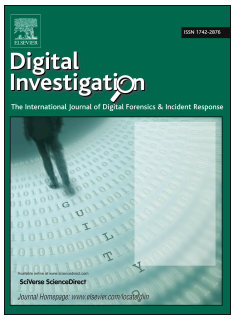
Accepted Manuscript



TREDE and VMPOP: Cultivating multi-purpose datasets for digital forensics – A Windows registry corpus as an example

Jungheum Park

PII: S1742-2876(17)30361-4

DOI: 10.1016/j.diin.2018.04.025

Reference: DIIN 777

To appear in: *Digital Investigation*

Received Date: 1 December 2017

Revised Date: 23 March 2018

Accepted Date: 26 April 2018

Please cite this article as: Park J, TREDE and VMPOP: Cultivating multi-purpose datasets for digital forensics – A Windows registry corpus as an example, *Digital Investigation* (2018), doi: 10.1016/j.diin.2018.04.025.

TREDE and VMPOP: Cultivating Multi-purpose Datasets for Digital Forensics - A Windows Registry Corpus as an Example

Jungheum Park[*]
Information Technology Laboratory, National Institute of Standards and Technology, Gaithersburg, MD 20899, USA

Abstract

The demand is rising for publicly available datasets to support studying emerging technologies, performing tool testing, detecting incorrect implementations, and also ensuring the reliability of security and digital forensics related knowledge. While a variety of data is being created on a day-to-day basis in; security, forensics and incident response labs, the created data is often not practical to use or has other limitations. In this situation, a variety of researchers, practitioners and research projects have released valuable datasets acquired from computer systems or digital devices used by actual users or are generated during research activities. Nevertheless, there is still a significant lack of reference data for supporting a range of purposes, and there is also a need to increase the number of publicly available testbeds as well as to improve verifiability as 'reference' data. Although existing datasets are useful and valuable, some of them have critical limitations on the verifiability if they are acquired or created without ground truth data. This paper introduces a practical methodology to develop synthetic reference datasets in the field of security and digital forensics. This work's proposal divides the steps for generating a synthetic corpus into two different classes: *user-generated* and *system-generated* reference data. In addition, this paper presents a novel framework to assist the development of system-generated data along with a virtualization system and elaborate automated virtual machine control, and then proceeds to perform a proof-of-concept implementation. Finally, this work demonstrates that the proposed concepts are feasible and effective through practical deployment and then evaluate its potential values.

Keywords: Forensic infrastructure, Dataset, Data corpora, Synthetic data, Reference data, Automated data generation.

1. Introduction

In recent years, the emergence and propagation of a broad range of information technologies are exploding, and they are being widely used for various purposes in our daily lives.

In a general accepted point of view, most software programs running on ICT (Information and Communications Technology) products are being imperfectly developed, and even when different developers implement program codes for identical operations, resultant codes and flows can be quite diverse. This is a result of a variety of factors such as the developers' level of skills and knowledge. These factors are more likely to cause serious issues in the field of information forensics and security, especially if software has potential vulnerabilities, errors or incorrect codes. For that reason, there have been demands for plausible mitigation strategies against the issues. As an example of such efforts, the digital forensics community has been trying to perform testing and validating different forensic tools, in order for examining if certain requirements for each tool are properly implemented, choosing a most suitable set among tools providing same functionalities, and encouraging continuous improvement.

In this circumstance, data corpora will play an important role in activities to meet the above-mentioned demands, if datasets are developed upon consideration of various use cases along with normal as well as abnormal user behaviors relating to each tool. That is, meaningful uses of these datasets may include; research, development, education, training, equipment check out, tool testing, and proficiency testing. Therefore, as a part of establishing a solid and fully functioning infrastructure, it is necessary to cultivate fine-grained datasets that can be used to support both academic and practical purposes. Although the necessity of systematically developing reference data has been increasing, activities in the field of security and forensics are still hobbled by the lack of available datasets since Garfinkel stated in 2007 (Garfinkel, 2007; Grajeda *et al*., 2017). Moreover, existing datasets were usually created at that point in time for particular purposes, and then they have rarely been updated (Grajeda *et al*., 2017). Continuous endeavors are required to develop various data corpora that also embrace emerging technologies, not an easy mission but necessary.

The primary purpose of this work is to develop a systematic and practical methodology to improve the efficiency of dataset development, through providing fundamental techniques as an infrastructure. The key idea behind the proposed strategy is that

+ Certain trade names and company products are mentioned in the text or identified. In no case does such identification imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that the products are necessarily the best available for the purpose.

[*] Corresponding author. The author is a visiting scientist at the National Institute of Standards and Technology.
E-mail address: junghmi@gmail.com (J. Park).