Chapter 1

A COST-EFFECTIVE DIGITAL FORENSICS INVESTIGATION MODEL

R.E. Overill, M. Kwan, K.P. Chow, P. Lai and F. Law

Abstract Computers operate at discrete points in time and hence digital traces are discrete events in temporal logic that reflect the occurrence of computer processes. From the perspective of a digital investigation, it is the duty of digital investigators or forensic examiners to retrieve digital traces so as to prove or to refute the alleged computer acts. Given the resource constraints of most organizations and the limited time-frame available for the examination, it is not always feasible or indeed necessary for forensic examiners to retrieve all the related digital traces and to conduct a thorough digital forensic analysis. It is therefore the aim of this paper to propose a model that can offer swift and practical digital examination in a cost-effective manner.

Keywords: Bayesian Network, investigation Model, digital forensics

1. Introduction

Digital forensics involves the application of a series of processes on digital evidence such as identification, preservation, analysis and presentation. In the analysis process, event reconstruction is an important phase. It is in this phase that digital forensic examiners have to evaluate the truthfulness of the forensics hypotheses of the crime or incident according to the identified and retrieved digital traces [3, 7]. Stemming from their inherent technological complexities, the identification and retrieval of digital traces covers a wide range of techniques such as cryptography, data carving, data reconstruction, etc. It is therefore reasonable to expect that the retrieval of digital traces at different levels of complexity will involve different levels of cost in terms of resources required such as expertise, time, tools, etc.

Fortunately, unlike physical events that are analogue and continuous, digital events are discrete and occur in temporal sequence [1]. Therefore, it is viable to determine the retrieval costs of individual digital traces. However, in the lack of model for digital forensics investigation, most digital forensics examiners tend to conduct a thorough retrieval of all the related digital traces, despite being aware of the costs of retrieving those digital traces.

We explain the aforementioned situation in a technical sense. Taking an example of an investigation that comprises $\mathbf{m}$ digital traces there is a total of $\mathbf{m}!$ permutations which represent all possible investigation paths. Not all investigation paths are equally cost-effective however. Typically, investigators may attempt to perform an exhaustive search to identify all $\mathbf{m}$ traces or they may conduct a random search for the traces.

However, since the investigation of different digital traces generally requires different resources (e.g. time, tools, expertise, etc.) in different amounts, the aforementioned approaches may be regarded as inefficient. Additionally, the limited time-frame available for an investigation also renders exhaustive search approaches impractical [2]. Consequently, in situations where digital traces that carry significant evidential weights could not be found, examiners would still endeavor to retrieve remaining traces. Those remaining traces, however, are not sufficient to prove the hypotheses, notwithstanding resources, which should have been deployed to other forensic cases, are wasted.

Based on the retrieval costs of digital traces and by permutation analysis, this paper aims to derive a cost-effective model to address the aforementioned problem in digital forensics investigation.

2. The Proposed Model

Using the collective experience and judgment of digital forensic examiners it is possible to rank the relative costs of investigating each of the $\mathbf{m}$ traces $\mathbf{T}_i$. The relative costs can be estimated in terms of their resource requirements (person-hours, access to specialist equipment, etc.) using standard business accountancy procedures, prior to ranking. Without loss of generality we can take the relative cost ranking to be: $\mathbf{T}_1 \leq \mathbf{T}_2 \leq \ldots \leq \mathbf{T}_{m-1} \leq \mathbf{T}_m$. As a direct consequence of this ranking, the minimum cost path for the overall investigation is immediately uniquely defined. It is worth to denote here that different organizations can adopt different relative costs to similar traces in order to meet with the organizational goals.

