Chapter 18

ANALYSIS OF THE DIGITAL EVIDENCE PRESENTED IN THE YAHOO! CASE

Michael Kwan, Kam-Pui Chow, Pierre Lai, Frank Law and Hayson Tse

Abstract The "Yahoo! Case" led to considerable debate about whether or not an IP address is personal data as defined by the Personal Data (Privacy) Ordinance (Chapter 486) of the Laws of Hong Kong. This paper discusses the digital evidence presented in the Yahoo! Case and evaluates the impact of the IP address on the verdict in the case. A Bayesian network is used to quantify the evidentiary strengths of hypotheses in the case and to reason about the evidence. The results demonstrate that the evidence about the IP address was significant to obtaining a conviction in the case.

Keywords: Yahoo! Case, digital evidence, Bayesian network, reasoning

1. Introduction

Scientific conclusions based on evidence have been used for many years in forensic investigations. In making their assessments, investigators consider the available facts and the likelihood that they support or refute hypotheses related to a case. Investigators recognize that there is never absolute certainty and seek a degree of confidence with which to establish their hypotheses [2].

A forensic investigation determines the likelihood of a crime through the analysis and interpretation of evidence. To this end, a forensic investigation focuses on the validation of hypotheses based on the evidence and the evaluation of the likelihood that the hypotheses support legal arguments [6, 10, 12, 14, 15]. The likelihood represents the degree of belief in the truth of the associated hypothesis. It is typically expressed as a probability and probabilistic methods may be used to deduce the likelihood of a hypothesis based on the available evidence [7, 9].

G. Peterson and S. Shenoi (Eds.): Advances in Digital Forensics V, IFIP AICT 306, pp. 241–252, 2009.

A crime and its associated digital evidence are usually linked by sub-hypotheses. This paper uses a Bayesian network [8] to analyze and reason about the evidence in the well-known Yahoo! Case [3].

In the Yahoo! Case, Yahoo! Holdings (Hong Kong) Limited (Yahoo! HHKL) supplied IP address information to Chinese authorities that led to the conviction of Shi Tao, a Chinese journalist, for sending confidential state information to foreign entities. Shi Tao received a 10-year sentence for his crime.

Shi Tao's authorized representative in Hong Kong subsequently lodged a complaint with the Office of the Privacy Commissioner for Personal Data. The complaint maintained that Yahoo! HHKL disclosed Shi Tao's "personal data" to Chinese authorities, which was a breach of the Hong Kong Personal Data (Privacy) Ordinance.

The investigation by the Privacy Commissioner concluded that an IP address, on its own, does not constitute personal data [13]. The conclusion was based on the position that an IP address is unique to a specific computer not a person and, therefore, does not meet the definition of personal data. The Privacy Commissioner also held that no safe conclusion could be drawn that user data corresponding to the IP address belonged to a living individual as opposed to a corporate or unincorporated body, or that it was related to a real as opposed to a fictitious individual.

We use Bayesian network inference to assess the evidentiary weight of the IP address in the Yahoo! Case. Four scenarios are evaluated:

- Yahoo! HHKL and the ISP participate in the investigation; all the digital evidence is available.
- Yahoo! HHKL participates in the investigation; digital evidence regarding the IP address is received from Yahoo! HHKL. However, the ISP does not participate in the investigation.
- Yahoo! HHKL does not participate in the investigation; digital evidence regarding the IP address is not received from Yahoo! HHKL. However, the ISP participates in the investigation.
- Yahoo! HHKL and the ISP do not participate in the investigation; no digital evidence regarding the IP address is available.

Although an IP address, by itself, is not viewed as personal data, our analysis shows that it carried significant evidentiary weight in the Yahoo! Case. Our analysis is based on the "Reasons for Conviction" [4], and the Administrative Appeals Board decision [1] regarding the Report of the Hong Kong Privacy Commissioner published under Section 48(2) of the Personal Data (Privacy) Ordinance (Chapter 486) [13].

