6th International Symposium on the Fusion of Science and Technologies (ISFT2017)
Jeju, S. KOREA 17th ~ 21st July, 2017

EMAIL HEADER ANALYSIS FOR AUTHOR IDENTIFICATION

Changhee Choi[1)*], Hwaseong Lee[1)], Ilhoon Jung[1)], Changon Yoo[1)], Hosang Yun[1)]
[1)] *Agency for Defense Development, Yuseong, Daejeon, Republic of Korea*

ABSTRACT: Undoubtedly, cyber-attacks are emerging as critical problems in modern society. Especially, phishing or scam using email has increased every year. It is helpful for tracking the origin of the attack that extracting valuable information for author identification. In previous research, email headers rarely used because they had only little information. In addition, they are easy to manipulate for a plaintiff or defendant.
In case of author identification of cyber incident, however, there is no reason to forge the email header. Under this assumption, we investigate the email database and extract the key information for author identification. Our analysis shows that there is more tracking information in the email header than we expected.

1. INTRODUCTION

Despite the remarkable development of Social Network Service(SNS), email is one of the most important communication tools in business, personal use, and military. Focusing on this point, Advanced Persistent Threat(APT) authors often use email as initial attack vectors such as spear-phishing. Duqu, one of top level APT attack, also uses email as a means of the first attack and infects the victims [1]. In the private sector, it is important to defend APT attack, but in military field, it is also important to find information about the author of a malicious email. There are many evidences such as EML file, document file, PE file, network logs, and system logs for tracking the origin of an attack.

Previous researches about APT attacker tracking are based on advanced analysis of the specialist. Although it is generally accurate, however, it is post analysis. Valuable information has already been destroyed by anti-forensics techniques. The specialist track the attacker based on little evidence which the attacker accidentally drops. In order to track the attacker using the maximum amount of information, it is necessary to catch the evidence in the act. We envision a system that analyses the email file as soon as received, and write log about tracking the author.

In this paper, we focus on the EML file which can be faced with the attacker firstly. A variety of email analysis techniques have been proposed for forensics. To find visible evidences of email, there exist many forensic tools [2]. Using these email forensic tools, it is possible to view the headers and contents of the email file. For more complex analysis, Guo *et al.* proposed the construction mechanism of the keywords commonly used in the header field [3]. Especially, they deeply analyzed the 'Received' field, and extracted the tracking information such as IP, name of the email program, version of the email program, domain name, and email sending time. They also analyzed the 'Message-ID' field, and identified mail server program and its version. Xie *et al.* analyzed the email contents and constructed the author network [4]. They calculated the similarity between emails based on the Latent Dirichlet Allocation model. Haggerty *et al.* constructed the framework for forensic investigation and concentrated on the triage and analysis of unconstructed email data for email networking [5].

Ampazis *et al.* proposed the author identification algorithm with Optimized Levenberg-Marquardt with Adaptive Momentum(OLMAM) trained feed forward neural networks [6]. They ignored the header information because they assumed that the header was already forged. Iqbal *et al.* focused on writing styles of an email body [7]. They extracted 419 features such as lexical, syntactic, structural, and domain-specific features. Three clustering methods (EM, K-means, Bisecting K-Means) were performed. Schmid *et al.* proposed the customized associative classification technique for authorship attribution [8]. They extracted unique writing style features of email text and produces an intuitive classifier. Previous researches were focused on fragmentary information of email header or text mining technique. In case of an infringement accident, there is no reason to manipulate an email header, so it will have admissibility of evidence.

In this paper, we investigate the email database focusing on the header. We also define four categories of email information for author tracking. For each category, valuable information for author information is analyzed. The remainder of the paper is structured as follows: In section 2, we investigate email header statistically and describe the author identification information. In section 3, we conclude our paper.

2. EMAIL ANALYSIS

Many previous researches ignored email header information due to risk of falsification. It is broadly true that there exists possibility of counterfeit [9]. Although APT authors try to clear their digital fingerprint, they will not able to fully manipulate all information. In case that attacker uses web mail service for spear-phishing, it is hard to hack the web mail server. Because hacking the web server requires a lot of extra effort, they will use VPN or Proxy to avoid tracing.
In addition, there is no reason for forging an email header in cyber incident from the standpoint of the evidence submitter. However, even if they use these detour techniques, it is almost impossible to distort all of the language, email sending time, system information, and habits of the author. First, we analyzed the email database in according to the email field number. Based on the major email fields, tracking information for author identification is investigated.

