RESEARCH ARTICLE OPEN ACCESS

Automated Analysis of Malicious Document Files

Ajinkya R. Bhosale [1], Swati S. Joshi [2]
PG Research Scholar [1], Assistant Professor [2]
Department of Computer Science and Enggineering
N. B. Navale Sinhgad COE, Solapur
Maharashtra-India

ABSTRACT

Malware is an important topic of security threat research. One challenge in malware analysis involves collecting useful data without risking experimenter's machines or systems. Static analysis of malware is valuable in providing insights on malware development mechanisms. Malicious code (or Malware) is defined as software that fulfils the deliberately harmful intent of an attacker. So, The process of determining the behaviour and purpose of a given Malware sample (such as Trojan horse, virus and worm) is called Analysis of Malware. This process is a important step to be able to develop effective detection techniques and removal tools. Now a days, Analysis of malware is a manual process that is tedious and time-intensive. To resolve this problem, a number of analysis tools have been proposed that automatically extract the behaviour of an unknown program by executing it in a restricted environment and recording the operating system calls that are invoked.

The approach we have proposed here can help security researchers to analyse these files get dump of its malicious contents and then proceed accordingly. Because of this approach malware researchers can write more generic solutions for these malware files. It will also reduce the time that is needed for analysing these files.

Keywords:- Malware, Internet Security, Virus, Worms, Trojan.

I. INTRODUCTION

An Automated Analysis of Malicious Document Files is an approach to analyze malicious contents in Office documents such as Word Files, Excel Files, Power Point Presentations, PDF Files etc. This analysis system is completely safe that it cannot harm any existing system or the system on which it is going to be analyzed. So, this approach is completely safe that anyone can use it on their existing system.

Now days, Internet has become our one of the basic needs, so it is quiet normal that many users download office documents or PDF Files from online sites. Because of all this malicious office documents are becoming more and more common in users. So, it increases the risk of user machines to get infected by these files. These malwares use vulnerabilities in PDF Readers or Office file Readers to infect user system. So, this is becoming more common approach for malware authors to infect user system.

The approach we have provided here can help security researchers to analyze these files get dump of its malicious contents and then proceed accordingly. Because of this approach malware researchers can write more generic solutions on these malware files. It will also reduce the time that is needed for analysing these files.

So, this approach not only helps us to analyze these types of malwares faster but also get generic solutions for that. It is really very useful way to do all this task.

The approach we have proposed here can help security researchers to analyze these files get dump of its malicious contents and then proceed accordingly. Because of this approach malware researchers can write more generic solutions for these malware files. It will also reduce the time that is needed for analysing these files.

This can prove that the proposed work can significantly improve the quality of work for Security Researchers and reduce time that is required by the process.

II. LITERATURE REVIEW

Following are some existing studies and approaches related for Automated Analysis of Malicious Document Files:

Lenny Zeltser published many articles on it and he also has tools and techniques that are useful for analysing malicious documents. These techniques are really very useful. We can get these tools and techniques for use in our approach. This cheat sheet outlines tips and tools for reverse-engineering malicious documents, such as Microsoft Office (DOC, XLS, PPT) and Adobe Acrobat (PDF) files.

Didier Stevens has also published all pdf analysing tools in his articles. These methods and techniques are also really good for analysing pdf files and objects in it. They have also explained all about PDF. This tool will parse a PDF document to identify the fundamental elements used in the analyzed file. It will not render a PDF document.

In http://forensicswiki.org/wiki/PDF, they have also published all the related articles about PDF and its format. It is also a good source of information for PDF files. They have explained all about PDF. The PDF is a document format from Adobe Inc. It is widely available on the web. Originally developed as a propriety format, version 1.7 was released as an open standard in 2008. The standard is published as ISO/IEC 32000-1:2008. Although an open standard, Adobe still owns patents and copyrights related to the PDF standard. Adobe has granted a worldwide royalty-free license to produce PDF software, but only if the software complies with the PDF standard.

In http://www.offensivecomputing.net/, they have published such a great tools called as PDF-Xray. This tool actually is very useful. It has almost all things that are needed to analyze pdf files. We can use it in our approach to analyze PDF files.

