



May 30th, 3:20 PM

## Cloud Forensics Investigation: Tracing Infringing Sharing of Copyrighted Content in Cloud

Yi-Jun He

*Department of Computer Science, The University of Hong Kong, yjhe@cs.hku.hk*

Echo P. Zhang

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

Lucas C.K. Hui


*Department of Computer Science, The University of Hong Kong, hui@cs.hku.hk*

Siu Ming Yiu

*Department of Computer Science, The University of Hong Kong, smyi@cs.hku.hk*

**K.P. Chow** and additional works at: <https://commons.erau.edu/adfsl>

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

 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

He, Yi-Jun; Zhang, Echo P.; Hui, Lucas C.K.; Yiu, Siu Ming; and Chow, K.P., "Cloud Forensics Investigation: Tracing Infringing Sharing of Copyrighted Content in Cloud" (2012). *Annual ADFSL Conference on Digital Forensics, Security and Law*. 13.

<https://commons.erau.edu/adfsl/2012/wednesday/13>

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](mailto:commons@erau.edu).

**EMBRY-RIDDLE**  
Aeronautical University™  
SCHOLARLY COMMONS

(c)ADFSL



# CLOUD FORENSICS INVESTIGATION: TRACING INFRINGING SHARING OF COPYRIGHTED CONTENT IN CLOUD

Yi-Jun He, Echo P. Zhang, Lucas C.K. Hui, Siu Ming Yiu, K.P. Chow

Department of Computer Science, The University of Hong Kong

Phone: +852-22415725; Fax: +852-25598447;

E-mail: {yjhe, pzhang2, hui, smyi, chow}@cs.hku.hk

## ABSTRACT

Cloud Computing is becoming a significant technology trend nowadays, but its abrupt rise also creates a brand new front for cybercrime investigation with various challenges. One of the challenges is to track down infringing sharing of copyrighted content in cloud. To solve this problem, we study a typical type of content sharing technologies in cloud computing, analyze the challenges that the new technologies bring to forensics, formalize a procedure to get digital evidences and obtain analytical results based on the evidences to track down illegal uploader. Furthermore, we propose a reasoning model based on the probability distribution in a Bayesian Network to evaluate the analytical result of forensics examinations. The proposed method can accurately and scientifically track down the origin infringing content uploader and owner.

**Keywords:** cloud forensics, peer to peer, file sharing, tracking, CloudFront

## 1. INTRODUCTION

With broadband Internet connection and with P2P programs such as Gnutella, FrostWire, BearShare, BitTorrent, and eMule, it takes very little effort for someone to download songs, movies, or computer games. But nowadays, people make use of it to share copyrighted files, or even images of child sexual exploitation. Since October 2009, over 300,000 unique installations of Gnutella have been observed sharing known child pornography in the US. Thus, many research works have been focused on criminal investigations of the trafficking of digital contraband on Peer-to-Peer (P2P) file sharing networks (Chow *et al.* 2009, Jeong *et al.* 2009, Jeong *et al.* 2010, Liberatore *et al.* 2010).

Recently, cloud computing has been dramatically developed and will soon become the dominant IT environment. It is a new computing platform where thousands of users around the world can access to a shared pool of computing resources (*e.g.*, storage, applications, services, *etc.*) without having to download or install everything on their own computers, and only requires a minimal management effort from the service provider. Several big service providers, Amazon, Apple, Google, *etc.* start to provide content sharing services in cloud which can offer the file sharing functionalities like what P2P networks can offer. With the cloud environment, file sharing becomes more convenient and efficient, since sharing can be done through web browser without requiring software installation, and cloud provides strong computation power and fast transmitting speed. Thus it is possible that cloud based content sharing will substitute the existing file sharing programs one day.

### 1.1 Can Existing Investigative Models be Applied to Cloud Computing?

Before cloud computing emerging, most forensics investigation models are built on P2P networks. When cloud based infringing content sharing happens, can existing investigative model for analyzing P2P network be applied to analyze cloud based file sharing network? The answer is NO, because cloud content sharing systems differentiate from P2P sharing systems in the following aspects:

1. Cloud Computing provides storage capacity at dedicate, and data is automatically geographically dispersed on edge servers on the cloud; while P2P file sharing systems support the trading of storage capacity between a network of 'peers'.

2. Typically, Cloud file sharing systems operate in a centralized fashion that end user is directed to an edge server that is near them to get data based on a complex load balancing process; but P2P services operate in a decentralized way that nodes on the Internet cooperate in an overlay network consisting of peers, where each peer is both a consumer and a producer of data and gets different piece of data from other peers.
3. Further, cloud file sharing services are paid services, where the user pays, for instance, per amount of data kept in storage and the amount of network traffic generated to upload or download files. Contrary to cloud storage services, P2P file sharing systems are typically not subscription based, but rather depend on group members that are part of a peer network to trade resources, primarily disk capacity and network bandwidth.

