Defence Science Journal, Vol. 62, No. 2, March 2012, pp. 112-116, DOI: 10.14429/dsj.62.1171

# A Note on Implementing Recurrence Quantification Analysis for Network Anomaly Detection

Ch. Aswani Kumar*, K. Bhargavi, and Garima Jalota

*Vellore Institute of Technology University, Vellore–632 014, India*

**E-mail: cherukuri@acm.org*

## ABSTRACT

This paper deal with the network anomaly detection, based on the analysis of non-stationary properties that occur in the aggregated IP traffic flows. We use recurrence quantification analysis (RQA), a mathematical nonlinear technique to achieve this task. The objective is to model the standard network traffic and report any deviation from it. We create a baseline from which we derive the RQA parameters. Using these parameters we explore the hidden recurrence patterns in the network traffic. Further, the detection is analysed using the support vector machine to classify the deviations from the regular traffic. Experiments are conducted on Vellore Institute of Technology University campus network traffic data to validate the model.

**Keywords:** Cross recurrence plots, network anomaly detection, non linear analysis, recurrence quantification analysis, support vector machines

## 1. INTRODUCTION

Anomaly detection is the problem of finding the malicious patterns in the data that deviate from the normal behavior[1]. In computer networks, this abnormal behavior generally occurs due to factors like denial of service (DoS), malfunctioning network devices, router mis-configuration, traffic overload, component failure and network intrusion. The deviation of the network traffic from the normal behavior happens due to: changes in path and routing loops, abrupt changes in traffic flow, variation in traffic delays. These properties are being identified and classified through the observation of traffic flow, packet delay, total number of packets sent, ratio of bytes sent in each direction, duration of connection, average size of packets and inter-arrival time. The present anomaly detection observes these transport layer statistics and the results are compared against a threshold value[2]. However it is observed that the network traffic features, characteristics of probability distributions of the IP packets, traffic density, traffic statistical characteristics change dynamically in the time domain[3]. Recently, recurrence quantification analysis (RQA) has been proposed to observe and study these non-stationary properties. Motivated by the analysis of Palmieri and Fiore[4], in this paper authors implement the RQA for anomaly detection.

## 2. RELATED WORK

Chandola[2], *et al.*have presented a good survey on anomaly detection. A survey on anomaly detection techniques in machine learning and statistical areas can be found in Hodge & Austin[5]. Several tools are available for network anomaly detection[6]. Some of these tools employ well defined rules to analyse the traffic patterns. Few other techniques include exponential smoothing and Holt Winters Forecasting, adaptive thresholding cumulative sum, maximum entropy estimation[7]. Many of these detectors works based on the transport layer statistics including the ratio of the bytes sent in each direction, the average size and mean inter-arrival time of the packets. Further the results are benchmarked against threshold values. These values are independent from the network utilisation and the number of users. Network anomaly can be detected using other models including Bayesian, support vector machine, neural networks, etc. Several data mining techniques are applied in the literature for solving the network traffic classification problem[8]. Mahoney[9] has proposed network traffic anomaly detection based on packet bytes. Lakhina[10], *et al*. have diagnosed the network traffic anomalies.

## 3. RECURRENCE QUANTIFICATION ANALYSIS

Zbilut and Webber[11] have introduced a deterministic way to identify non-stationary features in traffic flows called as recurrence quantification analysis (RQA). Traffic states propagation through time can be identified by analysing the evolution of some recurrence statistics. The analysis begins with choosing the parameters time delay and embedding dimension. The embedding dimension can be defined as the minimum number of dynamical variables that can describe the system attractor. The effect of these variables is further analyzed with the help of time delay[4]. A major step of RQA is the computation of statistical measures that enable us to identify the recurrent patterns. Generally these statistical measures are calculated at the successive fixed-length time windows which are known as epochs.

