An Intelligence Criminal Tracker for Industrial Espionage

Applying Digital Data Acquired Onsite to Target Criminals

Jieun Dokko[1,2(✉)], Michael Shin[1], and Soo Young Park[2]

[1] Texas Tech University, Lubbock, Texas 79409, USA
{Jieun.dokko, Michael.Shin}@ttu.edu,
maggie8482@gmail.com
[2] National Digital Forensic Center, Supreme Prosecutors' Office, Seoul, South Korea
{dkje8482, hilda01}@spo.go.kr

Abstract. The investigation of industrial espionage basically requires significant levels of expertise and a full data recovery on an entire device. In practice, an investigator cannot conduct an in-depth examination of every device, thereby inevitably collecting all the devices seemingly relevant to the crime. Such excessive collection leads to not only legal concerns about the data privacy but also a massive examination backlog in a lab. To alleviate the challenge, a field triage model enabling an accurate data processing and acquisition is proposed.

Keywords: Digital forensic field triage · Industrial espionage · Real-time data acquisition · Data reduction · Crime specific analysis · Intelligence analysis

1 Introduction

The volume and variety of data in many devices is very challenging to onsite triage investigations. Moreover, data security and privacy concerns oblige investigators to scrutinize every device to only acquire files pertinent to the allegation [6]. To improve the accuracy and speed for the investigation, we propose a filed triage, typical for industrial espionage. It primarily aims to acquire the minimum amount of relevant data from a device by performing on low latency analysis with interesting data related to the crime, and secondly aims to develop a reliable profile by putting together the findings from data processing.

2 Related Works

The authors in [1] propose a digital forensic investigation and verification model for industrial espionage (DEIV-IE). The model defines twelve crime features derived from the investigative line of questioning, a set of evidence file groups being examined for the crime features, and the analysis techniques quite often used to solve the crime in a laboratory. It presents a method for mapping the features, file groups and techniques

S. Goel et al. (Eds.): ICDF2C 2020, LNICST 351, pp. 224–230, 2021.
https://doi.org/10.1007/978-3-030-68734-2_12

and effectively detecting evidence data typical for the crime. The authors in [2] propose "selective imaging" which provides selective acquisition by locating key files, data processing, reviewing findings, and acquiring relevant files while missing their sources, thereby lacking verifiability in some cases. The authors in [3] aim to locate relevant data with improvement in retrieval and correlation analysis in Child Sexual Abuse cases by using existing several NLP techniques and semantic web technologies. The authors in [4] propose the digital forensic framework (D4I) compatible with cyber-attacks by mapping windows artifact categorization of SANS to the seven steps of cyber-attacks in CKC model, although it needs manual determinations.

3 Intelligence Criminal Tracker (ICT)

The ICT is a crime-targeted triage model for industrial espionage. It aims to acquire relevant data (minimal acquisition) by adjusting data acquisition according to its prior data processing results. The processing focuses on discovery of criminals based on the eleven hypotheses about criminal patterns shown in Table 1 and the corresponding forensic techniques borrowed from [1]. With the data processing intelligence, it also generates an ordered list of potential criminals and alleged criminal activities by tracing back the origins of data of the activities.

Table 1. The crime specific hypotheses applied to the ICT

No	Hypothesis
H1	Criminal activities occurring from four months before a suspect's resignation
H2	A suspect using the computer and accessing target files recently
H3	A suspect using the Internet to communicate with an accomplice
H4	Stolen files being commonly MS doc, text, csv, pdf, graphic (over 10 KB), video, and XML in order, albeit being dependent on the business of a victim
H5	Stolen files being likely compressed, archived, and encrypted during exfiltration
H6	Communicating or file sharing with accomplices using email IM, cloud service
H7	Searching crime relating information just before or after the crime occurred
H8	The use of portable storage devices involved in data exfiltration
H9	Unusual, suspicious activities as compared to prior usage
H10	Alternative ways of exfiltration e.g., a MS doc converting to a JPG
H11	A company monitoring employees' activities e.g., data loss prevention (DLP)

4 Investigation Methodology

As shown in Fig. 1, the ICT locates key files of the crime based on the mapping between the predefined eleven hypotheses and corresponding interesting data. In data processing, it performs the four tasks of extracting, ranking, grouping, and profiling to examine identifiers of criminals. Then, it conducts evidence acquisition, and generates a final report. To achieve more accurate findings, we form a methodology for data

processing depicted in Table 2. All the tasks use the adjacent date(s) of suspected employee(s) resignation as a default filtering option which was populated as common periods of these types of crime.

