Chapter 17

A COST-EFFECTIVE MODEL FOR DIGITAL FORENSIC INVESTIGATIONS

Richard Overill, Michael Kwan, Kam-Pui Chow, Pierre Lai and Frank Law

Abstract Because of the way computers operate, every discrete event potentially leaves a digital trace. These digital traces must be retrieved during a digital forensic investigation to prove or refute an alleged crime. Given resource constraints, it is not always feasible (or necessary) for law enforcement to retrieve all the related digital traces and to conduct comprehensive investigations. This paper attempts to address the issue by proposing a model for conducting swift, practical and cost-effective digital forensic investigations.

Keywords: Investigation model, Bayesian network

1. Introduction

A digital forensic investigation involves the application of a series of processes on digital evidence such as identification, preservation, analysis and presentation. During the analysis process, digital forensic investigators reconstruct events in order to evaluate the truth of the forensic hypotheses related to the crime or incident based on the digital traces that have been identified and retrieved [2]. Due to inherent technological complexities, the identification and retrieval of digital traces cover a variety of techniques such as cryptography, data carving and data reconstruction. Each technique has a different level of complexity and, therefore, a different resource cost (e.g., expertise, time and tools).

Unlike physical events that are continuous, digital events are discrete and occur in temporal sequence [3]. Because of the discrete nature, it is possible to quantify the retrieval costs of individual digital traces. However, in the absence of a suitable model for digital forensic investigations, most investigators attempt to conduct a comprehensive retrieval

G. Peterson and S. Shenoi (Eds.): Advances in Digital Forensics V, IFIP AICT 306, pp. 231–240, 2009.

of all related digital traces despite the substantial costs associated with retrieving all the traces.

A more effective technique is to focus only on the digital traces that can be extracted in a cost-effective manner. The reasons are that investigating different digital traces requires resources (e.g., expertise, time and tools) in different amounts, and that the traces found have different evidentiary weights with respect to proving a hypothesis. The limited resources available for an investigation renders exhaustive search approaches impractical [4]. Consequently, digital forensic investigators who endeavor to retrieve all the traces – especially those that are not sufficient to prove the hypotheses – waste valuable resources.

This paper describes a model for conducting swift, practical and cost-effective digital forensic investigations. The model considers the retrieval costs of digital traces and incorporates a permutation analysis.

2. Preliminaries

Using the collective experience and judgment of digital forensic investigators, it is possible to rank the relative costs of investigating each trace T_i $(i = 1 \ldots m)$. The relative costs may be estimated in terms of their resource requirements (person-hours, access to specialized equipment, etc.) using standard business accounting procedures. The relative costs can be ranked $T_1 \leq T_2 \leq \ldots \leq T_m$ without any loss of generality. As a direct consequence of this ranking, the minimum cost path for the overall investigation is uniquely identified.

Our focus is on digital traces residing on a hard disk. If the seized computer has sufficient storage, all the digital traces can be retrieved. If all the traces T_i $(i = 1 \ldots m)$ are retrieved, the minimum cost path is the permutation $[T_1 T_2 \ldots T_m]$. An example of a permutation path is shown in Figure 1.

The number of possible paths at each step is given by $m!$ This is a direct consequence of the fact that the problem of selecting the next available trace from an ordered permutation of m distinct traces is isomorphic to the problem of selecting the next object from a collection of m identical objects.

In order to save time and conserve resources, it is useful to determine early in an investigation whether or not the investigation should continue. This requires an estimation of the cumulative evidentiary weight associated with the investigation as $W = \sum_{i=1}^{m} w_i$, where the (scaled) relative fractional evidentiary weight w_i of each trace T_i is either assigned by an expert or, by default, is set to one. The weight assignment process has to be undertaken only once as a preprocessing step for each distinct

