Guidance Software

Hash Sets and Their Proper Construction

By Shawn McCreight and John Patzakis

I. EXECUTIVE SUMMARY

Analyzing a large set of files by identifying and matching the unique MD5 hash value of each file is an important part of the computer forensics process. The hash library feature of EnCase allows the investigator to import or custom build a library of hash sets, enabling the expedient identification of any file matches in the examined evidence. Computer forensics analysts often create different hash sets of known child pornography images, hacker tools, or non-compliant software to quickly isolate any files in an investigation that are included in that set.

When creating hash sets to identify suspect software, such as non-licensed software, steganography or counterfeiting utilities, it is important that examiners carefully create these hash sets to prevent false positives. This paper provides an overview of the benefits of hash set analysis and outlines the process for the proper creation of hash sets.

II. DEFINITION OF TERMS

MD5 HASH: MD5 is an algorithm that is used to generate a unique 128-bit fingerprint of a file of any length, (even an entire disk). Because there are 10^{38} different possible hash values, it is highly unlikely that any two files would have the same hash value. Further, it is "computationally infeasible" at this time to manufacture a file that generates a particular hash value. Thus, known files can be reliably identified through their MD5 hash value.

SAFE HASH SETS: These sets consist of hash values of files that are known to be innocuous. Safe hash sets are used to filter files from an investigation that are harmless

Copyright © 2001 Guidance Software, Inc. May not be copied or reproduced without the written permission of Guidance Software.

and simply get in the way. System files that have not been modified since installation are good examples of file hashes normally included in safe hash sets.

NOTABLE HASH SETS: Notable hash sets consist of hash values of known files that may be of interest to the examiner, thus warranting further attention. Hash values of known viruses, child pornography images, or classified company documents are examples of notable hash sets.

HASH CATEGORY: EnCase enables Hash sets to be organized into categories to provide more information about an identified file. For instance, hash sets for MS Office files could have the category of "Ignore" or "Safe," while the "Back Orifice" set could have the category "Hacker Tools."

BASE INSTALL: Many organizations configure their PC workstations with a uniform clean base installation of standard software and files, also referred to as a "gold disk."

III. SAFE HASH SETS

Safe Hash Sets are an important but underutilized resource for the forensic investigator. Using these sets can greatly increase the speed of a keyword search, as the search engine will then ignore innocuous files and not accumulate positive but irrelevant keyword hits. As generic system files are replete with terms such as "hacker" or "sex," search results can become flooded with irrelevant results. Notably, EnCase will ignore files with known hashes values, but will still search the slack areas of those files. Additionally, many known system files are small graphics files, which can be filtered out by their hash values to focus the investigation on image files that the user actually viewed or otherwise placed on their system.

When creating Safe Hash Sets, the examiner should be as inclusive as possible. For software applications, the examiner should include all the files in the installation. For instance, installing Windows, Office or Word Perfect and then hashing the entire folder structure is a good technique for generating these sets. Another effective measure for corporate examiners is to hash all the files on their organization's Base Install. That way, all known files can be filtered out at the outset of any internal investigation, leaving only the user-created and modified files for closer examination and text searches.

There is little downside in adopting an over-inclusive approach in creating Safe Hash Sets. If a hash of a file that is specific to the investigator's machine (such as an INI or CFG file, which are system files often uniquely modified by the exact configuration of the host computer) is included in a Safe set, this is not a problem since the odds of this same file being on the subject's computer are infinitesimal. If innocuous software, such as a word processing program, shares a generic dynamic link library (DLL) file with a hacking tool or other notable software application, the unique files in the notable application will be sufficient to positively identify the application. This important concept is discussed further in the next section.

Copyright © 2001 Guidance Software, Inc. May not be copied or reproduced without the written permission of Guidance Software.

IV. NOTABLE HASH SETS

Notable hash sets, which are often used to identify Child Pornography, Hacker Tools and perhaps stolen files from a company, are effective in identifying particular files for the presence of a file in a particular set. The presence of a file within a Notable Hash Set will quickly alert an examiner at the onset of his or her computer forensic examination. This is particularly important in investigations that involve large amounts of media, or in field investigations where an examiner must conduct an exam on-site. This is also an effective means to identify notable files that are deleted or renamed (but unaltered) or hidden in obscure folders.

EnCase features the ability to categorize different hash sets within the "Hash Category" field in the Case View. By giving a hash set a Hash Category, the examiner can better define particular sets. For instance, the "MS Office" hash set could have the category of "Ignore", while the "Child Porn 178" set could have the category "Possession = Felony." Most investigators will be more concerned with the Hash Category of a set at first. By sorting on the Hash Category, an examiner can see at a glance whether there are any nefarious or classified files in a case. This powerful technique will only be reliable if the author takes care when creating and labeling their sets.

Authors of Notable Hash Sets need to be extremely careful about which file hashes are included in these sets, since a single "false positive" can (at best) give the investigator a false lead, and may cause embarrassment later if not detected. It is not enough for a hash set author to simply install Stego Tools and hash all the folders. Here the user of the set relies on the expertise of the author in carefully choosing only those files whose presence would reliably indicate the presence of the set. In other words, the presence of "hacktool.exe" would indicate that a subject had installed the "Hacker Tools" application on their system. However, since "setup.dll" is potentially present in both MS Office and Hacker Tools, ***it must not be included*** in the Hacker Tools hash set. This could lead to the "discovery" of hacker tools on a system that only had MS Office installed. There are numerous files within the Hacker Tools folders, such as "ddosattack.exe" and so on that uniquely identify the presence of the hacker tools. It is crucial to include only these absolutely unique files.

The potential pitfall of false positives is only endemic to software tools, such as those used for counterfeiting or hacking, which share generic library files with innocuous applications or operating systems. There is no such concern of properly labeled hash sets of specific images or documents generating false positives.

V. CONCLUSION AND RECOMMENDATIONS

With investigations involving hundreds of gigabytes of media becoming routine, hash set analysis is becoming an absolute necessity in computer forensics investigations. Carefully created hash sets are essential to ensure the effectiveness of the technique and the confidence of the user.

Copyright © 2001 Guidance Software, Inc. May not be copied or reproduced without the written permission of Guidance Software.