Measuring Evidential Weight in Digital Forensic Investigations

Richard Overill, Kam-Pui Chow

Chapter 1

MEASURING EVIDENTIAL WEIGHT IN DIGITAL FORENSIC INVESTIGATIONS

Richard Overill and Kam-Pui Chow

Abstract This chapter describes a method for obtaining a quantitative measure of the relative weight of each individual item of evidence in a digital forensic investigation using a Bayesian network. The resulting evidential weights can then be used to determine a near-optimal, cost-effective triage scheme for the investigation in question.

Keywords: Bayesian network, evidential weight, triage, digital crime templates

1. Introduction

An inability to reliably quantify the relative plausibility of alternative hypotheses purporting to explain the existence of the totality of the digital evidence recovered in criminal investigations has hindered the transformation of digital forensics into a mature scientific and engineering discipline from the qualitative craft that originated in the mid 1980s [3]. A rigorous science-and-engineering-oriented approach can provide numerical results and also quantify the confidence limits, sensitivities and uncertainties associated with the results. However, there is a dearth of research literature focused on developing rigorous approaches to digital forensic investigations.

Posterior probabilities, likelihood ratios and odds generated using technical approaches such as Bayesian networks can provide digital forensic investigators, law enforcement officers and legal professionals with a quantitative scale or metric against which to assess the plausibility of an investigative hypothesis, which may be linked to the likelihood of successful prosecution or indeed the merit of a not-guilty plea. This approach is sometimes referred to as digital meta-forensics, some examples of which can be found in [4, 5].

A second and closely related issue involves the reliable quantification of the relative weight (also known as the probative value) of each of the individual items of digital evidence recovered during a criminal investigation. This is particularly important from the perspective of digital forensic triage – a prioritization strategy for searching for digital evidence in order to cope with the ever-increasing volumes of data and varieties of devices that are routinely seized for examination. The economics of digital forensics, also known as digital forensonomics [8], provides for the possibility of a quantitative basis for prioritizing the search for digital evidence during criminal investigations. This is accomplished by leveraging well-known concepts from economics such as the return-on-investment, or equivalently, the cost-benefit ratio.

In this approach, a list of all the expected items of digital evidence for the hypothesis being investigated is drawn up. For each item of digital evidence, two attributes are required: (i) cost, which is, in principle, relatively straightforward to quantify because it is usually measured in terms of the resources required to locate, recover and analyze the item of digital evidence (typically investigator hours plus any specialized equipment hire-time needed); and (ii) relative weight, which measures the contribution that the presence of the item of digital evidence makes in supporting the hypothesis (usually based on the informal opinions or consensus of experienced digital forensic investigators) [9].

The principal goals of this research are to: (i) demonstrate that a quantitative measure of the relative weight of each item of digital evidence in a particular investigation can be obtained in a straightforward manner from a Bayesian network representing the hypothesis underpinning the investigation; and (ii) demonstrate that the evidential weights can be employed to create a near-optimal, cost-effective evidence search list for the triage phase of the digital forensic investigation process.

It has been observed that a substantial proportion of digital crimes recorded in a particular jurisdiction at a particular epoch can be represented by a relatively small number of digital crime templates. It follows that, if each of these commonly-occurring digital crimes can be investigated more efficiently with the aid of its template, then the overall throughput of investigations may be improved with corresponding benefits to the criminal justice system as a whole.

2. Methodology

Bayesian networks were first proposed by Pearl [11] based on the concept of conditional probability originated by Bayes [2] in the eighteenth century. A Bayesian network is a directed acyclic graph (DAG) rep-

resentation of the conditional dependency relationships between entities such as events, observations and outcomes. Visually, a Bayesian network typically resembles an inverted tree.

In the context of a digital forensic investigation, the root node of a Bayesian network represents the overall hypothesis underpinning the investigation, the child nodes of the root node represent the sub-hypotheses that contribute to the overall hypothesis and the leaf nodes represent the items of digital evidence associated with each of the sub-hypotheses. After populating the interior nodes with conditional probabilities (likelihoods) and assigning prior probabilities to the root node, the Bayesian network propagates the probabilities using Bayesian inference rules to produce a posterior probability for the root hypothesis. However, it is the architecture of the Bayesian network together with the definition of each sub-hypothesis and its associated evidential traces that together define the hypothesis characterizing the specific investigation.

The first application of a Bayesian network to a specific digital forensic investigation appeared in 2008 [4]. Figure 1 presents a Bayesian network associated with a BitTorrent investigation, which will be employed later in this chapter.

The posterior probability produced by a Bayesian network when all the expected items of digital evidence are present is compared against the posterior probability of the Bayesian network when item i of the digital evidence is absent (but all the other expected evidential items are present). The difference between, and the ratio of, these two quantities provide direct measures of the relative weight of item i of the digital evidence in the context of the investigative hypothesis represented by the Bayesian network.

The relative weight RW_i of evidential item i satisfies the following proportionality equation:

$$RW_i \propto PP - PP_i \tag{1}$$

where PP is the posterior probability of the Bayesian network and PP_i is the posterior probability of the Bayesian network when evidential item i is absent.

