CERIAS Tech Report 2005-65

AUTOMATED DIGITAL EVIDENCE TARGET DEFINITION USING OUTLIER ANALYSIS AND EXISTING EVIDENCE

by Brian D. Carrier and Eugene H. Spafford

Center for Education and Research in
Information Assurance and Security,
Purdue University, West Lafayette, IN 47907-2086

Automated Digital Evidence Target Definition Using Outlier Analysis and Existing Evidence

Brian D. Carrier
carrier@cerias.purdue.edu

Eugene H. Spafford
spaf@cerias.purdue.edu

Center for Education and Research in
Information Assurance and Security - CERIAS
Purdue University
West Lafayette, IN 47907 USA

Abstract

Searching for digital evidence is a time consuming and error-prone process. In this paper, we introduce techniques to automate the searching process by suggesting what searches could be helpful. We also use data mining techniques to find files and directories created during the incident. The results from using these techniques on a compromised honeypot system are given and show that the data mining techniques detect a higher percentage of files than a random sampling would, but there are still many false positives. More research into the error rates of manual searches is needed to fully understand the impact of automated techniques.

Key Words: Digital Evidence, Digital Investigation, Digital Forensics, Data Mining, Spatial Outlier Analysis, Automated Target Definition

1. Introduction

One of the most time consuming tasks during a digital investigation is the process of searching for evidence. If evidence of an event does not exist, then the investigator can make only assumptions about what occurred. While it is common to hear people refer to current computer forensic analysis tools as being "automated," this paper introduces additional automation into the searching process by identifying for what should be searched.

This paper describes two approaches and four implementations of automated evidence searching. The first approach suggests new searches based on evidence that has been found and the second approach uses outlier analysis to find files and directories that were created or modified during the incident. These approaches can help to make investigations more thorough and accurate because the tool keeps track of what searches still need to be conducted. We implemented four analysis tools and one suggested additional searches based on existing evidence and the other three used different outlier analysis algorithms. The implementations were run on the file system data from a compromised Linux system.

Before we discuss how to search for digital evidence, we must define it. We define *digital evidence of an incident* as digital data that contain reliable information that support or refute a hypothesis about the incident being investigated. An object is evidence because it played a role in an event that supports or refutes an investigation hypothesis [CS04]. Only a subset of these objects will be considered legal evidence and presented in court.

The remainder of this paper is organized as follows. Section 2 provides a general process model for evidence searching and four phases are defined. These phases will be used to illustrate where automation can be incorporated. Section 3 describes our automated search technique that is based on evidence as it is found by an investigator. Sections 4 to 6 describe three approaches that use data mining techniques to find files and directories that are evidence and Section 7 concludes the paper.

2. The Search Process

The search phase of a digital investigation is where the digital crime scene is processed and evidence is recognized. The primary goal of this phase is to recognize objects that played a role in events related to either the incident or a hypothesis about the incident. The motivation for evidence searching could be to verify an incident report, to find evidence of a specific event, or to test a hypothesis.