



Annual ADFSL Conference on Digital Forensics, Security and Law

2012
Proceedings

May 31st, 10:30 AM

Multi-Parameter Sensitivity Analysis of a Bayesian Network from a Digital Forensic Investigation

Richard E. Overill

Department of Informatics, King's College London, richard.overill@kcl.ac.uk


Echo P. Zhang

Department of Computer Science, University of Hong Kong, pzhang2@cs.hku.hk

Kam-Pui Chow

Department of Computer Science, University of Hong Kong, chow@cs.hku.hk

Follow this and additional works at: <https://commons.erau.edu/adfsl>

 Part of the [Computer Engineering Commons](#), [Computer Law Commons](#), [Electrical and Computer Engineering Commons](#), [Forensic Science and Technology Commons](#), and the [Information Security Commons](#)

Scholarly Commons Citation

Overill, Richard E.; Zhang, Echo P.; and Chow, Kam-Pui, "Multi-Parameter Sensitivity Analysis of a Bayesian Network from a Digital Forensic Investigation" (2012). *Annual ADFSL Conference on Digital Forensics, Security and Law*. 10.

<https://commons.erau.edu/adfsl/2012/thursday/10>

This Peer Reviewed Paper is brought to you for free and open access by the Conferences at Scholarly Commons. It has been accepted for inclusion in Annual ADFSL Conference on Digital Forensics, Security and Law by an authorized administrator of Scholarly Commons. For more information, please contact commons@erau.edu.

EMBRY-RIDDLE
Aeronautical University™
SCHOLARLY COMMONS

(c)ADFSL



MULTI-PARAMETER SENSITIVITY ANALYSIS OF A BAYESIAN NETWORK FROM A DIGITAL FORENSIC INVESTIGATION

Richard E. Overill

Department of Informatics
King's College London
Strand, London WC2R 2LS, UK
+442078482833
+442078482588
richard.overill@kcl.ac.uk

Echo P. Zhang

Department of Computer Science
University of Hong Kong
Pokfulam Road, Hong Kong
+85222417525
+85225998477
pzhang2@cs.hku.hk

Kam-Pui Chow

Department of Computer Science
University of Hong Kong
Pokfulam Road, Hong Kong
+85228592191
+85225998477
chow@cs.hku.hk

ABSTRACT

A multi-parameter sensitivity analysis of a Bayesian network (BN) used in the digital forensic investigation of the Yahoo! email case has been performed using the principle of 'steepest gradient' in the parameter space of the conditional probabilities. This procedure delivers a more reliable result for the dependence of the posterior probability of the BN on the values used to populate the conditional probability tables (CPTs) of the BN. As such, this work extends our previous studies using single-parameter sensitivity analyses of BNs, with the overall aim of more deeply understanding the indicative use of BNs within the digital forensic and prosecutorial processes. In particular, we find that while our previous conclusions regarding the Yahoo! email case are generally validated by the results of the multi-parameter sensitivity analysis, the utility of performing the latter analysis as a means of verifying the structure and form adopted for the Bayesian network should not be underestimated.

Keywords: Bayesian network; digital forensics; multi-parameter sensitivity analysis; steepest gradient.

1. INTRODUCTION

The use of Bayesian networks (BNs) to assist in digital forensic investigations of e-crimes is continuing to increase [Kwan, 2008; Kwan, 2010; Kwan, 2011] since they offer a valuable means of reasoning about the relationship between the recovered (or expected) digital evidential traces and the

forensic sub-hypotheses that explain how the suspected e-crime was committed [Kwan, 2008].

One of the principal difficulties encountered in constructing BNs is to know what conditional probability values are appropriate for populating the conditional probability tables (CPTs) that are found at each node of the BN. These values can be estimated by means of a survey questionnaire of a group of experienced experts [Kwan, 2008] but as such they are always open to challenge. There are also a number of alternative methods for estimating CPT values, including reasoning from historical databases, but none are entirely exempt from the potential criticism of lacking quantitative rigour. In these circumstances it is important to know how sensitive the posterior output of the BN is to the numerical values of these parameters. If the degree of sensitivity can be shown to be low then the precise values adopted for these parameters is not critical to the stability of the posterior output of the BN.

In this paper we generalize the concept of *sensitivity value* introduced by Renooij and van der Gaag [Renooij, 2004] to a multi-parameter space and adapt the concept of *steepest gradient* from the domain of numerical optimization to define a *local multi-parameter sensitivity value* in the region of the chosen parameter set. This metric defines the steepest gradient at the chosen point in the parameter space, which is a direct measure of the local multi-parameter sensitivity of the BN. It should be emphasised that in this work we are not aiming to optimize either the conditional probabilities or the posterior output of the BN as was the case with the multi-parameter optimization scheme of Chan and Darwiche [Chan, 2004]; our objective here is to measure the stability of the latter with respect to simultaneous small variations of the former by determining the steepest local gradient in the multi-parameter space.