## 1.2 Contributions

This is the first paper providing accurate investigations of such content sharing networks in cloud. We analyze the functionality a typical cloud content sharing system: CloudFront, formalize an investigation procedure, and build an analysis model on it. Our research can help investigators:

1. Confidently state from where and how various forms of evidences are acquired in cloud;
2. Understand the relative strength of each evidence;
3. Validate that evidences from the fruits of a search warrant;
4. Assess the accuracy of the investigation result based on the evidences obtained using a Bayesian network.

In section 2, we give an overview of related works. In section 3, we introduce the background of the content sharing system. In section 4, we simulate the crime, and describe the investigation process. In section 5, we propose a Bayesian Network based model to obtain analytical results based on the evidences. In section 6, we analyze the proposed model in several aspects. Finally, we conclude the whole paper.

## 2. RELATED WORK

Many works have been proposed to solve security and privacy problems in the cloud field (Angin *et al.* 2010, Bertino *et al.* 2009, He *et al.* 2012). They focus on aspects for protecting user privacy and anonymous authentication. Works (Wang *et al.* 2010, Wang *et al.* 2011) improve data security in cloud using crypto technologies.

Also, traditional computer forensics have already had many works published on how to establish principles and guidelines to retrieve digital evidence (Grance *et al.* 2006). Works (Aitken *et al.* 2004, Kwan *et al.* 2007) show how to formulate hypotheses from evidence and evaluate the hypotheses' likelihood for the sake of legal arguments in court. In addition, the aspects of forensics in tracking infringing files sharing in P2P networks have been addressed by several works (Chow *et al.* 2009, Jeong *et al.* 2009, Jeong *et al.* 2010, Liberatore *et al.* 2010).

In contrast with the maturity of research on cloud security and privacy and traditional computer forensics, the research on forensic investigations in cloud computing are relatively immature. To the best of knowledge, our paper is the first work addressing the problem of tracking infringing sharing of copyrighted content in the cloud. There are other works (Birk *et al.* 2011, Marty 2011, Zafarullah *et al.* 2011) discussing the issues of forensic investigations in cloud, but (Marty 2011, Zafarullah *et al.* 2011) focus on how to log the data needed for forensic investigations, (Birk *et al.* 2011) gives an overview on forensic investigations issues without providing concrete solutions or investigation model.

## 3. BACKGROUND

In this section, we provide a technical overview of a content sharing system: CloudFront, which is a typical cloud based content sharing network.

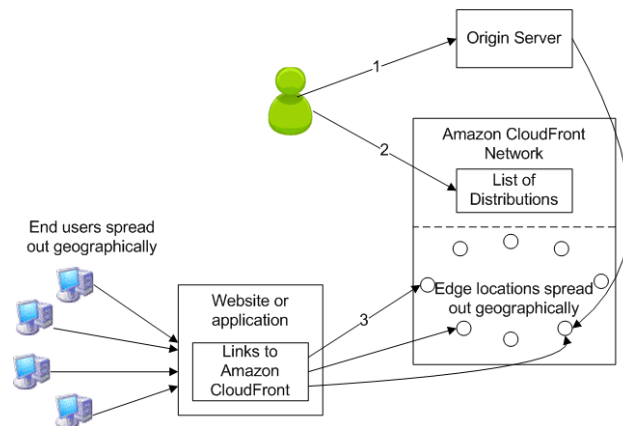
Amazon CloudFront (Amazon 2012) is a web service for content delivery. It delivers your content through a worldwide network of edge locations. End users are routed to the nearest edge location, so content is delivered with the best possible performance.

A CloudFront network is made up of four types of entities:

- *Objects* are the files that the file owner wants CloudFront to deliver. This typically includes web pages, images, and digital media files, but can be anything that can be served over HTTP or a version of RTMP.
- *Origin Server* is the location where you store the original, definitive version of your objects.
- *Distribution* is a link between your origin server and a domain name that CloudFront automatically assigns. If your origin is an Amazon S3 bucket, you use this new domain name in place of standard Amazon S3 references. For example, `http://mybucket.s3.amazonaws.com/image.jpg` would instead be `http://somedomainname.cloudfront.net/image.jpg`.
- *Edge location* is a geographical site where CloudFront caches copies of your objects. When an end user requests one of your objects, CloudFront decides which edge location is best able to serve the request. If the edge location doesn't have a copy, CloudFront goes to the origin server and puts a copy of the object in the edge location.

To share a file using CloudFront, the file owner first makes and publishes a distribution. The process is shown in Figure 1 and it includes:

1. Register an account on the origin server, and place objects in the origin server and make them publicly readable.
2. Create CloudFront distribution and get the distribution's domain name that CloudFront assigns. Example distribution ID: EDFDVBD632BHDS5, and Example domain name: `d604721fxaaqy9.cloudfront.net`. The distribution ID will not necessarily match the domain name.
3. Create the URLs that end users will use to get the objects and include them as needed in any web application or website. Example URL: `http://d604721fxaaqy9.cloudfront.net/videos/video.mp4`.



