

Sensitivity Analysis of a Bayesian Network for Reasoning about Digital Forensic Evidence

Richard E. Overill, Jantje A. M. Silomon

*Department of Computer Science, King's College London,
Strand, London WC2R 2LS*

{richard.overill, jantje.a.silomon}@kcl.ac.uk

Michael Y.K. Kwan, Kam-Pui Chow, Frank Y.W. Law, Pierre
K.Y. Lai

*Department of Computer Science, University of Hong Kong,
Pokfulam Road, Hong Kong*

{ykkwan, chow, ywlaw, kylai}@cs.hku.hk

Abstract—A Bayesian network representing an actual prosecuted case of illegal file sharing over a peer-to-peer network has been subjected to a systematic and rigorous sensitivity analysis. Our results demonstrate that such networks are usefully insensitive both to the occurrence of missing evidential traces and to the choice of conditional evidential probabilities. The importance of this finding for the investigation of digital forensic hypotheses is highlighted.

Keywords—Bayesian network; digital forensic investigation; digital evidence; sensitivity analysis.

I. INTRODUCTION AND BACKGROUND

Bayesian Belief Networks (BBNs) [1] have recently been applied to reasoning about available evidence in digital forensic investigations [2]. BBNs are used to quantify the evidential strengths of investigative hypotheses and hence enhance the reliability and traceability of the results produced by digital forensic investigations.

In [2], the BBN for a Hong Kong court case involving the use of a BitTorrent (BT) peer-to-peer (P2P) network to act as the initial 'seeder' for illegally uploading a copyright protected audio-visual file for subsequent distribution was constructed and examined. In Hong Kong only the uploading of copyright protected material is prohibited whereas in the UK any activity infringing copyright is forbidden. It was found that the BBN is a useful tool for quantifying and propagating the strengths of investigative hypotheses and their supporting evidence. However, there is an inherent subjectivity involved in assigning conditional probabilities to posterior evidence in the BBN.

This was alleviated to some extent through the use of a survey of expert digital forensic investigators and aggregating their responses.

In addition to the potential subjectivity of the conditional evidential probabilities, also known as likelihoods, a second issue to be addressed is that of missing evidence. It is often the case that not all of the expected evidential traces are actually recovered during a digital forensic investigation, and it is important to know the impact of one or more missing evidential traces upon the overall strength of the hypothesis under investigation.

In this paper, we address both of these issues by performing a rigorous sensitivity analysis on the BT BBN from [2], involving the systematic removal of evidential traces from this BBN and the systematic replacement of the aggregated likelihoods by their respective minimum and maximum values.

We detail the methodology used in section 2 below. In section 3 we discuss the results of the sensitivity analysis and in section 4 we present our conclusions.

II. SENSITIVITY ANALYSIS METHODOLOGY

As in [2], all BT BBN calculations were performed with the Microsoft MSBNx Editor [3].

A. Missing evidential traces

The structure of the BT BBN from [2] is given in Figure 1. It will be noted that this BBN contains a root hypothesis (H), five sub-hypothesis (H₁-H₅) and 18 associated evidential traces (E₁-E₁₈). It is the combination of these traces, rather than any individual subset, that is

required to formulate the prosecution's case. We have systematically removed the evidential traces according to the following schema:

- 1) each individual trace (Table A1);
- 2) all possible pairs of traces (Table A2);
- 3) a sample of n -tuples of traces; $3 \leq n \leq 9$ (Table A3).

B. Minimum and maximum evidential likelihoods

We have systematically replaced the aggregated likelihoods from the original survey [2] by the minimum and maximum values of the responses provided by the sample of 31 expert digital forensic investigators. We have removed 'outlier' values from this sample by discounting any single response lying at either extreme of the range.

The results of this study are given in Table A4. In addition, we have simultaneously set all the likelihoods to their respective minimum and maximum values in turn. The outcome of these tests can be found at the foot of Table A4. The result obtained using the original aggregated values, as reported in [2], is given at the head of Table A1 for purposes of comparison.