2. SENSITIVITY ANALYSIS

A number of types of BN sensitivity analysis have been proposed. The most straightforward, although tedious, is the direct manipulation method which involves the iterative variation of one parameter at a time, keeping all the others fixed. This single-parameter approach was used to demonstrate the low sensitivity of the BitTorrent BN in [Overill, 2010]. Three somewhat more sophisticated approaches to single-parameter sensitivity analysis, namely, *bounding sensitivity analysis*, *sensitivity value analysis* and *vertex likelihood analysis* were proposed in [Renooij, 2004], and were each applied to the Yahoo! Email BN to demonstrate its low sensitivity in [Kwan, 2011]. However, a valid criticism of each of these single-parameter approaches is that the effect of simultaneous variation of the parameters is not considered. The multi-parameter sensitivity analysis scheme proposed in [Chan, 2004] requires the determination of k^{th} -order partial derivatives with an associated computational complexity of $O(n \prod_{x_i} F(X_i) e^w)$ in order to perform a k -way sensitivity analysis on a BN of tree-width w with n parameters where $F(X_i)$ is the size of CPT i . In order to develop a computationally tractable approach to local multi-parameter BN sensitivity analysis, we have generalized the original concept of the single-parameter sensitivity value [Renooij, 2004] to multi-parameter space and then applied the steepest gradient approach from numerical optimization to produce a metric for the local (or instantaneous) multi-parameter sensitivity value at the selected point in parameter space.

3. YAHOO! CASE (HONG KONG)

3.1 Background of Yahoo! Case (Hong Kong)

On April 20, 2004, Chinese journalist Shi Tao used his private Yahoo! email account and sent a brief of Number 11 document which was released by the Chinese government that day, to an overseas web site called Asia Democracy Foundation. When the Chinese government found it out, Beijing State Security Bureau requested the e-mail service provider, Yahoo! Holdings (Hong Kong) to provide details of the sender's personal information, like identifying information, login times, and e-mail contents. According to Article 45 of the PRC Criminal Procedure Law ("Article 45"), Yahoo! Holding (Hong Kong) was legally obliged to comply with the demand. Mr. Shi was accused with the crime of

“providing state secrets to foreign entities.” After the course of investigation, Mr. Shi was convicted and sentenced to ten years in prison in 2005 [Case No. 29, 2005].

In [Cap.486, 2007], it mentioned that:

“In the verdict (the “Verdict”) delivered by the People’s Court on 27 April 2005, it stated that Mr. Shi had on 20 April 2004 at approximately 11:32 p.m. leaked information to “an overseas hostile element, taking advantage of the fact that he was working overtime alone in his office to connect to the internet through his phone line and used his personal email account (huoyan-1989@yahoo.com.cn) to send his notes. He also used the alias ‘198964’ as the name of the provider ...”. The Verdict reported the evidence gathered to prove the commission of the offence which included the following: “Account holder information furnished by Yahoo! Holdings (Hong Kong) Ltd., which confirms that for IP address 218.76.8.21 at 11:32:17 p.m. on April 20, 2004, the corresponding user information was as follows: user telephone number: 0731-4376362 located at the Contemporary Business News office in Hunan; address: 2F, Building 88, Jianxing New Village, Kaifu District, Changsha.”

3.2 Digital Evidence in Yahoo! Case (Hong Kong)

From the above verdict, we can tell that the information provided by Yahoo! Holdings (Hong Kong) Ltd. is taken as the digital evidences in the court. Since we utilize a Bayesian Network as our analysis model, we have to construct a Hypothesis-Evidence system. In this case, the main (or root) hypothesis is:

Hypothesis H_0 : “The seized computer has been used to send the material document as an email attachment via a Yahoo! Web mail account”

Under the main hypothesis, we have six sub-hypotheses and their corresponding evidences which are listed below:

Table 1 Sub-hypothesis H_1 : Linkage between the material document and the suspect’s computer.

ID	Description	Evidence Type
DE_1	The subject document exists in the computer	Digital
DE_2	The “Last Access Time” of the subject file lags behind the IP address assignment time by the ISP	Digital
DE_3	The “Last Access Time” of the subject file lags behind or closes to the sent time of the Yahoo! email	Digital

Table 2 Sub-hypothesis H_2 : Linkage between the suspect and his computer.

