Enhancing Malware Detection using Innate Immunization

Mohamed Ahmed Mohamed Ali[†,††] , **Mohd Aizaini Maarof**[††],

[†]Faculty of Mathematical Sciences , University of Khartoum, Khartoum, Sudan
[††]Faculty of Computing , Univirsiti Teknologi Malaysia, Skudai, Malaysia

Abstract
The massive amount of malware created everyday made the process of malware detection is a significant process to protect data and systems. The methods used are varying from signature based to behavior based, and from static to dynamic detection. Detection accuracy is the main obstacles facing the researchers in this field. Artificial immune system is one of the methods used frequently these days because of its ability to simulate the human immune system and take advantage of its strength in the detection of diseases. In this paper we introduce a dynamic hybrid signature-behavior base model by applying the innate immune system to enhance the detection accuracy. The proposed model is using the portable executable (PE) file representation and API call logs extracted from windows environment because of the wide spread of this type of files in different platforms. The results show that the proposed model accomplishes a better performance in detection of known malware, new unknown malware and polymorphic malware.

Key words:
Malware detection; artificial immune system; innate immune system

1. Introduction

Malware (MALicious softWARE) is a software designed to access secretly a computer system without the owner's consent which includes viruses, worms. Trojan hourses and other types of malicious software [1]. It represents the main threat facing the computer systems, networks and data. Creators of malware have different reasons to spread their product. Stealing important data, espionage on others systems, abuse of network resources are some examples of what malware can do [2],[3],[4]. The traditional method to detect malware depends on a stored database containing distinctive marks of these programs called signature has become ineffective due to the new types of malware that can hide itself from detection software, or change their shapes constantly (polymorphic malware). Malware writers also used malware packing technique by encrypting the executable malware file and hide the known signature to evade the detection process running by Anti-Malware or malware scanners [5]. Dealing with malware presents three types of processes, classification , analysis and detection. Malware classification is to categorize the malware depend on some criteria to make the dealing with it easier. Malware analysis is a multi-step process providing insight into malware structure and functionality, facilitating the development of an anti-malware. Analysis by behavior monitoring is used to observe malware interaction with respect to the system involve, (sandbox) is one of the tools used in this stage. Finally, the process malware detection, the detection process is to identify malware within the system by using a signatures of known malware or by detecting a suspicious behavior within the system. Malware detection techniques could be categorized to signature or behavior based depend on the heuristics to identify malware. Artificial Immune System (AIS) creates a new platform to solve some problems such as pattern recognition, data mining, intrusion detection and malware detection [6],[7],[8]. In the next section we address some of the malweare detection techniwues followed by the immune system, artificial immune system and finally we present the proposed model with its reulsts and findings.

2. Malware Detection

A malware detector is a system that attempts to identify malware. A virus scanner uses signatures and other heuristics to identify malware is an example of a malware detector [9]. Malware detection process consisting of two main approaches for the detection of malware: static analysis and dynamic analysis. Static analysis examines the binary code to determine properties of this program without running it. This technique was first used by compiler developers to optimize the code. On the other hand Dynamic analysis mainly concerned in monitoring the execution of a program to detect malicious behavior [10], [11].

3. The Immune System

The immune system is a vital, highly evolved biological system consists of many processes and rules within the organism and its core mission is to protect the organism from various diseases and injuries and trying to keep the neighborhood fabric. The main function of the immune system is to identify and eliminate the unfamiliar material such as pathogens and bacteria and different types of

Manuscript received October 5, 2013
Manuscript revised October 20, 2013

viruses. It does this largely without prior knowledge of the structure of these pathogens. It works by monitoring all cells and tissues and distinguish between normal cells that are part of the organism and the alien cells from the body until the defense process complete in a proper manner. The constant evolution of the viruses and pathogens makes this detection process is complex and very difficult and adaptation is nedded. The major types of the immune system involved in the protection process are the innate immune system and the adaptive immune stsem, the physical barriers like skin and mucus also take some responsabilties in the defence against pathogenes and viruses and considerd part of the defence mechanism [12].

3.1 The Innate Immune System

The innate immune system is so called because the body is born with the ability to recognize certain microbes and immediately destroy them. The cells of the innate system such as macrophages have receptors on its surfaces to bind with the patterns associated with the viruses called Pathogen Asssosiated Molecular Pattern PAMPs. Our innate immune system can destroy many pathogens on the first encounter [6].

