Technology Corner

Analysing E-Mail Headers for Forensic Investigation

M. Tariq Banday
University of Kashmir
India
sgrmtb@yahoo.com

ABSTRACT

Electronic Mail (E-Mail), which is one of the most widely used applications of Internet, has become a global communication infrastructure service. However, security loopholes in it enable cybercriminals to misuse it by forging its headers or by sending it anonymously for illegitimate purposes, leading to e-mail forgeries. E-mail messages include transit handling envelope and trace information in the form of structured fields which are not stripped after messages are delivered, leaving a detailed record of e-mail transactions. A detailed header analysis can be used to map the networks traversed by messages, including information on the messaging software and patching policies of clients and gateways, etc. Cyber forensic e-mail analysis is employed to collect credible evidence to bring criminals to justice. This paper projects the need for e-mail forensic investigation and lists various methods and tools used for its realization. A detailed header analysis of a multiple tactic spoofed e-mail message is carried out in this paper. It also discusses various possibilities for detection of spoofed headers and identification of its originator. Further, difficulties that may be faced by investigators during forensic investigation of an e-mail message have been discussed along with their possible solutions.

Keywords: E-mail Headers, E-mail Forensic, E-mail Analysis, E-mail spoofing, E-mail Investigation.

1. INTRODUCTION

E-mail is a highly distributed service that involves several actors which play different roles to accomplish end-to-end e-mail exchange (Crocke 2009). These actors fall under three groups namely *User Actors*, *Message Handling Service (MHS) Actors* and *ADministrative Management Domain (ADMD) Actors*. *User Actors* are *Authors*, *Recipients*, *Return Handlers* and *Mediators* which represent people, organizations or processes that serve as sources or sinks of messages. They can generate, modify or look at the whole message. *Message Handling Service (MHS) Actors* are *Originators*, *Relays*, *Gateways* and *Receivers* which are responsible for end-to-end transfer of messages. These Actors can generate, modify or look at only transfer data in the message. *ADministrative Management Domain (ADMD) Actors* are *Edges*, *Consumers* and *Transits* which are associated with different organizations and have their own administrative authority, operating policies and trust-based decision making. E-mail system is an integration of several hardware & software components, services and protocols, which provide interoperability between its users and among the components along the path of transfer. The system includes sender's client and server computers and receiver's client and server computers with required software and services installed on each. Besides, it uses various systems and services of the Internet. The sending and receiving servers are always connected to the Internet but the sender's and receiver's client connects to the Internet as and when required.

Cyber forensics deals with the collection and analysis of data from computer systems, networks, wired or wireless communication streams and storage media through scientifically proven techniques in a manner admissible in a court of law (Natarajan et al. 2009). It is also called digital forensics and computer forensics. Network forensics is the science that deals with capture, recording, and analysis of network traffic for investigation purpose and incident response (Emmanuel et al. 2010). E-mail forensics dealing with the investigation of e-mail message is a specialized type of network forensics which in turn is also a specialized type of cyber forensics. It refers to the study of source and content of e-mail message as evidence to identify the actual sender and recipient of a message, data/time of transmission, detailed record of e-mail transaction, intent of the sender, etc. This study involves investigation of metadata, keyword searching, port scanning, etc. for authorship attribution and identification of e-mail scams.

1. E-MAIL FORENSIC INVESTIGATION

E-mail forensics refers to the study of source and content of e-mail message as evidence, identification of the actual sender, recipient, date and time when it was sent, etc. Forensic analysis of an e-mail message aims at discovering the history of a message and identity of all involved entities. Besides message analysis, e-mail forensic also involves investigation of some client or server computer suspected of being used or misused for e-mail forgery. It may involve inspection of *Internet favorites, Cookies, History, Typed URL's, Temporary Internet Files, Auto-completion Entries, Bookmarks, Contacts, Preferences, Cache,* etc. Several Open Source software tools have also been developed to perform e-mail header analysis to collect evidence of e-mail fraud.