ID	Description	Evidence Type
PE_1	The suspect was in physical possession of the computer	Physical
DE_4	Files in the computer reveals the identity of the suspect	Digital

Table 3 Sub-hypothesis H_3 : Linkage between the suspect and the ISP

ID	Description	Evidence Type
DE_5	The ISP subscription details (including the assigned IP address) matches the suspect's particulars	Digital

Table 4 Sub-hypothesis H_4 : Linkage between the suspect and Yahoo! email account

ID	Description	Evidence Type
DE_6	The subscription details (including the IP address that sent the email) of the Yahoo! email account matches the suspect's particulars	Digital

Table 5 Sub-hypothesis H_5 : Linkage between the computer and the ISP

ID	Description	Evidence Type
DE_7	Configuration setting of the ISP Internet account is found in the computer	Digital
DE_8	Log data confirms that the computer was powered up at the time when the email was sent	Digital
DE_9	Web browsing program (e.g. Internet Explorer) or email user agent program (e.g. Outlook) is found activated at the time the email was sent	Digital
DE_{10}	Log data reveals the assigned IP address and the assignment time by the ISP to the computer	Digital
DE_{11}	ISP confirms the assignment of the IP address to the suspect's account	Digital

Table 6 Sub-hypothesis H_6 : Linkage between the computer and Yahoo! email account

ID	Description	Evidence Type
DE_{12}	Internet history logs reveal the access of the Yahoo! email account by the computer	Digital
DE_{13}	Internet cached files reveal the subject document has been sent as an attachment via the Yahoo! email account	Digital
DE_{14}	Yahoo! confirms the IP address of the Yahoo! email with the attached document	Digital

3.3 CPT values for Sub-Hypothesis and Evidence in Yahoo! Case (Hong Kong)

Before we set up the Bayesian Network, we have to obtain the CPT values of sub-hypothesis and evidence. In this study, all of the probability values are assigned by subjective beliefs based on expert professional opinion and experience in digital forensic analysis [Kwan, 2011]. From Table-7, we can

see there are two states for hypothesis H_0 – “yes” and “no”, and also for sub-hypotheses H_1 to H_6 . For the CPT values between sub-hypothesis and evidence, there are still two states for each sub-hypothesis – “yes” and “no”, but three states for each evidence – “yes”, “no” and “uncertain”. State “uncertain” means the evidence cannot be concluded to be either positive (“yes”) or negative (“no”) after examination of the evidence.

Table 7 Likelihood value for sub-hypothesis H_1 to H_6 given hypothesis H

State	H_1, H_5, H_6		H_2, H_3, H_4	
	“yes”	“no”	“yes”	“no”
$H = \text{“yes”}$	0.65	0.35	0.8	0.2
$H = \text{“no”}$	0.35	0.65	0.2	0.8

Table 8 Conditional probability values for DE_1 to DE_3 given sub-hypothesis H_1

State	DE_1			DE_2, DE_3		
	yes	no	u	yes	no	u
$H_1 = \text{yes}$	0.85	0.15	0	0.8	0.15	0.05
$H_1 = \text{no}$	0.15	0.85	0	0.15	0.8	0.05

Table 9 Conditional probability values for DE_4 given sub-hypothesis H_2

State	DE_4		
	yes	no	u
$H_2 = \text{yes}$	0.75	0.2	0.05
$H_2 = \text{no}$	0.2	0.75	0.05

Table 10 Conditional probability values for DE_5 given sub-hypothesis H_3

State	DE_5		
	yes	no	u
$H_3 = \text{yes}$	0.7	0.25	0.05
$H_3 = \text{no}$	0.25	0.7	0.05

Table 11 Conditional probability values for DE_6 given sub-hypothesis H_4

State	DE_6		
	yes	no	u
$H_4 = \text{yes}$	0.1	0.85	0.05
$H_4 = \text{no}$	0.05	0.9	0.05

Table 12 Conditional probability values for DE_7 to DE_{11} given sub-hypothesis H_5

State	DE_7, DE_8, DE_{10}			DE_9, DE_{11}		
	yes	no	u	yes	no	u
$H_5 = \text{yes}$	0.7	0.25	0.05	0.8	0.15	0.05
$H_5 = \text{no}$	0.25	0.7	0.05	0.15	0.8	0.05

Table 13 Conditional probability values for DE_{12} to DE_{14} given sub-hypothesis H_6

State	DE_{12}, DE_{13}			DE_{14}		
	<i>yes</i>	<i>no</i>	<i>u</i>	<i>yes</i>	<i>no</i>	<i>u</i>
$H_6 = \text{yes}$	0.7	0.25	0.05	0.8	0.15	0.05
$H_6 = \text{no}$	0.25	0.7	0.05	0.15	0.8	0.05

