ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 3, Issue 2, August 2013

FODA - A Fast Outlier Detection Algorithm for VPNs

Dr.Reji Kurien Thomas, Dr. T.Anand
Commander (Retd), CEO Techno Pilot, Assoc. Professor, Madha Group of Academic Institutions

Abstract*— A Virtual Private Network (VPN) is a computer network using public infrastructure to provide offices with a closed access to the organization's network, reducing expenses on leased lines. The E-business applications have prompted companies to manage processes with lowered operating costs and better customer satisfaction. The applications also need to be scalable in terms of voice, data or video traffic. Deploying VPNs requires planning and can be customized to suitable requirements with a broad range of VPN. The survival of many businesses is dependent on open access to network resources. VPNs provide security associated with private networks with different security technologies. Attackers try malicious activities with harmless-looking connections. Intrusion detection systems try and differentiate these attacks or dissimilar connections. Such dissimilar connections can also be outliers which are not real intrusions. Detection of outliers leads to identification of faults for administrators to take preventive measures. Outliers can also warn detection of new attacks. This paper suggests a new distance based Outlier Detection Algorithm, Fast Outlier Detection Algorithm (FODA) to detect outliers in VPNs.

***Index Terms*— Fast Outlier Detection Algorithm, Virtual Private Network, Outliers, Intrusion detection, Nearest Neighbor.**

I. INTRODUCTION

Wide Area Networks (WANs) based on private connections between two or more locations were developed 25 years back. The higher costs of leasing lines from Telecom Service Providers and reducing costs on the internet technologies made companies choose VPN, to implement their WAN access. Network management encompasses various operational task including hardware, software and configurations. Configuring a network involves enormous costs [1] and correctly configuring a network is very difficult [2]. Configuration settings have to be regularly updated to maintain the status quo [3], making it difficult for administrators who need to be adept in domain-specific knowledge, protocols, types of devices and configuration solutions [4]. The dynamic nature of networks is also an important factor in network management [5]. VPNs provide a virtual and easy infrastructure that connects branches, corporate offices, business partner sites and remote computers into a single corporate network. VPN devices create PVCs (Permanent Virtual Circuit) like a leased line, with tunnels allowing senders to encapsulate their data in IP packets hiding the routing and switching infrastructures from both the senders and receivers. The VPN device at the senders end encapsulates the outgoing packet or frame and forwards it and when the packet arrives on the receiving end, the receiving device retrieves the original packet. Packet authentication prevents data from being viewed and Packet authentication applies header to the IP packet to ensure its integrity. Outliers who do not conform to normal behavior and might occur in the network due to a variety of reasons like intrusion, breakdown of a system or terrorist activity. Outlier detection is discovering these exceptional behaviors. Outliers are an important concept of network analysis. Outlier is an observation and a subjective exercise. Outlier detection was researched within various application domains and knowledge disciplines. Anomalous transactions in a credit card connection can indicate usage of a stolen card, IP packets may indicate either a possible intrusion or attack, or a failure in the network. Outliers in military surveillance may indicate sudden troop movement. Many techniques have been developed for building outlier and anomaly detection. As the Network Databases list grows in size finding meaningful outliers becomes complex. The aim of this work is to propose an outlier detection mechanism, applicable to VPNs b irrespective of the Network Database Size. The approach consists of a identifying outliers based on the average values. The proposed mechanism is generic and can be easily implemented in VPN networks. The topology of a VPN is depicted in Fig. 1.

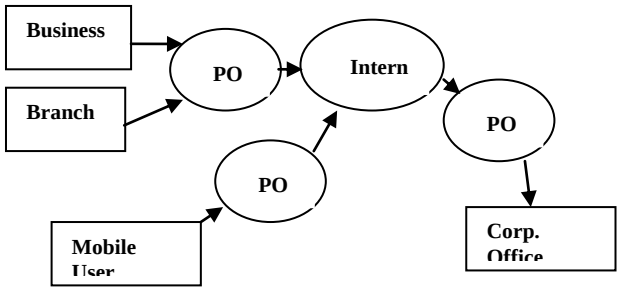

Fig. 1 A Virtual Private Network

II. OUTLIERS

An outlier is an observation that is numerically distant from the rest and as Grubbs defined an Outlier as an outlying observation that deviates from others in a sample. Outliers may contain important information and Outliers need to be investigated, since they contain valuable information about the process under investigation or the data gathering and recording process. Outliers can also indicate an erroneous data. Values derived statistically from such data that include outliers may be incorrect.Efficient detection of outliers

ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 3, Issue 2, August 2013

reduces the risk of making wrong assumptions based on the data, since data mining algorithms and statistical analysis produce wrong results. In statistics Outlier detection translates to clustering. Clustering algorithms detect outliers as by-products of any clustering processes. The importance of outlier detection is due to the fact that outliers in data translate to significant actionable information in a wide variety of application domains. Previous approaches have been proposed to detect outliers and can be classified into four major categories based on the techniques used [6], which are Distribution-based approach, Distance-based approach, Density-based approach and Clustering-based approach. Distribution-based approaches [7] develop statistical models from the given data and then apply a statistical test to determine if an object belongs to it and Objects with low probability of belonging to it are declared as outliers. A prior knowledge of the data distribution is required and distribution-based approaches cannot be applied to multidimensional data. This makes the distribution-based approaches difficult to be used in practical applications. Using variations In the distance-based approach outliers can be detected. Density-based approaches [8] compute the density of regions in the data and declare the objects in low dense regions as outliers. Any data point is treated an outlier, known as Local Outlier Factor (LOF) [9], depending on its distance from its local neighborhood. Clustering-based approaches consider clusters of small sizes and consider them outliers. A statistical model-based outlier detection was presented based on computational geometry [10] did not scale well as the number of dimensions increase. An approach for discovering outliers using distance metrics was first presented by Knorr et al. [11] E. Knorr and R. Ng. Finding intentional knowledge of distance-based outliers [11]. Fig. 2 depicts outliers.

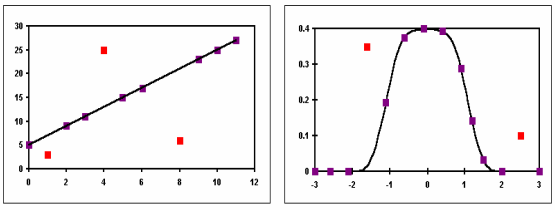

Fig. 2 Outliers

III. OUTLIER DETECTION

There is no fixed mathematical definition of what constitutes an outlier and an observation of an outlier is a subjective exercise. There are three fundamental approaches to the problem of outlier detection. First, determining outliers with no prior knowledge of the data stored. Second approach is when the data has both normality and abnormality. Third approach is eliminating the Outliers by retaining normal values. All the approaches used for identification assume that the data are from a normal distribution, and identify observations which are "unlikely" based on mean and standard deviation. For a normal distribution model, if outliers are expected for large samples the application should use a classification algorithm to detect outlier points. Deletion of an outlier is a controversial practice. It is acceptable in areas of practice where the underlying model of the process being measured and the usual distribution of measurement error are known precisely. There are numerous tests for identifying outliers. Common outlier tests for normal distributions are the Rosner test, Grubbs test, box plot rule and Dixon test which are based on hypothetical testing and non-regression based methods.

A. *Rosner's Test*

Rosner's Test for detecting up to k outliers can be used when the number of data points is above 25 and identifies both high and low and therefore always two-tailed [12]. The data is ranked in ascending order and the mean and standard deviation are determined. y, that is farthest from the mean are removed from the observations, test statistic R, is calculated as Ri+1 = |y(i) - ỹ(i)| / s(i). The R statistic is then compared with a critical value [13] Null hypothesis is tested, If R is less than the critical value it is concluded there are no outliers and if R is greater than the critical value the presence of k outliers is accepted.

B. *Dixon's Test*

Dixon's test is generally used for detecting a small number of outliers [12]. The results are sorted in an ascending order and with the sample size as the base, the tau statistic for the highest value or lowest value is computed as indicated in Table 1. This test can be used when the sample size is between 3 and 25 observations. The tau statistic is compared to a critical value at a chosen value of alpha [12]. If the tau statistic is lesser than the critical value it is concluded that no outliers are present.