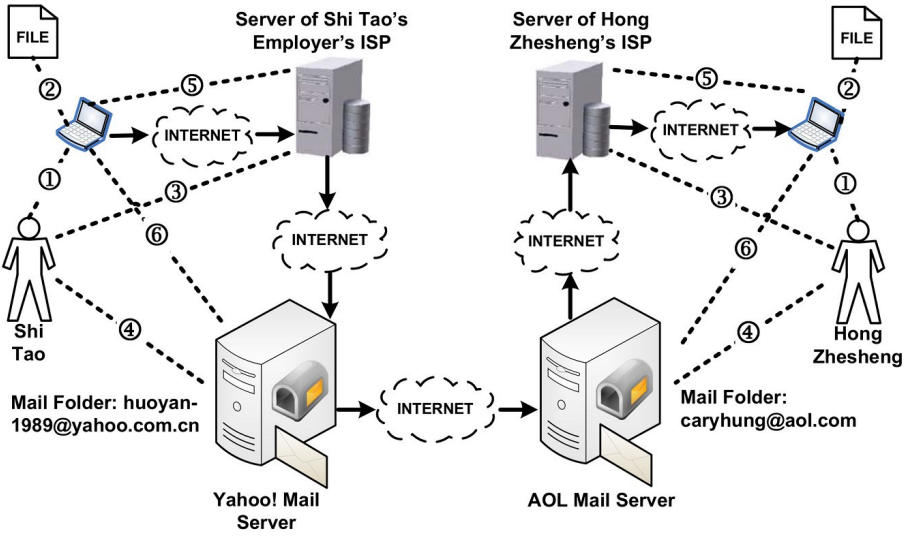

Figure 1. Entities and events in the Yahoo! Case.

2. Digital Evidence in the Yahoo! Case

In the Yahoo! Case, the Changsha Intermediate People's Court of Hunan Province convicted Shi Tao of providing state secrets to foreign entities. Based on the data provided by Yahoo! HHKL, the court determined that at approximately 11:32 pm on April 20, 2004, Shi Tao used a computer in his employer's office to access his personal email account (`huoyan-1989@yahoo.com.cn`) via the Yahoo! webmail interface and send some notes regarding a summary of a top-secret document issued by the Chinese Government to the email account of Hong Zhesheng (`caryhung@aol.com`) [13]. Shi Tao asked Hong Zhesheng, who resided in New York, to find a way to distribute the notes as quickly as possible without using Shi Tao's name [5].

Figure 1 shows the entities and events involved in the email transmission from Shi Tao to Hong Zhesheng. Based on this description, a digital forensic investigator would be required to ascertain the following facts:

1. Shi Tao had access to a computer connected to the Internet.
2. A copy of the electronic file was stored on the computer.
3. The computer had a web browser program.
4. To obtain Internet access, Shi Tao established a connection between the computer and the ISP. In this case, he used the dial-up account belonging to his employer. The ISP authenticated the account of Shi Tao's employer and assigned an IP address to Shi Tao's computer. Shi Tao's computer recorded the assigned IP address and used it for subsequent Internet access. Internet data originating from or destined to Shi Tao's computer went through the ISP.
5. Shi Tao launched the web browser program and entered the Yahoo! webmail URL in the browser window.
6. The web browser program sent an HTTP request to the Yahoo! mail server. When the requested web page was retrieved, it was displayed by the web browser program.
7. Shi Tao entered his user name and password to log into his email account. Based on the email subscription details, the Yahoo! mail server authenticated Shi Tao and allowed him to log into his email folder.
8. Shi Tao composed the email, attached the file and entered Hong Zhesheng's AOL email address. He then clicked the "Send" button to transmit the email along with the file attachment. Since Shi Tao used a web browser program to create the email, the email content was (possibly) cached in Shi Tao's computer.
9. The Yahoo! email server stored the email and the attachment, and placed it in the message queue for transmission to Hong Zhesheng's AOL email server via SMTP.

3. Evaluation of Digital Evidence

In general, an investigation must clarify a number of issues before a case can be brought to court. These issues include whether or not a crime was committed, how the crime was committed, who committed the crime and whether or not there is a reasonable chance of conviction.

We use a Bayesian network to quantify the evidentiary strengths of hypotheses and to reason about evidence. A Bayesian network is a directed acyclic graph whose edges indicate dependencies between nodes.

Each node is accompanied by a conditional probability table (CPT) that describes the dependencies between nodes. In our work, the nodes correspond to hypotheses and the digital evidence associated with hypotheses. The edges connect each hypothesis to the evidence that should be present if the hypothesis is valid.

4. Bayesian Network

The first step in constructing a Bayesian network for analyzing digital evidence in the Yahoo! Case involves the definition of the primary hypothesis (H), the main issue to be determined. In the Yahoo! Case, the primary hypothesis is: "The seized computer was used to send the material document as an email attachment using a Yahoo! webmail account."

The next step is to define the possible states of the hypothesis (Yes, No and Uncertain). Probability values are then assigned to each state. Each of these values represents the prior probability that the hypothesis is in the specific state. The prior probability of H, $P(H)$, is assumed to be equal to (0.333, 0.333, 0.333), i.e., all three states are equally likely.