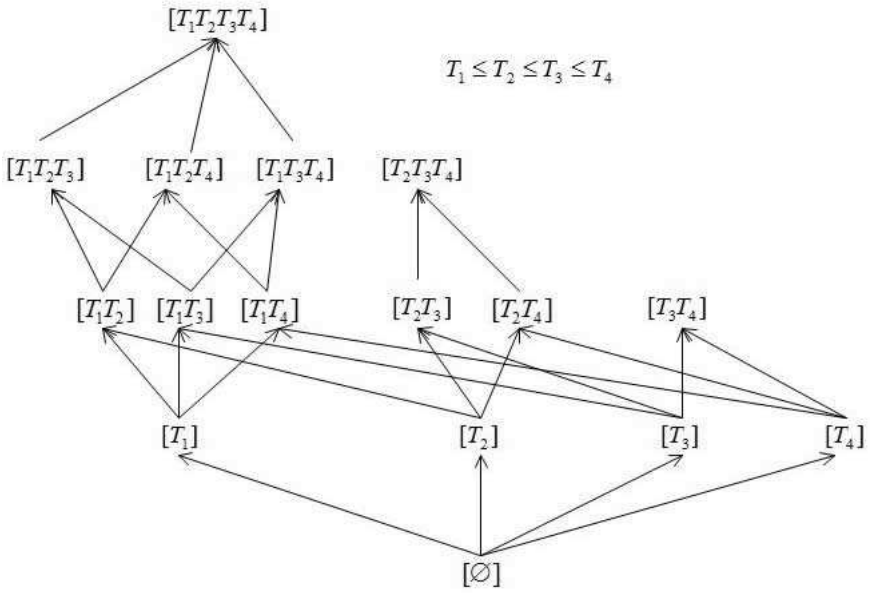


Figure 1. Example Path Diagram with 4 Traces

Our research focuses on digital traces on a hard disk. If the seized computer has sufficient storage, all the digital traces can in principle be preserved. If all the $\mathbf{T}_i$ (i=1...**m**) are preserved, the minimum cost path is simply the permutation $[\mathbf{T}_1\mathbf{T}_2 \ldots \mathbf{T}_{m-1}\mathbf{T}_m]$. Refer to the simple (**m**=4) example path diagram 1. It may be of some general interest to note here that the number of possible paths at each step is given by the corresponding binomial coefficient of **m**. This is a direct consequence of the isomorphism between the problem of selecting the next available trace from an ordered permutation of **m** distinct traces, and the problem of selecting the next object from a collection of **m** identical objects.

It is also valuable to acquire a prior indication as to whether the overall investigation should be proceeded with. We estimate the evidential weight associated with the investigation as $\mathbf{W}=\sum_{i=1}^{m} w_i$ where the relative fractional importance w_i of each trace $\mathbf{T}_i$ is either assigned by moderated independent expert peer review, or by default is set equal to $1/m$. We note in passing that this process only needs to be undertaken once as a pre-processing step for each distinct digital crime template. The estimate **W** should be compared with unity. If **W** is sufficiently close to unity, this signifies that 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.

In other words, the differential gap between $\mathbf{W}$ and unity can formulate a "cut-off" condition that can avoid identifying all traces exhaustively in a forensics investigation. This "cut-off" state can be illustrated by the following example. Suppose in an investigation where email exchanges between the culprit and the victim are essential. The forensic focus is therefore to confirm the computer, which was under the culprit's control at the material times, had been used to send and receive the material emails. Assume the evidential traces for the above premise are $\mathbf{T}_1$, $\mathbf{T}_2$, $\mathbf{T}_3$, $\mathbf{T}_4$, and $\mathbf{T}_5$ with evidential weights of 0.05, 0.10, 0.15, 0.2, and 0.35 respectively. Now, suppose the calculated evidential threshold value is 0.85. Therefore, if all the traces are found and identified, then the estimated total evidential weight is 0.85, which means a strong case. On the other hand, if trace $\mathbf{T}_1$ was not found the overall evidential weight will be 0.8, indicating 6% fall-off. If both $\mathbf{T}_1$ and $\mathbf{T}_2$ were missing, then the overall evidential weight will be 0.7, denoting 18% fall-off. Now, forensic investigators of this case should consider suspending the examination as the prospect of a successful prosecution is slim.