E-mail analysis begins from the recipient's mailbox which contains the e-mail message. The message is analysed to determine the source (*originator* and *author*). The analysis involves investigation of both control information (envelope and header) and message body. *Mailbox*, *domain name*, *message-ID* and *ENVID* are globally unique identities that are used in e-mail. The *Mailbox* is identified by an e-mail address and *domain name* is an identifier of an Internet resource. *Message-ID* is used for threading, aiding identification for duplications and

Domain Name System (*DNS*) tracking. The ENVelope Identifier (*ENVID)* is used for the purpose of message tracking. E-mail message comprises of envelope that contains transit-handling information used by the Message Handling Service (*MHS*) and message content which consists of two parts namely Body and Header. The Body is text but can also include multimedia elements in Hyper Text Markup Language (*HTML*) and attachments encoded in Multi-Purpose Internet Mail Extensions (*MIME*) (Resnick 2001). The Header is a structured set of fields that include *'From'*, *'To'*, *'Subject'*, *'Date'*, *'CC'*, *'BCC'*, *'Return-To'*, etc. Headers are included in the message by the sender or by a component of the e-mail system and also contain transit-handling trace information. Further, the message also contains special control data pertaining to Delivery Status (DS) and Message Disposition Notifications (MDN), etc. The control information i.e. envelope and headers including headers in the message body that contain information about the sender and/or the path along which the message has traversed represents the metadata of an e-mail message. The analysis of this metadata called header analysis can be used to determine genuineness of a message.

The e-mail address of sender responsible for submitting the message to the transfer service is specified in the *Sender* header field. This field is optional and needs to be specified only if the author of the message is different from the sender. The e-mail address of the author is contained in the mandatory *From* header field. Various other addresses related to author or sender of the message, are addresses specified in *Reply-To, MailFrom* and *Return-Path* fields. Address specified in *Reply-To* header overrides the *From* address for responses from recipients if specified. *MailFrom* specifies the address for receiving return control information like *MDN* and *DSN*. This address need not to be the same as that of author or sender responsible for submitting the message. *Return-Path* address is the address recorded by *MDA* from *MailFrom* control identifier. E-mail client programs and webmail interfaces add useful headers called *X-Headers* which besides other information contain useful information about the sender or author of the e-mail. A comparison of these addresses in all cases cannot be used to determine the genuineness of a sender or author because it is possible to spoof all of these addresses and thus not only hide one's address but also pretend to be somebody else by using another's valid e-mail address in these fields. Trace information in the form of *Received* header field recorded by originator, relay, mediator or destination may be used to validate only the domain part of the sender's or author's e-mail address. *Received-SPF* also cannot validate the e-mail address of a sender or author as it can only validate domain of address specified in *MailFrom* parameter. *DKIM-Signature* security field if present may be used to validate the domain part of the sender's or author's e-mail address if the sending and receiving servers are following *DKIM* signing protocol. Thus, on the basis of inconsistencies of sender's identity as may be revealed by various header fields, forensic experts can only make a wider guess about the genuineness of the sender's e-mail address and not a final decision. Originator, Relay including *MTA*, and Receiver add trace information at the beginning of the message which is in the form of *Received* and *Return-Path* header fields. *Received* header field specifies the address used in *MailFrom* parameter and may not be much useful for forensic investigation. The *Received* field contains vital information including names and IP addresses of originating host, relays or *MTA's*, Mediators and *MSA's*. However, it is also possible to spoof the IP address reported in this field by the use or misuse of various techniques mentioned in section 2.

Besides *header analys*is various other approaches that can be used for e-mail forensics include *bait tactics*, *server investigations*, and *network device investigation*. Custom and MIME headers appearing in the body of the message are also analysed for *sender mailer fingerprints* and *software embedded identifiers* (Marwan 2005).

Various software tools have been developed to assist e-mail forensic investigation. These include *eMailTrackerPro* (http://www.emailtrackerpro.com/), *EmailTracer* (http://www.cyber forensics. in), *Adcomplain* (http://www.rdrop.com/users/billmc/adcomplain.html), *Aid4Mail Forensic* (http://www.aid4mail.com/), *AbusePipe* (http://www.datamystic.com/ abusepipe.html), *AccessData's FTK* (www.accessdata.com/), *EnCase Forensic* (http://www.guidancesoftware.com), *FINALeMAIL* (http://finaldata2. com), *Sawmill-GroupWise* (http://www.sawmill.net), *Forensics Investigation Toolkit (FIT)* (http://www.edecision4u. com/FIT.html), *Paraben (Network) E-mail Examiner* (http://www.paraben.com/email-examiner.html), *etc.* These analyse headers of e-mail messages to detect the IP address of the originating machine. These tools often have abuse reporting features, e-mail classification option, support multiple encryption techniques like *Credant, SafeBoot, Utimaco, EFS, PGP, Guardian Edge, Sophos Enterprise* and *S/MIME*. Its current supported e-mail types are: *Lotus Notes NSF, Outlook PST/OST, Exchange EDB, Outlook Express DBX, Eudora, EML* (*Microsoft Internet Mail, Earthlink, Thunderbird, Quickmail, etc.), Netscape, AOL* and *RFC 833*. Some of these claim to be vetted by courts as standard digital investigation platforms.