4. LOCAL MULTI-PARAMETER SENSITIVITY VALUE

4.1 Conditional Independence

Before we proceed with the multi-parameter sensitivity value, we have to discuss the relationship between each sub-hypothesis and its set of evidences. In the Yahoo! case, the connections between sub-hypothesis and evidences belong to the class of diverging connections in a Bayesian network [Taroni, 2006] (see Figure 1). In a diverging connection model, E_1 and E_2 are conditionally independent given H . This means that: (1) with the knowledge of H , the state of E_1 does not change the belief about the possible states of E_2 . $P(E_1, E_2 | H) = P(E_1 | H)P(E_2 | H)$; (2) without the knowledge of H , the state of E_1 provides information about the possible states of E_2 .

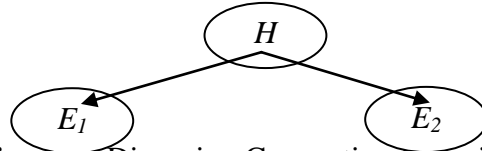


Figure 1: Diverging Connection Bayesian Network

In general, we cannot conclude that E_1 and E_2 are also conditionally independent given \bar{H} . However, for some special cases, if there are only two possible (mutually complementary) values for H , $H = \{H_1, H_2\}$, then given “ $H=H_1$ ”, E_1 and E_2 are conditionally independent, while, given “ $H=H_2$ ”, E_1 and E_2 are also conditionally independent. Hence we can conclude that E_1 and E_2 are also conditionally independent given \bar{H} .

4.2 Multi-Parameter Sensitivity Value

Under the standard assumption of proportional co-variation [Renooij, 2004] the theorems characterizing the algebraic structure of marginal probabilities [Castillo, 1997] permit the single-parameter sensitivity function of [Renooij, 2004] to be generalized to the multi-parameter case as follows:

If the parameters are given by $\mathbf{x} = (x_1, \dots, x_n)$ then the sensitivity function F is given by the multi-linear quotient:

$$F(\mathbf{x}) = \frac{a_0 + \sum a_i x_i + \sum_{i>j} a_{ij} x_i x_j + \dots + a_{12..n} x_1 x_2 \dots x_n}{b_0 + \sum b_i x_i + \sum_{i>j} b_{ij} x_i x_j + \dots + b_{12..n} x_1 x_2 \dots x_n} \equiv \frac{n(\mathbf{x})}{d(\mathbf{x})} \quad (1)$$

The components of the gradient vector are given by:

$$\nabla F_i = \frac{\partial F}{\partial x_i} = \frac{(a_i + \sum_{i>j} a_{ij} x_j + \dots + a_{12..n} x_1 \dots x_{i-1} x_{i+1} \dots x_n) \cdot d(\mathbf{x}) - (b_i + \sum_{i>j} b_{ij} x_j + \dots + b_{12..n} x_1 \dots x_{i-1} x_{i+1} \dots x_n) \cdot n(\mathbf{x})}{d(\mathbf{x})^2} \quad (2)$$

The multi-parameter sensitivity value is the value of the steepest gradient at the point \mathbf{x} and is given by the Euclidean norm of the gradient vector ∇F :

$$\|\nabla F\| = \sqrt{\sum_{i=1}^n \nabla F_i^2} \quad (3)$$

This last result follows from the first order Taylor expansion of F :

$$\delta F = F(x_1 + \delta x_1, \dots, x_n + \delta x_n) - F(x_1, \dots, x_n) = \frac{\partial F}{\partial x_1} \delta x_1 + \dots + \frac{\partial F}{\partial x_n} \delta x_n = \nabla F \cdot \delta \mathbf{x} \quad (4)$$

In principle the multi-linear forms in the numerator and the denominator of F are both of order n and contain 2^n terms leading to expressions for each ∇F_i containing 2^{2n} terms in both numerator and denominator. However, this analysis does not take into account the conditional dependencies implied by the structure of the BN. If the BN can be represented as a set of m sub-hypotheses H_j ($j=1, m$), each of which is conditionally dependent on a disjoint subset E_i ($i=1, n$) containing n_j ($j=1, m$) of the set of n evidential traces, so that $\sum_{j=1}^m n_j = n$, then the total number of parameters is given by $N = m + n$, but each H_j is conditionally dependent only upon its own subset of n_j evidential traces. This is known as the local Markov property of the BN. Although the conditional probabilities associated with sub-hypothesis H_j will influence those associated with sub-hypothesis H_k ($j \neq k$) via the process of Bayesian inference propagation through the network [Kwan, 2008], to a reasonable first approximation the sensitivity values for each sub-hypothesis may be evaluated by disregarding these indirect effects. Then, for the sub-hypotheses $F(\mathbf{x})$ is a multi-linear quotient of order m , while for the evidential traces associated with sub-hypothesis H_j $F(\mathbf{x})$ is a multi-linear quotient of order n_j . The Markov factorisation of the BN thus substantially reduces the number of terms involved in the expressions for $F(\mathbf{x})$ and ∇F . Nevertheless a symbolic algebraic manipulation program, such as MatLab [MathWorks, 2011], is required to perform the differentiations and computations reliably for a BN representing a real-world situation such as the Yahoo! email case.

