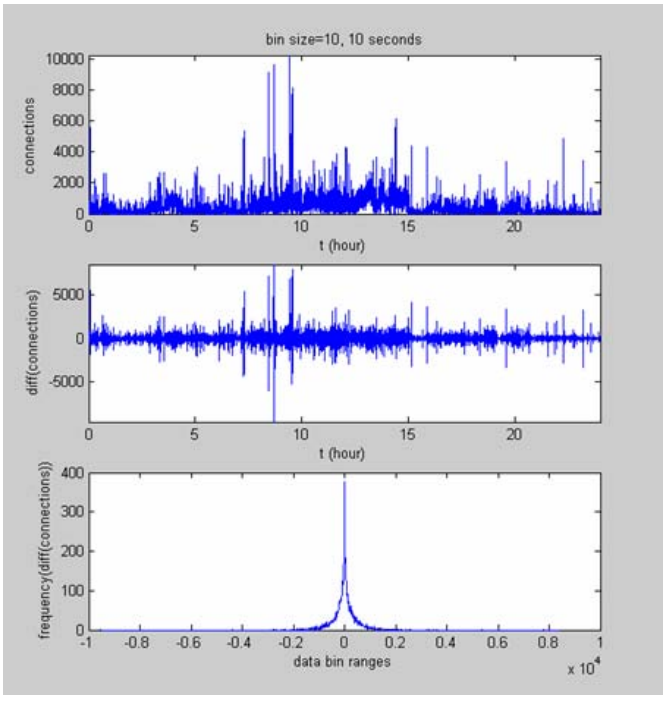

Figure 2. PLOTS OF DAILY CONNECTIONS, THE FIRST ORDERER DIFFERENCE ALONG CONNECTIONS AND HISTOGRAM OF HEAVY TAILS.

Note: diff(connections(t))=connections(t+1) - connections(t)).

.

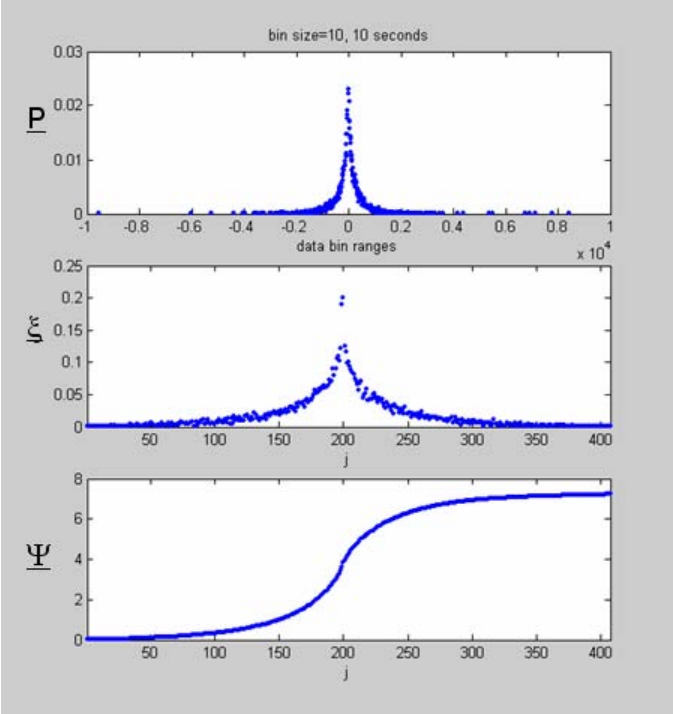

Figure 3. PLOTS OF PROBABILITY MASS FUNCTION (P), DIFFERENTIAL OF ENTROPY-BASED HEAVY TAILED DISTRIBUTION (ξ) AND ENTROPY-BASED HEAVY TAILED DISTRIBUTION (Ψ).

