Worm Detection by Combination of Classification With Neural Networks

Tawfiq Barhoom[1] and Hanaa Qeshta[2]
[1]The Islamic University – Gaza, Palestine
[2]College of Science and Technology , Palestine

Abstract *Security has become ubiquitous in every domain today as newly emerging malware pose an ever increasing perilous threat to systems. Worms are on the top of malware threats attacking computer system although of the evolution of worm's detection techniques. Early detection of unknown worms is still a problem. In this paper, we proposed and implemented a new approach for worm's detection. The proposed model uses combination of Decision Tree and Neural Network (NN) as classifying worm/ non worm traffic network. Our results showed that the detection rates of classification and detection known worms are at least 93.51% with NN, and 92.87% with Decision Tree, while the unknown worm detection rates was about 97.27% with NN , and 93.2% with Decision Tree. The detection rate of our proposed model in known worm was 95.59% while the unknown worms detection rate was 97.74%.*

Keywords: *Worms, Worm Detection, Classification, Data Mining, Decision Tree, Neural Network.*

Received August 21, 2011; Accepted November 6, 2012

1. Introduction

There is no doubt that while the internet is widely growing and more and more network services are being created, the number of the internet users is extremely increasing. The internet, as no other communication medium, has given an international dimension to the world. It has become the universal source of information for millions of people, at home, at school, and at work.

Few network security systems are able to cope with network-based attacks, such as worms. Their inability to detect zero-day attacks make them unfit to protect networks against the current proliferation of these attacks and the speed at which they propagate. The few systems that attempt to detect unseen network attacks are usually crippled by other factors such as the amount of information to be processed or the computational needs of their techniques. Worms are malicious processes that spread autonomously from one host to another they cause major problems in today's networks [10][13]. Current worm defense begins with manual worm detection followed by damage repair. Automatic detection is particularly challenging because it is difficult to predict what form the next worm will take. However, automatic detection and response is fast becoming an imperative because a newly released worm can infect millions of hosts in a matter of seconds. Worms are divided in two different kinds [1]: direct worms, which don't need a medium to propagate, because they use computer networks, exploiting operating systems bugs or weaknesses. Indirect worms, which spread in an "indirect" way, using deceitful means like peer to peer file sharing or, as already said, e-mails. The worm's life as consisting of many phases: target finding, transferring, activation, and infection. The first two phases cause network activities, worm behaviors in these two phases are critical for developing detection algorithms [1][8], in this paper we focus on direct worms.

1.1. Intrusion Detection System (IDS), and Worm Detection

The Purpose of the Intrusion Detection System (IDS) is to monitor network assets in order to detect misuse or abnormal behavior, which statistically analyzes input data (e.g., network traffic) [4]. The types of IDS have been proposed can be divided into two categories: network based (NIDS) and host based (HIDS). Network based (NIDS) tries to detect any attempt to subvert the normal behavior of the system by analyzing the network traffic. Host based (HIDS) to act as the last line of defense, which is seek to detect intrusions by analyzing the events on the local system where the IDS is being run. Generally host based IDSs classified into two categories: anomaly detection and misuse detection. Misuse detection try to identify behavior patterns that are characteristic of intrusions, but this can be difficult if an attack exhibits novel behavior, as it may when attackers develop new strategies. Anomaly detectors try to characterize the normal behavior of a system so that any deviation from that behavior can be labeled as a possible intrusion. Anomaly detection assumes that misuse or intrusions are strongly correlated to abnormal behavior exhibited

by either the user or by the system itself. Anomaly detection approaches must first determine the normal behavior of the object being monitored, and then use deviations from this baseline to detect possible intrusions [10].

Most studies has been proposed to detect worms as a misuse detection by using signatures of worms. This approach is unable to reach zero-day state and hence is not effective in protecting networks against the current deployment of worms and the speed at which they spread. On the other hand, another approach has been proposed to detect worms based on anomaly behavior detection where it is possible to detect abnormal behaviors and generate alarms, this approach is useful to detect unknown attacks. Also, current worm defense begins with manual worm detection followed by repairing the damage.