In order to derive the coefficients of the multinomial quotient for $F(\mathbf{x})$ we proceed as follows. Bayes' formula for the likelihood conditional probability is:

$$P(E|H) = \frac{P(H|E) \cdot P(E)}{P(H)} \quad (5)$$

Considering the conjunctive combination of the evidential traces $\{E_1, E_2, \dots, E_n\}$, (5) transforms into:

$$P(E_1 \wedge E_2 \wedge \dots \wedge E_n | H) = \frac{P(H|E_1 \wedge E_2 \wedge \dots \wedge E_n) \cdot P(E_1 \wedge E_2 \wedge \dots \wedge E_n)}{P(H)} \quad (6)$$

Then we obtain the multi-parameter posterior probability as:

$$P(H|E_1 \wedge E_2 \wedge \dots \wedge E_n) = \frac{P(E_1 \wedge E_2 \wedge \dots \wedge E_n | H) \cdot P(H)}{P(E_1 \wedge E_2 \wedge \dots \wedge E_n)} \quad (7)$$

Here,

$$P(E_1 \wedge E_2 \wedge \dots \wedge E_n) = P(E_1 \wedge E_2 \wedge \dots \wedge E_n | H) \cdot P(H) + P(E_1 \wedge E_2 \wedge \dots \wedge E_n | \bar{H}) \cdot P(\bar{H}) \quad (8)$$

As mentioned in Section 4.1, given H, E_1, E_2, \dots, E_n are conditionally independent of each other. Therefore,

$$P(E_1 \wedge E_2 \wedge \dots \wedge E_n | H) = P(E_1 | H) \cdot P(E_2 | H) \cdot \dots \cdot P(E_n | H) \quad (9)$$

According to the definition of conditional independence, we cannot in general assume that $E_1, E_2, \dots,$

E_n are conditionally independent of each other under \bar{H} . However, in the specific Yahoo! email case, there are only two possible (mutually complementary) states for each hypothesis. Therefore, we also have:

$$P(E_1 \wedge E_2 \wedge \dots \wedge E_n | \bar{H}) = P(E_1 | \bar{H}) \cdot P(E_2 | \bar{H}) \cdot \dots \cdot P(E_n | \bar{H}) \quad (10)$$

Denoting $P(E_1 | H), P(E_2 | H), \dots, P(E_n | H)$ by x_1, x_2, \dots, x_n , and $P(E_1 | \bar{H}), P(E_2 | \bar{H}), \dots, P(E_n | \bar{H})$ by c_1, c_2, \dots, c_n , we obtain the multi-parameter Sensitivity Function as:

$$P(H | E_1 \wedge E_2 \wedge \dots \wedge E_n) = \frac{P(H) \prod_{i=1}^n x_i}{P(H) \prod_{i=1}^n x_i + P(\bar{H}) \prod_{i=1}^n c_i} \quad (11)$$

As an un-biased pre-condition, it is usual to take the prior probabilities as: $P(H) = P(\bar{H}) = 0.5$ [Kwan, 2011], so (11) simplifies to:

$$P(H | E_1 \wedge E_2 \wedge \dots \wedge E_n) = \frac{\prod_{i=1}^n x_i}{\prod_{i=1}^n x_i + \prod_{i=1}^n c_i} \quad (12)$$

When finding the sensitivity value of the posterior of the root node H_0 , it is necessary to use a variant of (3), namely, the weighted Euclidean norm of the sensitivity values of the individual sub-hypotheses:

$$\|\nabla F\| = \sqrt{\sum_{j=1}^m \|\nabla F_j\| \|\nabla F\|_j^2} \quad (13),$$

where $\|\nabla F_j\|$ is the multi-parameter sensitivity value of sub-hypothesis is the multi-parameter sensitivity value of sub-hypothesis j as given by (3), reflecting the different contributions of the individual sub-hypotheses to the posterior of H_0 .