2. NEED FOR E-MAIL FORENSIC INVESTIGATION

Cybercriminals spoof e-mail messages to carry out various illegitimate activities through e-mail system and remain underground to evade any possible legal action against them. These include i) abuses like spamming, phishing, cyber bullying, child pornography, sexual harassment, racial vilification, etc., ii) misuse by transmitting viruses, worms, Trojan horses, hoaxes, and other malicious programs with an intent to spread them over Internet, and iii) carry out Internet infrastructure crimes through Denial of Services and Directory Harvesting Attacks. This injudicious use of e-mail cause many technological problems like misuse of storage space, wastage of computational resources, and network conjunction. Cybercriminals misuse SMTP to lie recipients about their true identities by not only spoofing one or more headers in the envelope or header of the message that somehow reveals their identity but also put misleading information in these headers. A highly technical spammer or phisher may also evade packet filters and spoof the source IP address of their packets to indicate that the message is from a trusted domain (Hastings et al. 1996).

Senders can lie about their true identities in various ways by using or misusing different techniques that include:

i. ***Spoofing***: (Radvanovsky 2006) It is an attempt to conceal the source of an e-mail message by placing false information in its headers. All possibilities to lie about true identities of sender listed above lead to spoofing. E-mail spoofing may be combined with IP spoofing to make its detection very difficult.

ii. ***Unauthorized Networks***: (Shunman et al. 2003) Wired or wireless networks that have been compromised by perpetrators to gain unauthorized access to the Internet can disguise their identities while sending an e-mail.

iii. ***Open Mail Relays***: (Shue et al. 2009) An open mail relay is a mis-configured mail relay that accepts mail form any computer and forwards it to another computer which otherwise should have accepted mail for and from specific computers. Such a relay becomes vulnerable to spammers and phishers who hide their identities behind these relays.

iv. ***Annomizers or re-mailers***: (Cherry, 2001) Re-mailers are websites that operate under the guise of protecting privacy of Internet users offering anonymous Internet surfing. They intentionally strip headers from e-mail and some even do not maintain server logs.

v. ***Open Proxy***: (Vivek et al. 2004) A proxy server is a machine that allows computers to connect through it to some other computer on the Internet. HTTP proxy server provided by ISPs, Corporate Proxy Server, transparent proxy server, and Open proxy server also called anonymous proxy server are different types of proxy servers which provide different levels of anonymity. Users connecting to Internet through these proxy servers share IP address. An Open proxy server does not maintain a strict log of user activities unlike others which maintain user logs synchronized with reliable time servers. Such open proxy servers provide anonymity and untraceable Internet activity.

vi. ***SSH Tunnel or Port-Redirector***: (Dusi 2008) A Tunnel in Internet means a secure data path through an un-trusted network. Depending upon the software and techniques used, tunneling can be accomplished through many ways. SSH also has a feature called SSH Port Forwarding, sometimes called SSH Tunneling, which allows establishing a secure SSH session and then tunneling arbitrary TCP connections through it. SSH Tunneling is an encrypted tunnel created using the SSH protocol connection. SSH Tunnels can be used by e-mail senders to hide their identities.

vii. ***Botnets***: (Banday et al. 2009) The term bot, derived from “ro-bot” in its generic form is used to describe a script or set of scripts or a program designed to perform predefined functions repeatedly and automatically after being triggered intentionally or through a system infection. Although bots originated as a useful feature for carrying out repetitive and time consuming operations but they are being exploited for malicious intent. Bots that are used to carry out legitimate activities in an automated manner are called benevolent bots and those that are meant for malicious intent are known as malicious bots. A botnet is a network of bots controlled by a botmaster. A botmaster can command its controlled bots (malicious bots) running on compromised computers across the globe to send e-mail to some designated addresses while concealing its identity and committing some e-mail fraud.

viii. ***Untraceable Internet Connections***: (Berthold et al. 2000; Landsiedel et al. 2005) Public and corporate Internet access points like cyber cafe, university campus, business organization, etc. provide Internet access to its users by shearing Internet connection. If a proper log of activity is not maintained, its users can easily conceal their identity to do illegal cyber activities including e-mail fraud without any fear of being traced.