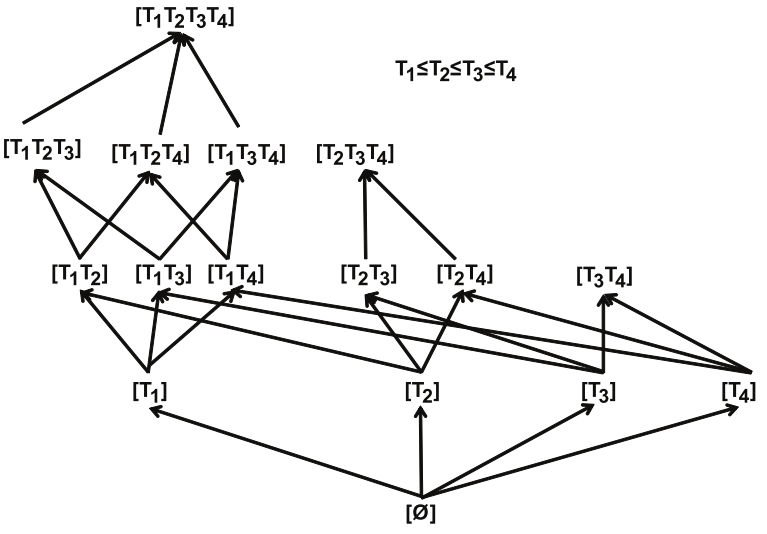


Figure 1. Path diagram with four traces.

digital crime template. If the cumulative estimate W is sufficiently close to one, the *prima facie* of the case can probably be established. Otherwise, it is unlikely that the available digital traces are sufficient to support the case.

The difference between W and one provides a "cut-off" condition that an investigator can use to avoid identifying all the traces exhaustively. The cut-off state is illustrated using the following example. Suppose that email exchanges between the culprit and victim are vital to an investigation. The forensic goal is to confirm that the computer, which was under the culprit's control, had been used to send and receive the emails in question. Assume that the evidentiary traces are T_1, T_2, T_3, T_4, T_5 with evidentiary weights 0.05, 0.10, 0.15, 0.2, 0.35, respectively; and the evidentiary threshold is 0.85. Therefore, if all the traces are retrieved, then the estimated total evidentiary weight is 0.85, which indicates a strong case. On the other hand, if trace T_1 is not found, the overall evidentiary weight is 0.8, which is a 6% falloff. If both T_1 and T_2 are missing, the overall weight becomes 0.7, an 18% falloff. At this point, the forensic investigator should consider suspending the investigation as the prospect of successful prosecution is slim.

3. Missing Traces

Since a computer may not have sufficient storage, there is always the chance that some traces may be missing or overwritten. Thus, it may

not be possible for an investigator to ascertain all the trace evidence pertaining to a case. Suppose a certain trace T_j $(1 \leq j \leq m)$ is not found. Then, all the investigative paths involving T_j are removed from the path diagram and the minimum cost path becomes $[T_1T_2 \ldots T_{j-1}T_{j+1} \ldots T_m]$. The estimate of the evidentiary weight is $W = \sum_{i \neq j}^{m} w_i$.

Similarly, if two traces T_j and T_k $(1 \leq j; k \leq m; j \leq k)$ are not found, then all the paths involving T_j or T_k must be deleted and the minimum cost path is $[T_1T_2 \ldots T_{j-1}T_{j+1} \ldots T_{k-1}T_{k+1} \ldots T_{m-1}T_m]$. The estimate of the evidentiary weight is $W = \sum_{i \neq j,k}^{m} w_i$. In general, if a total of k traces are not found $(1 \leq k < m)$, then all the investigative paths containing any of the k traces must be deleted from the path diagram.

It is important to consider the issue of the independence of digital traces T_i. While the observations of the traces are necessarily independent because they are performed individually *post mortem*, the digital traces must be created independently if the model is to retain its validity. Since it is possible in principle for one user action to create multiple digital traces T_i (which are not mutually independent), care must be taken to ensure the independence of the expected digital traces when selecting the set of traces.

4. Investigation Model

The model for conducting cost-effective digital forensic investigations has two phases.

Phase 1 (Preprocessing – Detecting Traces)

- Enumerate the set of traces expected to be present based on the type of crime suspected.
- Assign relative investigation costs to each expected trace.
- Rank the expected traces in order of increasing relative investigation costs.
- Assign relative evidentiary weights w_i to each ranked trace.
- Rank the expected traces within each cost band in order of decreasing relative evidentiary weight.
- Set the cumulative evidentiary weight estimate W to zero.
- Set the total of the remaining available weights W_{rem} to one.
- For each expected trace taken in ranked order:
  - Search for the expected trace.
  - Subtract the relative evidentiary weight w_i of the trace from W_{rem}.