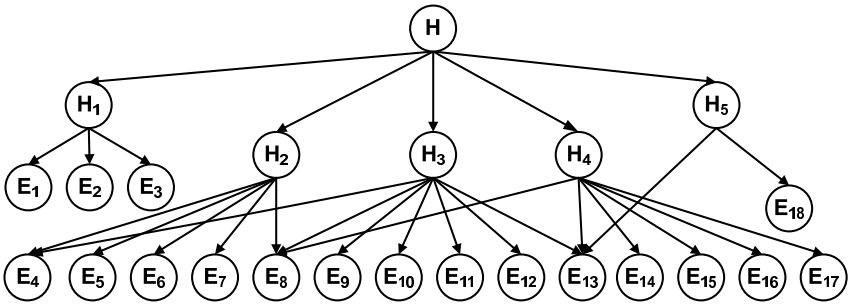
This equation can be written in normalized form as:

$$RW_i \propto 1 - \frac{PP_i}{PP} \tag{2}$$

or, alternatively as:

$$RW_i \propto \frac{PP}{PP_i} \tag{3}$$

From a ranking perspective, any of Equations (1), (2) or (3) could be used because, in each case, the relative weight of evidential item i



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, 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	Modification time of the destination file equals that of the source file
E_2	Creation time of the destination file is after its own 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 seized computer
E_5	File link for the shared file is created
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 about 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 about the publishing of the torrent file is found
E_{18}	Internet history record about the tracker server connection is found

Figure 1. Bayesian network for a BitTorrent investigation [4].

increases monotonically with the difference between the posterior probabilities. The remainder of this work will continue to employ Equation (1).

For a Bayesian network incorporating n items of digital evidence, it is necessary to perform $n + 1$ executions of the network. After all the relative evidential weights have been obtained in this manner using any of Equations (1), (2) or (3), the return on investment (RoI) and cost-benefit ratio (CBR) for item i of the expected digital evidence in the hypothesis satisfy the following proportionality equations [8]:

$$RoI_i \propto \frac{RW_i}{(EH_i \times HC) + EC_i} \tag{4}$$

$$CBR_i \propto \frac{(EH_i \times HC) + EC_i}{RW_i} \tag{5}$$

where EH_i is the examiner hours spent on evidential item i, HC is the hourly cost and EC_i is the equipment cost associated with evidential item i.

3. Results and Discussion

The real-world criminal case involving the illegal uploading of copyrighted material via the BitTorrent peer-to-peer network [4, 6] is used to illustrate the application of the proposed approach. The freely-available Bayesian network simulator MSBNx from Microsoft Research [7] was used to perform all the computations. The results were subsequently verified independently using the freeware version of AgenaRisk [1]. A previous sensitivity analysis performed on the Bayesian network for the BitTorrent case [10] demonstrated that the posterior probabilities and, hence, the relative evidential weights derived from them, are stable to within $\pm 0.5\%$.

The ranked evidential weights of the eighteen items of digital evidence shown in Figure 1 are listed in Table 1 along with their estimated relative costs [9] and their associated return-on-investment and cost-benefit-ratio values computed using Equations (4) and (5), respectively. The relative evidential recovery costs for the Bayesian network are taken from [9]; they were estimated by experienced digital forensic investigators from the Hong Kong Customs and Excise Department IPR Protection Group, taking into account the typical forensic examiner time required along with the use of any specialized equipment. The proposed approach assumes that the typical cost of locating, recovering and analyzing each individual item of digital evidence is fixed. However, it is possible that the cost could be variable under certain circumstances; for example, when evidence recovery requires the invocation of a mutual

Table 1. Attribute values of digital evidential items in the BitTorrent investigation.

Posterior Probability	Evidential Item	Relative Weight	Relative Cost	RoI	CBR
0.9255	–	–	–	–	–
0.8623	E_{18}	0.0632	1.5	4.214	0.237
0.8990	E_{13}	0.0265	1.5	1.767	0.566
0.9109	E_3	0.0146	1.0	1.459	0.685
0.9158	E_1	0.0097	1.0	0.968	1.033
0.9158	E_2	0.0097	1.0	0.968	1.033
0.9239	E_{11}	0.0016	2.0	0.082	12.20
0.9240	E_6	0.0015	1.0	0.151	6.622
0.9242	E_{16}	0.0013	1.0	0.127	7.874
0.9247	E_{12}	0.0008	1.5	0.050	20.00
0.9248	E_9	0.0007	2.0	0.036	27.78
0.9248	E_{10}	0.0007	1.5	0.047	21.28
0.9249	E_8	0.0006	1.0	0.062	16.13
0.9251	E_{15}	0.0004	1.0	0.040	25.00
0.9251	E_{17}	0.0004	1.5	0.027	37.04
0.9252	E_{14}	0.0003	1.5	0.021	47.62
0.9252	E_4	0.0003	2.0	0.013	76.92
0.9253	E_5	0.0002	1.0	0.015	66.67
0.9254	E_7	0.0001	1.5	0.007	142.90

legal assistance treaty with a law enforcement organization in another jurisdiction.