In http://www.reconstructer.org/, they have published really great articles about office files such as Word, Excel, and PowerPoint etc. The new version of the OfficeMalScanner suite introduces RTFScan. As you might know, there are several samples in the wild, using the RTF format as OLE and PE-File container. So here is a very first version of RTFScan. It currently is able to scan for malicious traces like shellcode, dumps embedded OLE and PE files and other data containers. Buffer decryption in RTFScan is not supported in this release, as OMS and RTFScan will be enhanced to a cryptanalysis feature to break keys up to 1024 bytes in seconds. The old brute force feature in OMS will be kicked then.

III. PROBLEM STATEMENT

Due to huge growth in number of malicious documents, it is a much tough job for Security Researchers to analyze them and give solution on them. Thus analyzing theses documents is main problem which they are facing now.

IV. OBJECTIVES AND SCOPE

The main objective is to analyze increasing number of malicious documents and get the best possible result through it.

Objective of this dissertation are:

1) Locate potentially malicious embedded code, such as shell code, VBA macros, or JavaScript.
2) Extract suspicious code segments from the file.
3) If relevant, disassemble and/or debug shell code.
4) If relevant, obfuscate and examine JavaScript, Action Script, or VB macro code.
5) Understand next steps in the infection chain.

Scope:

Proposed approach is used not only for reducing the time for analysis of malicious documents but also it improves the quality of output they get from the analysis. Thus this approach is really significant.

V. PROPOSED METHODOLOGY

The proposed approach can help security researchers to analyze these files get dump of its malicious contents and then proceed accordingly. Because of this approach malware researchers can write more generic solutions for these malware files. It will also reduce the time that is needed for analyzing these files.

Locate Potentially Malicious Embedded Code:

In this step we first try to locate potentially malicious embedded code. For this we have to analyze malicious document file and we have to finalize which content is malicious in that file. After that we can finally say that this part of file is malicious.

For this purpose we can use some tools that will help us to identify which part is malicious. Generally malicious contents are shellcode, VBA macros, or JavaScript etc. If we found some of these malicious contents in office file then we can say that the file is malicious.

So, If we found such as shellcode, VBA macros, or JavaScript in Office documents or PDF Files, then we first try to extract it and then tries to DE obfuscate if it is obfuscated or beautify to make it readable. These kinds of operations have been done on these malicious contents.

Extract Suspicious Code Segments:

Once we have decided which part of the file is malicious so we can finally get the dump of that malicious part. But it is much complicated job to decide which part is malicious. So, here Extracting Suspicious Code Segments means that all malicious code segments are dumped individually.

In this process we get dump of all malicious parts of current Office Document and then we proceed for further analysis.

If relevant, disassemble and/or debug Shellcode:

It is the most commonly found malicious contents in Office files. For analysis of shell code we have to disassemble or debug it so that we can find what is the actual purpose of that shell code.

So, when we get dump of shellcode, first we try to disassemble it or debug it so we can find its malicious activity. Once we find the malicious activity it would easy for us to understand the actual target of the malware.

If relevant, obfuscate and examine JavaScript, Action Script, or VB macro code:

It is also most commonly found malicious part in Office files. Most of the PDF Files have JavaScript included in them and also they can have Action Script.

Once we find any of them in dump then we can proceed further for analyzing them. First we obfuscate them or beautify them so that we can read them correctly and understand the code. Once the code is readable we can get the idea of its purpose and then we can find out the actual operation of the malware.

After getting the actual behaviour we get ready to understand the malware and its whole system so it is very important step for us.

Understand next steps in the infection chain:

Once we get our shellcode, JavaScript, ActionScript etc then we get completely ready to understand the total behaviour of the malware and what is its purpose.

This is really important for our whole malware analysis thing. So, it is very importance for us to get there.

VI. CONCLUSION

The approach we have proposed here can help security researchers to analyze these files get dump of its malicious contents and then proceed accordingly. Because of this approach malware researchers can write more generic solutions for these malware files. It will also reduce the time that is needed for analyzing these files.

This can prove that the proposed work can significantly improve the quality of work for Security Researchers and reduce time that is required by the process.

REFERENCES

[1] Practical Malware Analysis Book by Michael Sikorski and Andrew Honig

[2] Reversing: Secrets of Reverse-engineering by Eldad Eilam

[3] Practical Reverse Engineering by Bruce Dang