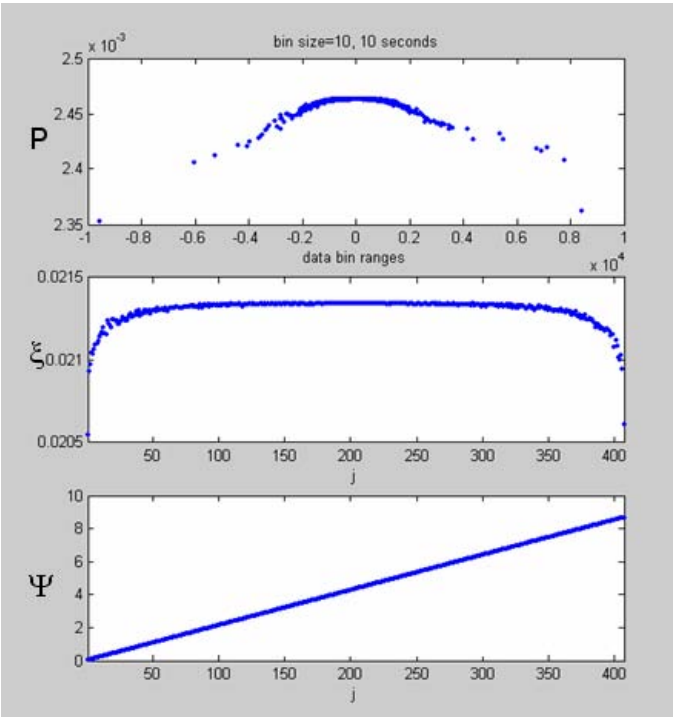

Figure 4. PLOTS OF TRANSFORMED PROBABILITY MASS FUNCTION (P), DIFFERENTIAL OF ENTROPY-BASED HEAVY TAILED DISTRIBUTION (ξ) AND ENTROPY-BASED HEAVY TAILED DISTRIBUTION (Ψ).

In general, network traffic data have heavy-tailed distributions. Power-law distributions are widely used for estimating packet interval time as well as in other networking applications and ε- contaminated (Gaussian-mixture) distributions are useful to detect anomaly network traffics. A power-law distribution is one-tailed and an ε- contaminated (Gaussian-mixture) distribution is two-tailed. For statistical anomaly detection, two-tailed distributions have been considered to derive the EHTDT.

Forward EHTDT

Figure 3 presents a probability mass function, a differential of entropy-based heavy tailed distribution, and entropy-based heavy tailed distribution. Experimental results indicate that these heavy tailed distributions are difficult to use to characterize network traffic. Hence, the Forward EHTDT has been developed for fast network traffic characterization.

A two-tailed probability mass function is defined as a probability vector

$$\hat{P} = (\hat{P}(1), \hat{P}(2), ..., \hat{P}(K-1), \hat{P}(K), \hat{P}(K+1), ..., \hat{P}(N)) \quad (1)$$

where $\hat{P}(K)$ is the unique maximum element of $\hat{P}$ and N is the maximum index number.

To characterize network traffic in a more compact form, the probability vector $\hat{P}$ is converted into a transformed probability vector P by the following two procedures:

The vector β is defined, such that

$$\beta(x) = \frac{1-\hat{P}(x)}{\lambda} \quad (2)$$

where $\lambda = \sum_{x=1}^{N}(1-\hat{P}(x))$ is the normalization factor, and $x = 1, 2, ..., N$.

The transformed probability vector P is then defined by

$$P = (\beta(K), \beta(K-1), ..., \beta(1), \beta(N), \beta(N-1), ..., \beta(K+1)) \quad (3)$$

where $P(1) = \beta(K)$ is the minimum element of P.

The Entropy-Based Heavy Tailed Distribution Ψ is defined by

$$\Psi(x) = \sum_{j=1}^{x} \xi(j) \quad (4)$$

where $\xi(j) = -P(j)\log_2 P(j)$ and $x = 1, 2, ..., N$.

Note that the last element of Ψ is the entropy of the transformed probability vector P.

$$\Psi(N) = \sum_{x=1}^{N} \xi(x) = -\sum_{x=1}^{N} P(x) \log_2 P(x) = H(P) \ . \qquad (5)$$

The main reason to transform data as part of a regression analysis is to achieve linearity. In practice, the proposed transformation provides approximate linearity as shown in Figure 4.

Inverse EHTDT

The Inverse Entropy-Based Heavy Tailed Distribution Transform can be determined by the following procedure. The first order differences of Ψ are used to determine ξ as follows:

$$\xi(x) = \Delta(\Psi) = \Psi(x) - \Psi(x-1) = -P(x) \log_2 P(x) \qquad (6)$$

where $\xi(1)$ is a stored parameter and $x = 2, 3, ..., N$.

It is emphasized that one can deal directly with P (or an estimate of P via an iteration technique), and then obtain the initial heavy tailed probability vector $\hat{P}$ from P. Inverting (3), the probability vector β is obtained as

$$\beta = (P(K), P(K-1), ..., P(1), P(N), ..., P(K+1)) \ . \qquad (7)$$

Then, $\hat{P}$ can be calculated via

$$\hat{P}(x) = 1 - \lambda \beta(x) \qquad (8)$$

where the normalization factor λ is a stored parameter and $x = 1, 2, ..., N$.

