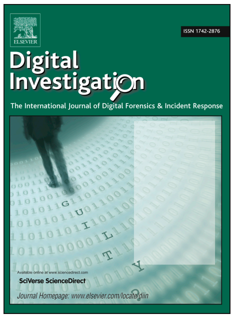
ISSN 1742-2876
Digital Investigation
The International Journal of Digital Forensics & Incident Response
SciVerse ScienceDirect
Journal Homepage: www.elsevier.com/locate/diin

Efficient monitoring and forensic analysis via accurate network-attached provenance collection with minimal storage overhead

Yulai Xie, Dan Feng, Xuelong Liao, Leihua Qin

PII: S1742-2876(18)30025-2

DOI: 10.1016/j.diin.2018.05.001

Reference: DIIN 778

To appear in: *Digital Investigation*

Received Date: 17 January 2018

Revised Date: 20 March 2018

Accepted Date: 6 May 2018

Digital Investigation

Digital Investigation 00 (2018) 1–11

Efficient monitoring and forensic analysis via accurate network-attached provenance collection with minimal storage overhead

Yulai Xie[a,*], Dan Feng[a,b], Xuelong Liao[a], Leihua Qin[a]

[a] *School of Computer, Huazhong University of Science and Technology, Wuhan, People's Republic of China*
[b] *Wuhan National Laboratory for Optoelectronics, Huazhong University of Science and Technology, Wuhan, People's Republic of China*

Abstract

Provenance, the history or lineage of an object, has been used to enable efficient forensic analysis in intrusion prevention system to detect intrusion, correlate anomaly, and reduce false alert. Especially for the network-attached environment, it is critical and necessary to accurately capture network context to trace back the intrusion source and identify the system vulnerability. However, most of the existing methods fail to collect accurate and complete network-attached provenance. In addition, how to enable efficient forensic analysis with minimal provenance storage overhead remains a big challenge.

This paper proposes a provenance-based monitoring and forensic analysis framework called PDMS that builds upon existing provenance tracking framework. On one hand, it monitors and records every network session, and collects the dependency relationships between files, processes and network sockets. By carefully describing and collecting the network socket information, PDMS can accurately track the data flow in and out of the system. On the other hand, this framework unifies both efficient provenance filtering and query-friendly compression. Evaluation results show that this framework can make accurate and highly efficient forensic analysis with minimal provenance storage overhead.

Keywords: Provenance, forensic analysis, provenance filtering, provenance compression

1. Introduction

A variety of security mechanisms (e.g., encryption and access control) has been adopted to protect against the intrusion and data leak. However, there are always a variety of vulnerabilities (e.g., no rational configuration of firewall rules, weak passwords, etc.) that are likely to be exploited as the computer system gets more and more complicated. Due to hacker attacks, insider leaks, the abuse of administrator privileges and other reasons, the computer system is easily compromised, leading to the loss or leakage of data. For instance, in April 2010, the account information of over six million Internet users was leaked due to the weak cryptography used by the China Software Developer Network company (Daily, 2010); in April 2014, the Heartbleed security bug found in OpenSSL also engulfed about half a million of web servers in the wild (Wakefield, 2014).

After intrusion or data leakage occurred, a big challenge is to investigate how data leakage or intrusion happens. The existing methods typically develop tools (King and Chen, 2005; King et al., 2005) or systems (Goel et al., 2005; Xie et al., 2016) to explore the causality-based context in the system or disk log. The causality-based context, which we term as provenance, describes how data come to its present status and can be used to enable monitoring and forensic analysis by capturing the data flow and dependency relationship between different data objects. Provenance has been widely used in recording experimental details, debugging, optimizing search (Shah et al., 2007), and data rebuild (Xie et al., 2013a). Provenance-based methods have also been used in both local (Pohly et al., 2012) and distributed environments (Zhou et al., 2011; Tariq et al., 2011; Gehani et al., 2010) to trace back the intrusion source. However, two challenges remain to be addressed. First, how to collect network socket accurately and completely? In a networked environment, any miss of network intrusion information can result in a severe problem. For instance, inter-host viruses propagate can be completed promptly, and the miss of capturing such information can be a disaster to the computer system. However, existing methods either do not record the IP and port information (Pohly et al., 2012) or cannot capture (Gehani and Tariq, 2012) the provenance of the short-lived socket connection. Second, since the size of provenance is consistently increasing, how to enable efficient provenance stor-

[*] corresponding author

Email addresses: `ylxie@hust.edu.cn` (Yulai Xie), `dfeng@hust.edu.cn` (Dan Feng), `xlliao@hust.edu.cn` (Xuelong Liao), `lhqin@hust.edu.cn` (Leihua Qin)