**Fig. 1.** The working protocol of Amazon CloudFront.

To download a shared file from CloudFront, the end user needs to do the following steps. For simplicity, we assume the end user resides in Hong Kong.

1. After clicking the URL from the web application or website, CloudFront determines which edge location would be best to serve the object. In this case, it is the Hong Kong location.
2. If the Hong Kong edge location doesn't have a copy of *video.mp4*, CloudFront goes to the origin server and puts a copy of *video.mp4* in the Hong Kong edge location.
3. The Hong Kong edge location then serves *video.mp4* to the end user and then serves any other requests for that file at the Hong Kong location.
4. Later, *video.mp4* expires, and CloudFront deletes *video.mp4* from the Hong Kong location. CloudFront doesn't put a new copy of *video.mp4* in the Hong Kong location until an end user requests *video.mp4* again and CloudFront determines the Hong Kong location should serve the image.

#### 4. INVESTIGATION PROCESS

The objective of investigation is to obtain evidences through observation of data from the Internet and other possible parties, such as service providers or seized devices. In this section, we discuss techniques and methods for collecting evidences from Amazon CloudFront.

As the cloud based content sharing is new, we do not have a real case in Hong Kong to study. Thus we have to suppose there is a crime and simulate the whole crime process based on the most common and regular criminal behaviors, and find out what evidences the criminal may leave. Our paper gives a good guidance for collecting evidences if a real case happens in the future.

In this part, we simulate a suspect intends to share infringing files using Amazon CloudFront. A general case is that the suspect has a movie in his computer and wants to share the movie publicly. The suspect needs to do following steps:

1. Subscribe to Amazon CloudFront service, providing email address, user name and credit card information to service provider.
2. Register an origin server, providing email address and user name to the server administrator.
3. Upload infringing files from local disk to the origin server. This step may involve installing a FTP/FTPS/SFTP client software in order to do the uploading.
4. Login Amazon CloudFront to create a distribution for the infringing file. The distribution is the URL link to the file in the CloudFront network.
5. Register a forum, and publish the distribution to the forum. When end users click the distribution, CloudFront will retrieve the files from the nearest edge location.

Once such an illegal sharing happens, we can follow the guidance below to track the suspect:

First, we can trace the suspect file link uploader's IP address through four steps ( $E_i$  represents evidence  $i$ ):

1.  $E_1$ : The suspect posted the distribution, the posted message is found.
2.  $E_2$ : The suspect has a forum account and he is logged in. So, the suspect's forum account is found.
3.  $E_3$ : The IP address must be recorded by the forum. Check with the forum administrator for the IP address of the user who created the posts.
4.  $E_4$ : Check with Internet Service Providers for the assignment record of IP address to get its subscribed user.

Through the above four steps, we are only sure that the suspect has posted a link on the forum, but not sure about whether it is the same suspect who created the link. Thus the second step is to check with

the CloudFront provider for four issues:

1.  $E_5$ : The suspect has an Amazon CloudFront account, and logged in the CloudFront with the tracked IP address.
2.  $E_6$ : The origin server domain name is found under the suspect's CloudFront account.
3.  $E_7$ : The infringing file distribution creation record is found under the suspect's CloudFront account.
4.  $E_8$ : The registered credit card holder of that CloudFront account is the suspect.

The third step is to check with the origin server administrator that the suspect is the infringing file owner. The evidences include

1.  $E_9$ : The suspect has an origin server account, and logged in the origin server with the tracked IP address.
2.  $E_{10}$ : The infringing file exists under the suspect origin server account.

The last step is to find out the devices that the suspect used to do the infringing file sharing. The evidences include

1.  $E_{11}$ : Hash value of the infringing file matches that of the file existing on the suspect's devices (including Tablet PC, Laptop, Desktop Computer or Mobile Phone).
2.  $E_{12}$ : A FTP/FTPS/SFTP client software is installed on the devices.
3.  $E_{13}$ : Origin server connection record is found in FTP/FTPS/SFTP client software.
4.  $E_{14}$ : Infringing file transmission record is found in FTP/FTPS/SFTP client software.
5.  $E_{15}$ : Cookie of the Origin server is found on the devices.
6.  $E_{16}$ : Internet history record on Origin server is available.
7.  $E_{17}$ : URL of Origin server is stored in the web browser.
8.  $E_{18}$ : The distribution origin name is the origin server.
9.  $E_{19}$ : Credit card charge record of CloudFront is found.
10.  $E_{20}$ : Cookie of the CloudFront website is found.
11.  $E_{21}$ : CloudFront service log-in record is found.
12.  $E_{22}$ : Distribution creation record is found.
13.  $E_{23}$ : Removing the file in origin server will affect the distribution validity.
14.  $E_{24}$ : Web browser software is available on the devices.
15.  $E_{25}$ : Internet connection is available.
16.  $E_{26}$ : Internet history record on publishing forum is found.
17.  $E_{27}$ : The distribution link posted on the forum is as the same as what is created in CloudFront.
18.  $E_{28}$ : Cookie of the publishing forum is found.