TABLE I: DIXON'S TEST VALUE RANGES

Observations	Highest Value	Lowest Value
3 – 7	$\tau = x_n - x_{n-1} / x_n - x_1$	$\tau = x_2 - x_1 / x_n - x_1$
8 – 10	$\tau = x_n - x_{n-1} / x_n - x_2$	$\tau = x_2 - x_1 / x_{n-1} - x_1$
11 – 13	$\tau = x_n - x_{n-2} / x_n - x_2$	$\tau = x_3 - x_1 / x_{n-1} - x_1$
14- 20-30	$\tau = x_n - x_{n-2} / x_n - x_3$	$\tau = x_3 - x_1 / x_{n-2} - x_1$

C. *Boxplot Rule*

The boxplot rule is a visual test to inspect for outliers. The inter-quartile range is included into a box and the 5% and 95% confidence intervals are indicated with error bars outside of the box and Values outside the confidence interval are probable outliers [14]. Q1 – x / Q3 – Q1 -> k and with 95% confidence interval limit and 5% x-Q3 / Q3 – Q1 -> k and 5% confidence interval limit: assigning to a residence the point.

ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 3, Issue 2, August 2013

IV. THE FAST OUTLIER DETECTION ALGORITHM

The Fast Outlier Detection Algorithm is a simple query oriented technique where the number of iterations used by the Nearest Neighbour classifier is reduced by taking the Average of the Parameters as the main factor to classify outliers. FODA is applied as an experiment to ATM Transactions to detect outliers, based on the expenditure pattern. Financial Institutions and Banks perform millions of transactions every day with thousands of users. Banking systems require authenticity and validity. Most of the banks use verification software to authorize data. Any customer who withdraws money from an ATM requires an authorization like the ATM pin which is a secret number. ATM cards can be stolen or misused or forged and needs identification. Te objective of this work is to identify transactions which are anomalous or outlier to the regular spending pattern using nearest neighbour algorithm. The nearest-neighbour method is assumed to be slow when the dimensionality with decrease in accuracy. The reason is that in a high-dimensional space all points tend to be far away from each other, so nearest neighbours are not meaningfully similar. An ATM transaction is composed of the following to list a few important details like (1) ATM Number, (2) Account Number. (3) Card Number, (4) Date of Withdrawal, (5) Time of Withdrawal and (6) Withdrawal amount. The first objective is to build the Training examples D. D Consists of all the ATM transaction of a given card. Each Transaction and specifically the transaction amount is an instance of example E to identify the nearest neighbour.

A. NN Algorithm to Identify Nearest Neighbor in ATM List

Calculating the distance between E and all examples in the training set

Let the Amount withdrawn be called the WithAmt.

Let the MaxAmt in the Range be defined as WithAmt + 20% of WithAmt

Let the MinAmt in the Range be defined as WithAmt - 20% of WithAmt

TestAmt = WitrhAmt //The First Amount is taken as an Example E

For i = 1 to N //Number of Records of Withdrawals from ATM's within a Quarter

MaxAmt = TestAmt + 20% of TestAmt

MinAmt = TestAmt - 20% of TestAmt

For j=i+1n to N

If (WithAmt > TestAmt and WithAmt <= MaxAmt)

"Store"

End if Cond.

If (WithAmt < TestAmt and WithAmt >= MaxAmt)

"Store"

End if Cond.

If (WithAmt > MaxAmt or WithAmt < MaxAmt)

"Outlier "

End if Cond.

End for (j)

TestAmt = WitrhAmt //Each Withdrawal Amount is taken as the next Example E

End for (i)

For Example, the transaction occurred for Six Months is as follows

TABLE 2: WITHDRAWALS IN ATM

.

Mon 1	Mon 2	Mon 3	Mon 4	Mon 5	Mon 6
2000	5000	7000	1000	2000	4000
1000	4000	3000	1000	1000	1000
3000	1000		1000	3000	2000
1000			1500	1000	1000
3000			2000	3000	2000
			1000		
			1000		
			1000		
			500		

The Total Transactions is 29 – D

The Total iterations for classifying Outliers would be 29+28+27+26+…….1. Taking the First Amount 2000 as an Example, the distances from other transactions would be as indicated in Table 3