2.1 Top 10 frequency in email header

Table 1. Top 10 frequency of the field in email header

Rank	Header name	Count	Count/ email
1	Received	134,455	3.71
2	Content-Type	65,517	1.81
3	Content-Transfer-Encoding	50,397	1.39
4	Subject	31,963	0.88
5	Date	31960	0.88
6	From	31,958	0.88
7	To	31,703	0.87
8	Mime-version	30,349	0.84
9	Message-id	26,379	0.73
10	Delivered-to	25,945	0.72

Fields of the email header were defined by standard such as RFC 822, 1123, 2156, 2078 and so on [10-13]. However, various email server software and client software define their own email-field such as 'X-Originating-IP'. To cover this irregularity, we investigated CSDMC2010 spam email database [14] and personally received spam email. Fields of email headers were extracted and sorted with frequency of appearance by descending order as shown in Table 1. Total number of spam mail is 36,247. For smooth explanation, we have prepared an example according to Table 1 as shown in Fig. 2 [14]. To preserve privacy of the spammer and email receiver, we replaced the personal information with meaningless information.

Because of internal policy of web mail service such as a distributed server and a spam checker, there are several transportations. Whenever Mail Transfer Agent(MTA) received the email, 'Received' field is added in email header. Since the field is written from bottom to top according to the order of passage, information of the real sender can be found at 'Received' field of the bottom in the documents. Since it contains a lot of information of a sender and most of email file contain this field, it is very important field for author identification. In our investigation as shown in Table 1, one mail includes average 3.71 of 'Received' fields per mail. It means that mail has passed an average of 3.71 mail servers. In 'Received' field of Fig. 1, we can intuitively know that email comes from domain of 'demo.hec.hr' with IP address of '2x3.2x2.1x4.1x4'. This email was sent to domain of 'abc.abcmining.org' with Extended Simple Mail Transfer Protocol(ESMTP). Receiver address is 'hievery@abcminig.org' and received time is 'Tue, 2 Jun 2009 18:01:48'. We also know that the time zone of the sender is +09:00(JST). 'Delivered-To' field is used for email loop detection. There are various reasons such as auto responders, email bounces, misconfigured email server. 'Date' field is send time at client side. In the example, the send time is 'Tue, 2 Jun 2009 18:01:47', MTA server received the email at 1 second later ('Tue, 2 Jun 2009 18:01:48'). 'message-ID' field is globally unique ID for email. From this field, we can detect the email client program and version [3]. 'From' field is email address of the sender. Since this is very obvious information, sender will forge this part first. There are several methods to maintain anonymity. For example, there are many web email service that supports anonymous accounts [15]. However, encoding information like 'UTF-8' attached to nickname can be useful for author identification. 'To' field is email address of receiver, and it can include encoding information. 'Subject' field is subject of email, and it can also include encoding information. 'MIME-Version' field is related with transfer protocol. Currently, it uses almost version 1.0. 'Content-Type' field describes the email body. In example, this content type is text and content is html documents. Noteworthy, there is a 'charset' of body content for decoding. 'Content-Transfer-Encoding' field is related with encoding method of transmission. There are 7bit, 8bit, quoted-printable, base64, and so on.

```
Received: from demo.hec.hr (demo.hec.hr [2x3.2x2.1x4.1x4])
        by abc.abcmining.org with ESMTP id n5291lUO028314
        for <hievery@abcmining.org>;
        Tue, 2 Jun 2009 18:01:48 +0900 (JST)
Delivered-To: m0620212@mail.csmining.org
Date: Tue, 2 Jun 2009 18:01:47 +0900 (JST)
Message-ID:
<373628939195799.BDWQBSSHLWFROXO@demo.hec.hr>
From: '=?UTF-8?BxxZWc6xxM7JiBxx+s7Jxx7JuQxxyk?=' <hievery@abcmining.org>
To: hievery@abcmining.org
Subject: [SPAM]    he broke his leg
MIME-Version: 1.0
Content-Type: text/html; charset='utf-8'
Content-Transfer-Encoding: 7bit
```