The hypothesis H is the root node in the Bayesian network. Sub-hypotheses that are causally dependent on the hypothesis assist in proving the hypothesis. The sub-hypotheses and the associated evidence and events are represented as child nodes in the Bayesian network.

Figure 1 lists six sub-hypotheses that support the primary hypothesis H in the Yahoo! Case. The six sub-hypotheses are:

- $\boldsymbol{H_1}$: Linkage between the material document and the suspect's computer (Table 1).
- $\boldsymbol{H_2}$: Linkage between the suspect and the computer (Table 2).
- $\boldsymbol{H_3}$: Linkage between the suspect and the ISP (Table 3).
- $\boldsymbol{H_4}$: Linkage between the suspect and the Yahoo! email account (Table 4).
- $\boldsymbol{H_5}$: Linkage between the computer and the ISP (Table 5).
- $\boldsymbol{H_6}$: Linkage between the computer and the Yahoo! email account (Table 6).

The evidence and events for the six sub-hypotheses are listed in Tables 1–6.

The states of the various sub-hypotheses are dependent on the state of H. Each sub-hypothesis, which is a child node of H, has an associated conditional probability table (CPT). The CPT contains the prior

Table 1. H_1: Linkage between the material document and the suspect's computer.

ID	Evidence Description	Type
DE1	Subject file exists on the computer	Digital
DE2	Last access time of the subject file is after the IP address assignment time by the ISP	Digital
DE3	Last access time of the subject file matches or is close to the sent time of the Yahoo! email	Digital

Table 2. H_2: Linkage between the suspect and the computer.

ID	Evidence Description	Type
PE1	Suspect was in physical possession of the computer	Physical
DE4	Files on the computer reveal the identity of the suspect	Digital

Table 3. H_3: Linkage between the suspect and the ISP.

ID	Evidence Description	Type
DE5	ISP subscription details match the suspect's particulars	Digital

Table 4. H_4: Linkage between the suspect and the Yahoo! email account.

ID	Evidence Description	Type
DE6	Subscription details of the Yahoo! email account match the suspect's particulars	Digital

Table 5. H_5: Linkage between the computer and the ISP.

ID	Evidence Description	Type
DE7	Configuration settings for the ISP Internet account are found on the computer	Digital
DE8	Log data confirms that the computer was powered up at the time the email was sent	Digital
DE9	Web program or email user agent program was found to be activated at the time the email was sent	Digital
DE10	Log data reveals the assigned IP address and the assignment time by the ISP to the computer	Digital
DE11	Assignment of the IP address to the suspect's account is confirmed by the ISP	Digital

Table 6. H_6: Linkage between the computer and the Yahoo! email account.

ID	Evidence Description	Type
DE12	Internet history logs reveal that the Yahoo! email account was accessed by the computer	Digital
DE13	Internet cache files reveal that the subject file was sent as an attachment via the Yahoo! email account	Digital
DE14	Yahoo! confirms the IP address of the Yahoo! email with the attached document	Digital

probabilities of the sub-hypothesis based on the state of the hypothesis. The probability values are typically assigned by digital forensic experts based on their subjective beliefs.

Table 7. Conditional probabilities of $H_1 \ldots H_6$.

	$H_1 \ldots H_6$		
H	**Yes**	**No**	**Uncertain**
Yes	0.60	0.35	0.05
No	0.35	0.60	0.05
Uncertain	0.05	0.05	0.90

We assume that all the sub-hypotheses ($H_1 \ldots H_6$) have the CPT values shown in Table 7. For example, an initial value of 0.6 is assigned for the situation where H and H_1 are Yes. This means that when the seized computer was used to send the material document as an email attachment using a Yahoo! webmail account, the probability that a linkage existed between the material document and the seized computer is 0.6. Additionally, there may be instances where it is not possible to confirm a Yes or No state for H_1 from the evidence although the seized computer was used to send the document. This uncertainty is modeled by assigning a probability of 0.05 to the Uncertain state.

After assigning conditional probabilities to the sub-hypotheses, the observable evidence and events related to each sub-hypothesis are added to the Bayesian network. For reasons of space, we only discuss Hypothesis H_1 (Linkage between the material document and the seized computer) in detail to demonstrate the use of a Bayesian network.

The evidence for H_1 that establishes the linkage between the material document and the seized computer includes: (i) the subject file exists on the computer; (ii) the last access time of the subject file is after the IP address assignment time by the ISP; and (iii) the last access time of

Table 8. Conditional probabilities of E_1, E_2, E_3.