### 3.1 Recurrence Plots

Recurrence plots (RP) describe the nonlinear characteristics of time series data. By using these plots, we can approach to the non-stationary variation patterns of time-series data[12]. The

Received 21 August 2011, revised 6 February 2012, online published 13 March 2012

idea behind these plots is to reconstruct the system dynamics. Then the distances between all pairs of embedded vectors are computed which will generate a symmetric two-dimensional square matrix. The RP visualizes this distance matrix. The proper choice of the time delay *s* and threshold ε together with the correct estimation of an embedding dimension *m*, is fundamental for achieving satisfactory results. Figure 1 shows a sample recurrence plot of distance coded matrix plot without normalizing the source data on x-axis and test data on y-axis. The values at the right side of the figure indicate the distance to the next recurrence point. The visible diagonal line in Fig. 1 is called line of identification (LOI) and consists entirely of recurrence points. These plots are symmetrical with regard to the LOI and the LOI acts as a reference line for identifying the other recurrence points.

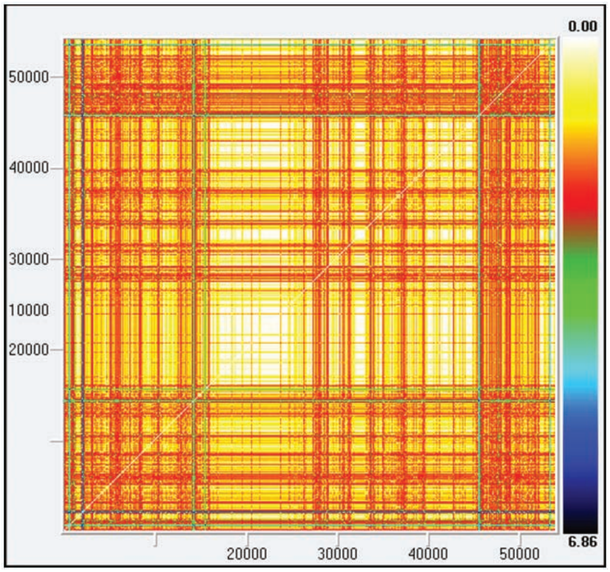


**Figure 1. Plot for chaotic signals (non linear).**

### 3.2 Support Vector Machines

Presently support vector machines (SVM) are known for their flexible data-mining classification. In its operation, an SVM model non-linearly transforms the original input space into a higher dimensional feature space and employs a linear hyperplane in the feature space to separate points. The training samples ($x_i$, $y_i$) constitute the hyperplane vector $w$[13]. The vectors located closer to the hyperplane will be the support vectors. The support vectors referring to this study focus on the anomalous traffic and the negative vectors as the normal traffic.

## 4. SYSTEM ARCHITECTURE

Figure 2 explains the architectural design to detect anomalies using RQA. The study starts with analyzing the key network traffic parameters out of which the inter-arrival time series is found out to be the main one. Next we create a baseline in order to compare the test data with the baseline known as a knowledge base. The input to the RQA is the inter-arrival time series observed in different time periods. The threshold value is set up as the baseline parameter for that respective time period. The resultant RQA derivatives are set up as epochs and these are stored in the database. Thus finally these derivatives are classified using LIBSVM. Each epoch is classified as either regular or anomalous. The Regular and Anomalous classes are classified in the form of confusion matrix using WEKA[14].

## 5. EXPERIMENTAL RESULTS AND ANALYSIS

This section provides experiments we have conducted using the proposed architecture and analysis of the results obtained from the experiments.

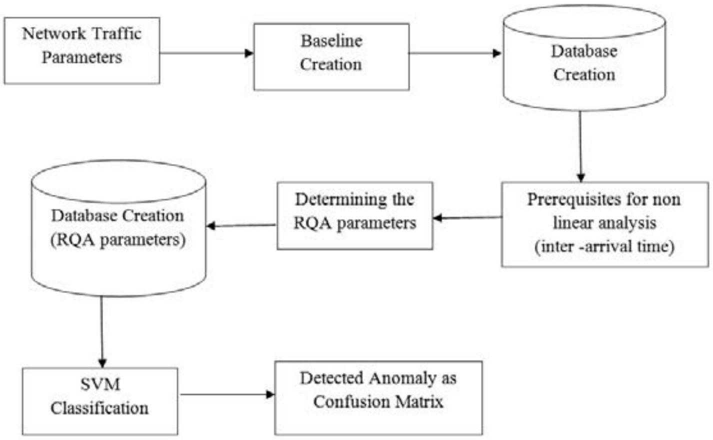

Network Traffic Parameters
Baseline Creation
Database Creation
Prerequisites for non linear analysis (inter -arrival time)
Determining the RQA parameters
Database Creation (RQA parameters)
SVM Classification
Detected Anomaly as Confusion Matrix