## 5. THE PROPOSED MODEL FOR ANALYZING EVIDENCES

In this part, we construct a Bayesian model to analyze the evidences found above and calculate the probability of guilt. We begin the construction with the set up of the top-most hypothesis, which is

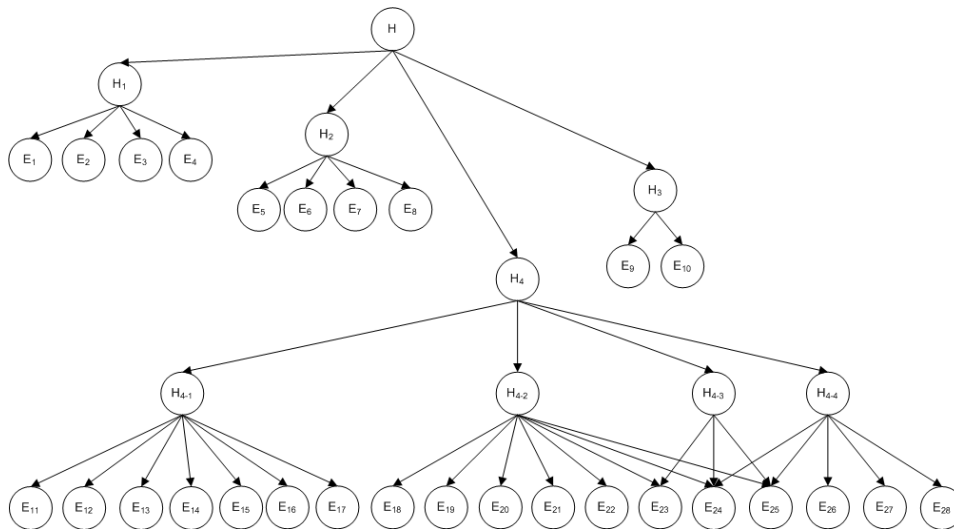
Hypothesis  $H$ : "The suspect is the origin file owner and the uploader"

Usually, this hypothesis represents the main argument that the investigator wants to determine. It is the root node. It is an ancestor of every other node in the Bayesian Network, hence its state's probabilities are unconditional. To support the root hypothesis, sub-hypotheses may be added to the Bayesian Network, since they are useful for adjusting the model to a more clearly structured graph. As show in Table 1, in the CloudFront model, four sub-hypotheses are created for the root node and four sub-sub-hypotheses are created for the hypothesis  $H_4$ .

**Table 1.** Hypotheses with CloudFront

$H_1$	The suspect posted a link to the forums.
$H_2$	The suspect created the link using CloudFront.
$H_3$	The suspect is the file owner.
$H_4$	The seized devices (including Tablet PC, Laptop, Desktop Computer, Mobile Phone) have been used as the initial sharing machine to share infringing file on Internet.
$H_{4-1}$	Has the pirated file uploaded from the seized devices to the origin server?
$H_{4-2}$	Has the distribution on the origin server been created using CloudFront?
$H_{4-3}$	Has the connection between the seized devices and the CloudFront been maintained?
$H_{4-4}$	Has the distribution been posted to newsgroup forum for publishing?

The built Bayesian model is shown in Figure 2:



**Fig. 2.** Bayesian Network Diagram for Amazon CloudFront.

### 5.1 How To Use the Model for Assessment

**Initialization** Take the Bayesian network model for Amazon CloudFront as an example. The possible states of all evidences are: “Yes, No, Uncertain”. All hypotheses are set to be “Unobserved” when there is no observation made to any evidence. The initial prior probability  $P(H)$  is set to be Yes: 0.3333, No: 0.3333, Uncertain: 0.3333. The initial conditional probability value of each evidence is set as shown in Figure 3. Take  $E_9$  as an example, we assign an initial value of 0.85 for the situation when  $H_3$  and  $E_9$  are both “Yes”. That means when the suspect is the file owner, the chance that the suspect has an origin server account, and logged in the origin server with the tracked IP address is 85%. The resulting posterior probability  $P(H_i)$  and  $P(H_{ij})$ , that is the certainty of  $H_i$  and  $H_{ij}$  based on the initialized probability values of evidences, should be evenly distributed amongst their states, as show in Table 2.

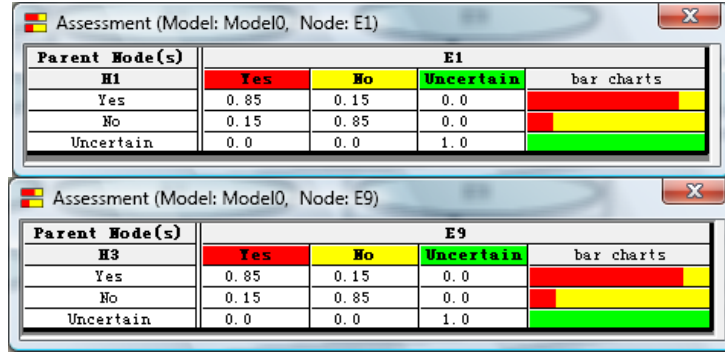


Fig. 3. The Initial Conditional Probabilities.

