D. Adware

Recent malware application tries to be legalized and hard to detect by antivirus companies. Adware method has come up for that reason. The Advertisement is a legal operation for all over the world to announce or mention about a brand or a product. Therefore advertisers want to show their brand or product to as much as people using advertising services like television, radio, websites and the services which people can integrate into their solution to gain money. Although adware applications are not malicious applications, they are highly disturbing applications and they are mostly categorized as potentially unwanted applications(PUP).

III. DETECTION MECHANISM

It clearly needs to understand the Android application structure to detect them and block them because Android has its own unique application structure that only runs on the Android operating system. Android has permission system to restrict access to the critical components like the internet, SMS, browser utilities, camera etc [6]. However, detecting malicious applications using permissions or combination/sequence of permissions could not succeed in the wild. That is why recent malicious applications(adware applications) requires the permission that only normal applications use such as internet permission. Therefore, detection mechanism has been evolving to string and method features which are stored in the dex file format [7].

In the beginning, there were two sets of applications; malicious and benign. They are needed to extract string and method features to cluster them into multiple malware families. Because of application replication, which means appending malicious code to the legitimate application to replicate them with the malicious version, it is hard to cluster malicious application using generic features like strings and methods. Therefore, the strings and methods which belong to the malicious application need to be subtracted from all strings and methods which belong to clean applications. That will cause to have pure strings and methods which identify malicious behavior. That process will help further processes just by decreasing strings and method count and processing time as well.

Clustering mechanism relies on simple techniques like a threshold to achieve to join the cluster. A malware sample compared with all clusters and the sample will join the most similar cluster. However, in this research, there is another invention to make the clustering better and more accurate. The simple threshold mechanism does not fit into that research because as far as a cluster increases file number it should become more stable and more characteristic cluster that identify that malware family specifically. As you can see in "Fig. 1." the problem has been solved by logarithmic addiction technique.

Therefore, addiction method helps to solve the problem. In this research initial threshold set to 1 and as long as cluster become larger addictive value compound with the initial threshold to become harder to join a cluster. For that reason the first files in the clusters should be fit into the malware family characteristics otherwise some of the family may be divided into multiple clusters unnecessarily and that will cause the noisy data for further processes like signature creation.

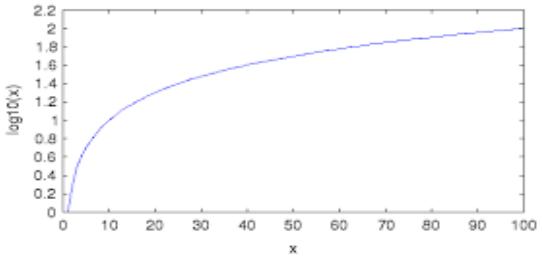

Fig. 1. Addictive function log10

The Similarity of a feature or feature sets was another challenge for this research topic. There are numerous algorithms to calculate the similarity of two files or even two text files. In this project, it has been solved by another simple solution like the function as shown in "Fig. 2.". It consists of two parts; string similarity function and method similarity function. Both functions have the same formula as the length of the intersection of the file which attempts to join the cluster and the cluster itself divided by the length of clusters. After calculating the values of those function, they are multiply by preset values respectively, 1.0 and 1.5 and at the end, the sum of those two values is the similarity of two files.

$$f_{string}() = \frac{|(cluster \cap strings)|}{|cluster|}$$

$$f_{method}() = \frac{|(cluster \cap methods)|}{|cluster|}$$

$$fsimilarity() = fstrings \times 1.0 + f_{method} \times 1.5$$

Fig. 2. Similarity formula

Those techniques which are mentioned in the above just for clustering malware samples into separate multiple groups to characterize them better. The next process to detect a malicious application is the generating a signature for a malware sample or for a malware family. Generating signature fully depends on the malware engine, which helps to scan the files and detect malicious files using the virus signature database because that engine will extract the features and try