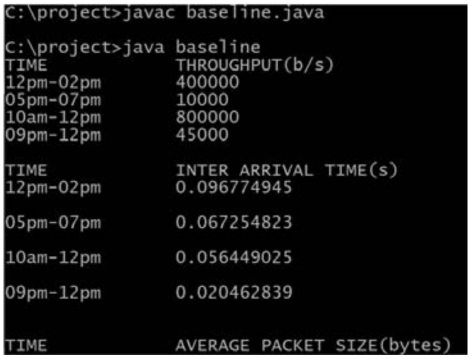
**Figure 2. Architecture specification.**

### 5.1 Results

We have conducted the experiments using VIT University (www.vit.ac.in) network. The university campus has a secure WiFi network with Pronoto UniFi OSS and service gateway. The bandwidth offered by the network is 54 Mbps. We have captured the data packets of the network at different time intervals using the Wireshark network tool (www.wireshark.org) which gives us the data like the source address, destination address, amount of time the packet is transferred which is used for further analysis to classify between regular and anomalous. From these packets we have extracted timing and volume information like average packet size, inter-arrival time, throughput. We have collected incoming traffic only. The baseline creation is the process of measuring and rating the performance of a network. This process is the first step in anomaly detection. It is used to analyze the normal traffic flow and to set the threshold value for different time. Average inter-arrival time between packets, is the most selective feature and is calculated with at least microsecond precision and accuracy. The assumption for RQA is the scalar time series. The patterns of the recurrence in the network traffic are analysed by studying the variable parameters: recurrence, determinism, entropy and laminarity. Figure 3 shows the network baseline. The average throughput and inter arrival time values are considered as a threshold.

Figure 4 shows the recurrence obtained in the peak hours. A denser region in the Fig. 4 explains more recurrence points at that particular epoch. Figure 5 shows the determinism obtained in the peak hours. High value in the Fig. 5 represents more recurrence in the particular epoch. Figure 6 shows the entropy obtained during peak hours. The entropy indicates the complexity of the recurrence structure. Chaotic behavior is indicated by the high entropy and periodic behavior is indicated by low entropy. Laminarity indicates the percentage of recurrence points comprising vertical line structures. High values in the Fig. 7 represent strong intermittent. Figure 8 indicates the sample data values for Recurrence Rate, Determinism, Entropy, Laminarity and Trapping time.

By using the descriptors of RQA, SVM classification has been performed. While building the knowledge base, creating the discriminating feature vectors is an important task. The set of obtained RQA descriptors are combined for each of the training set. Then the feature vectors are associated with the inter-arrival time and average packet length measurements respectively. To determine the most revealing features for binary classification, all the RQA features are subjected to



```
C:\project>javac baseline.java

C:\project>java baseline
TIME             THROUGHPUT(b/s)
12pm-02pm        400000
05pm-07pm        10000
10am-12pm        800000
09pm-12pm        45000

TIME             INTER ARRIVAL TIME(s)
12pm-02pm        0.096774945

05pm-07pm        0.067254823

10am-12pm        0.056449025

09pm-12pm        0.020462839


TIME             AVERAGE PACKET SIZE(bytes)
```