Fig. 1. Example of a common email header

2.2 Important information for author identification

For author identification, it is necessary to inspect standard and non-standard headers in email. In our inspection of an email database, 1222 email headers were found. We analyzed the contents of the various email headers and identified the useful fields for author identification. Totally 103 fields are identified and we scored 1 to 5 on each field according to the tracking usefulness. For lack of space, we cannot list all the fields. We categorize these fields into four group and select most important fields. We also consider the frequency of appearance. It was also analyzed whether it was a continuous type or a categorical type considering that it would be used for machine learning in the future.

Area of author Area is key information for identifying the author group. This information can be found in 'Received', 'X-Received', and 'Date' field.

- IP: Sender IP can be found in 'Received', 'X-Received', 'X-Originating-IP', and 'X-ClientIP'. 'Received' fields are attached in order from bottom to top as email pass through the MTA, the bottom one is from sender. If there is no IP in 'Received' field, it would have been written in other fields. Since IP is just a set of numbers and it is not perfectly continuous mapping with author area [16]. Therefore, it is inappropriate that it defines as continuous type variable. However, it cannot be used as a categorical type because range of IPv4 is about 4.3 billion. Our solution is to extract other information such as country, region, city, latitude, longitude and time zone from IP. In 2017, available number of countries is 195 and it can be used as

a sufficiently categorical type. Region and city is important data, but it is inefficient that these features are used in machine learning due to many kinds of categories. Latitude and longitude can be used as continuous feature. The range of latitude and longitude is -90~+90 and -180~+180, respectively.

• Time zone: Time zone information can be extracted in 'Received', 'X-Received', and 'Date' field. Since the time zone varies from country to country, it can be good feature for locating the author. It can be used as both continuous -12:00~+14:00 and categorical type with 40 items. Using minimum resolution 15 minutes, it is possible to map time zone from 1 to 108(=27 × 4).

Language of author Language information can be obtained from 'Content-Type', 'Encoding', 'Accept-Language', 'From', 'To', 'Subject', and so on. Language information is focused on the victim rather than an author. If a content of the text is awkward, the attack can be failed. Therefore, the author or his/her assistant is good at that language.

• Content-Type:Charset: 'Charset' is assignment set between a character and a code such as natural numbers. For example, 'us-ascii' is for alphabet, and 'euc-kr' is for Hangeul which is Korean alphabet. This field indicates the available language for the author. Keyboard layout also can be inferred by this field. According to the RFC2978, there are currently 257 charsets in common use [13].

• Encoding: After assignment with 'Charset', it needs to be processed again for storage or transmission. This is called encoding. There are '7bit','8bit', 'base 64', and 'quoted printable'.

• Accept-Language, Content-Language: 'Accept-Language' field seems to be originally from HTTP header. It indicates which language and locale variant the receiver is preferred. For example, 'en-US' shows that it is in the United States and uses English. 'Content-Language' refers to the language and local of the content, similar to 'accept-language'.

• Other encoding items: The fields shown to receiver, such as 'From', 'To', 'Subject', may contain encoding information. 'From' and 'To' field has an email address and a nick name for it. Unlike email address that consist of alphabet and numbers, a nick name is more familiar with the users. To support various language of nick names, encoding information is attached in front of email address and subject.

Date&time of author 'Date' and 'Time' of sending time is important data for estimating author's life pattern and work environment.

• Date: Author generally send a bunch of email for limited period of time. After a certain period of time, the attack vector will be useless and is discarded. This period is called a campaign, and the 'date' field can be a good feature to group them [17]. In addition, multiple pieces of information can be extracted from date. First, whether it is weekday/weekend can be decided. For example, if a country is well protected by labor laws, it is likely to have sent mail during the week. Second, it can be determined whether the sent date is in holiday of specific country.