Table 2. Bayesian Network Initial Posterior Probability

Hypotheses	Initial Posterior Probability
$H_1$	Yes: 0.3333, No: 0.3333, Uncertain: 0.3333
$H_2$	Yes: 0.3333, No: 0.3333, Uncertain: 0.3333
$H_3$	Yes: 0.3333, No: 0.3333, Uncertain: 0.3333
$H_4$	Yes: 0.3333, No: 0.3333, Uncertain: 0.3333
$H_{4-1}$	Yes: 0.3333, No: 0.3333, Uncertain: 0.3333
$H_{4-2}$	Yes: 0.3333, No: 0.3333, Uncertain: 0.3333
$H_{4-3}$	Yes: 0.3333, No: 0.3333, Uncertain: 0.3333
$H_{4-4}$	Yes: 0.3333, No: 0.3333, Uncertain: 0.3333

**Assessment** In the investigation process, if an evidence  $E_i$  is found, then the state of that evidence should be changed to “Yes”, and the prior probability  $P(E_i)$  should be set to 1. On the other hand, if the evidence is not found, then the state of that evidence should be changed to “No” or “Uncertain”. If “No”, the prior probability of that evidence should be 0; if “Uncertain”, the prior probability of that evidence should be set subjectively between 0-1. If we assume all evidences are found, and switch all the entailing evidences to state “Yes”, the propagated probability values of the hypotheses are shown in Table 3. According to the Bayesian Network calculation, the posterior probability of  $H$  at state “Yes” reaches the highest value 99.8819% under this circumstance, which means that there is a maximum chance of 99.8819% that the suspect is the origin file owner and uploader.

Table 3. Propagated Probability Values of the Hypothesis

Hypotheses	Posterior probability when all evidences are found
$H$	Yes: 0.998819, No: 0.00118103, Uncertain: 0
$H_1$	Yes: 0.999823, No: 0.000177368, Uncertain: 0
$H_2$	Yes: 0.999823, No: 0.000177368, Uncertain: 0
$H_3$	Yes: 0.994364, No: 0.00563629, Uncertain: 0
$H_4$	Yes: 0.998967, No: 0.00103306, Uncertain: 0
$H_{4-1}$	Yes: 0.999999, No: 9.7077e-007, Uncertain: 0
$H_{4-2}$	Yes: 0.999994, No: 5.50123e-006, Uncertain: 0
$H_{4-3}$	Yes: 0.849277, No: 0.150723, Uncertain: 0
$H_{4-4}$	Yes: 0.999025, No: 0.000975118, Uncertain: 0



However, in reality, some evidences may not be found, so we should correspondingly amend the absent evidences states from “Yes” to “No”. For example, we assume the evidences  $E_{17}$ ,  $E_{10}$ ,  $E_7$  are not found, thus we change the states of them from “Yes” to “No”. As a result, the posterior probability of  $H$  at state “Yes” will be reduced from 99.8819% to 99.3379%. The result shows that the posterior probability of  $H$  would be reduced if some evidences are absent.

We used the software “MSBNx” (Microsoft, 2001) to calculate the probability of the hypotheses. The above analysis result of the root hypothesis tells the judge the probability that a hypothesis is true. In our experiment, if all evidences are found, the probability would be more than 99%. The numerical result is a good scientific reference to the judge. If real cases happen, investigators can use this model to evaluate the digital forensics findings, and adjust the evidences found, thus get a quantitative probability of the crime.

## 6. ANALYSIS OF THE MODEL

Though Bayesian Network has been used for a while in security and forensics, our analysis below shows that its construction and usage in cloud has its own characteristics.

### 6.1 Difficulties in Building the Model

Following the guidances in section 4, it is not difficult to find out the origin file link uploader and file owner. However, the nature of the cloud causes some difficulties when following our guidances to do investigation.

- First, in the cloud environment, file sharing is centrally controlled. Inevitably, investigators need assistance from cloud service providers (SPs), such as Amazon, in order to lock-in the suspect fast. However, many of the cloud SPs are international organizations, so there are a number of restrictions placed on connecting evidences from foreign organizations. Some of these restrictions are the decision of the foreign Government, while others are the result of international organizations being unwilling to leak customers information. Thus, one may wonder if the cloud SPs do not collaborate with forensics investigators, does the model still work?
- Fortunately, the following evaluation result of our model shows that, it is still possible to track the suspect with a high probability even without the help from the cloud SP. If without the help of cloud SP,  $E_5...E_8$  would be absent. Thus, we set the states of  $E_5...E_8$  to be “No”, and keep the states of other evidences to be “Yes”. As a result,  $P(H_2)$  is dramatically reduced to 0.53%, and  $P(H)$  is reduced from 99.8819% to 96.38%. The result shows that, though the cloud SP is an important third party to provide valuable evidences, its absence would not affect much of the final result when other evidences are all found.
- The second difficulty is that the suspect can use mobile phones or other persons’ computers to upload the files. Also, cloud computing allows using web based technology, which makes content sharing easy from any device that supports a web browser. As a result, investigators may not be able to get any evidence if just investigate suspect’s personal computers, or may miss some important evidences existing in other devices. Thus it expands the scope of the investigation. For example, when finding the evidences supporting  $H_4$  in our model, investigators must investigate all devices with browsers, including Tablet PC, laptop, desktop computer, mobile phone.