	E_1			E_2			E_3		
H_1	**Y**	**N**	**U**	**Y**	**N**	**U**	**Y**	**N**	**U**
Y	0.85	0.15	0	0.85	0.15	0	0.85	0.12	0.03
N	0.15	0.85	0	0.15	0.85	0	0.12	0.85	0.03
U	0	0	1	0	0	1	0.03	0.03	0.94

the subject file matches or is close to the sent time of the Yahoo! email. Each node has the states: Yes (Y), No (N) and Uncertain (U).

The next step is to assign conditional probability values to the evidence. Table 8 shows the conditional probability values of evidence E_1, E_2 and E_3 given specific states of Hypothesis H_1.

After conditional probabilities are assigned to the entailing evidence, it is possible to propagate probabilities within the Bayesian network. In particular, the likelihood of H_1 is computed based on the observed probability values of evidence E_1, E_2 and E_3. The well-known MSBNX program [11] was used to propagate probabilities in the Bayesian network developed for the Yahoo! Case.

If evidence E_1, E_2 and E_3 have Yes states, then the digital forensic investigator can confirm that there is a likelihood of 99.6% that Hypothesis H_1 (Linkage between the material document and the suspect's computer) is true. Furthermore, based on the 99.6% likelihood for H_1, the investigator can also conclude that there is a 59.9% likelihood that H (The seized computer was used to send the material document as an email attachment using a Yahoo! webmail account) is true. Figure 2 shows the Bayesian network when E_1, E_2 and E_3 all have Yes states.

The same methodology is used to compute the likelihoods of the other five sub-hypotheses based on the probability values of the associated evidentiary nodes. Finally, the likelihoods of the six sub-hypotheses are used to compute the overall likelihood of the primary hypothesis.

5. Impact of the IP Address

In order to assess the evidentiary weight of the IP address in the Yahoo! Case, we identify four scenarios that involve differing amounts of evidence provided to the Chinese authorities by Yahoo! HHKL and the ISP.

- **Scenario 1:** In this scenario, Yahoo! HHKL and the ISP participate in the investigation. When all the evidence (DE1–DE14

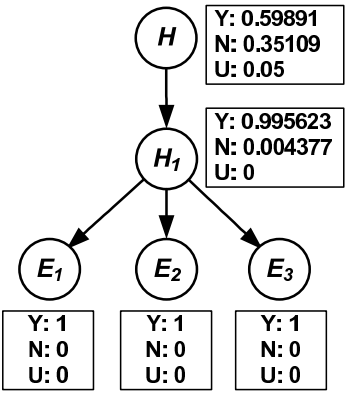

Figure 2. Probability distributions with E_1, E_2, E_3 = Yes.

and PE1) in Tables 1–6 is available and is true, the likelihood of Hypothesis H is 90.5%.

- **Scenario 2:** In this scenario, the ISP does not participate in the investigation. The evidentiary items DE5 (Table 3) and DE11 (Table 5) are missing. The corresponding likelihood of Hypothesis H is 88.1%.

- **Scenario 3:** In this scenario, Yahoo! HHKL does not participate in the investigation. The evidentiary items DE6 (Table 4) and DE14 (Table 6) are missing. The corresponding likelihood of Hypothesis H is 83.0%.

- **Scenario 4:** In this scenario, Yahoo! HHKL and the ISP do not participate in the investigation. Evidentiary items DE5 (Table 3), DE6 (Table 4), DE11 (Table 5) and DE14 (DE14) are missing. The corresponding likelihood of Hypothesis H is 78.7%.

Table 9 lists the four scenarios and their likelihoods. Note that the availability of the IP address affects the likelihood by 11.7%. In particular, the likelihood is 90.5% (very likely) when all the evidence is available, but it drops to 78.7% (probable) when evidence related to the IP address is not available. Although the IP address by itself does not reveal the identity of a specific user, it provides additional information that can further confirm the identity of the user.

The Reasons for Verdict [5] in the Yahoo! Case identified six primary facts:

- **Fact 1:** Shi Tao attended the press briefing and obtained the information.

Table 9. Likelihood of Hypothesis H.