Protocols offering security and anti-spam filters that are capable to perform mail categorization have been developed to secure e-mail service against sender-spoofing. Security protocols that add privacy to SMTP either create encrypted secure channel between the sender and the receiver during SMTP transactions or use end to end symmetric or asymmetric cryptographic schemes. Further, various domain validation anti-spoofing standards using either IP addresses or digital signatures to validate sending domain have also been developed that help an e-mail system at the receiving end to detect the spoofing of addresses and as such enables it to decide how to handle incoming e-mails. A detailed record of e-mal security protocols and procedures along with the functioning of some prominent security protocols is given in (Banday 2010b). Very limited numbers of e-mail users use these protocols to secure their e-mails due to either their limited technical skill or unawareness about their existence (Banday 2010a)]. Further, their use has not been made mandatory and as such unwillingness of some ESP’s does also limit their use. Furthermore, spammers constantly change spam sending techniques and its structure to evade security procedures and protocols, leaving scope for e-mail forgery and e-mail crime raising the need for e-mail forensic analysis. Forensic investigation of an e-mail message can be carried out by the use of various techniques and software tools. However, these techniques and tools may prove to be ineffective to identify the e-mail forgery or the actor responsible for it. This is because analysis through various techniques discussed above can halt due to lack of co-operation between different service providers. Further, invention of new means and changing tactics of the cybercriminals make e-mail forensic an active area of research thereby making it necessary to analyse e-mail messages for any possible forgery.

3. ANALYZING AN E-MAIL MESSAGE

A sample header set of an e-mail message sent by *tariq@tariq.com* pretending to be *alice@alice.com* and sent to *bob@bob.com* is shown in table 1. In this e-mail, the sender's address, date e-mail was sent, reply-to address, and various other fields have been spoofed. The identification identities like domain name, IP address, etc. which could have revealed servers used in the process of sending the e-mail have been suitably edited. This header set is used to demonstrate the information contained in various headers of the message.

Table 1: Sample header set of an e-mail message

Para No	*Header*	*Value*
1	***X-Apparently-To:***	*bob@bob.com via a4.b4.c4.d4; Tue, 30 Nov 2010 07:36:34 -0800*
2	***Return-Path:***	*< alice@alice.com >*
3	***Received-SPF:***	*none (mta1294.mail.mud.bob.com: domain of alice@alice.com does not designate permitted sender hosts)*
4	***X-Spam-Ratio:***	*3.2*
5	***X-Originating-IP:***	*[a2.b2.c2.d2]*
6	***X-Sieve:***	*CMU Sieve 2.3*
7	***X-Spam-Charsets:***	*Plain='utf-8' html='utf-8'*
8	***X-Resolved-To:***	*bob@bob.com*
9	***X-Delivered-To:***	*bob@bob.com*
10	***X-Mail-From:***	*alice@alice.com*
11	***Authentication-Results:***	*mta1294.mail.mud.bob.com from=alice.com; domainkeys=neutral (no sig); from=alice.com; dkim=neutral (no sig)*
12	***Received:***	*from 127.0.0.1 (EHLO mailbox-us-s-7b.tariq.com) (a2.b2.c2.d2) by mta1294.mail.mud.bob.com with*

Para No	Header	Value
		SMTP; Tue, 30 Nov 2010 07:36:34 -0800
13	***Received:***	*from MTBLAPTOP (unknown [a1.b1.c1.d1]) (Authenticated sender: tariq@tariq.com) by mailbox-us-s-7b.tariq.com (Postfix) with ESMTPA id 8F0AE139002E for <bob@bob.com>; Tue, 30 Nov 2010 15:36:23 +0000 (GMT)*
14	***From:***	*"Allice" <Alice@a.com>*
15	***Subject:***	*A Sample Mail Message*
16	***To:***	*"Bob Jones" <bob@bob.com>*
17	***Content-Type:***	*multipart/alternative; charset="utf-8"; boundary="KnRl8MgwQQWMSCW6Q5=_HgI2hwAdah5NLY"*
18	***MIME-Version:***	*1.0*
19	***Content-Transfer-Encoding:***	*8bit*
20	***Content-Length:***	*511*
21	***Reply-To:***	*"Smith" <smith@smith.com>*
22	***Organization:***	*Alices Organization*
23	***Date:***	*Tue, 28 Nov 2010 21:06:22 +0530*
24	***Return-Receipt-To:***	*smith@smith.com*
25	***Disposition-Notification-To:***	*jones@jones.com*
26	***Message-Id:***	*<20101130153623.8F0AE139002E@mailbox-us-s-7b.tariq.com>*