Very Low-Bit Rate EHTDT

The Entropy-Based Heavy Tailed Distribution Ψ can be decomposed as

$$\Psi = \Phi + \eta \ . \qquad (9)$$

The sum of a linear function Φ and a nonlinear function η where Φ is taken as

$$\Phi(x) = \xi(1) + \frac{\Psi(N) - \Psi(1)}{N-1}(x-1) \qquad (10)$$

for $x = 1, 2, ..., N$.

Φ will contain most of the energy of the heavy tailed information; the linear distribution of the heavy tailed approximation can be estimated with the entropy of the inverse distribution $\Psi(N)$ and $\xi(1)$. The nonlinear function η is then

$$\eta = \Psi - \Phi \ . \qquad (11)$$

The differential vector is given by

$$\varepsilon(x) = \eta(x) - \eta(x-1) \qquad (12)$$

where $x = 2, 3, ..., N$.

Data reduction and feature selection are two reasons to compress the nonlinear function η. Fourier coefficients, wavelet coefficients, and principal components are commonly selected for features. In this approach, the principal component analysis (PCA) will be applied to select features and analyze anomaly detection data sets. The estimated nonlinear function $\hat{\eta}$ can be determined with a few principal components.

The reconstructed function $\hat{\Psi}$ can be expressed as

$$\hat{\Psi} = \Phi + \hat{\eta} \ . \qquad (13)$$

The estimated heavy tailed probability vector $\hat{P}$ can also be determined by the Inverse Entropy-Based Heavy Tailed Distribution Transform. A very high compression ratio can be achieved by the proposed methods. A few principal components, $\Psi(N)$ and $\xi(1)$ are selected for features.

LOSSY COMPRESSION

Principal component analysis (PCA) is a popular technique in many areas of multivariate analysis. There are various generalizations of PCA such as multiple correspondence analysis (MCA), non-metric principal component analysis (NCA) and ordinary metric PCA. In correspondence analysis, the variables are linearly transformed to provide orthogonal solutions. First, we will briefly describe ordinary metric PCA and Singular Value Decomposition (SVD)-Based Coding. Then, the Rate Controlled Eigen-Base Coding for EHTDT is introduced. The proposed coding system will reduce the dimensionality of the data enormously and capture the effective feature structure.

Principal Component Analysis (PCA)

Suppose that $f_1, f_2, ..., f_M$ are $N \times 1$ observation vectors. Let μ be the mean vector of the observation vectors $f_1, f_2, ..., f_M$. Zero mean observation vectors are given by

$$\phi_x = f_x - \mu \qquad (14)$$

where $x = 1, 2, ..., M$.

The empirical covariance matrix S is computed as

$$\mathrm{S} = \frac{1}{M} \sum_{x=1}^{M} \phi_x \phi_x^T \ . \qquad (15)$$

The unique set of M orthonormal eigenvectors of S, $Q_M = [q_1, q_2, ..., q_M]$, and their associated eigenvalues, $\lambda_1, \lambda_2, ..., \lambda_M$ are computed. Linear combinations of the first L eigenvectors $Q_L = [q_1, q_2, ..., q_L]$ corresponding to the L largest eigenvalues (e.g., $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_L$) span the space of the zero mean observation vector to capture most of the relevant information in the input data. The projection of the vector ϕ_x

onto the lines spanned by the orthonormal basis $Q_L = [q_1, q_2, ..., q_L]$ is given by the following operation

$$p_x = Q_L^T \phi_x = (p_{1x}, p_{2x}, ..., p_{Lx})^T \quad (16)$$

where $1 \le x \le N$.

The elements of vector p_x are called the principal components. The $N \times 1$ principal component vector p_x contains compact information for f_x. The reconstructed vector $\hat{f}_x$ can be computed as

$$\hat{f}_x = Q_L p_x + \mu \quad (17)$$

where $1 \le x \le M$.

Singular Value Decomposition (SVD)

The singular value decomposition (SVD) is relevant to principal component analysis in several respects. The basic concept is to represent a given data matrix X of size $N \times K$. SVD is then applied to this matrix to obtain U, S, and V matrices. This compression operation is expressed in the following equations. Singular Value Decomposition is given by

$$\mathrm{X} = USV^T \quad (18)$$

where the dimensions of X, U, S, and V are $N \times K$, $N \times K$, $K \times K$, and $K \times K$, respectively.

Reconstructed data is computed by

$$\hat{X} = \hat{U}\hat{S}\hat{V}^T \quad (19)$$

where $L \le K$ and the dimensions of $\hat{X}, \hat{U}, \hat{S}$, and $\hat{V}$ are $N \times K$, $N \times L$, $L \times L$, and $K \times L$, respectively.

