UNITED STATES MILITARY ACADEMY

HOMEWORK 1

CS483: DIGITAL FORENSICS

SECTION I1

MAJ ADAM DUBY

BY

CADET ZACHARY BOLEN, CO C2, ‘22

WEST POINT, NEW YORK

20 January 2022

____ MY DOCUMENTATION IDENTIFIES ALL SOURCES USED AND ASSISTANCE RECEIVED IN COMPLETING THIS ASSIGNMENT.

____ I DID NOT USE ANY SOURCES OR ASSISTANCE REQUIRING DOCUMENTATION IN COMPLETING THIS ASSIGNMENT.

SIGNATURE____

Non-Total Recall: A Review of “Comparison of Fuzzy Hashing Algorithms for Binary Analysis”

Zachary T. Bolen
United States Military Academy
Dept. of Electrical Engineering and Computer Science
West Point NY, United States
zachary.bolen@westpoint.edu

***Abstract*—Pagini et al. present a consideration of the suitability of fuzzy hashing algorithms for detecting similarities between programs, describe a test method for understanding the intricacies behind varying results between algorithms, and draw conclusions about the right choice of algorithm for a given binary analysis task. This review first summarizes the paper and the authors’ findings and then assesses the effectiveness of the paper as a reference and guide post for future work in the field of digital forensics.**

***Index Terms*—binary analysis, fuzzy hash, approximate matching, malware, review**

I. Introduction

Fuzzy hashing (sometimes referred to as approximate matching) is a fundamental tool in the belt of forensic investigators and cyber security personnel. No defined standard exists for which algorithm or procedure to use, but a de facto standard has emerged in the form of `ssdeep`. [1] One objective of the reviewed paper was to outline the ways in which different fuzzy hashing algorithms respond to specific scenarios that make minor changes to the use case.

II. Analysis of Paper

A. Paper Strengths

The reviewed paper has an excellent introduction and a thorough background section that properly places the topic at hand in context. For readers with a background in computer science, cryptography, or cyber security the background provides more than enough information to allow those not actively involved in digital forensics to understand the fundamentals of the algorithms and how they might be measured against one another.

A major strength of the paper is the sections describing the test scenarios and experimental setup, sections 4 through 7. Section 4 is likely the most important regarding the methodology of the study since it describes the reasoning behind the chosen scenarios and the real-life situations they derive from. The clarity of the approach, as well as the supporting logic, serves to increase confidence in the results and conclusions of the paper. The results should be repeatable for independent verification, and the authors made it clear that they want these types of experiments to underpin future algorithm development. They also stress the importance of considering proper use cases for the analyzed algorithms, as each one is suited for different situations.

B. Paper Weaknesses

In light of the above section, the reviewed paper is incredibly informative and well-written. However, the scope of the paper detracts from its overall effectiveness. The combination of a significant background and history section with an extensive testing methodology and results portion divides attention and could subtract from a reader’s overall takeaway of the paper. Those in the community that the reviewed paper is primarily for will likely be able to skim or skip those sections and focus on the results, but students and newcomers will need to spend time digesting the halves separately to best appreciate the work done by the authors.

III. Applications to Digital Forensics and Future Work

Digital forensic investigators have been seeking consensus on the applicability and use of fuzzy hashing since `ssdeep` was published in 2007. The reviewed paper goes a long way toward establishing the fundamentals of what that consensus needs to take into consideration regarding use cases, likely scenarios, and the ever-accelerating arms race between bad actors and cyber security personnel. Future designers of fuzzy hashing algorithms and tools that make use of them will benefit from adapting the tests described in sections 4-7 for benchmarking purposes to ensure that their products stand up to scrutiny and the real-world situations that investigators will be facing.

IV. Conclusions

In conclusion, Pagini et al. designed a set of experiments that will guide the forensic community toward a better understanding of the tools and concepts at play when using fuzzy algorithms. The quality of the background section makes the paper accessible to students and people new to the field, while the detailed methodology and results sections are beneficial to other security researches hoping to improve the quality and credibility of their work.

Acknowledgment

Z.T.B. thanks Senator Jerry Moran of Kansas for his appointment to the United States Military Academy, as well as the Army Cyber Institute for its continued support of his education.

He also apologizes for the Total Recall joke in the title, but once he thought of it his mind was made up.

REFERENCES

[1] F. Pagani, M. Dell'Amico, and D. Balzarotti, "Beyond Precision and Recall: Understanding Uses (and MIsuses) of Similarity Hashes in Binary Analysis," In *CODASPY '18, Eighth ACM Conference on Data and Application Security and Privacy*. 12 pages. March 2018. ACM, New York, NY, USA. https://doi.org/10.1145/3176258.3176306