The Header *X-Apparently-To* shown in Para 1 is relevant when mail has been sent as a BCC or to recipients of some mailing list. This field in most of the cases contain the address as in To field. But if mail has been sent to a BCC recipient or a mailing list, *X-Apparently-To* is different from *TO* field. Some may show *TO* while others may not show it. Thus *X-Apparently-To* always shows the e-mail address of recipient regardless of whether mail has been sent using *TO*, BCC, CC addresses or by the use of some mailing list.

The *Return-Path* header is the e-mail address of the mailbox specified by the sender in the *MailFrom* command. This address can also be spoofed, if no authentication mechanism is in place at the sending server as has been done in the sample e-mail shown at Para 2 in Table 1. It is not possible to determine genuineness of Return-Path header through header analysis alone.

The header shown in Para 3 is the *Received-SPF,* the value of which specifies that the mail has come from a domain which either does not have a SPF record or is not yet a designated permitted sender.

The spam score calculated by the spam filtering software of the receiving server or MUA is contained in *X-Spam-Ratio* field. This value for the e-mail under study is *3.2* as shown in Para 4. If this ratio exceeds certain pre-defined threshold, e-mail will be classified as spam. Different receiving servers and MUA's used different X-Header fields to indicate spam score and classification decision taken with regard to the current message. These include X-Spam-Flag, X-Spam-Checker-Version, X-Spam-Level, X-Spam-Status, etc.

X-Originating-IP specified the IP address of the last MTA of the sending SMTP Server, which has delivered the e-mail to the server of *bob@bob.com*. In the sample e-mail it is *[a2.b2.c2.d2]* as shown in Para 5. This address is also contained in the *Received* header field.

X-Sieve header specifies the name and version of message filtering system. This pertains to the scripting language used to specify conditions for message filtering and handling. In the sample e-mail the name of the message filtering software is *CMU Sieve* and its version is *2.3.*

X-Spam-Charsets header specifies the character set used for filtering the messages. The value for this field in sample e-mail at Para 7 indicates that 8-bit Unicode Transformation Format (*UTF*) has been used by bob's server. UTF is a variable length character set having a special property of being backward-compatible to ASCII.

X-Resolved-To address is the e-mail address of the mailbox to which the mail has been delivered by MDA of bob's server. In most cases, it is the same as X-Delivered-To field. *X-Delivered-To* is the address of the mailbox to which the mail has been delivered by MDA of bob's server. In the sample e-mail both *X-Resolved-To and X-Delivered-To* addresses are *bob@bob.com* as in Paras 8 and 9.

X-Mail-From header specifies the e-mail address of the mailbox specified by the sender in the *MailFrom* command which in the sample e-mail is *alice@alice.com*.

The *Authentication-Results* header in Para 11 indicates that *mta1294.mail.mud.bob.com* received mail from *alice.com* domain which neither has DomainKeys signature nor DKIM signature.

Para 12 is the second *Received* header field containing the trace information indicating *127.0.0.1* as the IP address of the machine that send the message. This machine is actually named *mailbox-us-s-7b.tariq.com* and has IP address *a2.b2.c2.d2*. It has used *EHLO* SMTP command to send the mail. The mail was received by *mta1294.mail.mud.bob.com* using *SMTP*. The message has been received on *Tue, 30 Nov 2010* date at *07:36:34* time. The clock is 8 hrs behind Greenwich Mean Time.

Para 13 is the first *Received* header field representing the trace information indicating *MTBLAPTOP* as the names of the machine that send the message. This machine is not known to the receiver but has an IP address *a1.b1.c1.d1* and *tariq@tariq.com* is the owner of the mailbox who has sent the message. The MTA must follow some authentication mechanism to identify its mailbox users otherwise it is not possible to include authenticated sender's mailbox address with the *Received* field. The message has been received by *mailbox-us-s-7b.tariq.com* using *ESMTPA* protocol which has been running a program called *Postfix*. The message is for *bob@bob.com* and has an ID of *8F0AE139002E*. The message has been received on *Tue, 30 Nov 2010 at 15:36:23*. The clock is set according to Greenwich Mean Time.