3. Missing Traces

Given sufficiently large storage may not be available in every computer, there exists the chance that some of the traces are missing or overwritten. In other words, there is no way that an examiner can fully ascertain the trace evidence of the case. Suppose that a single trace $\mathbf{T}_j$ $(1 \leq j \leq \mathbf{m})$ is not found. All investigative paths involving $\mathbf{T}_j$ must be deleted from the path diagram and the minimum cost path becomes $[\mathbf{T}_1\mathbf{T}_2 \ldots \mathbf{T}_{j-1}\mathbf{T}_{j+1} \ldots \mathbf{T}_{m-1}\mathbf{T}_m]$. The estimate of the evidential weight is given by $\mathbf{W}{=}\sum_{i\neq j}^{m} w_i$.

Similarly, if any two traces, $\mathbf{T}_j$ and $\mathbf{T}_k$ $(1 \leq j; k \leq \mathbf{m}; j \leq k)$ are not found then all investigative paths involving $\mathbf{T}_j$ or $\mathbf{T}_k$ must be deleted and the minimum cost path is $[\mathbf{T}_1\mathbf{T}_2 \ldots \mathbf{T}_{j-1}\mathbf{T}_{j+1} \ldots \mathbf{T}_{k-1}\mathbf{T}_{k+1} \ldots \mathbf{T}_{m-1}\mathbf{T}_m]$; the estimate of the evidential weight is $\mathbf{W}{=}\sum_{i\neq j,k}^{m} w_i$. More generally, if a total of any k traces are not found $(1 \leq k < \mathbf{m})$, then all investigative paths containing any of these k traces must be deleted from the path diagram.

We consider here briefly the issue of the independence of the digital traces $\mathbf{T}_i$. While the observations of the traces are necessarily independent because they are performed individually *post mortem*, it should be noted that the digital traces must also be created independently if the model is to retain its validity. Since it is possible in principle for one

user action **A** to result in the creation of multiple digital traces $\mathbf{T}_i$ which are consequently not mutually independent, care must be taken when selecting the set of expected digital traces to ensure their independence.

4. The Schema

We can now set out a two-phase schema for performing a minimum cost path digital forensic examination as follows: ***Phase 1 (pre-processing - detecting the traces)***:

- Enumerate the set of traces that are expected to be present in the seized computer based on the type of computer crime that is suspected of having been committed.
- Assign relative investigation costs to each of the expected traces.
- Rank the expected traces in order of increasing relative investigation costs.
- Assign relative importance weights w_i to each of the ranked traces.
- Rank the expected traces within each cost band in order of decreasing relative importance weight.
- Set **W**, the cumulative evidential weight estimate, equal to zero.
- Set W_{rem}, the remaining total of available weights, to 1.
- For each expected trace, taken in ranked order:
  - Search for the expected trace.
  - Subtract the relative importance weight w_i of the expected trace from W_{rem}.
  - If the expected trace is present add its relative importance weight w_i to **W**.
  - If **W** is sufficiently close to 1 then proceed immediately to ***Phase 2***.
  - If (**W**+W_{rem}) is insufficiently close to 1 then abandon the forensics investigation.

Phase 2 (Bayesian network - analyzing the traces):

- Set up a full Bayesian Network model for the hypothesis of the digital crime and run and analyze the Bayesian Network model for the hypothesis of the digital crime as described previously in [3].