## 6.2 Sensitive Data Analysis

From the model, we can find some evidences which will have the greatest impact or the minimal impact to the result, and some evidences which have the most interconnection with other evidences.

1. As show in Figure 3, if  $P(E_i/H_j)$  which is the initial posterior probability of each evidence  $E_i$  caused by the hypothesis  $H_j$  is the same, for example,  $P(E_1/H_1) = P(E_9/H_3) = 0.85$ , then  $E_9$  and  $E_{10}$  would have the most significant effect to the posterior probability of  $H$ . For example, if we change the state of  $E_9$  from “Yes” to “No”, the  $P(H)$  will be reduced from 99.8819% to 99.43%. If changing any other evidence such as  $E_1$ , the  $P(H)$  will be reduced from 99.8819% to 99.86%.
2. If adopting the same initial posterior probability above, then  $E_{24}$  and  $E_{25}$  have the minimal impact to  $P(H)$ . If we change the state of  $E_{24}$  from “Yes” to “No”, the  $P(H)$  will be reduced from 99.8819% to 99.8818%.
3. The hypothesis  $H_4$  is in a diverging connection with  $H_{4-1}$ ,  $H_{4-2}$ ,  $H_{4-3}$ ,  $H_{4-4}$ , and  $H_{4-1}$  is in a diverging connection with  $E_{11}...E_{17}$ , hence, provided the states of  $H_{4-1}$  and  $H_4$  are unobservable, change in  $E_{11}...E_{17}$  will also change the probability values of  $H_{4-1}$  and  $H_4$ . When  $H_{4-1}$  or  $H_4$  changes, the likelihood of  $E_{11}...E_{17}$  will change also.
4. Similarly, since  $H$ ,  $H_4$ ,  $H_{4-1}$  and  $E_{11}$  are in serial connection, hence change in  $E_{11}$  will also propagate the variation to  $H$  if  $H_4$  and  $H_{4-1}$  remain unobservable.
5. Nodes  $E_{24}$  and  $E_{25}$  are common nodes for hypotheses  $H_{4-2}$ ,  $H_{4-3}$ ,  $H_{4-4}$ . In other words, there is a converging connection to  $E_{24}$  and  $E_{25}$  from hypothesis nodes  $H_{4-2}$ ,  $H_{4-3}$ ,  $H_{4-4}$ . According to the rules of probability propagation for converging connection in Bayesian network, when the states of  $E_{24}$  and  $E_{25}$  are known, the probabilities of  $H_{4-2}$ ,  $H_{4-3}$ ,  $H_{4-4}$  will influence each other. Therefore, change in the state of  $E_{24}$  or  $E_{25}$  will change the probability of these three hypotheses. Further, since  $H_{4-1}$ ,  $H_{4-2}$ ,  $H_{4-3}$ ,  $H_{4-4}$  are in divergent connection with parent hypothesis  $H_4$ , hence changes in  $H_{4-2}$ ,  $H_{4-3}$ ,  $H_{4-4}$  will also influence the probability of  $H_{4-1}$ .

## 6.3 Initial Probability Assignment

As show in Table 2 and Figure 3, for simplicity, in the initialization phase of assessment, we set the initial prior probability  $P(H)$  to be Yes: 0.3333, No: 0.3333, Uncertain: 0.3333, and set the initial posterior probability of each evidence  $E_i$  caused by the hypothesis  $H_j$  to be the same, such as  $P(E_1/H_1) = P(E_9/H_3) = 0.85$ . However, in reality, the assignment of initial prior probability and initial posterior probability may not be like this. Such assignments often rely on subjective personal belief which is affected by professional experience and knowledge. Also individual digital forensic examiner’s belief may not represent the general and acceptable view in the forensic discipline. Thus, the assignment needs to be done among a group of forensic specialists. In order to help investigators to perform a more accurate assignment of the initial probability, we first classify the evidences into two levels in section 6.4.

## 6.4 Critical Evidence Set of Evidences

The investigation found out 28 evidences. Each of them will affect  $P(H)$  in varying degrees. Thus we classify the evidences into two levels, L0, L1 due to the degree of importance to  $H$ . L0 is the lowest degree and L1 is the highest degree. The higher the more important. Please note that such classification needs intensive understanding of each evidence, and it needs to be done by a group of forensic specialists.