III. DISCUSSION OF RESULTS

The results of our sensitivity analysis are collected in the Tables in the Appendix.

A. Missing evidential traces

Table A1 demonstrates that the output of the BT BBN is remarkably stable towards any single missing evidential trace. The mean percentage deviation of these values from the case where no traces are missing is -0.8%. Similarly, the results in Table A2 show that any pair of missing evidential traces does not greatly impact the probability of the BT BBN hypothesis. Here the mean percentage deviation is -2.0%. These deviations are negative since the absence of a trace decreases the probability of the root hypothesis. It will however be noted that the entries in Tables A1 and A2 involving E_1 , E_2 and E_3 (belonging to H_1) as well as E_{13} and E_{18} (belonging to H_5) are somewhat smaller in magnitude than the remainder. This can be attributed to the fact that, relative to the other hypotheses, H_1 and H_5 have fewer associated evidential traces and therefore the absence of any of these evidential traces has a proportionately greater impact on the BT BBN.

Even when more than two evidential traces are not recovered, the data in Table A3 suggests that in many cases the BT BBN can still yield a reasonably high

probability from a prosecutorial perspective. Indeed, we have recorded (Table A3, penultimate line) one extreme case in which 7 of the 18 evidential traces are missing but the BT BBN still maintains a probability in excess of 0.8.

B. Minimum and maximum evidential likelihoods

In Table A4 it will be seen that the BT BBN is also remarkably stable towards variation of the values of individual likelihoods: for the individual minimum and maximum values of the likelihoods the mean percentage deviations are -0.2% and +0.05% respectively. However, when all likelihoods are set to their respective minimum values simultaneously, the BT BBN experiences a much larger downward probability shift from the aggregated values (-25.5%) than the corresponding upward shift when all likelihoods are set to their respective maximum values simultaneously (+0.7%). This phenomenon can be explained by consulting the histogram of responses to the original survey [2], which exhibits a pronounced skewness towards the higher end of the range, due to the level of expert knowledge of the great majority of the respondents.

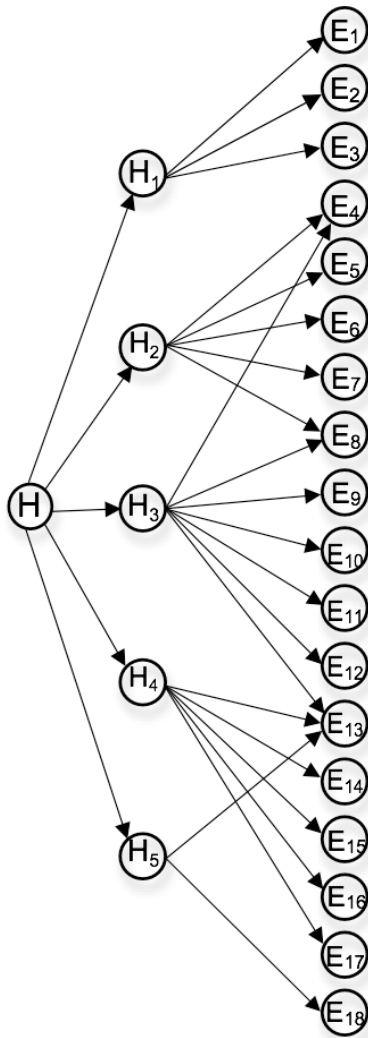
IV. SUMMARY AND CONCLUSIONS

The sensitivity analysis reported in this paper demonstrates that the BT BBN used in [2] is insensitive to the occurrence of missing evidence and also to the choice of evidential likelihoods to an unexpected degree.

However, as would be anticipated, in very extreme cases of missing evidential traces (Table A3), or improbable choices of values for the likelihoods (Table A4, final line), much lower probabilities for the investigatory hypothesis are found.