**Figure 3. Network baseline values.**

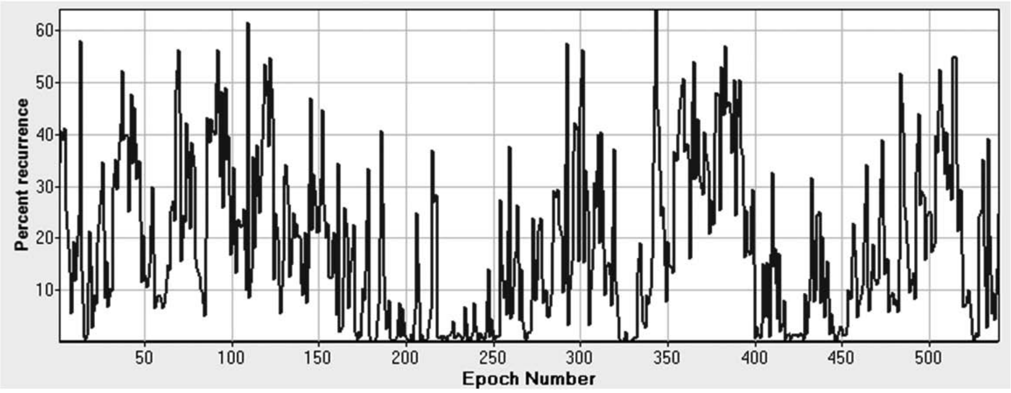


**Figure 4. Recurrence obtained in the peak hours.**

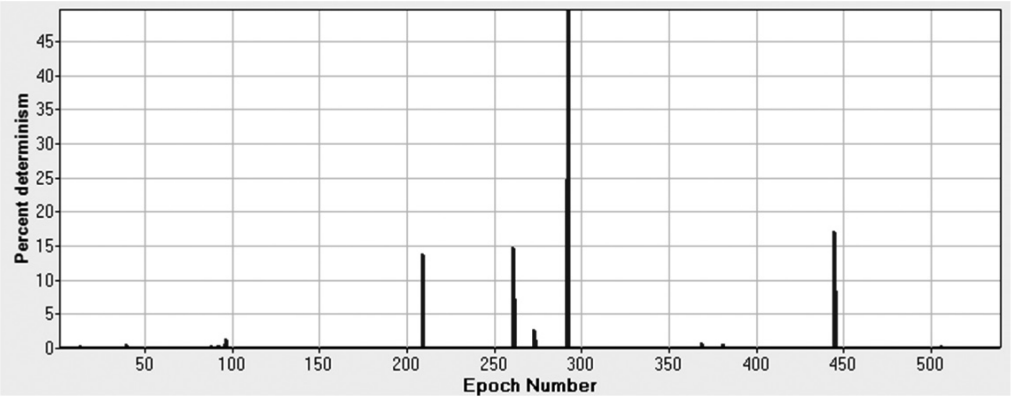


**Figure 5. Determinism obtained in the peak hours.**

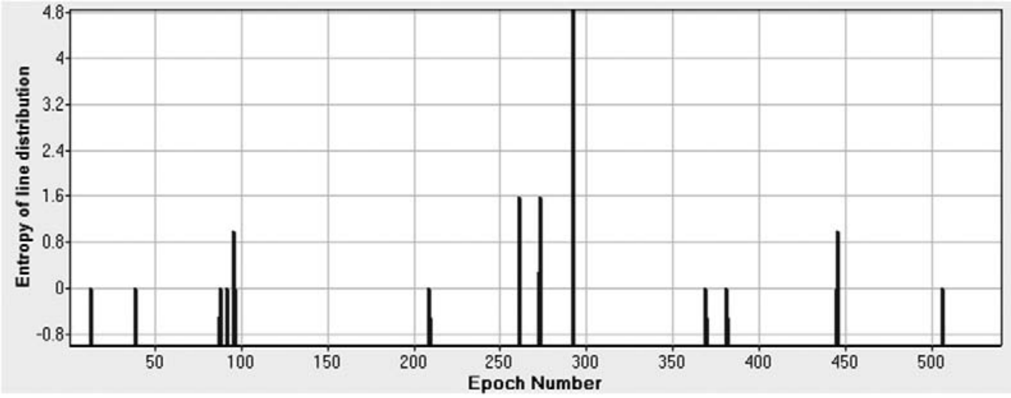


**Figure 6. Entropy obtained in the peak hours.**

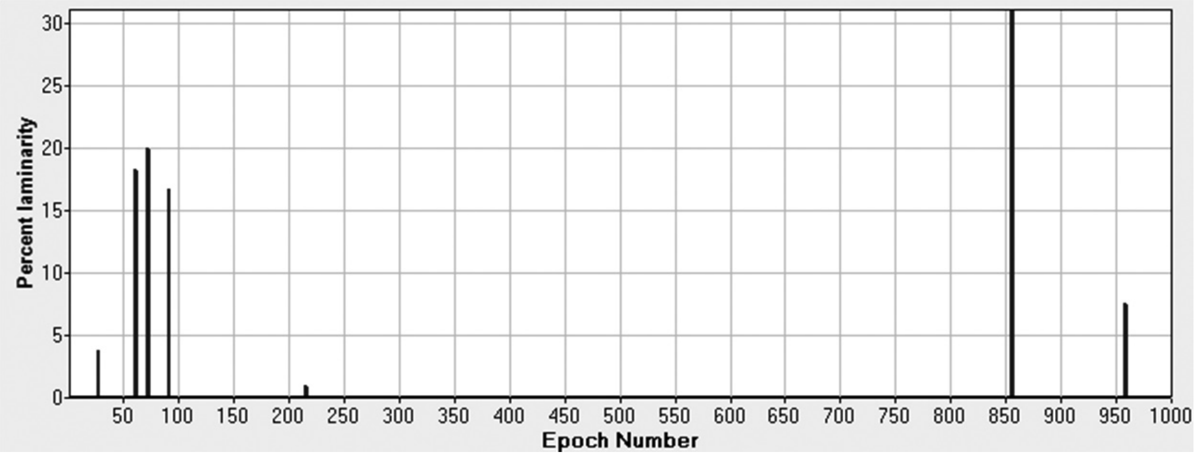
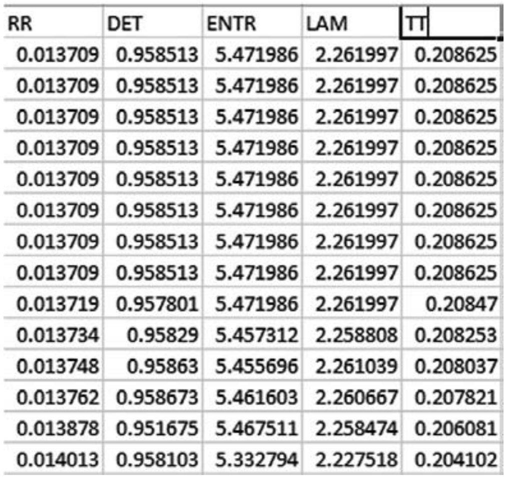


**Figure 7. Laminarity obtained in the peak hours.**

selection by calculating their mutual information gain. Then we have ranked them to select the best ones for building the final binary classification model implemented through SVMs. The WEKA 3.5.8 has the information gain algorithm along with the rankers method and 10-fold cross validation is used to spot them. Based on the best information gain, the attributes are ranked as %REC, ENT and %DET. The ultimate classification model has been built with these three best discriminating features which are, calculated on the inter-arrival time feature vectors. For this purpose, we have used the WEKA LIBSVM wrapper library and the C-SVC (regularized support vector classification) algorithm with a pure sigmoid kernel ($C = 1$, £ = 0.3, $r = 0$) and 10-fold cross validation. The RQA features are given as input to the wrapper as a CSV file.