3.2 The Adaptive Immune System

Adaptive immunity is the use of antigen-specific receptors on T and B cells to drive targeted effecter responses in two stages. First, the antigen is presented to and recognized by the antigen specific T or B cell leading to cell priming, activation, and differentiation, which usually occurs within the specialized environment of lymphoid tissue. Second, the effecter response takes place, either due to the activated T cells leaving the lymphoid tissue and homing to the disease site, or due to the release of antibody from activated B cells (plasma cells) into blood and tissue fluids, and hence to the infective focus [13].

4. Artificial Immune System

Artificial immune system is a leading area of research over the past two decades. It is a part of the Bio-inspired computing which is using computer modeling nature. At the same time studying the nature improved the usage of computer, combine the power of computer artificial intelligenc, machine learning with the biological immune system to resolve various issues does not solve yet in the computer system environment and one of these issues the problem of malware recognition in the field of computer security [14], [15], [16]. Different algorithms were applied in this research area to solve some chronic problems in comuter systems such as pattern recognition, data mining, malware detection.

5. Dynamic Innate Immune System Model

This section presents the proposed novel model for malware detection. The model is using a portable executable file representation and API call logs extracted from windows environment because of the wide spread of this type of files in different platforms. The data set used in this research was downloaded from a well known research group website in the computer security field [17],[18]. The use of this data set in many research projects in malware detection will make it a suitable way to evaluate our results with others models. The dataset was downloaded from www.vxheaven.org [19]. The data set has a total of 514 files in a text format showing the windows function and the parameters used with these functions. The data set contains 98 benign files obtained from Windows 7, 117 Trojan file, 165 virus file and 134 worm file. The file size is ranged from 100KB to 15000KB. Some of malware families also shown clearly in the data set. Table I shows the file distribution in the data set. Firstly, we execute the portable executable (PE) files and extract the API call log using API monitor v2 and save the extracted API calls as a text file. The commands extracted can be categorized into tow parts, the function command and the parameters paced to the function. Fig.1 shows a captured sample from the text file of a Trojan called AVKILL running on Win32 environment. The malware sorted in families depend on the malware name as Trojan, virus, worm. Another sorting is made within the same type is by name of the file to create more relation between the families. At the end of this sorting process we come with a group of families each of which contains from two to three related malware. API calls devided into Functions (F), parameters (P), based on the document frequency feature selection method the dynamic model build a matrix data structure to store the functions, parameters of the API calls extracted.

TABLE I. File distributions of the dataset

Malware type	Number of samples
Trojan	117
Virus	165
Worm	134
Benign	98
Total	514

```
explorer*0x194*IsBadReadPtr*lp:0xDEEAA0,
ucb:0x10*0x0*SUCCESS*0
explorer*0x194*IsBadReadPtr*lp:0xDEECF8,
ucb:0x10*0x0*SUCCESS*0
explorer*0x194*IsBadWritePtr*lp:0xDEED70,
ucb:0x10*0x0*SUCCESS*0
explorer*0x194*IsBadReadPtr*lp:0xDEECF8,
ucb:0x10*0x0*SUCCESS*0
```

Fig. 1. API calls sample

The matrix stores the functions and parameters of the API calls and do the mapping between them. The matching between malware families is made by analyzing the functions appeared in each file of the same family. Three types of matching were made to enhance the detection ability and to reduce errors. The first matching is from type to type for example Trojan to virus or virus to worm, the second matching made from family to family within the same type of malware such as Win32.Newbiero Win32.Netsp worm families, lastly within the same family within the same type of malware like Win32.HLLP virus family. The definition of the matrix could be shown as:
F= functions in the text file extracted. .
P= parameters called by the functions.
M= (F, P) the matrix represent the relation between the function and its parameters.

6. Experiment Results

The dynamic innate malware detection shows a promising results in detection known malware samples from the same families with a the capability to detect other malware not in the same malware families with some other unknown malware as we define earlier as a polymorphic malware. Table II shows the detection results.

TABLE II. Detection results

Malware type	Detection accuracy
Trojan within family	99.5%
not in family	98.1%
Virus within family	97.3%
not in family	96.8%
Worm within family	99.3%
not in family	98.8%

7. Limitations and future work

Even the malware detection shows a promising results in detection but the high storage needed to store the matrices of (F,P) for the malware samples was a problem, small systems with very limited resources could not run the model. The future work is to make this model work in a small amount of resources such as storage and processing time.

8. Conclusion

The innate immune system is a highly complicated and powerful protection system. It protects the living organism from evading viruses. The majority of the innate immune components has not utilized yet. In this work we are trying to take advantage of some of that system. Malware detection accuracy is a major goal for the researchers dealing with detection of malware. Many models and frameworks proposed during the last two decades, but have their limitations because of the accuracy. In this paper we try to address the problem of detection accuracy by applying the innate immune system. The results and findings enhanced the detection accuracy with some limitations of the high storage resources must be resolved in the future.