Our overall finding is gratifying because it implies that the exact choice of values for the inherently subjective evidential likelihoods is not as critical as might have been expected. Values falling within the consensus of experienced expert investigators are sufficiently reliable to be used in the BBN model. Furthermore, our results imply that the inability to recover one or more evidential traces in a digital forensic investigation is not generally critical for the probability of the investigatory hypothesis under consideration.

Since the actual BT BBN used in this study is typical in form and structure of the BBNs employed to investigate other digital crime hypotheses (for example, the Internet auction fraud investigations [4]), this study offers reassurance that those BBNs will also exhibit similarly stable behaviour towards the non-recovery of evidential traces and the values chosen for the evidential likelihoods.



HYPOTHESES:

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

H₁ The pirated file was copied from the seized optical disk to the seized computer

H₂ A torrent file was created from the copied file

H₃ The torrent file was sent to newsgroups for publishing

H₄ The torrent file was activated, which caused the seized computer to connect to the tracker server

H₅ The connection between the seized computer and the tracker server was maintained

EVIDENCE:

E₁ Modification time of the destination file equals that of the source file

E₂ Creation time of the destination file is after its own modification time

E₃ Hash value of the destination file matches that of the source file

E₄ BitTorrent client software is installed on the seized computer

E₅ File link for the shared file is created

E₆ Shared file exists on the hard disk

E₇ Torrent file creation record is found

E₈ Torrent file exists on the hard disk

E₉ Peer connection information is found

E₁₀ Tracker server login record is found

E₁₁ Torrent file activation time is corroborated by its MAC time and link file

E₁₂ Internet history record about the publishing website is found

E₁₃ Internet connection is available

E₁₄ Cookie of the publishing website is found

E₁₅ URL of the publishing website is stored in the web browser

E₁₆ Web browser software is available

E₁₇ Internet cache record about the publishing of the torrent file is found

E₁₈ Internet history record about the tracker server connection is found

Figure 1: The BT BBN [2]

REFERENCES

[1] J. Keppens and J. Zeleznikow, "A model based reasoning approach for generating plausible crime scenarios from evidence", Proceedings of the Ninth International Conference on Artificial Intelligence and Law (ICAIL'03), Edinburgh, Scotland, 24-28 June 2003, New York: ACM, 2003, pp. 51 – 59.

[2] M. Kwan, K.-P. Chow, F. Law and P. Lai, "Reasoning about evidence using Bayesian networks", Chapter 12, Advances in Digital Forensics IV, I Ray and S Sheno, Eds. Berlin: Springer, 2008, pp. 142 - 155.

[3] Microsoft Research, MSBNx: Bayesian Network Editor and Tool Kit, Microsoft Corporation, Redmond, Washington, USA, 2001.
<http://www.research.microsoft.com/adapt/MSBNx>

[4] M. Kwan, R. E. Overill, K.-P. Chow, J. A. M. Silomon, H. Tse, F. Law and P. Lai, "Internet auction fraud investigations", Proceedings of the Sixth Annual IFIP WG 11.9 International Conference on Digital Forensics, Hong Kong, 3-6 January 2010, Advances in Digital Forensics VI, S Sheno, Ed. Berlin: Springer, 2010, in press.

APPENDIX

TABLE A1
Single Missing Traces

Missing Evidence	BT BBN $p(H)$
none [2]	0.9255
E ₁	0.9158
E ₂	0.9158
E ₃	0.9109
E ₄	0.9252
E ₅	0.9253
E ₆	0.9240
E ₇	0.9254
E ₈	0.9249
E ₉	0.9248
E ₁₀	0.9248
E ₁₁	0.9239
E ₁₂	0.9247
E ₁₃	0.8990
E ₁₄	0.9252
E ₁₅	0.9251
E ₁₆	0.9242
E ₁₇	0.9251
E ₁₈	0.8623