5. RESULTS AND DISCUSSION

In Table 14 the multi-parameter sensitivity values for each sub-hypothesis of BN for the Yahoo! case [Kwan, 2011] are set out and compared with the corresponding single-parameter sensitivity values from [Kwan, 2011]. Before proceeding to discuss these results, however, it should be mentioned at this point that while validating the MatLab code for the multi-parameter sensitivity values, a number of numerical discrepancies were noted when reproducing the previously reported single-parameter sensitivity values ([Kwan, 2011], Table 4). All but one of these discrepancies are not numerically significant in terms of the conclusions drawn; however, in the case of digital evidential trace DE_6 the correct sensitivity value is actually 2.222, not 0.045, which implies that the effect of DE_6 on the posterior of H_4 should be significant. This revised result brings the single-parameter sensitivity value for DE_6 into line with the vertex proximity value for DE_6 ([Kwan, 2011], Table 5) which indicated that the posterior of H_4 is indeed sensitive to variations in DE_6 , thereby resolving the apparent disagreement between the two sensitivity metrics for the case of DE_6 noted in [Kwan, 2011]. The corrected results for the single-parameter sensitivity values are given in Table 14 below. However, a review of sub-hypothesis H_4 reveals that it is not critical to the prosecution case since its associated evidence DE_6 is only weakly tied into the case; the fact that Mr Shi registered with Yahoo! for a webmail account at some time in the past cannot be assigned a high probative value and this is reflected in the ‘non-diagonal’ structure of the corresponding CPT (see Table 11), unlike all the other CPTs in this BN.

It is reasonable to interpret the significance of a sensitivity value by comparing it with unity [Renooij, 2004; Kwan, 2011]; a value below unity implies a lack of sensitivity to small changes in the associated conditional probability parameters, and *vice versa*. In other words, a sensitivity value less than unity implies that the response of the BN is smaller than the applied perturbation. It cannot be assumed that single-parameter sensitivity values as computed in [Kwan, 2011] are necessarily either smaller or

larger than the corresponding multi-parameter sensitivity values computed here, since the form of the sensitivity function being used is not identical. In the Yahoo! email case, three of the sub-hypotheses, namely H_2 , H_3 and H_4 , are associated with only a single evidential trace so each of their sensitivity values is unchanged in the multi-parameter analysis. Of the three remaining sub-hypotheses, H_1 and H_6 both have three associated evidential traces whereas H_5 has five.

Table 14: Single- and multi-parameter sensitivity values for $H_1 - H_6$ of the Yahoo! email case

Sub-hypothesis	Digital Evidence	Single-parameter Sensitivity value	Component j $\ \nabla F\ _j$ of Multi-parameter Sensitivity value	Multi-parameter Sensitivity value
H_1	DE ₁	0.1500	0.0125	0.0225
	DE ₂	0.1662	0.0134	
	DE ₃	0.1662	0.0134	
H_2	DE ₄	0.2216	0.2216	0.2216
H_3	DE ₅	0.2770	0.2770	0.2770
H_4	DE ₆	2.2222	2.2222	2.2222
H_5	DE ₇	0.2770	0.0051	0.0110
	DE ₈	0.2770	0.0051	
	DE ₉	0.1662	0.0045	
	DE ₁₀	0.2770	0.0051	
	DE ₁₁	0.1662	0.0045	
H_6	DE ₁₂	0.2770	0.0565	0.0939
	DE ₁₃	0.2770	0.0565	
	DE ₁₄	0.1662	0.0494	

From Table 14 it will be noted that, with the exception of the somewhat anomalous case of H_4 discussed earlier, the largest multi-parameter sensitivity value is an order of magnitude smaller than the smallest of the single-parameter sensitivity values reported previously [Kwan, 2011]. This finding is at first sight somewhat surprising in that it might be expected that permitting many parameters to vary simultaneously would produce the opportunity for greater sensitivity values to be found. However, it must be remembered that, unlike the case of classical numerical optimization, the parameters of the BN are not completely independent due to the conditional independence referred to in Section 4.1 as well as the interdependence produced by the propagation of belief (posterior probabilities) through the BN [Kwan, 2008]. In certain circumstances, these forms of co-variance can result in smaller multi-parameter sensitivity values than might otherwise have been anticipated, as explained below.

In Table 15 we give the single- and multi-parameter sensitivity values for the root node H_0 . The multi-parameter sensitivity value is computed using the weighted Euclidean norm given in (12).