The relative evidential weights in Table 1 can be used to create an evidence search list, with the evidential items ordered first by decreasing relative weight and, second, by decreasing return-on-investment or, equivalently, by increasing cost-benefit ratio. This search list can be used to guide the course of the triage phase of the digital forensic investigation in a near-optimal, cost-effective manner. Specifically, it would ensure that evidential "quick wins" (or "low-hanging fruit") are processed early in the investigation. Evidential items with low relative weights that are costly to obtain are relegated until later in the investigation, when it may become clearer whether or not these items are crucial to the overall support of the investigative hypothesis.

An advantage of this approach is that, if an item of evidence of high relative weight is not recovered, then this fact is detected early in the investigation; the investigation could be de-prioritized or even abandoned before valuable resources (time, effort, equipment, etc.) are expended unnecessarily. In addition, it may be possible to terminate the investigation without having to search for an item of evidence of low relative

weight with a high recovery cost (e.g., having to use a scanning electron microscope to detect whether or not a solid state memory latch or gate is charged) as a direct consequence of the law of diminishing returns.

In the BitTorrent example, if evidential item E_{18} cannot be recovered, the impact on the investigation would probably be serious and may well lead to the immediate de-prioritization or even abandonment of the investigation. On the other hand, the absence of evidential items E_5 or E_7 would make very little difference to the overall support for the digital forensic investigation hypothesis.

The approach can be refined further by considering the roles of potentially exculpatory (i.e., exonerating) items of evidence in the investigative context. Such evidence might be, for example, CCTV footage that reliably places the suspect far from the presumed scene of the digital crime at the material time. The existence of any such evidence would, by definition, place the investigative hypothesis in jeopardy. Therefore, if any potentially exculpatory evidence could be identified in advance, then a search for the evidence could be undertaken before or in parallel with the search for evidential items in the triage schedule. However, since, by definition, the Bayesian network for the investigative hypothesis does not contain any exculpatory evidential items, the network cannot be used directly to obtain the relative weights of any items of exculpatory evidence. Therefore, it is not possible to formulate a cost-effective search strategy for exculpatory items of evidence on the basis of the Bayesian network itself.

4. Conclusions

This chapter has described a method for obtaining numerical estimates of the relative weights of items of digital evidence in digital forensic investigations. By considering the corresponding return-on-investment and cost-benefit ratio estimates of the evidential items, near-optimal, cost-effective digital forensic triage search strategies for the investigations can be constructed, eliminating unnecessary utilization of scarce time, effort and equipment resources in today's overstretched and under-resourced digital forensic investigation laboratories. The application of the method to evidence in a real case involving the illegal uploading of copyright protected material using the BitTorrent peer-to-peer network demonstrates its utility and intuitive appeal.

References

[1] Agena, AgenaRisk 7.0, Bayesian Network and Simulation Software for Risk Analysis and Decision Support, Cambridge, United Kingdom (www.agenarisk.com/products), 2018.

[2] T. Bayes, An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, F.R.S. communicated by Mr. Price, in a letter to John Canton, A.M.F.R.S., *Philosophical Transactions (1683-1775)*, vol. 53, pp. 370–418, 1763.

[3] F. Cohen, *Digital Forensic Evidence Examination*, ASP Press, Livermore, California, 2010.

[4] 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.

[5] M. Kwan, R. Overill, K. Chow, J. Silomon, H. Tse, F. Law and P. Lai, Evaluation of evidence in Internet auction fraud investigations, in *Advances in Digital Forensics VI*, K. Chow and S. Shenoi (Eds.), Springer, Heildelberg, Germany, pp. 121–132, 2010.

[6] Magistrates' Court at Tuen Mun, Hong Kong Special Administrative Region v. Chan Nai Ming, TMCC 1268/2005, Hong Kong, China (www.hklii.hk/hk/jud/en/hksc/2005/TMCC001268A_2005.html), 2005.

[7] Microsoft Research, MSBNx: Bayesian Network Editor and Tool Kit, Microsoft Corporation, Redmond, Washington (msbnx.azurewebsites.net), 2001.

[8] R. Overill, Digital forensonomics – The economics of digital forensics, *Proceedings of the Second International Workshop on Cyberpatterns*, 2013.

[9] R. Overill, M. Kwan, K. Chow, P. Lai and F. Law, A cost-effective model for digital forensic investigations, in *Advances in Digital Forensics V*, G. Peterson and S. Shenoi (Eds.), Springer, Heidelberg, Germany, pp. 231–240, 2009.

[10] R. Overill, J. Silomon, M. Kwan, K. Chow, F. Law and P. Lai, Sensitivity analysis of a Bayesian network for reasoning about digital forensic evidence, *Proceedings of the Third International Conference on Human-Centric Computing*, 2010.

[11] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, Morgan Kaufman, San Mateo, California, 1988.