- If the expected trace is retrieved, add its relative evidentiary weight w_i to W.
- If W is sufficiently close to one, proceed to Phase 2.
- If $W + W_{rem}$ is not close enough to one, abandon the forensic investigation.

Phase 2 (Bayesian Network – Analyzing Traces)

- Run and analyze the Bayesian network model for the crime hypothesis using the retrieved traces as evidence (as described in [6]).

5. BitTorrent Case Study

This section uses a BitTorrent case study [6] to demonstrate the cost-effective digital forensic investigation model.

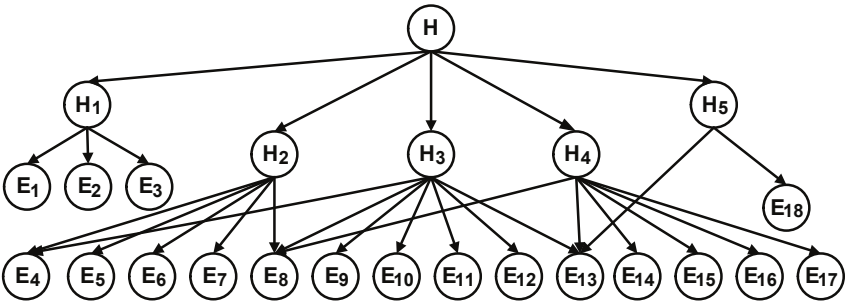
Figure 2 shows a Bayesian network with eighteen expected evidence traces (E_i) and their relationships to the five hypotheses (H_i). The Bayesian network is constructed by enumerating every path through which an evidentiary trace could have been produced and assigning it a probability.

The ideal case, in which all eighteen evidence traces are retrieved, is shown in Table 1. Note that each piece of trace evidence E_i is ranked according to its cost T_j.

The actual case, corresponding to a situation where two of the expected traces (E_8 and E_{14}) are missing, is shown in Table 2.

A potential complication involving the proposed investigation model should be noted. A trace could initially be assigned a low cost; however, upon further consideration, it could be determined that the cost is much higher. Examples of such a situation are a file that turns out to be protected by encryption or a partition that turns out to be deleted. In such cases, the cost of investigating the trace must be revised and all the traces must be ranked again based on the revised costs. This is necessary to maintain the minimum cost strategy for the investigation.

Constructing a Bayesian network model corresponding to an investigation requires the definition of the overall structure of the network, including the hierarchy of hypotheses and the associated posterior digital evidence (or traces) whose presence or absence determines the prior probabilities of the corresponding hypotheses. Next, numerical values are assigned to the prior probabilities. Traditionally, forensic investigators assign the prior probabilities based their expertise and experience.



HYPOTHESES:

H − The seized computer was used as the initial seeder to share the pirated file on a BitTorrent network

H_1 − The pirated file was copied from the seized optical disk to the seized computer

H_2 − A torrent file was created from the copied file

H_3 − The torrent file was sent to newsgroups for publishing

H_4 − The torrent file was activated and the computer connected to the tracker server

H_5 − The connection between the seized computer and the tracker was maintained

EVIDENCE:

E_1 − Modification time of the destination file is the same as that of the source file

E_2 − Creation time of the destination file is after its modification time

E_3 − Hash value of the destination file matches that of the source file

E_4 − BitTorrent client software is installed on the computer

E_5 − File link for the shared file is found

E_6 − Shared file exists on the hard disk

E_7 − Torrent file creation record is found

E_8 − Torrent file exists on the hard disk

E_9 − Peer connection information is found

E_{10} − Tracker server login record is found

E_{11} − Torrent file activation time is corroborated by its MAC time and link file

E_{12} − Internet history record of the publishing website is found

E_{13} − Internet connection is available

E_{14} − Cookie of the publishing website is found

E_{15} − URL of the publishing website is stored in the web browser

E_{16} − Web browser software is found

E_{17} − Internet cache record of the publishing of the torrent file is found

E_{18} − Internet history record of the tracker server connection is found

Figure 2. Bayesian network model for the BitTorrent case.

