Short Paper Review of:

peHash: A Novel Approach to Fast Malware Clustering

Citation:


Author(s) Affiliation(s):

- Malware Collect .org
- RWTH Aachen University

Claims:

- Present a non cryptographic fast calculating hash function for binaries in a PE format that can transform structural information about the sample into the hash value.
- Evaluation of the tool on different malware sets to show that the results allow for a significant reduction in sample counts.
- The authors claim that they will show that Anti-Virus is too slow and inaccurate to scan each malware sample, thus blacklisting based on this is not possible.

Likes:

- peHash is now a part of Dionaea, as well as many other collecting systems as an alternative to sha-1 and md5sum. However I was unable to at least doing a shallow search to find much more about it.
- The algorithm is straight forward and spelled out in the paper, even if you didn't have the source code you could easily reproduce it from what's given.
- I was glad to see that the followed up on the potentially broken clusters by doing manual analysis of all 282 + 322 malware samples that were in those groups.

Dislikes:

- Normally I wouldn't care but since this is a published proceeding I do... I found some spelling errors.
I thought the first statement of the introduction regarding the existing systems would have benefited from more than just one reference to a single example. I would have preferred a paper analyzing the different types of systems that exist. The same holds true for the second statement about automated analysis.

In my external research I found that pehash is used in a number of other automated analysis software tools but I was unable to find any stand alone OSS that did nothing but calculate the pehash of a binary.

**Repeat / Add to Work?**

First let me start off by saying that I'm a mwcollect.org user, contributor, etc. and know the admins very well so I'm slightly biased. I found this paper to be of particular interest because I do keep a database of almost 300,000 malware samples. Many of which may be duplicates of the same malware. The basic premise of this paper revolves around the idea that a hash will be generated the same for all variants of a specific instance of malware no matter what it has polymorphed into. They do this by creating a hash function for PE's (Portable Executable) that generates hashes based on the structural data and information about the malware. This information is then clustered and groups of the malware variants emerge. This method was tested on the mwcollect.org alliance database which is publicly accessible if you're a member so the database used for testing is readily available.

This is a system that I would be interested in tweaking and potentially releasing my own code on as well as results. I could see using my own malware dataset as well as the newer mwcollect.org and openinfosecfoundation.org datasets as case studies. I really liked this paper thought I think that it would have benefited from more analysis and representational statistics/graphs.

PeHash focused on portable executables but I've seen a lot of other malware types out in the wild and not all polymorphic malware is in that container format.