Table 15: Single- and multi-parameter sensitivity values for H_0 of the Yahoo! email case

Root hypothesis	Sub-hypothesis	Single-parameter Sensitivity value	Component $\ \nabla F\ _j$ of Multi-parameter Sensitivity value	Weight of Component $\ \nabla F\ _j$	Multi-parameter Sensitivity value
H_0	H_1	0.3500	0.0064	0.0225	0.0052
	H_2	0.2000	0.0030	0.2216	
	H_3	0.2000	0.0030	0.2770	
	H_4	0.2000	0.0030	2.2222	
	H_5	0.3500	0.0037	0.0110	
	H_6	0.3500	0.0037	0.0939	

The multi-parameter sensitivity results in Tables 14 and 15 raise a number of important issues for discussion. Firstly, it appears that a multi-parameter sensitivity analysis should be performed in every case in which a BN is used to reason about digital evidence. Since a multi-parameter sensitivity analysis can often yield substantially different sensitivity values from a single-parameter sensitivity analysis, it is necessary to assess the sensitivity of the posterior output of the BN as rigorously as possible before putting forward forensic conclusions based on that BN. In the present Yahoo! email case we found consistently smaller multi-parameter sensitivity values than the corresponding single-parameter results, but further investigations have shown that this is by no means a universal trend. By creating CPTs with different ratios of ‘diagonal’ to ‘off-diagonal’ values we have been able to explore the circumstances under which the multi-parameter sensitivity analysis is likely to yield larger values than the corresponding single-parameter analysis. Table 16 offers a few such examples. In particular we find that when the structure of one or more CPT tables deviates significantly from the typical ‘diagonal’ form, as was the case with H_4 -DE₆ here, then the single- and multi-parameter sensitivity values will increase towards and quite possibly exceed unity. This observation can be understood as follows: A typically ‘diagonal’ CPT signifies that the truth or falsehood of the sub-hypothesis strongly predicts the presence or absence of the associated evidence; they are logically tightly-coupled, which is a highly desirable property. An anomalously ‘non-diagonal’ CPT, however, signals that there is little correlation between the truth of the sub-hypothesis and the presence or absence of the evidence. In other words, the evidence is a very poor indicator of the truth of the sub-hypothesis, and does not discriminate effectively. This manifests itself as a steep gradient on the associated CPT parameters’ hyper-surface indicating the direction of a ‘better’ choice of parameters. Similarly, a combination of ‘diagonal’ and ‘non-diagonal’ CPTs associated with the same sub-hypothesis can create a numerical tension or balance between the belief reinforcing effects propagated by the former and the belief weakening effects propagated by the latter, resulting in an increased sensitivity value.

Secondly, the above observations lead us to propose that single- and multi-parameter sensitivity analyses can be effectively employed to test an existing BN for logical consistency between its sub-hypotheses and their associated evidences, through the CPT values. If the single- or multi-parameter sensitivity analysis results suggest that the BN’s posterior output is sensitive to the values of one of more of the BN’s CPTs, it is necessary to review those CPT values critically with a view to possibly revising them, after which the single- or multi-parameter sensitivity analysis should be repeated. It should be noted, however, that this is not equivalent to optimizing the posterior output BN with

respect to the CPT values. Rather, it is using the sensitivity analysis to highlight possible issues in the process by which the original CPT values were generated. If repeated reviews of the CPT values do not lead to a stable posterior output from the BN, this would suggest that the structure of the BN or of the underlying sub-hypotheses and their associated evidential traces concerning the way in which the digital crime was committed, may be faulty and require revision. We recall that this was the case with H4-DE6 in the present study. Note, in particular, that the single-parameter sensitivity results in Table 16 would not necessarily cause concern, while the corresponding multi-parameter results clearly indicate a serious sensitivity problem. Hence we contend that in general single-parameter sensitivity analyses are not sufficient in and of themselves and we recommend that multi-parameter sensitivity analyses should be undertaken as a matter of course.

A final, and more general, consideration is whether there are any special characteristics of digital forensic analysis which would dictate that the requirements of a sensitivity analysis should differ from traditional forensics. Because digital forensics is a much more recent discipline than most of traditional forensics there has been less time for it to establish a corpus of verified knowledge and a ‘track record’ of demonstrably sound methodologies. This means that digital forensics needs to strive to establish itself as a mature scientific and engineering discipline, and routinely performing sensitivity analyses is one means of helping to achieve this goal. Indeed, it may result in the analytical methods of traditional forensics being required to proceed in a similar fashion in order to avoid the imputation of ‘reasonable doubt’ by wily defence lawyers.

