A SURVEY ON DETECTION TECHNIQUES OF MALICIOUS SOFTWARE OR MALWARE

[1]PANKAJ DEOSKAR, [2]ANJU SINGH, [3]DIVAKAR SINGH

[1]MTech Scholar CSE Deptt. BUIT, [2]Astt. Prof. of IT Deptt.,
[3]HOD of CSE Deptt. BUIT, Barkatullah University, Bhopal
Email- pankaj_78600@yahoo.com, [2]asingh0123@rediffmail.com, [3]divakar_singh@rediffmail.com

Abstract: Malware or sometimes also called malicious softwares are serious threats to the personal computer as well as for cyber security. These are basically software which is being written in some of the programming language such as c,c++ or any other conventional programming language. Many antivirus software (AVS) has been developed for their deletion but is possible only when keys of malware must be identified, but by this it would be to late to protect the system. The aim of this paper is to study of the detection of malware by using ADT (anomaly detection technique) by identifying the critical features.

Keywords: Bioinformatics, Anomaly Detection Technique, MSA, Data Mining.

I. INTRODUCTION

Bioinformatics is the combination of biology and Computer science in which biologist has taken the help of computer experts to develop a software for sequence analysis. Sequence analysis was basically used in late 1970's for the identification of genes information. They have used in the sequence of two spices of DNA and amino acid to understand the relationship between them. Advantages of sequence analysis and sequence alignments were used to identify the conserved regions of the biological data in order to identify common genes that would be applicable two or more spices.

It was the time of early 21'st century that computer experts tried to detect the malware (Malicious + Software) by applying the similar techniques which has been used in the identification of genes. This was done by pattern matching in order to identify the critical features of malware. Critical features of malware can be identified by applying the powerful data mining tools and techniques.

Malware is also known as malicious software there are mainly two types of malwares:

Dependent malicious software.

Independent malicious software.

Examples of there malwares are virus, worms, spyware, logic bombs, Trap Doors and Trojans etc.

There are several methods for alignment and many algorithms used in the study of biology in the sequence analysis areas. In general, an alignment is basically an adjustment of a sequence as compared to other sequences. The aim is to arrange two (pairwise alignment) or more (multiple sequence alignment) possibly variable length sequences of DNA or protein in such a way that regions of similarity across sequences (rows of a matrix) fall in the same successive columns of the matrix, where such similarity signifies functional, structural or evolutionary commonality. Global alignment tries to align every item in every sequence and tends to work best when the sequences are of roughly similar length, such as the Needleman-Wunsch techniques. Local alignment, on the other hand, tries to align regions of the sequences even if the sequences are not similar overall, such as the Smith- Waterman technique. A Multiple sequence alignment (MSA) is a sequence alignment of three or more biological sequences which is generally used in the alignment of proteins, DNA or RNA Multiple sequence alignment also refers to the process of aligning such a sequence set. Because three or more sequences of biologically relevant length can be difficult and are almost always time consuming to align by hand. Therefore different computational algorithms are used to produce and analyze the alignment.

Alignment of Malware

There are several types of alignment methods which are used to detect computer virus. The aim of sequence alignment is to convert variable length alignment into fixed length. The reason behind this aim is to detect a small pattern malware instead of detecting the entire file. Also it will be applicable in that situation when a group of different virus pattern is being given at a time, they must be classified according to their behaviour and structure. The aim of SA is to verify that whether it is applicable for the detection of different types of malwares.

II. LITERATURE SURVEY

Literature survey of the paper is based on detection of malicious software by applying anomaly techniques which is different from the concept of viral signature data base which is using in traditional antivirus soft wares (AVS)

Lin Chen et al had propose there work for the host-based systems. They have tried to overcome fundamental limitations of traditional host based anti-malware systems, which are likely to be deceived and

attacked by malicious codes. An layered detection model proposed to overcome these two issues.

Currently many works are going on the detection of malware which are adversely affecting in the different areas of computer and their applications. They are as follows;

Yi Chen et al. in "Malicious Software Detection Using Multiple Sequence Alignment and data Mining tools" has commented on precision, accuracy and sensitivity. But they had not given any clear picture on the classification of the different malware.

" A layered malware detection model using VMM" proposed by Lin chen et. Al was based on only host based system. But this model was not secured for the client side systems

"Polymorphic Malware Detection using hierarchical Hidden Markov Model" proposed by Fahad Bin et. AL which was restricted to only a class of computer virus he had not commented on the classification of other malwares.