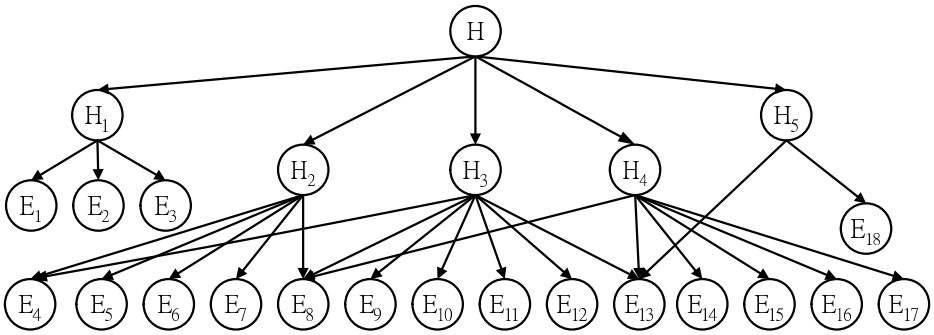
5. BitTorrent Example Case Study

In this section we use the BitTorrent (BT) example from our previous work to demonstrate how our proposed cost-effective digital forensics investigation methodology can be applied in practice. The Bayesian Network model diagram (1.5) is taken from Figure 5 of [3]. The 18 expected evidences **E** in 1 are ranked by cost as the 18 expected traces **T** for the ideal case in 1. The actual case, where two of the expected traces are in fact missing, is similarly described in 2. A potential complication of our proposed schema should be highlighted at this juncture. It is possible that a trace which is initially assigned a low cost may subsequently be discovered to entail a significantly cost. Typical examples would include a file which turns out to be protected by encryption, or a partition which turns out to be deleted, etc. In such cases, the cost of investigating that trace must be revised and the traces must be re-ranked according to the revised cost. This procedure is necessary in order to maintain the minimum cost strategy for the investigation.

6. Analytic Assignment of Prior Probabilities

For any given type of digital crime, the Bayesian network model corresponding to the investigation of that crime requires two kinds of input. Firstly, the overall structure of the Bayesian network itself has to be defined; this comprises the hierarchy of hypotheses and the associated posterior digital evidence (or traces) whose presence or absence determines the prior probabilities of the corresponding hypothesis. Secondly, numerical values have to be assigned to the prior probabilities. Traditionally, forensic examiners have assigned the prior probabilities semi-quantitatively based on a consensus of past experience and professional expertise. Recently, however, there have been a number of challenges to the qualitative assessments of forensic examiners acting as expert witnesses in judicial proceedings, primarily on the grounds that these assessments are non-rigorous and/or subjective.

These challenges can be effectively countered if a rigorous analytic procedure is developed to assign the prior probabilities quantitatively. We propose here that the application of complexity theory in its various manifestations [4] can be used to address this issue directly. Essentially, every route by which each evidential trace could have been produced is enumerated, and the probability associated with each route is evaluated using the tools and techniques of complexity theory. We illustrate our contention with an example from the BitTorrent (BT) case [3]: we evaluate the prior probability that hypothesis H_2 is true given that trace evidence E_8 (i.e. T_5) is found.



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 disc 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, which caused the seized computer to connect to the tracker server

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

EVIDENCE:

E_1 - Modify time of the destination file equals to that of the sourced file

E_2 - Creation time of the destination file lags behind its own Modify time

E_3 - Hash value of the destination file matches that of the sourced file

E_4 - BitTorrent client software is installed on the computer

E_5 - File link for the shared file created

E_6 - File being shared exists in the hard disk

E_7 - Torrent file creation record is found

E_8 - Torrent file exists in the hard disk

E_9 - Peer connection information is found

E_{10} - Tracker server log-in record is found

E_{11} - Torrent file activation time reflected from MAC time of Torrent file and its link file

E_{12} - Internet history record on 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 available

E_{17} - Internet cache record on publishing of Torrent file is found

E_{18} - Internet history record on Tracker server connection is found

Figure 2. Bayesian Network Model for the BT Case

Table 1. Traces, Costs and Weights for the Ideal BT Case

	Traces	Relative Cost	Relative Importance Weight	**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 after its own modification time	1	1/18	3/18	15/18
$T_3(E_2)$	Creation time of the destination file is after its own 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 created	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 about 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 about the publishing of the torrent file is found	1.5	0.5/18	14/18	4/18
$T_{15}(E_{18})$	Internet history record about 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 seized 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, Costs and Weights for the Actual BT Case