TABLE A3
Multiple Missing Traces

Missing Traces	BT BBN $p(H)$
E ₄ , E ₈ , E ₁₃	0.8928
E ₁ , E ₄ , E ₈ , E ₁₃	0.8794
E ₉ , E ₁₀ , E ₁₁ , E ₁₂ , E ₁₄ , E ₁₅	0.6613
E ₄ , E ₆ , E ₈ , E ₉ , E ₁₁ , E ₁₅ , E ₁₆	0.6994
E ₁ , E ₂ , E ₅ , E ₇ , E ₈ , E ₁₀ , E ₁₄	0.8042
E ₄ , E ₆ , E ₈ , E ₉ , E ₁₁ , E ₁₃ , E ₁₅ , E ₁₆	0.4645

TABLE A4
BT BBN using Minimum and Maximum Likelihoods

Evidence	BT BBN $p(H)$ min	BT BBN $p(H)$ max
E ₁	0.9244	0.9256
E ₂	0.9250	0.9256
E ₃	0.9249	0.9256
E ₄	0.9254	0.9255
E ₅	0.9255	0.9255
E ₆	0.9252	0.9258
E ₇	0.9255	0.9255
E ₈	0.9252	0.9255
E ₉	0.9255	0.9256
E ₁₀	0.9255	0.9256
E ₁₁	0.9255	0.9256
E ₁₂	0.9254	0.9256
E ₁₃	0.9236	0.9255
E ₁₄	0.9253	0.9255
E ₁₅	0.9254	0.9255
E ₁₆	0.9254	0.9255
E ₁₇	0.9253	0.9255
E ₁₈	0.9012	0.9320
E ₁ -E ₁₈	0.6896	0.9316

TABLE A2

BT BBN $p(H)$ with all Pairs of Missing Traces																	
	E_1	E_2	E_3	E_4	E_5	E_6	E_7	E_8	E_9	E_{10}	E_{11}	E_{12}	E_{13}	E_{14}	E_{15}	E_{16}	E_{17}
E_2	0.8383																
E_3	0.8297	0.8297															
E_4	0.9155	0.9155	0.9106														
E_5	0.9156	0.9156	0.9107	0.9220													
E_6	0.9141	0.9141	0.9091	0.8982	0.9080												
E_7	0.9157	0.9157	0.9108	0.9228	0.9239	0.9115											
E_8	0.9151	0.9151	0.9102	0.9231	0.9239	0.9149	0.9241										
E_9	0.9150	0.9150	0.9101	0.9245	0.9246	0.9232	0.9247	0.9201									
E_{10}	0.9150	0.9150	0.9101	0.9245	0.9246	0.9232	0.9247	0.9201	0.9174								
E_{11}	0.9140	0.9140	0.9090	0.9236	0.9237	0.9223	0.9237	0.9146	0.9092	0.9092							
E_{12}	0.9150	0.9150	0.9100	0.9244	0.9246	0.9232	0.9246	0.9198	0.9170	0.9170	0.9085						
E_{13}	0.8863	0.8863	0.8799	0.8986	0.8988	0.8970	0.8988	0.8952	0.8942	0.8942	0.8882	0.8939					
E_{14}	0.9155	0.9155	0.9105	0.9249	0.9250	0.9237	0.9251	0.9227	0.9245	0.9245	0.9235	0.9244	0.8969				
E_{15}	0.9154	0.9154	0.9104	0.9248	0.9249	0.9236	0.9250	0.9221	0.9244	0.9244	0.9235	0.9243	0.8963	0.9191			
E_{16}	0.9144	0.9144	0.9094	0.9239	0.9241	0.9227	0.9197	0.9165	0.9235	0.9235	0.9226	0.9235	0.8907	0.9081	0.9040		
E_{17}	0.9154	0.9154	0.9104	0.9248	0.9249	0.9236	0.9250	0.9221	0.9244	0.9244	0.9225	0.9243	0.8963	0.9191	0.9174	0.9040	
E_{18}	0.8457	0.8457	0.8375	0.8618	0.8620	0.8597	0.8621	0.8617	0.8618	0.8618	0.8611	0.8618	0.8408	0.8620	0.8619	0.8610	0.8619