The columns of U are called the *left singular vectors*. The rows of V^T contain the elements of the *right singular vectors*. The elements of S are only nonzero on the diagonal, and are called the *singular values*. For example, if $rank(\mathrm{X}) = L$, then

$$diag(S) = (s_1, s_2, ..., s_L) \quad (20)$$

where $s_1 > s_1 > , ..., > s_L > 0$.

Note that for a square and symmetric matrix, the singular value decomposition is equivalent to diagonalization, or solution of the eigenvalue problem. The SVD-based compression method is popular to compress large data matrices.

The fundamental concept of the SVD-based compression scheme is to use a smaller number of dimensions to approximate the original matrix. The SVD does not provide a computationally efficient method of compression. However, the importance of using the SVD for principal component analysis is that SVD provides the standardized versions of principal component scores. Component scores are useful for correspondence analysis.

Rate Controlled Eigen-Based Coding for EHIDT

The classes of admissible transformations in SVD are different for different types of data. Admissible transformations should be found to minimize the appropriate loss function. It is common to calculate the principal components using a covariance matrix. The reason is that eigenvectors of a covariance matrix may provide admissible transformations. There are other ways of computing principal components. In one method, eigenvectors of a correlation matrix are used to compute principal components with standardized variables. However, the principal component based EHTDT coding is simply implemented. Even though the loss function is not minimized in the coding scheme, the scheme does yield a reduction in the computational complexity and misclassification rate.

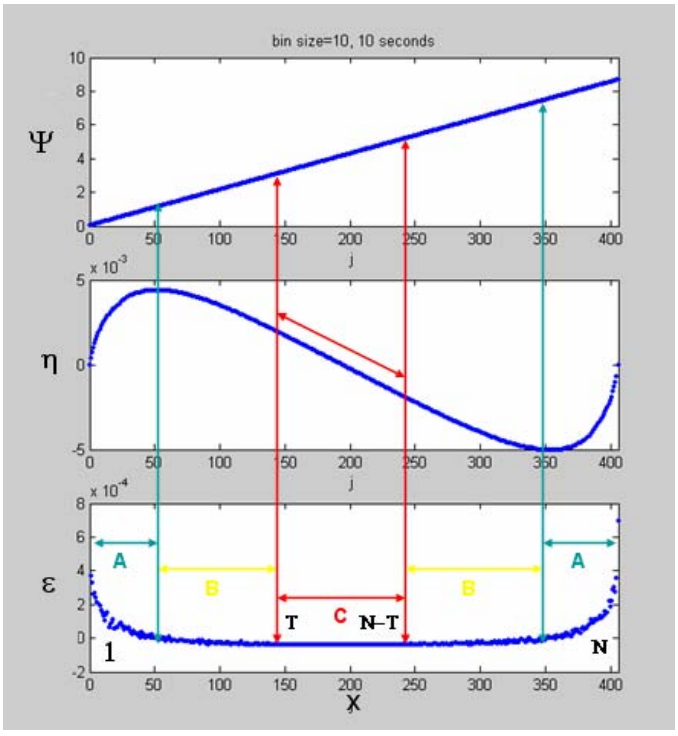

Figure 5. RATE CONTROLLED EIGEN-BASED CODING REGIONS.

Suppose that $f_1, f_2, ..., f_M$ are $N \times 1$ *nonzero mean observation* vectors

$$f(x) = \varepsilon(x) \quad (21)$$

where $x = \{2, 3, ..., \mathrm{T}, N-T, ..., N\}$ and $\varepsilon(x) = \eta(x) - \eta(x-1)$. T is a threshold determined from the $\varepsilon(x)$ plot as indicated in Figure 5. For x in between T and N-T, $\varepsilon(x)$ can be taken as

zero for all practical purposes; this gives $f(x)$ equal to zero in this range of x between T and N-T.

The *empirical covariance matrix* of the nonzero mean observation vectors is defined for computational simplicity.

The $N \times N$ *empirical covariance matrix* R is defined by

$$\mathrm{R} = \frac{1}{M} \sum_{x=1}^{M} f_x f_x^T \qquad (22)$$

where the correlation matrix R is almost always *nonsingular* and *symmetric*.

Instead of using the empirical covariance matrix S or ordinary empirical correlation matrix, the *empirical covariance matrix* of the nonzero mean observation vectors is defined for computational simplicity. The empirical covariance matrix S provides that the first principal component corresponds to a line that passes through the mean, and minimizes the mean square error of approximating the data. On the other hand, the *empirical covariance matrix* R of the nonzero mean observation vectors R provides an effective cluster separation for each streaming network traffic datasets. For anomaly detection, the *empirical covariance matrix* R minimizes computational complexity and maximize detection rate.