	Traces	Relative Cost	Relative Importance Weight	**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 after its own modification time	1	1/18	3/18	15/18
$T_3(E_2)$	Creation time of the destination file is after its own 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 created	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 about 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 about 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 about 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 seized 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

Evidence E_8 is that the torrent file is present on the hard disk of the seized computer. The various scenarios resulting in the presence of the torrent file are as follows:

- it was placed there by a covert malware process (e.g. a Trojan horse)
- it was copied or downloaded there from some other source
- it was created there from the pirated file

A state-of-the-art anti-malware scan would reveal the presence of a suitable vector Trojan with a probability of approximately 0.98 [5]; the efficacy of heuristics and behaviour blocking against zero-day malware attacks are the principal sources of uncertainty in this figure. A thorough, careful inventory of local networked drives and portable storage media would reveal the presence of any source copy of the torrent file with a probability in excess of 0.95. A high-quality search engine would detect the presence of any downloadable copy of the torrent file with a similar probability [6]. As a result, the probability that the torrent file was created in situ on the hard disc of the seized computer would be assigned at least 0.88. Furthermore, it is also possible to derive error bars for the assigned probabilities assuming that the errors are normally distributed. In the present case we obtain 0.94 ± 0.06.

7. Summary and Conclusions

The proposed two-phase digital investigation methodology is intended to achieve its twin goals of reliability and cost-effectiveness through the use of what is essentially a pre-processing and pre-screening phase which runs in parallel with the usual data collection phase, referred to here collectively as Phase 1.

The cost ranking and importance weighting of the expected traces, which only has to be undertaken once for all investigations of a similar type, enables the lowest cost traces to be examined first. This means that both 'best case' and 'worst case' scenarios can be efficiently processed. The combined use of importance weights and ranked costs means that an ultimately futile investigation may be detected early, using only low cost traces, and abandoned. By the same token, an investigation which will ultimately prove unsuccessful may be halted before the most high cost traces are investigated.

The model will perform best in cases where the distribution of importance versus cost is skewed towards low cost, and worst in cases where the distribution of importance versus cost is skewed towards high cost.

In the average case, where this distribution is essentially unskewed, or even uniform, the model will exhibit an intermediate performance. However, it should be noted that even in the most pathological cases, the model's performance should not be significantly worse than the current exhaustive search or random search for traces.

One of the advantages of the model is that it offers the possibility of creating templates of expected traces and their associated costs and importance weights for each distinct type of digital crime. Taken together with the systematic schema (section 4) for the investigation process, it offers less experienced investigators a benchmark by which to calibrate their own investigative procedures, while at the same time providing trainee investigators with a system for adoption in its entirety.

Given the current worldwide under-resourcing by governments of public sector law enforcement agencies, cost-effective utilization of scarce resources is essential and our minimum cost path approach appears well-matched to attaining this objective.

References

[1] E. Casey, Digital Evidence and Computer Crime: Forensic Science, Computers, and the Internet, 2nd ed. London, UK: Academic Press, 2004.

[2] House of Lords, House of Commons, 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.

[3] M. Kwan, K.P. Chow, F. Law & P. Lai. Reasoning About Evidence using Bayesian Network, *Advances in Digital Forensics IV*, International Federation for Information Processing (IFIP) January 2008, Tokyo, pp.141-155.

[4] S. Lloyd. Measures of Complexity: A Nonexhaustive List, *IEEE Control Systems Magazine*, Vol. 21 Issue 4 August 2001, pp. 7-8.

[5] Take the Kaspersky Challenge: See what your current antivirus is missing. (www.kaspersky.com/virusscanner/)

[6] S. Brin & L. Page, The Anatomy of a Large-Scale Hypertextual Web Search Engine, *Computer Networks and ISDN Systems*, Vol. 30, No. 1–7. (1998), pp. 107-117.

[7] B Carrier, E Spafford, Defining Event Reconstruction of Digital Crime Scenes, *Journal of forensic sciences*, Vol. 49, No. 6. (2004), pp. 1291-8.