"Data Mining for Malicious Code Detection and Security Applications" proposed by Bhavani Thurai Singham is a great innovation in the field of machine learning approach.

All of their work were mostly based on anomaly technique. Further they have used different data mining algorithms in order to classify whether the new upcoming data set is malware or not. They all are working in different areas of computer and there applications.

In the above techniques they usually train & test the data because sometimes virus and worms are not being classified differently.

Most of the previous work was done on the detection of malicious program but non of them had clearly commented on classification of existing patterns of computer virus and worms. However Bhavani Thurai Singham has proposed her paper on the detection of malware. In her paper she suggested that anomaly technique is quiet efficient for the detection in case to detect unusual patterns and behaviors further she concluded that we also need to build models in real time for real time intrusion detection. Data mining is also being applied for credit card fraud detection and biometrics related applications some progress has been made on topics such as stream data mining.

As it is known that computer viruses and other forms of malware pose a threat to virtually any software system. A common technique that virus writers use to avoid detection is to enable the virus to change itself by having some kind of self modifying code. This kind of virus is commonly known as metamorphic virus and can be particularly difficult to detect. Out of entire virus family, metamorphic virus can be treated as simple sequences of op-codes then sequence analysis techniques used in other field of study like bioengineering could be used to develop Hidden Markov Model (HMM). This profile would then be used to score an arbitrary op-code sequence. If the output score exceeds a designated threshold it could be concluded that he input sequence was likely to have been from that same virus family. One of the most common technique to detect viruses is called signature detection, which involves an analysis of known viruses to find signatures or strings of bytes, which are found in viruses and not in most non-malicious code.

Later on A. Narayanan, X. Wu and Z. R. Yang had research on an understanding of viral protease specificity may help in the development of future anti-viral drugs involving protease inhibitors by identifying specific features of protease activity for further experimental investigation according to their research and experiments their results show that artificial neural network and symbolic learning techniques captures some fundamental and new substrate attributes but neural network outperform their symbolic counterpart.

Later a new method T-coffee (Tree-based Consistency Objective Function For Alignment Evaluation) is a software which has used a progressive approach for multiple sequence alignment that provides a dramatic improvement in accuracy. M-coffee expression and 3D- coffee are the different variant of T-coffee. Last but not least ANN accuracy precision and sensitivity has been calculated on unaligned sequences and doubly aligned sequences.

CONCLUSION

Detection of malware by using pattern matching and applying data mining tools is quiet new approach in this field of research area. Anomaly detection technique proved to be a boom in the field of classification of malwares. At first detection is based on classification of computer virus and computer worms because they have the similar functionalities. Then it would be applicable for different malwares. Also it may be useful in the detection of the new upcoming dataset in order to identify whether the patterns of new dataset consists malware or not.

REFERENCES

[1] Yi Chen, Ajit Narayanan,Ban Tao and Shaoning Pang "Data Mining for Malicious Code Detection and Security Applications", 26th IEEE International Conference,2012, pp- 8-14.

[2] Nawar Malhis, Kelvin Xi Zhang and BF Francis Ouellette, "Detecting Protein-domains DNA-Motifs Association in Saccharomyces cerevisiae Regulatory Networks", 21st International Conference IEEE-2012,

[3] Xin Li, Xuefeng Zheng, Jingchun Li, and Shaojie Wang "Frequent Itemsets Mining in Network Traffic Data", Fifth International Conference, 2012, pp-394-397.

[4] E. Corchado, and Á. Herrero, "Neural visualization of network traffic data for intrusion detection", Appl. Soft Comput., vol. 11, March 2011, pp. 2042-2056.

[5] D.M. Mount, Bioinformatics: Sequence and Genome Analysis, 2nd ed.. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press Cold Spring Harbor, NY

[6] Y. Tang, B. Xiao and X. Lu, "Signature Tree Generation for Polymorphic Worms", IEEE Transactions on Computers, vol. 58, no. 4, 2011, pp. 565-579.

[7] Bhavani Thuraisingham "Data Mining for Malicious Code Detection and Security Applications" IEEE 2011.

[8] Hogeweg, P. Searls, B. David. ed. "The Roots of Bioinformatics in Theoretical Biology". PLoS Computational Biology **7** (3), 2011. www.ncbi.nlm.nih.gov/pmc/articles/PMC3068925/.