Several types of machine learning techniques were used in the field of intrusion detection in general and in detecting worms in particular. Data Mining has an important role and is essential in worms detection systems, where several of worms detection is mainly based on data mining techniques [7]. But many researches have shown that standalone anomaly classifiers used by anomaly detection systems are unable to give acceptable accuracies in real-world deployments [5]. Therefore, the combination method is particularly useful for difficult problems, and is more likely to achieve higher accuracies [2].

In this paper we propose techniques to detect and classify internet unknown worm by using data mining and neural network approaches.

2. Worms Detection Problem

With increasing in challenges of security issues, and increasing of thread and attacks ways such as worms attack, many important issues remain: First, detection systems must be more effective, detecting a wider range of attacks with fewer false positives. Second, detection system must keep pace with modern networks' increased size, speed, and dynamics. The worms still one of the most infection malware codes dangerous. To detect the internet worm, many approaches were proposed and based on signatures for misuse detection that can't detect unknown/new worms. Most commercial programs are normally based on the signature approach where the worm signature is available after the attack to the network system for several days or weeks. Moreover, signature extraction must be considered by using expert knowledge. Thus, network anomaly detection and real times detection was become a great challenge.

3. Related Works

Many recent researches in few last years were proposed and present Worms Detection domain based on artificial neural network, and data mining as an efficient ways to increase the security of networks.

Several researches tried to suggest methods by using data mining technique [3, 10, 11, 1, and 12]. Farag et-al [3], proposed a model for detecting unknown worms based on local victim information. Produced model is used to identify worm traffic from normal traffic, also it can predict the infection percentage in the network. Their proposed system uses Artificial Neural Network (NN) for classifying worm/ non worm traffic. Sarnsuwan et-al [10], produced techniques to detect and classify many types of internet worm at network end-point by using data mining approaches which are Bayesian network, C4.5 Decision tree and Random forest. They use port and protocol profiles to train and test their detection models. Siddiqui et-al [11], presented the idea of using sequence of instructions extracted from the disassembly of worms and clean programs using data mining techniques as the primary classification feature. The analysis is facilitated by the program control flow information contained in the instruction sequences, they formulated the problem as a binary classification problem and built tree based classifiers including decision tree, bagging and random forest. Aiello et-al [1], proposed a new technique to detect internet worm. Their research based on the fact that an indirect worm needs to spread quickly and so it sends a lot of e-mail in a short while, producing an anomalous behavior. They found stealthy worms through detecting traffic anomalies. They worked on a mail-server log of a real network and the results obtained drove us to detect indirect worm with different approaches based on various parameters. They focus their attention on one mail server. Stopel, et-al [8, 11], proposed the models to detect malicious activity of worms by looking at the attributes derived from the computer operation parameters such as memory usage, CPU usage, traffic activity etc, using Artificial Neural Networks ANN, and two other known classifications techniques, Decision Tree and k-Nearest Neighbors. Stopel et-al [12], proposed an approach for detecting the presence of computer worms based on ANN using the computer's behavioral measures. The identification of the most prominent features to capture efficiently the computer behavior in the context of worm activity and compared three different feature selection techniques for the dimensionality reduction in order to evaluate the different techniques, several computers were infected with five different worms and 323 different features of the infected computers were measured. The evaluation each technique by preprocessing the dataset according to each one and training the ANN model with the preprocessed data. Then evaluated the ability of the model to detect the presence of a new computer worm, in particular, during heavy user activity on the infected computers. The best accuracy was achieved by using only five attributes selected by the Fisher's score