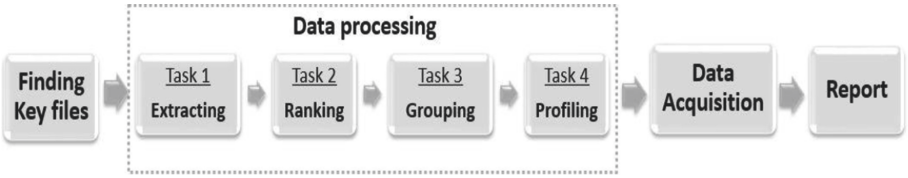

Fig. 1. Investigation methodology of the ICT

Table 2. A methodology for the data processing

No	Method in each step
1	Develop the hypotheses about criminal features of industrial espionages
2	Decide what actions a criminal might have taken based on the hypotheses
3	Identify the specific linked activities in a system to be analyzed
4	Examine the files to prove the alleged linked activities in a system from key files
5	Extract the identifiers creating or relating to the activities and count them
6	Rank the identifiers by its frequency of occurrences in the outputs of the ranking tests
7	Group the ranked identifiers to find a criminal who may use several identifiers
8	Create a profile describing the top ranked criminal-candidate's behaviors temporally by tracing back the data identifiers in the ranked group examined from

4.1 Extracting

The steps from No 1 to 5 in Table 2 depict how to extract alleged identifiers in the crime. The ICT examines files storing the identifiers possibly relating to potential criminal activities [8], and extracts the ten types of identifiers as follows: a) OS user account, b) user name registered in app installation, c) volume label and machine name, d) last modifier, e) account of email and instant message, f) login ID of web-site, g) password of email, instant message, and web-site, h) user name, friendly name or alias used for email, instant message, and web-site or services and, i) phone number linked to any account. Next the ICT decides the scope of potential stolen files and lists them for the ranking task. They are initially defined as the files recently created, accessed or modified, uploaded, downloaded, backed-up, copied and deleted by a user and can be updated.

4.2 Ranking

The ICT assesses a suspect-candidate rate and an accomplice-candidate rate on detected identifiers by the different measures. It sets the eight suspect testing conditions (SC#)

and three accomplice testing conditions (AC#) described in Table 3. The ranking is conducted separately under each condition in sequence.

Table 3. The test conditions for the ranking of criminal-candidates

No	Description of an identifier getting one score per each action under the condition
SC1	Recently or within the alleged crime time, using a file in the stolen file list
SC2	Using an encrypted file or a compressed file containing a stolen file in the list
SC3	Sending an email or IM attaching a file in H4 and H5, attaching a file in the list, having a clickable link in its contents, uploading a file in the list to a server
SC4	Searching Internet for info e.g., the crime, or a new job before or after the crime
SC5	Recently using an external media, additionally having a file in the list stored in it
SC6	Showing suspicious activities e.g., using an unusual account, URL, server, USB, unusual data backups, downloading, installing, running an anti-forensic app
SC7	Converting a file, screen capturing, recording, printing a file in the list
SC8	Having a behavior being monitored, leading to an alert notification
AC1	An correspondent account to an email or IM account detected as a potential suspect
AC2	Receiving an email or IM with an attachment(s) from an alleged local account, additionally the attachment falling in the list
AC3	Sharing a web service with a local account

Under one condition, an identifier accumulates one score whenever the identifier itself is returned as the output from the testing, and according to the score they achieved, all identifiers are ranked under the condition. As a final point, the ranks of identifiers achieved from all the conditions are summed, and prioritized by the sum from small to large sequentially, formulated as below.

$$Identifier.Condition1.RankNum = Rank.DescendingOrder\,(Count\,(identifier1.ScoreOfoccurrence)\ldots Count\,(identifierN.ScoreOfoccurrence)) \tag{1}$$

$$Identifier.RankNum = Rank.AscendingOrder\,(Sum\,(identifier1.allConditions.RankNum)\ldots Sum\,(identifierN.allConditions.RankNum)) \tag{2}$$

Upon the completion, the ICT creates a list of the identifiers achieving at least one score. The list contains the rank number of an identifier with the total scores, the frequency of the occurrence, and the associated sources of each occurrence detected as shown in Fig. 2. The file an identifier is extracted from, is associated to the identifier, so each identifier can be traceable by looking at the file. The source reference on the file is classified into two types: file attributes in the file system or a file header, and entry attributes in the records of Registry [5], Windows artifacts [7], email, IM and the like, and referred to respectively.