However, these assessments have been challenged in judicial proceedings primarily on the grounds that they are non-rigorous and subjective.

These challenges can be countered if a rigorous analytic procedure is used to quantitatively assign the prior probabilities. A promising

Table 1. Traces, relative costs and weights for the ideal BitTorrent case.

	Trace	**Rel. Cost**	**Rel. Wt.**	W	W_{rem}
	Initial Values			**0**	**1**
T_1 (E_6)	Shared file exists on the hard disk	1	2/18	2/18	16/18
T_2 (E_1)	Modification time of the destination file is the same as that of the source file	1	1/18	3/18	15/18
T_3 (E_2)	Creation time of the destination file is after its modification time	1	1/18	4/18	14/18
T_4 (E_3)	Hash value of the destination file matches that of the source file	1	1/18	5/18	13/18
T_5 (E_8)	Torrent file exists on the hard disk	1	1/18	6/18	12/18
T_6 (E_{16})	Web browser software is found	1	1/18	7/18	11/18
T_7 (E_5)	File link for the shared file is found	1	0.5/18	7.5/18	10.5/18
T_8 (E_{15})	URL of the publishing website is stored in the web browser	1	0.5/18	8/18	10/18
T_9 (E_7)	Torrent file creation record is found	1.5	2/18	10/18	8/18
T_{10} (E_{13})	Internet connection is available	1.5	2/18	12/18	6/18
T_{11} (E_{10})	Tracker server login record is found	1.5	0.5/18	12.5/18	5.5/18
T_{12} (E_{12})	Internet history record of the publishing website is found	1.5	0.5/18	13/18	5/18
T_{13} (E_{14})	Cookie of the publishing website is found	1.5	0.5/18	13.5/18	4.5/18
T_{14} (E_{17})	Internet cache record of the publishing of the torrent file is found	1.5	0.5/18	14/18	4/18
T_{15} (E_{18})	Internet history record of the tracker server connection is found	1.5	0.5/18	14.5/18	3.5/18
T_{16} (E_4)	BitTorrent client software is installed on the computer	2	2/18	16.5/18	1.5/18
T_{17} (E_{11})	Torrent file activation time is corroborated by its MAC time and link file	2	1/18	17.5/18	0.5/18
T_{18} (E_9)	Peer connection information is found	2	0.5/18	1	0

Table 2. Traces, relative costs and weights for the actual BitTorrent case.

	Trace	**Rel. Cost**	**Rel. Wt.**	W	W_{rem}
	Initial Values			**0**	**1**
T_1 (E_6)	Shared file exists on the hard disk	1	2/18	2/18	16/18
T_2 (E_1)	Modification time of the destination file is the same as that of the source file	1	1/18	3/18	15/18
T_3 (E_2)	Creation time of the destination file is after its modification time	1	1/18	4/18	14/18
T_4 (E_3)	Hash value of the destination file matches that of the source file	1	1/18	5/18	13/18
T_5 (E_8)	Torrent file exists on the hard disk *(missing)*	1	1/18	5/18	12/18
T_6 (E_{16})	Web browser software is found	1	1/18	6/18	11/18
T_7 (E_5)	File link for the shared file is found	1	0.5/18	6.5/18	10.5/18
T_8 (E_{15})	URL of the publishing website is stored in the web browser	1	0.5/18	7/18	10/18
T_9 (E_7)	Torrent file creation record is found	1.5	2/18	9/18	8/18
T_{10} (E_{13})	Internet connection is available	1.5	2/18	11/18	6/18
T_{11} (E_{10})	Tracker server login record is found	1.5	0.5/18	11.5/18	5.5/18
T_{12} (E_{12})	Internet history record of the publishing website is found	1.5	0.5/18	12/18	5/18
T_{13} (E_{14})	Cookie of the publishing website is found *(missing)*	1.5	0.5/18	12/18	4.5/18
T_{14} (E_{17})	Internet cache record of the publishing of the torrent file is found	1.5	0.5/18	12.5/18	4/18
T_{15} (E_{18})	Internet history record of the tracker server connection is found	1.5	0.5/18	13/18	3.5/18
T_{16} (E_4)	BitTorrent client software is installed on the computer	2	2/18	15/18	1.5/18
T_{17} (E_{11})	Torrent file activation time is corroborated by its MAC time and link file	2	1/18	16/18	0.5/18
T_{18} (E_9)	Peer connection information is found	2	0.5/18	16.5/18	0

approach is to use complexity theory [7]. Essentially, every path by which an evidential trace could have been produced is enumerated, and the probability associated with each path is evaluated using techniques from complexity theory.