TABLE 3: DISTANCES OF TRANSACTIONS

Dataset D	Distance From E(2000))
2000.00	
1000.00	1000.00
3000.00	1000.00
1000.00	1000.00
3000.00	1000.00
5000.00	3000.00
4000.00	2000.00
1000.00	1000.00
7000.00	5000.00
3000.00	1000.00
1000.00	1000.00
1000.00	1000.00
1000.00	1000.00
1500.00	500.00
2000.00	0.00
1000.00	3000.00
1000.00	1000.00
1000.00	1000.00
500.00	1500.00
2000.00	0.00
1000.00	1000.00
3000.00	1000.00
1000.00	1000.00
3000.00	1000.00
4000.00	2000.00
1000.00	1000.00
2000.00	0.00
1000.00	1000.00

The graph for NN in the Dataset D as depicted in Fig. 3. Would be

ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 3, Issue 2, August 2013

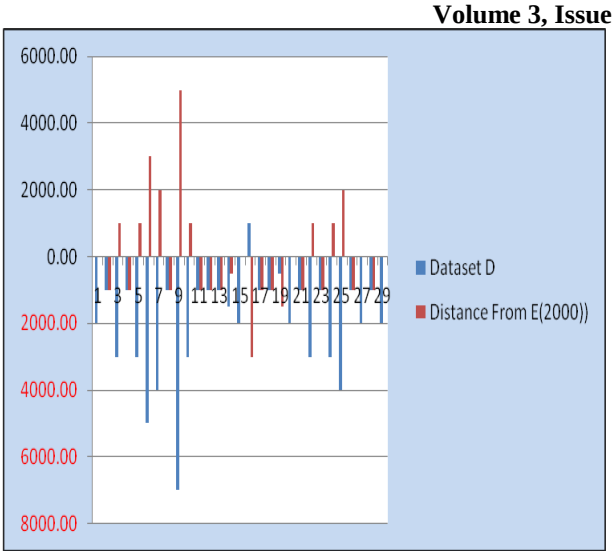

Fig. 3 Graph of Distance of Data Elements

For this type of outlier the density of the neighbours of a given instance plays a key role. Furthermore an instance is not explicitly classified as either outlier or non- outlier instead for each instance a local outlier factor (LOF) is computed which will give an indication of how strongly an instance can be considered an outlier. There is a weakness of the distance-based method in identifying certain type of outliers according to Breuning

B. FODA Algorithm to Trace Outliers in ATM Transactions within a Specified Period

Let the Average of all the amounts withdrawn be the BaseAmt. .

BaseAmt.= Average(Amounts withdrawn from ATM for Six Mon s)

Let the Upper Range in the Range be defined as BaseAmt. + 20% of BaseAmt.

Let the Lower Range in the Range be defined as BaseAmt. - 20% of BaseAmt.

The Record set D = Average(Amounts in Mon), Count(Trans(in Mon)

For i = 1 to N //Number of Records in D
MaxAmt = maximum (Amount withdrawn) in Mon
MinAmt = maximum(Amount withdrawn) in Mon
If (MaxAmt > UpperRange)
"Outlier for the Present Month "
End if Cond.
If (MinAmt < LowerRange)
"Outlier for the Present Month "
End if Cond.
End for (j)
End for (i)

The ATM transactions with FODA would be the graph for Outliers would result in Fig. 4

TABLE 4:ATM AMOUNT WITHDRAWALS WITH NN CLASSIFIERS -1

2000 1000 3000 1000 3000 Avg=2000, Sum=10000 Highest=3000, Trans=5	5000 4000 1000 Avg=3333,Sum=10000 Highest=5000, Trans=3	7000 3000 Avg=5000,Sum=10000 Highest=7000,Trans=2
1000 1000 1000 1500 2000 000 1000 1000 500 Avg=1100,Sum=10000 Highest=2000,Trans=9	2000 1000 3000 1000 3000 Avg=2000,Sum=10000 Highest=3000,Trans=5	4000 1000 2000 1000 2000 Avg=2000,Sum=10000 Highest=4000,Trans=5

TABLE 5:ATM AMOUNT WITHDRAWALS WITH NN CLASSIFIERS -2

Accountno	Avg	trns	highest
	0	0	0
1	2000	5	3000
2	3333	3	5000
3	5000	2	7000
4	1100	9	2000
5	2000	5	3000
6	2000	5	4000