method. The average accuracy of new worm detection using these attributes is 0.90. These five attributes were related to memory management, and number of system context switches. Stopel et-al [13], presented a new approach based on ANN for detecting the presence of computer worms based on the computer's behavioral measures, and used two other known classifications techniques, Decision Tree and k-Nearest Neighbors, to test their ability to classify correctly the presence, and the type of the computer worms even during heavy user activity on the infected computers. By comparing these three approaches, the ANN approach has computational advantages when real-time computation is needed, and has the potential to detect previously unknown worms. Addition, ANN may be used to identify the most relevant, measurable features, and thus reduce the feature dimensionality. Rasheed et-al [9], proposed intelligent early system detection mechanism for detecting internet worm. The average of failure connections by using Artificial Immune System (AIS) is the main factor that their technique depends on in detecting the worm. Their paper shows that intelligent Failure Connection Algorithm (IFCA) operation is faster than traditional algorithm in detecting worms. Their results show that the IFCA detects the worm faster than traditional Failure Connection algorithm, and the algorithm can detect different types of worms. Wang et-al [14], proposed a new worm detection approach based on mining dynamic program executions that captures dynamic program behavior to provide accurate and efficient detection against both seen and unseen worms. The execution on a large number of real-world worms and benign programs, and trace their system calls. They applied two classifier-learning algorithms which are Naive Bayes (NB) and Support Vector Machine (SVM)) to obtain classifiers from a large number of features extracted from the system call traces. The learned classifiers are further used to carry out rapid worm detection with low overhead on the end-host. The experimental results clearly demonstrate the effectiveness of proposed approach to detect new worms in terms of high detection rate and a low false positive rate.

4. Research Objective

The main objective of this research is to propose model which is an adaptive worms detection model based on anomaly-behavior detection that can detect known and unknown worms by using combination of classification with artificial neural networks in order to achieve an acceptable accuracy.

There are many specific objectives extracted from the main objective:

- Identifying the different types of worms, their way of spreading, the phases of the worms' life, and the extent of its effect on the behavior of network traffic. These factors help us to extract important characteristics and are important for building our model.
- Applying worms' detection model based on anomaly-behavior detection using supervised learning machine technique and by combination of classification in data mining with neural networks, so that to be able to detect both known and unknown worms.
- Training our model on normal network behavior, to be able to predict the behavior of non-normal, and to be effective in protecting the network and detect known/ unknown worms' attacks.
- Testing our model on new untested network behavior, to observe the system's ability to detect this behavior, so that we can prove that this model is able to adapt and to detect known/ unknown worms.
- Trying to test various behaviors of our model and evaluating the results.
- Reduce the false positive and negative rate, and improve the detection rate through the measurement and evaluation by using programs and tools.
- Improve network security and protecting them from threats of worms.
- Introducing a new way of detecting zero-day worms' attacks.

5. Worms List

We describe several information of various worms that are used in our experiments. Many characteristics of each worm including Port profiles, and rate of scan per second used by worm to infect new hosts [1 and 14], are shown in Table 1.

Table1. Worm characteristics.

Worm	Scan per second	Port
CodeRedII	4.95	TCP 80
Rbot-AQJ	0.68	TCP 139,769
Zotob.G	39.34	TCP 135,445,UDP 137
SoBig.E	21.57	TCP 135,UDP 53
Sdbot-AFR	28.26	TCP 445
Rbot.CCC	9.7	TCP 139,445
Forbot-FU	32.53	TCP 445
Blaster	10.5	TCP 135,444 UDP 69

- Code red II uses a buffer overflow to exploit vulnerability on Microsoft IIS web servers. After the worm propagates itself to any host, it sends DOS attack and provides backdoors to attackers. Then, this worm will find new hosts to infect with port 80 on TCP.
- Rbot.AQJ worm provides backdoors and allows attackers to remotely access on the vulnerable computer via IRC channels on Windows platform with ports 139 and 769 on TCP protocol.