| RR | DET | ENTR | LAM | TT |
|---|---|---|---|---|
| 0.013709 | 0.958513 | 5.471986 | 2.261997 | 0.208625 |
| 0.013709 | 0.958513 | 5.471986 | 2.261997 | 0.208625 |
| 0.013709 | 0.958513 | 5.471986 | 2.261997 | 0.208625 |
| 0.013709 | 0.958513 | 5.471986 | 2.261997 | 0.208625 |
| 0.013709 | 0.958513 | 5.471986 | 2.261997 | 0.208625 |
| 0.013709 | 0.958513 | 5.471986 | 2.261997 | 0.208625 |
| 0.013709 | 0.958513 | 5.471986 | 2.261997 | 0.208625 |
| 0.013709 | 0.958513 | 5.471986 | 2.261997 | 0.208625 |
| 0.013719 | 0.957801 | 5.471986 | 2.261997 | 0.20847 |
| 0.013734 | 0.95829 | 5.457312 | 2.258808 | 0.208253 |
| 0.013748 | 0.95863 | 5.455696 | 2.261039 | 0.208037 |
| 0.013762 | 0.958673 | 5.461603 | 2.260667 | 0.207821 |
| 0.013878 | 0.951675 | 5.467511 | 2.258474 | 0.206081 |
| 0.014013 | 0.958103 | 5.332794 | 2.227518 | 0.204102 |

**Figure 8. Sample values for the parameters.**

### 5.2 Analysis and Discussion

Confusion matrix shown in Table 1 summarizes the classification results of 15797 feature vector instances. From this matrix, we can infer that 14538 epochs are the true positive regular and 598 epochs are the true positive anomalies. Similarly 82 epochs are classified as the false negative regular and 579 are classified as the false negative anomalies. Table 2 shows the accuracy of the classification with regard to different other methods. In confusion matrix, the true positive (TP) rate is obtained from the diagonal element divided by the sum over the relevant row, i.e.14538/(14538+82)=0.994 for class Regular and 598/(598+579)=0.508 for class Anomalous. In the matrix shown in Table 1, the false positive (FP) rate is calculated by subtracting the column sum of class x from the diagonal element and dividing the rows sum of all other classes; i.e. 579/(579+82)=0.492 for class Regular and 598/(14538+598)=0.006 for class Anomalous. In the confusion matrix, precision is obtained by dividing the diagonal element by the sum over the relevant column, i.e. 14538/(14538+579)=0.962 for class Regular and 598/(82+598)=0.879 for class anomalous. F-Measure combines precision and recall.

Table 3 reports the most significant metrics that have been used to assess the effectiveness and accuracy of the technique. The evaluation has also concentrated on the measurements derived from the class probabilities. Precisely the evaluation produces the mean absolute error and the root mean squared error of the probability estimates. The Kappa statistic is also used for evaluating the classifier. The precision value shows the errors occurred due to False positives while the recall scores indicate the errors caused due to True negatives. Further extension of this study includes the investigation of cross recurrence quantification which enables a nonlinear analysis of bivariate data. It measures the degree of coupling between multiple combined traffic features. Another interesting direction would be the application of RQA in group

**Table 1. Confusion matrix (on Epochs)**

| Classified as | Regular | Anomalous |
|---|---|---|
| Regular | 14538 | 82 |
| Anomalous | 579 | 598 |

**Table 2. Accuracy**

| TP Rate | FP Rate | Precision | Recall | F-Measure | ROC Area | Class |
|---|---|---|---|---|---|---|
| 0.994 | 0.492 | 0.962 | 0.994 | 0.978 | 0.9 | Regular |
| 0.508 | 0.006 | 0.879 | 0.508 | 0.644 | 0.9 | Anomalous |

**Table 3. The detection performance matrix on epochs**

| | | |
|---|---|---|
| Correctly classified instances | 15136 | 95.8157% |
| Incorrectly classified instances | 661 | 4.1843% |
| Kappa statistic | 0.6235 | - |
| Mean absolute error | 0.0698 | - |
| Root mean squared error | 0.1922 | - |
| Relative absolute error | 50.5975% | - |
| Root relative squared error | 73.1997% | - |
| Total number of instances | 15797 | - |

communications[15]. It is to be noted that inter-arrival time variance appears to be a good feature to be used for building traffic profiles. We believe that the high errors can be minimized by measuring the multiple traffic features.

## 6. CONCLUSIONS

Authors have analyzed the problem of network traffic classification and anomaly detection. For this purpose we have used a non-linear mathematical model called recurrence quantification analysis. Work has concentrated on the nonlinear dynamics of some specific traffic flow types to accomplish this task. The approach proved to be effective to evaluate and recognize the complex nature of network traffic.