Scenario	Likelihood
Scenario 1: Yahoo! HHKL and the ISP participate in the investigation	90.5%
Scenario 2: Yahoo! HHKL participates in the investigation and confirms the IP address of the Yahoo! email with the attached document; the ISP does not participate in the investigation	88.1%
Scenario 3: Yahoo! HHKL does not participate in the investigation; the ISP participates in the investigation	83.0%
Scenario 4: Yahoo! HHKL and the ISP do not participate in the investigation	78.7%

- **Fact 2:** Shi Tao was present in the office of his employer at the material time.
- **Fact 3:** Shi Tao was the only employee who knew the information.
- **Fact 4:** The office of the employer was the registration address for the IP address.
- **Fact 5:** The IP address was assigned to the employer at the time the email was sent.
- **Fact 6:** The email was sent from the material IP address.

We developed a Bayesian network modeling these facts to evaluate the hypothesis: "Shi Tao sent the material email at the material time from the office of his employer." Experiments with the Bayesian network indicate that when all six facts are completely supported, the likelihood of Hypothesis H is 99.9%. However, when the IP address is missing (i.e., Facts 4–6 relating to the IP address are Uncertain), the overall likelihood drops to 14.9%, a reduction of 85.0%. This drop underscores the importance of the IP address in obtaining a conviction in the Yahoo! Case.

6. Conclusions

Bayesian networks provide a powerful mechanism for quantifying the evidentiary strengths of investigative hypotheses and reasoning about evidence. The application of a Bayesian network to analyze digital evidence related to the Yahoo! Case demonstrates that the IP address was significant to obtaining a conviction. Investigators and prosecutors can use this technique very effectively to evaluate the impact of specific evidentiary items before a case is brought to court.

References

[1] Administrative Appeals Board, Shi Tao v. The Privacy Commissioner for Personal Data, Administrative Appeal No. 16 of 2007, Hong Kong (www.pcpd.org.hk/english/publications/files/Appeal_Yahoo.pdf), 2007.

[2] C. Aitken and F. Taroni, *Statistics and the Evaluation of Evidence for Forensic Scientists*, John Wiley and Sons, New York, 2004.

[3] R. Bascuas, Property and probable cause: The Fourth Amendment's principled protection of privacy, *Rutgers Law Review*, vol. 60(3), pp. 575–645, 2008.

[4] Changsha Intermediate People's Court of Hunan Province, Criminal Verdict, First Trial Case No. 29, Changsha Intermediate Criminal Division One Court, Changsha, China (www.globalvoicesonline.org/wp-content/ShiTao_verdict.pdf), 2005.

[5] Changsha Intermediate People's Court of Hunan Province, Reasons for Verdict, First Trial Case No. 29, Changsha Intermediate Criminal Division One Court, Changsha, China (www.pcpd.org.hk/english/publications/files/Yahoo_annex.pdf), 2005.

[6] R. Cook, I. Evett, G. Jackson, P. Jones and J. Lambert, A model for case assessment and interpretation, *Science and Justice*, vol. 38, pp. 151–156, 1998.

[7] P. Dawid, Statistics and the Law, Research Report No. 224, Department of Statistical Science, University College London, London, United Kingdom, 2004.

[8] F. Jensen, *An Introduction to Bayesian Networks*, Springer-Verlag, New York, 1996.

[9] 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.

[10] R. Loui, J. Norman, J. Altepeter, D. Pinkard, D. Craven, J. Linsday and M. Foltz, Progress on Room 5: A testbed for public interactive semi-formal legal argumentation, *Proceedings of the Sixth International Conference on Artificial Intelligence and Law*, pp. 207–214, 1997.

[11] Microsoft Research, MSBNx: Bayesian Network Editor and Tool Kit, Microsoft Corporation, Redmond, Washington (research.microsoft.com/adapt/MSBNx).

[12] J. Mortera, A. Dawid and S. Lauritzen, Probabilistic expert systems for DNA mixture profiling, *Theoretical Population Biology*, vol. 63(3), pp. 191–206, 2003.

[13] Office of the Privacy Commissioner for Personal Data, Report Published under Section 48(2) of the Personal Data (Privacy) Ordinance (Chapter 486), Report No. R07-3619, Hong Kong (www.pcpd.org.hk/english/publications/files/Yahoo_e.pdf), 2007.

[14] H. Prakken, C. Reed and D. Walton, Argumentation schemes and generalizations in reasoning about evidence, *Proceedings of the Ninth International Conference on Artificial Intelligence and Law*, pp. 32–41, 2003.

[15] D. Walton, Argumentation and theory of evidence, in *New Trends in Criminal Investigation and Evidence – Volume II*, C. Breur, M. Kommer, J. Nijboer and J. Reijntjes (Eds.), Intersentia, Antwerp, Belgium, pp. 711–732, 2000.