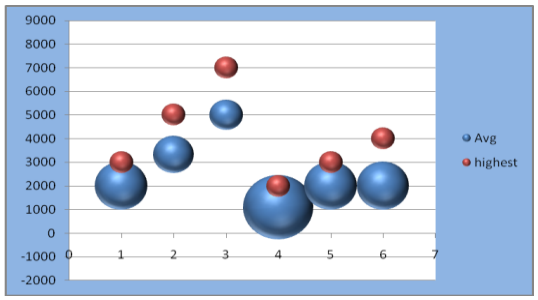

Fig. 4 Graph of Distance Based on Average and High Values

The Distances based on Euclidean Distance would translate to Table 6. The Values are Commonly Divided by the Sum Amount to keep the Calculation Simpler using the formula

$$D(X,Y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 3, Issue 2, August 2013

Average	Highest
0.2	0.3
0.01776889	0.04
0.232546843	0.04
0.015017729	0.25
0.034218441	0.01
0.027483525	0.01

TABLE 6: Euclidean Distances between the Samples

And the resulting graph displaying the Outliers in a more pronounced way would translate to Fig. 5

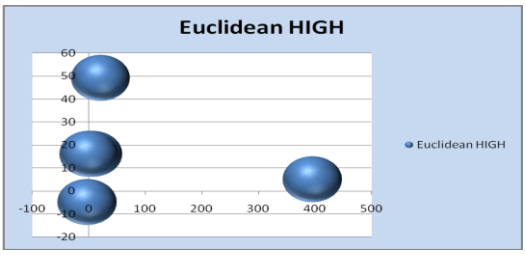

Fig. 5 Graph of Distance Based on FODA

C. *Application of FODA to Network Intrusion in VPNs*

Defining a representative normal region in a VPN is challenging and the boundary between normal and outlying behaviour is difficult to define. Availability of training set or validation is difficult to presume. Also Data might contain noise and the Normal behaviour keeps evolving. Anomaly detection depends on, Nature of input data, Availability of supervision, Type of anomaly: point, contextual, structural. Table 7 describes the listing of network traffic. FODA identifies record numbers five, eight and ten as outliers, since the average bytes in network traffic is between 170 to 200 and the average is 176.6

TABLE 7: VPN Traffic Listing

Source IP	St. Time	Dest. IP	Dest. Port	Bytes	Outlier
201.165.39.85	01:07:20	163.93.178.225	135	192	No
201.165.39.85	01:13:56	163.93.178.218	135	195	No
201.165.39.85	01:14:29	163.93.178.217	135	180	No
201.165.39.85	01:14:30	163.93.178.255	135	199	No
201.165.39.85	**01:14:32**	**163.93.178.253**	**135**	**19**	**Yes**
201.165.39.85	01:14:34	163.93.178.254	135	177	No
201.165.39.85	01:14:35	163.93.178.222	135	172	No
201.165.39.85	**01:14:37**	**163.93.178.221**	**135**	**285**	**Yes**
201.165.39.85	01:14:42	163.93.178.250	135	195	No
201.165.39.85	**01:14:43**	**163.93.178.249**	**135**	**152**	**Yes**

FODA-VPN to detect outliers in Network Traffic

Let BaseTransBytes.= Average(Average of the bytes received from an IP Address)
Let the UpperRange in the Range be defined as BaseTransBytes. + 2% of BaseBytes.
Let the LowerRange in the Range be defined as BaseTransBytes. - 2% of BaseBytes.
The Record set D = Source IP Address and No of Transmitted to Destination IP and Destination IP grouped by Source IP Address, Bytes, Dest. IP Address
For i = 1 to N //Number of Records in D
TransBytes = Bytes
If (TransBytes > UpperRange)
"VPN Outlier"
End if Cond.
If (TransBytes < LowerRange)
"VPN Outlier"
End if Cond.
End for (j)
End for (i)

V. CONCLUSION

This paper has presented a new algorithm FODA for VPNs. VPN is designed to meet the demands for information access in a secure, cost-effective environment. More and more businesses demand a higher level of network access, expanding their network and using the Internet as the backbone to create Virtual Private Networks (VPN). The strength of FODA is that, it is simple to implement and use. FODA prediction is robust to noisy data since it applies averaging k-nearest neighbours. Statistical tests are used to determine experimental observations for outliers. The tests on normal data sets are easy to use but tests on non-normal data are more complex [14]. Some of these tests are included in Barnett and Lewis [15]. In many situations, the data can be transformed to an approximate a normal distribution and be analysed using the techniques presented above [14]. Almost all statistical analyses assume that data follows a normal distribution. FODA achieves the objective of new directions from related adversarial mining outliers /applications intrusion detection, fraud detection. Knowledge and experience from these adversarial domains can be interchangeable and helps prevent repetitions of common mistakes and reinventions. The technique is comprehensible and easy to explain. The paper demonstrates FODA decreases the number of iterations in any applicable domain.