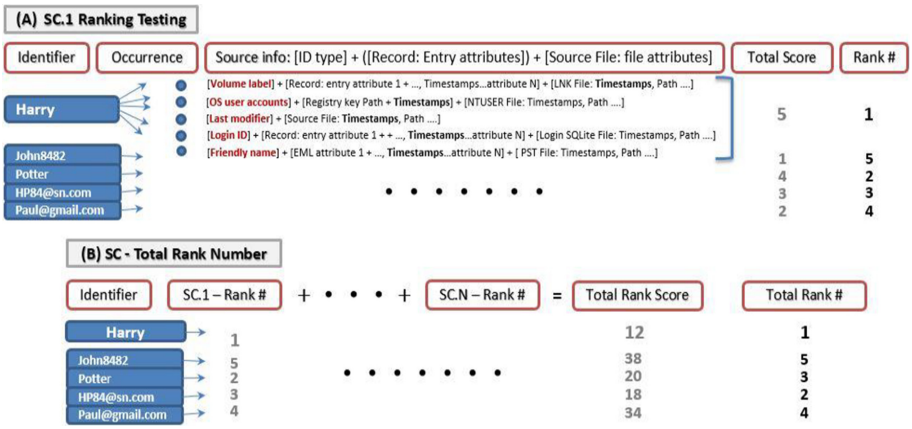

Fig. 2. (A) Ranking under a condition and, (B) Ranking under all conditions of the alleged identifiers

4.3 Grouping

The ICT works by grouping the ranked identifiers by three criteria: commonness, resemblance and connection, to find a criminal who may use several distinct identifiers. The rules for grouping are described below. Identifiers in the rules are not case-sensitive.

Commonness. Check if an identifier includes a white space. If yes, split the identifier by the white space. Next, from the beginning of it, divide an identifier or a segment split by a white space into 'letter string, 'number string, and 'the others'. Discard any segment that is less than three characters. If any segment is a substring of another segment, group the identifiers of each segment into one.

Resemblance. Check if the two identifiers contain the same characters the number of which is more than a half of the characters of them, except for the common components of identifiers like "@domain", or "country code". If yes, extract the same characters in the order from each identifier and determine if they are in the same order. If yes, group the identifiers into one. For instance, "h@rry" is added to the group of "harry" because the number of the same characters extracted in the order from "h@rry" and "harry" which is "hrry" is more than half of the characters of them both.

Connections. If various identifiers in multiple sources are created from a single activity, group them into one, except for common user accounts e.g., administrator, guest, and user. For instance, "Harry" from the Skype subfolder of NTUSER.DAT, and the Skype id of "potter8482@gmail.com" are grouped into one. In addition, if there are different types of paired personal identifiers in the same source, e.g., user ID and password for a website, classify each pair into each single group.

Identifiers in a group sorted above can be put into one of the existing groups by applying the aforementioned algorithms of the commonness and resemblance. Finally, the ICT lists the ranked groups of identifiers, and the formulation is as below.

$$Group\,A = [identifier1.GroupA,\ identifier3.GroupA \ldots .identifierN.GroupA] \quad (3)$$

$$RankScore.GroupA = Sum\,(RankNum.identifier1.GroupA, RankNum.identifier3.GroupA \ldots .RankNum.identifierN.GroupA) \quad (4)$$

$$Groups.RankNum = Rank.AscendingOrder\,(RankScore.GroupA,\ RankScore.GroupB \ldots RankScore.GroupN \quad (5)$$

4.4 Profiling

The ICT traces the activities gaining a score in ranking tests, of all identifiers belonging to the first ranked group and generates timelines describing the activities of the prime suspect and accomplice candidates. As shown in Fig. 3, the descriptive timeline is arranged with identifiers, and their source information as 'Linking time', 'Rank condition #', 'Identifier', 'ID Type', 'Activity', 'Entry attributes' and 'File attributes' in order. But each artifact has a different set of attributes like various timestamps, paths, and counts. The ICT classifies various attributes into several types based on the similarity between theirs roles, and decides time attributes comparable to one another for sorting. The 'Linking time' as a comparable timestamp, is decided on for the time when events observed in the records of artifacts occurred, thereby the timeline ordered by the linking time can reflect the order of a user's activities related to the events. The linking time order may not be exact, due to different mechanism timestamp updates between artifacts, but effectively close.