- Zotob.G exploits buffer over flow vulnerability on MS Windows Plug and Play and provides backdoors to attackers with ports 135, 445 on TCP protocol and port 137 on UDP protocol.
- Sobig.E worm is attached with email or spam mail from bil@Microsoft.com and upport@yahoo.com. If any user opens this file, the worm will start its process. This worm spreads to other hosts with port 135 on TCP protocol and port 53 on UDP protocol.
- Sdbot-AFR worm exploits a buffer overflow vulnerability of Windows and provides a backdoor to attackers with port 445 on TCP protocol. this worm has a higher rate of scan per second.
- Rbot.CCC worm also provides backdoors and allows attackers to remotely access on the vulnerable computer via IRC channels on Windows platform, this worm propagates itself with ports 139 and 445.
- Forbot-FU propagates itself to other hosts with Trojan/Optix on Windows. This worm exploits buffer overflow vulnerability of Windows and provides backdoor to attackers with port 445 on TCP protocol.
- Blaster worm exploits a buffer overflow vulnerability of DCOM RPC on Windows XP and Windows 2000 by connecting to ports 135 and 444 on TCP protocol and port 69 on UDP protocol. This worm can download and operate itself. After that, the worm sends DOS attacks to prevent patch update by sending SYN flood to the destination port 80.

6. The Proposed Methodology

6.1. Data Set

The dataset was download from [15], which were collected from 13 different network endpoints. Each network end-point has different behavior from each other. Input dataset includes number of features selected from many features that help in identifying worm behavior from non-worm network behavior. Most of these inputs are numerical values. Each end host was installed with actual worm and simulated worm.

The reason behind the selection of these datasets is the datasets were collected from the network end points such as homes, offices and universities. In addition, some end points run peer to peer applications. Each instance of dataset has 7 attributes as follows in table2.

The datasets were separated into several categories in terms of normal data and type of worm. In addition, datasets from each end point were collected into different groups, for example, 13 end points have 13 groups. Example of datasets is shown in Table 3.

From Table3, the direction column has one byte flag represented by an integer where 1 represents "incoming broadcast packets", 2 represents "outgoing broadcast", 3 represents "outgoing unicast" and 4 represents "incoming unicast". The protocol column represents transport-layer protocols using an integer such as 6 represents "TCP" and 17 represents "UDP". The Key code column is one byte virtual key code that identifies the data types such as "d9" represents worm behavior and others represent normal behavior. There are many port numbers used in normal class data such as port numbers 22, 53, 80, 123, 135, 137, 138, 443, 445, 993 and 995 which are known ports (0:1023) and registered ports (1024:65535) for specific applications such as on-line Games and peer to peer applications.

Table2. Description of dataset attribute.

Attribute	Description
Session id	20-byte SHA-1 hash of the concatenated hostname and remote IP address.
Direction	one byte flag indicating as outgoing unicast, incoming unicast, outgoing broadcast or incoming broadcast packets.
Protocol	Transport-layer protocol of the packet.
Source port	Source port of the packet.
Destination port	Destination port of the packet
Timestamp	Millisecond-resolution time of session initiation in UNIX time format.
Virtual key code	One byte virtual key code that identifies the data if it is normal data or worm.

Table3. Example of a data profile.

Session ID	Direction	Protocol	Src Port	Des port	Time Stamp	Key code
Sha-1 code	3	6	2025	445	1130861747.125	d9
Sha-1 code	4	17	1026	53	1130863119.917	0

6.2. Prepare And Preprocessing Phase

- We used the datasets from [15], and selected some attributes are Direction, Protocol, Src Port, Des Port, Time Stamp and Key Code columns to be used as our datasets.
- Replacement the data key code from "d9" to "Worm", and replacement normal data key code to "Normal", in order to facilitate conducting experiments practical.
- Select all attributes of dataset expect the "Session ID" which are (Direction, Protocol, Src Port, Des Port, Time Stamp and Key Code) to be used as our datasets in the experiments.
- The collection of dataset from all 13 endpoints by sampling method and then cluster it, from 2 causes of experiments, in all experiments we divided the sample of dataset into 70% to training phase, and

30% to testing phase we explain the processed of each case:

- First case: the experiment done in this case into 2 phases, the first is training with 5000 profiles(3500 worms and 1500 normal profiles), and second is testing with 2000 profiles (1200 worms and 800 normal profiles).
- Second case: the experiment done in this case one 2 datasets, one dataset contained on all type of worms with sampling 6500 profiles (2500 normal and 4000 worm profiles), and the other dataset contained all types expect one of types as unknown with sampling 6000 profiles(2500 normal and 3500 worms profile). The training phase was on the first dataset, and the testing phase was on second dataset. The details of these experiments show in the Table3.