REFERENCES

[1] Demystifying Opex and Capex Budgets - Feedback from Operator Network Managers, 2007. http://www.researchandmarkets.com/reports/448691/demystifying opex and capex budgets feedback.

[2] W. Enck, T. Moyer, P. McDaniel, S. Sen, P. Sebos, S. Spoerel, A. Greenberg, Y. Sung, S. Rao, andW. Aiello, "Configuration Management at Massive Scale: System Design and Experience," IEEE J. Selected Areas in Communications,

ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 3, Issue 2, August 2013

Special Issue on Network Infrastructure Configuration, April 2009.

[3] Z. Kerravala, "Configuration management delivers business resiliency." The Yankee Group, November 2002.

[4] H. Ballani and P. Francis, "CONMan: A Step towards Network Manageability," in Proc. ACM SIGCOMM, 2007.

[5] A. Greenberg, P. Lahiri, D. A. Maltz, P. Patel, and S. Sengupta, "Towards a Next Generation Data Center Architecture: Scalability," in Proc. of PRESTO Workshop at SIGCOMM, 2008.

[6] N. Beckmann, H. P. Kriegel, R. Schneider, and B. Seeger. The Rtree: an efficient and robust access method for points and rectangles. ISIGMOD, 1990.

[7] M. Bawa, T. Condie, and P. Ganesan. Lsh forest: self-tuning indexes for similarity search. In WWW, 2005.

[8] C. B¨ohm and F. Krebs. The k-nearest neighbor join: Turbo charging the kdd process. Knowl. Inf. Syst., 6(6):728–749, 2004.

[9] T. M. Chan. Approximate nearest neighbor queries revisited. In SoCG, 1997.

[10] R. Fagin, R. Kumar, and D. Sivakumar. Efficient similarity search and classification via rank aggregation. In SIGMOD, 2003.

[11] A. Guttman. R-trees: a dynamic index structure for spatial searching. In SIGMOD, 1984.

[12] Gibbons, R. D. 1994. Statistical Methods for Groundwater Monitoring. John Wiley & Sons, Inc.

[13] Gilbert, R.O. (1987). Statistical Methods for Environmental Pollution Monitoring. Van Nostrand Reinhold, New York.

[14] Iglewicz, B., Hoaglin, D. How to detect and handle outliers. ASQC Quality Press, 1993.

[15] Barnett, V. (1983). Principles and methods for handling outliers in data sets. Statistical Methods and The Improvement of Data Quality, pp. 131-166S. Chen, B. Mulgrew, and P. M. Grant, "A clustering technique for digital communications channel equalization using radial basis function networks," IEEE Trans. on Neural Networks, vol. 4, pp. 570-578, July 1993.

AUTHOR BIOGRAPHY

Dr. Reji Kurien Thomas is a highly commended former military Commander who specializes in IT Security. An avid Aviator and Gold Medalist in every course undertaken, he has been educated from the best of institutions including Stanford University and Kent State University in the United States of America. He has presently founded a firm specializing in IT Security and Energy Healing and doing extensive research in the fields of Nanotechnology, Scalar Energy, Quantum Mechanics, Cryptology and Network Communication. He is a member of Project Management Institute (PMI), USA since 2013 and has attended several international conferences and publications on wireless sensor networks and cryptography. He also has several certifications in the computer field like cloud computing, network security to name a few.

Dr.T.Anand is presently working as an Associate Professor, Dept. of Computer Science and Engineering in Madha Engineering College, Chennai, Tamil Nadu. He has 10 years of teaching experience. He has presented research papers in more than 15 national and international conferences and published 3 papers in national and international journals. His research areas include Networking and Cloud Computing. He is a member of Computer Society of India and life Time Member of ISTE. He has guided more than 20 Post Graduate Engineering Students. He finished his Master of Computer Education in University of Madras. He completed his M.Phil (Computer Science) in Periyar University then he was awarded Master of Engineering in computer science and Engineering by Vinayaka Missions University.