[9] Jozsef Hegedus, Yoan Miche, Alexander Ilin and Amaury Lendasse "Methodology for Behavioral-based Malware Analysis and Detection using Random Projections and K-Nearest Neighbors Classifiers", 5th IEEE conference- 2011, pp 1016-1025.

[10] J. Kinable and O. Kostakis "Malware classification based on call graph clustering", Journal in Computer Virology, 2011, pp. 1–13.

[11] Fahad Bin Muhaya, Muhammad Khurram Khan and Yang Xiang "Polymorphic Malware Detection Using Hierarchical Hidden Markov Model", IEEE 9th International Conference, 2011 pp – 151-155.

[12] S. Cesare and Y. Xiang, "A Fast Flowgraph Based Classification System for Packed and Polymorphic Malware on the Endhost", in IEEE 24th International Conference on Advanced Information Networking and Application (AINA 2010), pp. 721-728.

[13] Y. Liu, L. Zhang, J. Liang, S. Qu and Z. Ni "Detecting trojan horses based on system behavior using machine learning method", in Machine Learning and Cybernetics (ICMLC), 2010 International Conference on, vol. 2, July 2010, pp. 855 –860.

[14] I. Firdausi, C. Lim, A. Erwin, and A. Nugroho, "Analysis of machine learning techniques used in behavior based malware detection", in advances in Computing, Control and Telecommunication Technologies (ACT), 2010 Second International Conference on, December 2010, pp. 201 –203.

[15] L. Sun, S. Versteeg, S. Boztaş, and T. Yann, "Pattern recognition techniques for the classification of malware packers", in Information Security and Privacy, ser. Lecture Notes in Computer Science, R. Steinfeld and P. Hawkes, Eds. Springer Berlin / Heidelberg, , vol. 6168, 2010, pp. 370–390.

[16] A. Srivastava and J. Giffin, "Automatic discovery of parasitic malware", In Recent Advances in Intrusion Detection (RAID'10), ser. Lecture Notes in Computer Science, S. Jha, R. Sommer, and C. Kreibich, Eds. Springer Berlin / Heidelberg, vol. 6307, 2010, pp. 97–117.

[17] T. Blasting "An Anriod Applications sandbox of suspicious detection", 5th international conference on malicious software, 2010 pp- 860-864.

[18] J Cuff and G. Barton "Evaluation and Improvement of Multiple Sequence Methods for Protein Secondary Structure Prediction, Proteins: Structure. Function. Genetic", IEEE conference-34, 2009, pp-508-519.

[19] T. Xinguang, D.Miyi, S. Chunlai and L. Xin. "Detecting network intrusions by data mining and variable length sequence pattern matching", J Sys Eng Electr, 20 (2), 2009, pp 405-411.

[20] E. Menahem, A. Shabtai, L. Rokach, and Y. Elovici, "Improving malware detection by applying multi inducer ensemble", Computational Statistics & Data Analysis, vol. 53, no. 4, 2009, pp. 1483 – 1494.

[21] Lin Chen, Huaping Hun, Bo Liu, and Qianbing Zheng "A layered malware detection model using VMM" , IEEE 11th International Conference, 2009, pp 1259-1264.

[22] Symantec "Symantec internet security threat report.pdf: Volume XII,"Symantec 2008."

[23] B. Payne, M. Carbone, M. Sharif, and W. Lee. Lares "An architecture for secure active monitoring using virtualization", In Proceedings of the IEEE Symposium on Security and Privacy, 2008.

[24] X. Jiang, D. Xu, and X. Wang. "Stealthy malware detection through vmm-based "out-of-the-box" semantic view reconstruction", In Proceedings of the ACM Conference on Computer and Communications Security, 2007.

[25] S. H. Oh and W. Lee. "A clustering-based anomaly intrusion detection for a host computer", IEICE Trans. on Information and Systems, E87-D(8), 2004, pp-2086–2094.

[26] T. Garfinkel and M. Rosenblum, "A Virtual Machine Introspection Based Architecture for Intrusion Detection Proceedings of Network and Distributed System" Security Symposium (NDSS'03), San Diego, California, USA. 2003, pp- 1-16.

[27] T.F. Smith and M.S. Waterman, "Identification of Common Molecular Subsequences", Journal of Molecular Biology 147, 1981, pp. 195–197.

[28] S.B. Needleman and C.D. Wunsch, "A general method applicable to the search for similarities in the amino acid sequence of two proteins", Journal of Molecular Biology 48 (3), 1970, pp. 443–453.

★★★