- L1:  $E_1, E_3, E_5, E_7, E_{10}, E_{11}, E_{18}, E_{22}, E_{23}, E_{27}$
- L0:  $E_2, E_4, E_6, E_8, E_9, E_{12}, E_{13}, E_{14}, E_{15}, E_{16}, E_{17}, E_{19}, E_{24}, E_{20}, E_{21}, E_{25}, E_{26}, E_{28}$

According to the critical set classification, we define the following 7 deductions, which helps investigators to understand the logic relation among the evidences, and provides a basic knowledge to investigators when assigning initial posterior probabilities to each evidence.

1. If  $E_1$  and  $E_3$  are found,  $H_1$  would have a high probability to be true, no matter  $E_2$  or  $E_4$  is found or not.
2. If  $E_5$  and  $E_7$  are found,  $H_2$  would have a high probability to be true, no matter  $E_6$  or  $E_8$  is found or not.
3. If  $E_{10}$  is found,  $H_3$  would have a high probability to be true, no matter  $E_9$  is found or not.
4. If  $E_{11}$  is found,  $H_{4-1}$  would have a high probability to be true, no matter  $E_{12}$ ,  $E_{13}$ ,  $E_{14}$ ,  $E_{15}$ ,  $E_{16}$  or  $E_{17}$  is found or not.
5. If  $E_{18}$ ,  $E_{22}$ ,  $E_{23}$  are found,  $H_{4-2}$  would have a high probability to be true, no matter  $E_{19}$ ,  $E_{20}$ ,  $E_{21}$ ,  $E_{24}$  or  $E_{25}$  is found or not.
6. If  $E_{23}$  is found,  $H_{4-3}$  would have a high probability to be true, no matter  $E_{24}$  or  $E_{25}$  is found or not.
7. If  $E_{27}$  is found,  $H_{4-4}$  would have a high probability to be true, no matter  $E_{24}$ ,  $E_{25}$ ,  $E_{26}$ , or  $E_{28}$  is found or not.

Take deduction 1 as an example, it means that if the posted distribution and the poster’s IP address are found, it is most likely that the suspect posted a distribution with this IP address. The existence of  $E_2$  or  $E_4$  will not affect much of the probability of  $H_1$ . The basis of making such a deduction is that some forums support anonymous posting or unregistered user posting, so  $E_2$  may not exist even if  $E_1$  and  $E_3$  are found; also the IP subscriber may not be the suspect because he can use public network for posting, so  $E_4$  may not exist. Thus  $E_2$  and  $E_4$  are just supplementary evidences for  $H_1$ , but not the critical ones. Finally, according to deduction 1, investigators should assign higher initial probabilities to  $E_1$  and  $E_3$ , and lower posterior probabilities to  $E_2$  and  $E_4$ . The exact posterior probabilities should be carefully decided among specialists. Here, we just give an example to demonstrate the importance of initialization.

Parent Node(s)	E1				Parent Node(s)	E2			
H1	Yes	No	Uncertain	bar charts	H1	Yes	No	Uncertain	bar charts
Yes	0.85	0.15	0.0		Yes	0.15	0.15	0.0	
No	0.15	0.85	0.0		No	0.15	0.85	0.0	
Uncertain	0.0	0.0	1.0		Uncertain	0.0	0.0	1.0	

**Fig. 4.** The Different Initial Posterior Probability to Elements in L1 and L0.

**Example:** We assign each L1 element the same posterior probabilities as  $E_1$  show in Figure 4, and assign each L0 element the same posterior probabilities as  $E_2$  show in Figure 4. To prove deduction 1, we set four situations as show in Table 4. We found that if  $E_1$  and  $E_3$  are found, and  $E_2$  and  $E_4$  are not found,  $P(H_1)$  is 0.9498, which is not much different from situation 1; but if  $E_1$  and  $E_3$  are not found, and  $E_2$  and  $E_4$  are found,  $P(H_1)$  is just 0.0935. Thus, it proves the deduction 1. Similarly, if we assume all evidences in L1 are found and all evidences in L0 are not found,  $P(H)$  is 0.9935, which is still a high value; but if all evidences in L1 are not found and all evidences in L0 are found,  $P(H)$  is only 0.0937. Thus it proves the importance of L1 evidences.

**Table 4.** The Impact of Critical Data

Situations	$E_1, E_3$	$E_2, E_4$	$P(H_1)$
1	Yes	Yes	0.9907
2	No	No	0.0180
3	Yes	No	0.9498
4	No	Yes	0.0935

### 6.5 Other Cloud Content Sharing Networks

Actually there exist many other cloud content sharing networks, such as Seagate FreeAgent GoFLEX Net Media Sharing (Engines), which represents the technologies that are hardware assisted, and

supports transforming user's storage device into personal cloud storage. In such GoFlex network, there is no such origin server in CloudFront, but investigators need to investigate the user's storage device such as USB storage instead, because origin files exist in the USB storage. Other than that, investigators can follow investigation process of CloudFront to find other evidences in GoFlex, and build the Bayesian Network model.