We illustrate the approach using an example from the BitTorrent case. In particular, we evaluate the prior probability that hypothesis H_2 is true given that trace evidence E_8 (i.e., T_5) is found (Figure 2). E_8 is the evidence that the torrent file is present on the hard disk of the seized computer.

Three scenarios that result in the presence of the torrent file are:

- The file was placed on the seized computer by a covert malware process.
- The file was copied or downloaded to the seized computer from some other source.
- The file was created on the seized computer from the pirated file.

Assume that a state-of-the-art, anti-malware scan reveals the presence of a Trojan with a probability of approximately 0.98 [5]. Additionally, a thorough, careful inventory of the local networked drives and portable storage media reveals the presence of a source copy of the torrent file with a probability greater than 0.95. Furthermore, a high-quality search engine detects the presence of a downloadable copy of the torrent file with a similar probability [1]. As a result, the probability that the torrent file was created *in situ* on the hard disk of the seized computer is at least 0.88. The error bars for the assigned probabilities are derived assuming that the errors are normally distributed. Based on these assignments, we obtain a probability value of 0.94 ± 0.06.

6. Conclusions

The proposed two-phase digital forensic investigation model achieves the twin goals of reliability and cost-effectiveness by incorporating a pre-processing phase, which runs in parallel with the data collection phase. The evidentiary weighting and cost ranking of the expected traces, which are undertaken only once for all similar investigations, enable the lowest cost traces to be examined first. This means that the "best-case" and "worst-case" scenarios can be processed efficiently. The combined use of evidentiary weights and ranked costs enables an ultimately futile investigation to be detected early and abandoned using only low cost traces. By the same token, an investigation that would ultimately prove to be unsuccessful could be halted before any high cost traces are investigated.

This model performs best in cases where the distributions of evidentiary weights versus costs are skewed towards low costs, and it performs the worst when the distributions are skewed towards high costs. In the average case, where the distribution is essentially unskewed (or even uniform), the model exhibits intermediate performance. However, it should be noted that, even in the most pathological cases, the performance would not be significantly worse than the current exhaustive or random search for traces.

One of the main advantages of the model is that it offers the possibility of creating templates of expected traces and their associated costs and evidentiary weights for every type of digital crime. These templates can provide investigators with benchmarks for calibrating their investigative procedures, and also offer novice investigators with an investigative model that can be adopted in its entirety.

References

[1] S. Brin and L. Page, The anatomy of a large-scale hypertextual web search engine, *Computer Networks and ISDN Systems*, vol. 30(1-7), pp. 107–117, 1998.

[2] B. Carrier and E. Spafford, Defining event reconstruction of digital crime scenes, *Journal of Forensic Sciences*, vol. 49(6), pp. 1291–1298, 2004.

[3] E. Casey, *Digital Evidence and Computer Crime: Forensic Science, Computers and the Internet*, Academic Press, London, United Kingdom, 2004.

[4] Joint Committee on Human Rights, Counter-Terrorism Policy and Human Rights: Terrorism Bill and Related Matters, Third Report of Session 2005-06, HL Paper 75-I, HC 561-I, House of Lords, House of Commons, London, United Kingdom, 2005.

[5] Kaspersky Lab, Free online virus scanner, Woburn, Massachusetts (www.kaspersky.com/virusscanner).

[6] M. Kwan, K. Chow, F. Law and P. Lai, Reasoning about evidence using Bayesian networks, in *Advances in Digital Forensics IV*, I. Ray and S. Shenoi (Eds.), Springer, Boston, Massachusetts, pp. 275–289, 2008.

[7] S. Lloyd, Measures of complexity: A non-exhaustive list, *IEEE Control Systems*, vol. 21(4), pp. 7–8, 2001.