The unique set of N orthonormal eigenvectors is computed with the correlation matrix R and the corresponding L eigenvectors to form an $N \times L$ eigenvector matrix Q_L. The first L eigenvectors are $Q_L = [q_1, q_2, ..., q_L]$ and their associated eigenvalues are $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_L$.

The simple coding pairs are given by

$$P = Q_L^T X \qquad (23)$$

and

$$\hat{X} = Q_L P \qquad (24)$$

where $N \geq L$ and $X = [f_1, f_2, ..., f_M]$.

Note that the columns of $L \times M$ matrix $P = [p_1, p_2, ..., p_M]$ are the L *-dimensional* principal component vectors $p_1, p_2, ..., p_M$ for the N *-dimensional* vectors $\hat{f}_1, \hat{f}_2, ..., \hat{f}_M$ and the columns of $N \times M$ matrix $\hat{X} = [\hat{f}_1, \hat{f}_2, ..., \hat{f}_M]$ are the reconstructed N *-dimensional* vectors $\hat{f}_1, \hat{f}_2, ..., \hat{f}_M$.

This coding scheme is simpler than the *Karhunen-Lòeve* transform and other principal component analysis techniques.

CONCLUSION

Power-law distributions are widely used for estimating packet interval time as well as in other networking applications and ε - contaminated (Gaussian-mixture) distributions are useful to detect anomaly network traffics. The estimation of important tail characteristics is directly linked to the interpretation of the underlying network traffic. Since there is a limitation to estimate parameters of various heavy tailed distributions because of mixture distribution characteristics of heavy tailed network traffic, an efficient and practical parameter estimation technique has not been derived. For statistical anomaly detection, heavy-tailed probability distributions of network traffic data have been proposed to mitigate the limitation of parameter estimation. In this work, the EHTDT transform converts such a heavy tailed distribution into a transformed probability distribution more amenable for lossy compression of network traffic data.

Experimental results indicate that a compact characterization of heavy tailed network traffic data can be achieved by the EHTDT transform and the Rate Controlled Eigen-Based Coding approaches. Efficient intrusion detection data modeling can be developed by the proposed approaches using various network traffic features.

REFERENCES

[1] Baiardi, F., Telmon, C., and Sgandurra, D., 2009, "Modeling and managing risk in billing infrastructures". In Critical Infrastructure Protection III, IFIP Advances in Information and Communication Technology, Vol. 311, C. Palmer and S. Shenoi, eds., Boston: Springer, pp. 51-64.

[2] Crovella, M., 2001, "Performace Evaluation with Heavy Tailed Distributions". In Lecture Notes in Computer Science; Vol. 2221. London: Springer-Verlag, pp. 1-10.

[3] Dasgupta, A. Hopcroft, J., Kleinberg J., and Sandl, M., 2005, "On learning mixtures of heavy-tailed distributions". In Proceedings of the 46th Annual IEEE Symposium on Foundations of Computer Science, pp. 491-500.

[4] Dainotti, A., Pescapè, A., and Ventre, G., 2006, "A packet-level characterization of network traffic". In Proceedings of the 11th IEEE International Workshop on Computer-Aided Modeling, Analysis and Design of Communication Links and Networks.

[5] Elleithy, K., and Al-Suwaiyan, A., 2001, "Network traffic characterization for high-speed networks supporting multimedia". In IEEE Proceedings of the 34th Annual Simulation Symposium, pp. 200.

[6] Kornexl, S. , Paxson, V., Dreger, H., Feldmann, A., and Sommer, R., 2005, "Building a time machine for efficient recording and retrieval of high-volume network traffic". In Proceedings of the 5th ACM SIGCOMM Conference on Internet Measurement, pp 267-272.

[7] Maillart1, T., and Sornette1, D., 2009, "Heavy tailed distribution of cyber-risks". URL http://arxiv.org/PS_cache/arxiv/pdf/0803/0803.2256v2.pdf.

[8] Rezaul, K., and Grout, V., 2009, "An approach for characterising heavy-tailed internet traffic based on EDF statistics," in Intelligent Engineering Systems and Computational Cybernetics. Netherlands: Springer, pp. 173-184.

[9] Vempala, S., and Wang, G., 2004, "A spectral algorithm for learning mixture models". Journal of Computer and System Sciences archive, vol. 68, Issue 4, Special issue on FOCS, pp. 841-860.

[10] Meza, J. , Campbell, S., and Bailey, D., 2009, "Mathematical and Statistical Opportunities in Cyber Security". URL http://arxiv.org/PS_cache/arxiv/pdf/0904/0904.1616v1.pdf.