Table4. Training and testing datasets

Case	Training phase		Testing phase		Output
	Worms	Normal	Worms	Normal	
1	3500 All types	1500	1200 All types	800	2 classes
2	3500 All types expect one	1500	4000 All types	2500	2 classes

6.3. The Proposed Models

We proposed our models based on classification in data mining which is Decision Tree and Neural Network (NN), as classification. The train of these models to detect and classify worms by using dataset, as shown in Figure 1.

6.4. Data Mining Model

Using Decision Tree technique which is proposed by a training/learning dataset and built from rules that are created during the training. These rules are used to predict and classify datasets. To classify an unknown instance, the Decision tree will start at the root and traverse to a leave node. The result of classification and prediction occurs at the leave node, as shown in Figure 2.

6.5. Neural Network Model.

Neural Network is an adaptive system that changes its structure based on external or internal information that flows through the network during the learning phase. Modern neural networks are non-linear statistical data modeling tools. They are usually used to model complex relationships between inputs and outputs or to find patterns in data, as shown in Figure 3.

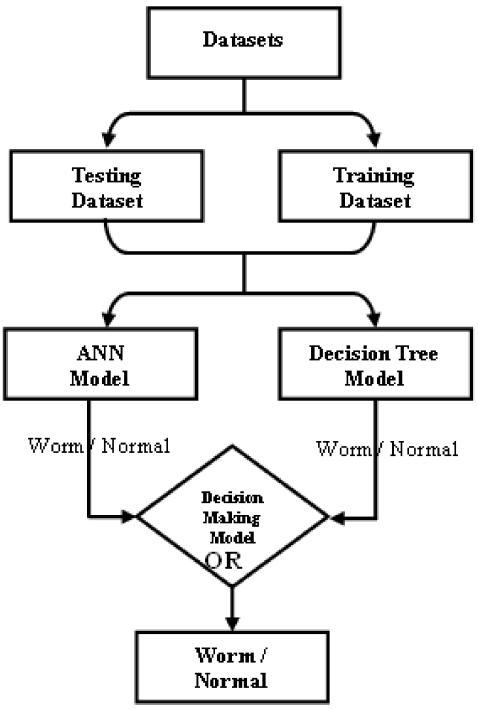

Figure 1. Proposed worm detection model.

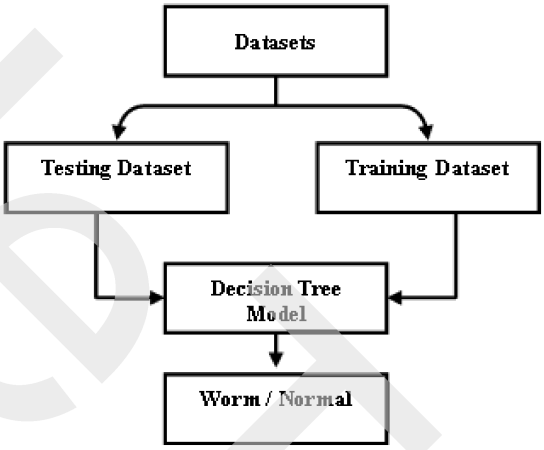

Figure 2. Decision tree model.