Linking Time	RC	Identifier	I.D. type	Activity	Entry attributes (name, path, time, count, other..)	File attributes: [type] name, path, time
3/23/2020 16:01	SC7	Harry	OS account	Rename a S-file	Price.xlsx, C:\Users\Harry\PR\..., Happy.jpg...	[Journal] $Usnjrnl, C:\$Extend\$Usnjrnl....
3/23/2020 17:04	SC3	HP84@sn.com	Outlook ID	Send an email	State (Send), subject, body, attachment (Happy.jpg)	[Email] HPmailbcx.pst, path, timestamps, size,....
3/23/2020 18:19	SC1	Potter	Last modifier	Modify a file	.... Blank	[S-file] price.xlsx, C:\Users\Harry\PR\....
3/23/2020 18:20	SC2	Harry	OS account	Archive a file	.... Blank	[S-file] PR.zip, C:\Users\Harry\Desktop\PR.zip.....
3/23/2020 21:11	SC5	Potter	Volume label	Use an USB	Price.xlsx F:\project\PR.zip\Price.xlsx, timestamps...	[LNK] price.xlsx.lnk, path, timestamps, size,.....
3/24/2020 09:01	SC3	HP84@gmail.com	Google Drive	Use a G-drive	Price.jpg, C:\Users\Harry\Google drive\ Happy.jpg...	[IE] WebCacheVC1.dat, path, timestamps ...
3/24/2020 12:01	SC4	Harry	OS account	Internet search	http://...search%20disk%20wiping/....	[IE] WebCacheVC1.dat, path, timestamps ...
3/24/2020 16:23	SC5	HarryCD	Volume label	Burn a CD	Price.xlsx, E:\PR.zip\Price.xlsx, timestamps, HarryCD	[Jumplist] 1b4dd67f29cb1962, path times...
3/24/2020 23:01	SC6	Harry	OS account	Download a file	http://download.eraser.com/eraser.exe,...	[IE] WebCacheVC1.dat, path, timestamps...
3/24/2020 23:19	SC6	Harry	OS account	Execute a file	PROGRAM FILES\ERASER\ERASER.EXE..., 2...	[Prefech] ERASER.EXE-BE552234.pf, time...

Fig. 3. Illustration of the descriptive timeline to events regarding a user's certain activities

5 Data Acquisition and Reporting

After processing, the ICT generates a forensic image of the pertinent files referred to during data processing and a final report (detailing the outputs of each task, acquisition information, and a descriptive timeline of the criminal's activities).

6 Conclusions

The ICT is designed to acquire only relevant data with precision from devices on site, reflecting the investigative thinking process with digital forensic techniques through the four processing tasks. It helps improve the decision-making for taking relevant devices to a lab, so that pointless analysis in a lab can be avoided. The approach can be useful for deriving typical criminal behaviors in other crimes and suggesting a crime-based triage model as an adaptive strategy for criminal activities differed and complicated over time. Currently, the tool only works on industrial espionage related to insider threats, thus further studies are needed to focus on developing new hypotheses, criminal patterns for espionage related to cyber-attacks, and corresponding forensic techniques built in it.

References

1. Dokko, J., Shin, M.: A digital forensic investigation and verification model for industrial espionage. In: Breitinger, F., Baggili, I. (eds.) ICDF2C 2018. LNICST, vol. 259, pp. 128–146. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-05487-8_7
2. Quick, D., Choo, K.-K.R.: Big forensic data reduction: digital forensic images and electronic evidence. Cluster Comput. **19**(2), 723–740 (2016)
3. Amato, F., Castiglione, A., Cozzolino, G., Narducci, F.: A semantic-based methodology for digital forensics analysis. J. Parallel Distrib. Comput. **138**, 172–177 (2020)
4. Dimitriadis, A., Ivezic, N., Kulvatunyou, B., Mavridis, I.: D4I-Digital forensics framework for reviewing and investigating cyber attacks. Array **5**, 100015 (2020)
5. Roussev, V., Quates, C., Martell, R.: Real-time digital forensics and triage. Digit. Invest. **10** (2), 158–167 (2013)
6. Korea, S.: Criminal Procedure Act, December 2017. http://www.law.go.kr
7. Singh, B., Singh, U.: Program execution analysis in windows: a study of data sources, their format and comparison of forensic capability. Comput. Secur. **74**, 94–114 (2018)
8. Rowe, N.C.: Finding and rating personal names on drives for forensic needs. In: Matoušek, P., Schmiedecker, M. (eds.) ICDF2C 2017. LNICST, vol. 216, pp. 49–63. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-73697-6_4