Acknowledgment

This research is supported by the Faculty of Computing, Universiti Teknologi Malaysia, Malaysia, Faculty of Mathematical Sciences, University of Khartoum, Sudan.

References

[1] S. Kramer and J. Bradfield, "A general definition of malware," Journal in Computer Virology, vol. 6, pp. 105-114, 2010.[2]N. Idika and A. P. Mathur, "A Survey of Malware Detection Techniques," Department of Computer Science,Purdue University, West Lafayette, IN 47907.USA2007.
[2] P. Vinod, et al., "Survey on Malware Detection Methods," Malaviya National Institute of Technology, 2009.
[3] Y. Zhang, et al., "Biologically inspired model for computer virus detection," in System of Systems Engineering (SoSE), 2012 7th International Conference on, 2012, pp. 66-69.
[4] L. Sun, et al., "Pattern Recognition Techniques for the Classification of Malware Packers," in Information Security and Privacy. vol. 6168, R. Steinfeld and P. Hawkes, Eds., ed: Springer Berlin / Heidelberg, 2010, pp. 370-390.
[5] L. N. D. Castro and F. J. Von Zuben, "Artificial Immune Systems:Part I - Basic Theory and Applications," Technical Report1999.
[6] R. Tian, et al., "Function length as a tool for malware classification," in Malicious and Unwanted Software, 2008. MALWARE 2008. 3rd International Conference on, 2008, pp. 69-76.
[7] H. Yin, et al., "Panorama: capturing system-wide information flow for malware detection and analysis," presented at the Proceedings of the 14th ACM conference on Computer and communications security, Alexandria, Virginia, USA, 2007.
[8] M. Christodorescu, et al., "Semantics-aware malware detection," in Security and Privacy, 2005 IEEE Symposium on, 2005, pp. 32-46.
[9] C. Cifuentes and A. Fraboulet, "Intraprocedural static slicing of binary executables," 1997, p. 188.
[10] J. Bergeron, et al., "Static detection of malicious code in executable programs," Int. J. of Req. Eng, pp. 184–189, 2001.
[11] N. K. Jerne, "Towards a network theory of the immune system," Annales d'immunologie, vol. 125C, pp. 373-389, 1974.
[12] J. Parkin and B. Cohen, "An overview of the immune system," The Lancet, vol. 357, pp. 1777-1789, 2001.

[13] D. Dasgupta, et al., "Recent Advances in Artificial Immune Systems: Models and Applications," Applied Soft Computing, vol. 11, pp. 1574-1587, 2011.
[14] T. Burczyński, et al., "Immune Computing: Intelligent Methodology and Its Applications in Bioengineering and Computational Mechanics," in Computer Methods in Mechanics. vol. 1, M. Kuczma and K. Wilmanski, Eds., ed: Springer Berlin Heidelberg, 2010, pp. 165-181.
[15] M. A. M. Ali and M. A. Maarof, "Malware Detection Techniques Using Artificial Immune System Proceedings of the International Conference on IT Convergence and Security 2011." vol. 120, K. J. Kim and S. J. Ahn, Eds., ed: Springer Netherlands, 2012, pp. 575-587.
[16] S. Manzoor, et al., "A Sense of 'Danger'for Windows Processes," Artificial Immune Systems, pp. 220-233, 2009.
[17] M. Z. Shafiq, et al., "Improving accuracy of immune-inspired malware detectors by using intelligent features," presented at the Proceedings of the 10th annual conference on Genetic and evolutionary computation, Atlanta, GA, USA, 2008.

Mohamed Ahmed Mohamed Ali received his B.Sc. in Computer Science from Univesity of Khartoum, Sudan in 1999. M.Sc in Computer Science from University of Khartoum, Sudan in 2006. Now he doing his PhD at Universiti Teknologi Malaysia. His main research interests are in the areas of Computer Security, Malware Detection, Artificial Immune Systems, Bioinformatics. Member of Information Assurance & Security Research Group (IASRG) at UTM.

Mohd Aizaini Maarof received his B.Sc (Computer Science) from WMU - USA, M.Sc (Computer Science) from CMU - USA, and PhD (IT Security) degree from Aston University, Birmingham, UK. He is a Professor at Faculty of Computer Science & Information System, UTM. His research interest is in Information System Security. He is also a member of Information Assurance & Security Research Group (IASRG) at UTM.