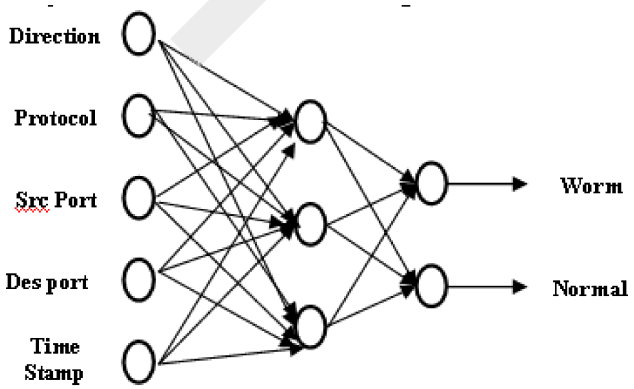

Figure 3. NN Model.

6.6. Experimental Setup

We procedure the experimental by using the Rapid Miner program to implement our model, using data mining and artificial neural network techniques.

In our experiments, in the first step, the performance of each detection and classification model is measured and compared by using the detection rates which are True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN). The performance parameters (TP, FP, TN, and FN) are described as follows:

- True Positive (TP): an algorithm classifies Worm according to the actual data (Worm).
- False Positive (FP): an algorithm classifies Worm opposite from the actual data (Normal).
- True Negative (TN): an algorithm classifies Normal according to the actual data (Normal).
- False Negative (FN): an algorithm classifies Normal opposite from the actual data (Worm).

$$Overall\ Accuracy = \frac{|TP| + |TN|}{|TP| + |FP| + |TN| + |FN|}$$

To calculate the detection rate, let N_{Worm} be the total number of Worm profiles and N_{Normal} be the total number of Normal profiles in the testing dataset. Thus, we have

$$Detction\ Rate = \frac{N_{worm} \times |TP| + N_{normal} \times |TN|}{N_{worm} + N_{normal}}$$

From experiments, our models can classify and detect known worms and unknown worms with high detection rates without feature extraction. The results from case 1 are shown in table 5, where the detection rates of Decision Tree and Neural Network models are over 93%. Table 6 presents the detection results with unknown worms that we consider in case 2, that can detect unknown worms with overall detection rate over 94%. In particular, the Decision tree can detect unknown worms with the average detection rate over 93%, while NN have average detection rates from all experiments over 94%. Table 7 showed the results of our proposed model, as it achieved with average detection rate more than 97%,

Table 5. Results of case 1 (known worm detection).

Model	Detection rate (%)	TP (%)	TN (%)	FP (%)	FN (%)
Decision Tree	92.87	96.39	85.61	15.1	3.28
NN	93.51	97.42	86.73	14.3	2.91

Table 6. Results of case 2 (unknown worm detection)

Model	Detection rate (%)	TP (%)	TN (%)	FP (%)	FN (%)
Decision Tree	93.2	96.84	86.57	14.41	2.69
NN	94.27	97.35	88.07	14.1	2.04

Table 7. Results of proposed model

Case #	Detection rate (%)	TP (%)	TN (%)	FP (%)	FN (%)
1	95.59	97.16	90.05	11.89	1.17
2	97.74	98.68	92.57	10.37	1.02

From table 5, and 6 we comparing Decision Tree and NN models, in case 1, and case 2, we observed the detection rate of NN was the best result, while the results from our proposed model was the best, as it achieved the results more than 97% in table 7.

7. Conclusion

In this paper we proposed new approach based on combination of Data Mining which is Decision Tree and Neural Network (NN). The proposed model produced good results in worm detection. The advantage of the NN and Decision Tree methods over other techniques were their ability to classify correctly a worm not used in the training. The proposed model produced perfect results with accuracy of 95.59% which using dataset in case 1 that detecting known worms, and 97.74% which using dataset in case 2 that detecting unknown worms.

Table 8. Comparing between detection rates of all models.

Detection rates / Model	Case # 1	Case # 2
Neural Network	93.51	94.27
Decision Tree	92.87	93.2
Propose model	95.59	97.74

8. Future Work

Future research could work to used multi classifiers, and other kinds of malware. Try to using clustering methods to build the model to detect known and unknown worms. Also, try to find a new method to detect worms by using collaborative methods classification and clustering in conjunction with one another.

