ABSTRACT

As the primary vehicle for most organized cybercrimes, malicious software or malware has become one of the most serious threats to computer systems and the Internet. With the recent advent of automated malware development toolkits, there is no need for amateur cyber-offenders to be familiar with the intricacies of coding or programming. This has led to a surge in the number of new malware threats and has created several major challenges for the Anti-virus industry.

Malware has affected desktop computers and mobile devices alike. Microsoft Windows is the most widely used consumer operating system worldwide, mostly used in desktop computers. Hence a vulnerability identified in one system could compromise a large number of systems. On the other hand, the ubiquity of mobile devices has led to a growing adoption of mobile applications in a variety of application areas and 98.05% of all malware detected in 2013 targeted the Android platform, confirming both the popularity of this mobile OS and the vulnerability of its architecture. Hence, in this thesis the study and analysis of the malware affecting the above mentioned operating systems is done to find better solutions to mitigate the various threats caused by malware.

The traditional methods of protecting a system through signature based detection are beset with drawbacks justifying the need to look for better and efficient solutions. This thesis aims to address the problems related to traditional methods and arrive at solutions that are more effective. The work

in the thesis centers around investigating Data Mining approaches to design efficient malware detection mechanisms and to overcome the drawbacks of traditional methods both in Windows and Android based systems.

Analysis of the various malware is a crucial step to understand it better in order to build effective defense mechanisms for protecting the systems. The two approaches of malware analysis widely used in literature are Static and Dynamic analysis. Although understanding malware using dynamic analysis can provide a comprehensive view, it is still subjected to high cost in environment deployment and manual efforts in investigation. In this thesis, we do a thorough static feature-based analysis of two types of malware, one affecting the world's widely used operating system-Windows, and the other affecting the recent most popular mobile operating system-Android. Multiple features are extracted from executable files and the most discriminating subset of features is used for further detection methods.

Malware belonging to families of Backdoors, Trojans, Flooders, Viruses, Worms, Constructors, DoS in Windows and Android have been analyzed and the discriminating features are identified. For Windows based malware, the analysis is done based on PE header features and API calls. For Android based malware, the Permissions declared in the manifest file and the API calls are used for further analysis.

Decision trees are one of the simplest and commonly used classification systems. Decision trees have been used in malware detection due to their simplicity and ability to form generalized rules. The performance

of detectors using decision trees can be improved by reducing the average height of the tree. The standard approach to decision tree induction is a top-down, greedy algorithm that makes locally optimal, irrevocable decisions at each node of a tree. In this thesis, an alternative approach based on Theil index is proposed, in which the algorithm uses limited lookahead to decide what test to use at a node while constructing the tree so that you get shorter and better trees. The major limitation of this method is the time it takes due to lookahead.

In an ensemble of classifiers, the challenge is to choose the minimal number of classifiers that achieve the best performance to reduce the processing time which is a crucial metric for malware detection. We make use of Harmony search, a music inspired algorithm to prune the ensemble and find the best subset of classifiers for the detection task. Two algorithms are proposed. HS_ENSEM binary uses binary solutions that aim to include or exclude a classifier from the ensemble. HS_ENSEM weighted uses real number solutions which are tuned with the help of harmony search to become the weights of classifiers. Classifiers falling below a particular threshold are excluded from the final ensemble.

A scalable detection mechanism for Android based malware is designed using Multifeature collaborative decision fusion (MCDF) method. We combine heterogeneous features of a malicious file like the Permission based features and the API call features in order to provide a better detection

by training an ensemble of classifiers and combining their decisions using collaborative approach based on probability.

Ensemble techniques are mainly divided into two categories: multiexpert systems and multilevel systems. In multiexpert systems, the classifiers work in parallel whereas in multistage systems they use a serial approach where the next classifier is trained for patterns only if they are rejected by the previous classifiers. Multilevel combination is more interesting in the sense that it is possible not to use costlier classifiers, unless they are actually needed, i.e., when previous simpler classifiers' predictions are not confident and have a high probability of error. In this thesis, a multilevel (multistage) system is proposed where features that are easier to extract and low cost classifiers are made use of in the first level. If the confidence of the level 1 classifier falls below a particular threshold, level 2 is considered for detection. Since our concern is not only accuracy but also cost, combining several classifiers may not be feasible due to the increased memory and computation needed. Besides, the time to output is very important for a malware detection application to be usable in real life. Therefore we opted for a multilevel ensemble system with a small number of classifiers, balancing between accuracy and cost.

The various works mentioned above tries to find a rational solution to the important problem of malware detection. Data mining techniques have been used to identify interesting patterns due to the large volume and diversity of malware files.