• Time: Since campaign often last on a date unit, it is inappropriate to use the 'time' field in campaign group formation. Also, it is not improper to use 'time' field directly as a feature. Instead of direct use, 'time' can be divide into the life pattern such as sleep, morning shift, lunch time, afternoon shift, dinner time, night shift, graveyard shift.

```
Mozilla/5.0 (Macintosh; Intel Mac OS X 10.9; rv:24.0)
Gecko/20100101 Thunderbird/24.2.0
```

Fig. 2. Example of 'User-Agent' field

System of author Actually, it is not essential part of email header. However, many email systems insert the system and client program (ex. internet browser) into header for convenience.

• User-agent: Although the database we surveyed cannot represent a population, 2138 emails had this field. 'User-Agent' field contains a lot of information about system and program of a user. As shown in figure~\ref{fig:user-agent}, the user's computer is Macintosh, and the operation system is OS X ver(X Mavericks) 10.9. This OS was released after October 23, 2013, so we can see that the email was written later. The user uses free email client software Thunderbird ver. 24.2.

3. CONCLUSION

In this paper, we analyzed the email header for author identification. In previous researches, email headers rarely used because they had only little information. In addition, they are easy to manipulate for plaintiff or defendant. In case of author identification in cyber incident, however, there is no reason to forge the email header. Under this assumption, we investigate the email database and extract the key information for author identification. In our analysis, we can see that there is more trace information in the email header than we thought. In future, additional email databases should be inspected for reliability. It is also a further study to divide the author group by applying various clustering algorithm based on the proposed feature set.

4. REFERENCES

[1] Bencsáth, B., Pék, G., Buttyán, L. and Félegyházi, M., 2012, Duqu: Analysis, detection, and lessons learned, *Proc. of ACM European Workshop on System Security (EuroSec),* Bern, Switzerland.

[2] Devendran, V.K., Shahriar, H. and Victor, C., 2015, A comparative study of email forensic tools, Int. J. of Information Security, Vol. 6, pp.111-117.

[3] Guo, H., Jin, B. and Qian, W., 2013, Analysis of email header for forensics purpose, *Proc. of Communication Systems and Network Technologies(CSNT)*, Gwalior, India, pp. 340-344

[4] Xie, L., Liu, Y. and Chen, G., 2015, A forensic analysis solution of the email network based on email contents, *Proc. of Fuzzy Systems and Knowledge Discovery (FSKD)*, Zhangjiajie, China, pp. 1613-1619

[5] Haggerty, J., Karran, A., Lamb, D. and Taylor, M., 2011, A framework for the forensic investigation of unstructured email relationship data, *Int. J. of digital crime and forensics*, vol. 3, pp. 1-18

[6] Ampazis, N., Iakovaki, H. and Dounias, G., 2007, *Proc. of 19th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2007)*, Patras, Greece, vol. 2, pp. 413-417

[7] Iqbal, F., Binsalleeh, H., Fung, B.C. and Debbabi, M., 2010, Mining write-prints from anonymous e-mails for forensic investigation, *Int. J. Digital Investigation*, vol. 7, pp. 56-64

[8] Schmid, M.R., Iqbal, F., and Fung, B.C., 2015, E-mail authorship attribution using customized associative classification, *Int. J. Digital Investigation*, vol. 14, pp. S116-S126

[9] Lin, E., Aycock, J., and Mannan, M., 2012, Lightweight client-side methods for detecting email forgery, *Prog. of Information Security Applications: 13th International Workshop*, pp. 254-269,

[10] Crocker, D. H., 1982, Standard for the format of arpa internet text messages, *RFC822,*

[11] Braden, R., 1989, Requirements for internet hosts-application and support, *RFC 1123*

[12] Kille, S., 1998, MIXER (Mime internet X.400 enhanced relay): Mapping between X.400 and RFC 822/mime, *RFC 2156*

[13] Linn, J., 1997, Generic security service application program interface, version 2. *RFC 2078*

[14] Group, C., 2010, CSDMC2010 spam corpus, http://csmining.org/index.php/spam-email-datasets-.html

[15] Anonymous, 2012, Anonymousemail, https://anonymousemail.me

[16] IP2Location, Ip2location lite[region-city-latitude-longitude-zipcode-timezone], http://lite.ip2location.com /database/ip-country

[17] Symantec, 2016, Internet security threat report. Symantec