References

[1] Aiello M., & et al, "Worm Detection Using E-mail Data Mining", *Primo Workshop Italiano su PRIvacy e Security*, 2006.

[2] Ashfaq A., Javed M., & Khayam S., *"An Information-Theoretic Combining Method for Multi-Classifier Anomaly Detection Systems", IEEE Communications Society subject matter experts for publication in the IEEE ICC*, 2010.

[3] Farag I., & et al, "Intelligent System for Worm Detection", *International Arab Journal of e-Technology*, Vol.1, No. 1, 2009.

[4] Holst, A., Ekman, J., & Larsen, S., "Abnormality Detection in Event Data and Condition Counters on Regina Trains". *The Institution of Engineering and Technology International Conference on Railway Condition Monitoring*, pp. 53 – 56, 2006.

[5] Ismail I., MarsonoM., and Nor S., "Detecting Worms Using Data Mining Techniques: Learning in the Presence of Class Noise", *Sixth International Conference on Signal-Image Technology and Internet Based Systems*, 2010.

[6] Kruegel Ch., Valeur F., Vigna G., *Intrusion Detection and Correlation, Challenges and Solutions. Book on Advances in Information Security*, Springer Science and Business Media, Inc., USA, Vol. 14, 2005.

[7] Law K. & Kwok L., IDS False Alarm Filtering Using KNN Classifier. *Lecture Notes in Computer Science, Springer Berlin / Heidelberg,* pp. 114-121, 2005.

[8] Li P., Salour M., & Su X., "A Survey of Internet Worm Detection and Containment". *Communications Surveys & Tutorials, IEEE*, 2008.

[9] Rasheed M., & et al, "Intelligent Failure Connection Algorithm for Detecting Internet Worms", *International Journal of Computer Science and Network Security*, Vol. 9, No.5, 2009.

[10] Sarnsuwan N., Wattanapongsakorn N., and Charnsripinyo Ch., "A New Approach for Internet Worm Detection and Classification", *Networked Computing (INC), 6th International Conference, 2010, and The 13th National computer science and engineering conference: green computing technology*, 2009.

[11] Siddiqui M., & Wang M., "Detecting Internet Worms Using Data Mining Techniques", *Cybernetics and Information Technologies, Systems and Applications (CITSA),* 2008.

[12] Stopel D., & et al, "Improving Worm Detection with Artificial Neural Networks through Feature Selection and Temporal Analysis Techniques", *International Journal of Mathematical and Computer Sciences*, 2005.

[13] Stopel, D., & et al, "Application of Artificial Neural Networks Techniques to Computer Worm Detection", *International Conference on Neural Networks ICNN International Journal of Applied Mathematics and Computer Sciences*, 2006.

[14] Wang X., & et al, "Detecting Worms via Mining Dynamic Program Execution", *Authorized licensed use limited to: The Ohio State University*, 2008.

[15] Wireless and Secure Networks (WiSNet) Research Lab at the NUST School of Electrical Engineering and Computer Science (SEECS), http://wisnet.seecs.edu.pk/

Tawfiq Barhoom received his Ph.D degree from ShangHai Jiao Tong University (SJTU), in 2004. This author is the Dean of faculty of IT, Islamic University-Gaza. His current interest research include Design Patterns, Secure Software, Modeling, XMLs security, Web services and its Applications and Information retrieving.

Hanaa Qeshta received her B.Sc. degree in computer science in 2002. M.Sc. degree in information technology in 2012. Both B.Sc. and M.Sc. degrees from The Islamic University-Gaza, Palestine. She part-time lecturer in IT Dept, College of Science and Technology-KhanYounis, and part-time lecturer in IT Dept, Al Aqsa University- KhanYounis. Her research of interest includes computer networks, security systems, data mining, and intelligent systems.