## **7. CONCLUSION**

Performing forensic to crime based on cloud content sharing network is new. We analyzed a typical cloud content sharing network, and proposed guidances to track the origin file uploader and owner if illegal sharing happens in such network. A Bayesian Network model is also built (section 5) for analyzing the collected evidences to obtain a scientific evaluation result of the probability of a crime. Analyses of model construction difficulties, initialization, sensitive data and critical data set are done. One interesting result we found is that though the cloud SP is an important third party for providing valuable evidences, its absence would not affect much of the probability of tracking the suspect when other evidences are all found. If following the proposed guidances, there would be a chance of more than 99% to track the origin file uploader and owner.

## **ACKNOWLEDGEMENTS**

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**

- C.G.G. Aitken and F. Taroni (2004), *Statistics and the evaluation of evidence for forensic scientists*, John Wiley.
- Amazon (2012), 'Amazon cloudfront', <http://aws.amazon.com/cloudfront/>, Accessed in April, 2012.
- P. Angin, B. K. Bhargava, R. Ranchal, N. Singh, M. Linderman, L. B. Othmane, and L. Lilien (2010). 'An Entity-Centric Approach for Privacy and Identity Management in Cloud Computing'. 29<sup>th</sup> IEEE Symposium on Reliable Distributed Systems (SRDS). October 31 - November 3. New Delhi, Punjab, India.
- E. Bertino, F. Paci, R. Ferrini, and N. Shang (2009), "Privacy-preserving Digital Identity Management for Cloud Computing", *IEEE Data Engineering Bulletin*, Vol 32(Issue 1): Page 21–27.
- D. Birk and C. Wegener (2011), 'Technical issues of forensic investigations in cloud computing environments'. *IEEE Sixth International Workshop on Systematic Approaches to Digital Forensic Engineering (SADFE)*. May 26. Oakland, CA, USA.
- Yi-Jun He, Sherman S. M. Chow, Lucas C.K. Hui, S.M. Yiu (2012), 'SPICE – Simple Privacy-Preserving Identity-Management for Cloud Environment', 10<sup>th</sup> International Conference on Applied Cryptography and Network Security (ACNS), June, Singapore.
- K. P. Chow, R. Jeong, M. Kwan, P. Lai, F. Law, H. Tse, and K. Tse (2009), 'Security analysis of foxy peer-to-peer file sharing tool'. *HKU Technical Report TR-2008-09*.
- C. Engines. 'Pogoplug', <http://www.pogoplug.com/>. Accessed in April, 2012.
- R. S. C. Jeong, P. K. Y. Lai, K. P. Chow, M. Y. K. Kwan, and F. Y. W. Law (2010). 'Identifying first seeders in foxy peer-to-peer networks. 6<sup>th</sup> IFIP International Conference on Digital Forensics. January 4-6. Hong Kong, China.

- R. S. C. Ieong, P. K. Y. Lai, K. P. Chow, F. Y. W. Law, M. Y. K. Kwan, and K. Tse (2009). 'A model for foxy peer-to-peer network investigations'. 5<sup>th</sup> IFIP WG 11.9 International Conference on Digital Forensics. January 26-28. Orlando, Florida, USA.
- M. Y. Kwan, K. Chow, F. Y. Law, and P. K. Lai (2007), 'Computer forensics using Bayesian network: A case study'. <http://www.cs.hku.hk/research/techreps/document/TR-2007-12.pdf>.
- M. Liberatore, B. N. Levine, and C. Shields (2010). 'Strengthening forensic investigations of child pornography on P2P networks'. Proceedings of the 2010 ACM Conference on Emerging Networking Experiments and Technology (CoNEXT). November 30 - December 03. Philadelphia, PA, USA.
- R. Marty (2011). 'Cloud application logging for forensics'. 26th ACM Symposium On Applied Computing (SAC). March 21 – 24. TaiChung, Taiwan.
- Microsoft (2001). 'Microsoft bayesian network editor'. <http://research.microsoft.com/en-us/um/redmond/groups/adapt/msbnx/>. Accessed in April, 2012.
- K. K. T. Grance, S. Chevalier and H. Dang (2006), 'Guide to computer and network data analysis: Applying forensic techniques to incident response'. National Institute of Standards and Technology.
- C. Wang, Q. Wang, K. Ren, and W. Lou (2010). 'Privacy-preserving public auditing for data storage security in cloud computing'. The 29<sup>th</sup> IEEE International Conference on Computer Communications (INFOCOM). March 15-19. San Diego, CA, USA.
- Q. Wang, C. Wang, K. Ren, W. Lou, and J. Li (2011), 'Enabling public auditability and data dynamics for storage security in cloud computing'. IEEE Transactions on Parallel and Distributed Systems, 22(5): Page 847–859.
- Zafarullah, F. Anwar, and Z. Anwar (2011). 'Digital forensics for eucalyptus'. 9<sup>th</sup> International Conference on Frontier of Information Technology (FIT). December 19-21. Islamabad, Pakistan.