Table 16: Some examples of other CPT values yielding large multi-parameter sensitivity values

Sub-hypothesis	Digital Evidence	$P(E H)$	$P(E \bar{H})$	Single-parameter Sensitivity value	Multi-parameter Sensitivity value
H_1	DE ₁	0.7	0.3	0.300	2.4448
	DE ₂	0.3	0.7	0.300	
	DE ₃	0.3	0.65	0.7202	
	DE ₄	0.7	0.25	0.2770	
H_2	DE ₅	0.6	0.3	0.3704	1.8680
	DE ₆	0.3	0.65	0.7202	
	DE ₇	0.4	0.55	0.6094	

6. CONCLUSIONS

In this paper we have described a multi-parameter sensitivity analysis on the BN from the Yahoo! email case. . As a result of the present analysis, it can be concluded that one the Yahoo! email sub-hypotheses (H_4) exhibits a significant degree of single- and multi-parameter sensitivity and hence its associated evidence and conditional probabilities should be reviewed. That review indicated that while the CPT was in fact quite accurate, it represented a poor choice of evidence and sub-hypothesis which should either be discarded or revised.

More generally, we have shown that the definition and computation of local multi-parameter sensitivity values is made feasible for BNs describing real-world digital crimes by the use of a symbolic algebra system such as Matlab [MathWorks, 2011] to evaluate the steepest gradient analytically.

Finally, we should reiterate that the principal aim of this work is to devise a metric to assess the

instantaneous (or local) stability of a BN with respect to the values chosen to populate its CPTs. The value of such a metric is that it enables the digital forensic examiner to know whether or not the set of conditional probabilities chosen to populate the CPTs of the BN lies on a flat or a steep part of the hyper-surface in parameter space. The problem of attempting to find local (or global) optima on the parametric hyper-surface of a BN is a separate issue which has been addressed elsewhere [CD04].

ACKNOWLEDGEMENT

The work described in this paper was partially supported by the General Research Fund from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. RGC GRF HKU 713009E), the NSFC/RGC Joint Research Scheme (Project No. N_HKU 722/09), HKU Seed Fundings for Applied Research 201102160014, and HKU Seed Fundings for Basic Research 201011159162 and 200911159149.

REFERENCES

- Castillo, E., Gutierrez, J.M. and Hadi, A.S. (1997), "Sensitivity Analysis in Discrete Bayesian Networks", *IEEE Trans. Systems, Man & Cybernetics, Pt. A*, 27 (4) 412 – 423.
- Chan, H. and Darwiche, A., (2004) 'Sensitivity Analysis in Bayesian Networks: From Single to Multiple Parameters', 20th Conference. on Uncertainty in Artificial Intelligence, July 7-11, Banff, Canada
- Report published under Section 48(2) of the Personal Data (Privacy) Ordinance (Cap.486) (2007), issued by Office of the Privacy Commissioner for Personal Data, Hong Kong, http://www.pcpd.org.hk/english/publications/files/Yahoo_e.pdf (accessed on 14 March 2007)
- First Instance Reasons for Verdict of the Changsha Intermediate People's Court of Hunan Province (2005), delivered by the Changsha Intermediate Criminal Division One Court, in Case No. 29 of 2005, China. http://lawprofessors.typepad.com/china_law_prof_blog/files/ShiTao_verdict.pdf (accessed on 9 October 2010).
- Kwan, M., Chow, K., Law, F. and Lai, P., (2008) 'Reasoning about evidence using Bayesian networks', *Advances in Digital Forensics IV*, Springer, Boston, Mass., pp.275-289.
- Kwan, Y.K., Overill, R.E., Chow, K.-P., Silomon, J.A.M., Tse, H., Law, Y.W. and Lai, K.Y., (2010) 'Evaluation of Evidence in Internet Auction Fraud Investigations', *Advances in Digital Forensics VI*, Springer, Boston, Mass., pp.121-132,
- Kwan, M., Overill, R., Chow, K.-P., Tse, H., Law, F. and Lai, P., (2011) 'Sensitivity Analysis of Digital Forensic Reasoning in Bayesian Network Models', *Advances in Digital Forensics VII*, Springer, Boston, Mass., pp.231-244.
- MathWorks (2011) MatLab: <http://www.mathworks.co.uk/> (accessed on 11 November 2011).
- Overill, R.E., Silomon, J.A.M., Kwan, Y.K., Chow, K.-P., Law, Y.W. and Lai, K.Y. (2010), 'Sensitivity Analysis of a Bayesian Network for Reasoning about Digital Forensic Evidence', 4th International Workshop on Forensics for Future Generation Communication Environments, August 11-13, Cebu, Philippines.
- Renooij, S. and van der Gaag, L.C. (2004), 'Evidence-invariant Sensitivity Bounds'. 20th Conference on Uncertainty in Artificial Intelligence, July 7-11, Banff, Canada.
- Taroni, F., Aitken, C., Garbolino, P., and Biedermann, A. (2006), *Bayesian Network and Probabilistic Inference in Forensic Science*, John Wiley & Sons Ltd., Chichester, UK.