## ACKNOWLEDGEMENTS

The authors gratefully acknowledge the financial support of the National Board of Higher Mathematics, Dept of Atomic Energy, Govt. of India under the grant number 2/48(11)2010-R&D 11/10806. Also, authors sincerely thank the anonymous reviewers for their highly useful insights.

## REFERENCES

1. Zhani, M. & Elbiaze, H. Analysis and prediction of real network traffic. *Journal of networks*, 2009, **4**(9), 855-65.
2. Chandola, V.; Banerjee, A. & Kumar, V. Anomaly detection: A survey. *ACM Computing Surveys*, 2009, **41**(3), 1-58.
3. Palmeieri, F. & Fiore, U. A nonlinear, recurrence-based approach to traffic classification. *Computer Networks*, 2009, **53**(6), 761–73.
4. Palmeieri, F. & Fiore, U. Network anomaly detection through non-linear analysis. *Computers Security*, 2010, **29**(7), 737-55.
5. Hodge, V. & Austin, J. A survey of outlier detection methodologies. *Art. Intel. Revi.*, 2004, **22**(2), 85-126.
6. Yu, G.; McCallum, Y. & Don, T. Detecting anomalies in network traffic using maximum entropy estimation. *In* Proceedings of the 5th ACM SIGCOMM Conference on internet measurement, 2005, pp. 32.
7. Siris, V.A. & Papagalou, F. Application of anomaly detection algorithms for detecting SYN flooding attacks. *In* Proceedings of IEEE Global telecommunications conference, 2004, pp. 2050-054.
8. Nguyen, T.T.T. & Armitage, G. A survey of techniques for internet traffic classification using machine learning. *J. IEEE Comm. Surveys Tutorials*, 2008, **10**(4), 56-76.
9. Mahoney, M. Network traffic anomaly detection based on packet bytes. *In* Proceedings of ACM Symposium on Applied Computing, 2003, pp. 346-50**.**
10. Lakhina, A.; Crovella, M. & Diot, C. Diagnosing network-wide traffic anomalies. *ACM SIGCOMM Comp. Comm. Rev.*, 2004, **34**(4), 219-30.
11. Zbilut, J.P & Webber, C.L. Embeddings and delays as derived from recurrence quantification analysis. *Physics Letters A*, 1992, **171**(3-4), 199-203.
12. Masugi, M. Applying a recurrence plot scheme to analyze non-stationary transition patterns of IP-network traffic. *J. Comm. Nonlinear Sci. Numerical Simul.*, 2009, **14**(4), 1418-430.
13. Cristianini, N. & Shawe-Taylor, J. Support vector machines and other kernel based learning methods, Cambridge University Press, 2000.
14. Hall, M.; Frank, E.; Holmes, G.; Pfahringer, B.; Reutemann, P. & Witten, I.H. The WEKA data mining software: An update. *ACM SIGKDD Explorations Newsletter*, 2006, **11**(1), 10-18.
15. Aswani Kumar, Ch.; Sri Lakshmi, R. & Preethi, M. Implementing secure group communications using key graphs. *Def. Sci. J.*, 2007, **57**(2), 279-86.

### Contributors

**Dr Ch. Aswani Kumar** received PhD (Computer Science) from Vellore Institute of Technology (VIT) University, India. Currently working as a Associate Professor at School of Information Technology and Engineering, VIT University, Vellore, India. He has published 35 refereed research papers so far in various national, international journals and conferences. His research interests include: Data mining, formal concept analysis, information security, and machine intelligence.

**Ms Bhargavi** is a BTech (IT) from VIT University, Vellore, India. Presently she is working as Database Administrator in Ford Technology Services India, Chennai. Her research interests include: Databases and networks.

**Ms Garima** is a BTech (IT) from VIT University, Vellore, India. Presently she is working as Associate Software Engineer in Robert Bosch Engineering & Business Solutions, Bangalore. Her research interests include: Database management systems and networks.