The *From*, *Subject* and *To* lines respectively are the e-mail address of the author, subject of the message, and the e-mail address of the intended recipient. *Subject* and *To* are specified by the sender, and the *From* address is taken by the system from the current logged in user. However, *From* header can very easily be spoofed as has been dome in this sample e-mail. The paragraphs 14, 15 and 16 in the sample e-mail show the values of these three fields. The *From* address has been spoofed to carry an address *Alice@a.com* with a user friendly name *Alice*.

Content-Type, MIME-Version, Content-Transfer-Encoding and Content-length in paragraphs 17, 18, 19 and 20 are the MIME headers describing the type of MIME content, transfer encoding, its version and length so that the MUA's can perform proper decoding to render the message successfully on client.

This is the address, sender of this e-mail wants recipient to use for sending reply in response to this e-mail. Normally, this is used by the senders to send replies. Carefully crafted sender spoofing combined with fake *Reply-To* e-mail address can lead to serious information leaks. The *Reply-To* address *"Smith" smith@smith.com* in Para 21 is an arbitrary address that may belong to some user who may not be related to the sender in any way.

Organization header field indicates that the organization of claimed sender is *Alices Organization*. *Organization* header field is an information field representing the organization of a sender. It can be misused by the spammer to give a false impression about a sender as has been done in this e-mail.

Date header indicates that the e-mail was composed and submitted for delivery on *Tue, 28 Nov 2010 21:06:22 +0530*, which is not in conformity with the date in the *Received* field of Para 23.

Return-Receipt-To field indicates the e-mail address, MSA, MTA and MDA must use for sending delivery notifications such as successful or failure notifications. The address mentioned for this field in Para 24 is again an arbitrary address that may belong to some user who may not be related to the sender in any way.

Disposition-Notification-To field indicates an e-mail address, MUA must use when submitting a message indicating that the message has been displayed. This address specified in Para 25 is also an arbitrary address that does belong to some user who may not be related to the sender in any way.

Para 26 contains the *Message-Id* of the message which is *20101130153623.8F0AE139002E@mailbox-us-s-7b.tariq.com*. Generally, a domain name is appended with a unique number by the sending server to form the *Message-Id*.

4. DISCUSSIONS

In the above sample e-mail message, several fields have been spoofed which can be detected easily because the first *Received* field shows the address of authenticated sender which is different from the sender of the message. However, address of authenticated sender may not be always included with the authentication results (in case no authentication mechanism is adhered to or annomizers strip this line). Further, date is also inconsistent as can be noted from the comparison of timestamp in *Received* headers and the date field. Some header fields with context to authentication and above analysed e-mail message are discussed further hereunder:

SPF mechanisms can be used to describe the set of hosts which are designated outbound mailers for the domain. The test besides success or failure may also result into *softfail*, *neutral*, *none*, *permerror* or *temperror*. For example, a successful *Received-SPF* entry could be as follows:

Received-SPF: pass (mta1104.mail.mud.tariq.com: domain of
tariq@tariq.com designates a2.b2.c2.d2 as permitted sender)

Here, the *mta1104.mail.mud.tariq.com MTA* notifies its recipient through *Received-SPF* that domain of *tariq@tariq.com* i.e. *tariq.com* which has an *IP* address *a2.b2.c2.d2* is a permitted sender designated by Sender Policy Framework (Wong 2006).

In case, the domain *alice.com* had been DomainKeys and DKIM complaint and had passed these tests, it could have been as follows:

Authentication-Results: mta1294.mail.mud.bob.com from=alice.com;
domainkeys=pass (ok); from=a.com; dkim=pass (ok)

In this case, it could have included DKIM-Signature (Allma et al. 2007) and/or DomainKey-Signature fields as follows:

DKIM-Signature: v=1; a=rsa-sha1; c=simple; d=alice.com;
h=from:to:subject:date:message-id:content-type q=dns/txt; s=s512;
bh=XX………..=; b=XXX………==;

This is the DKIM Signature signed with SHA1 algorithm. DKIM uses the email headers and body to generate a signature. If the headers are rewritten or text is appended to the message body after it has been signed, the DKIM verification fails. DKIM is backward compatible with the DomainKeys system. When an e-mail message is signed with DKIM, it will include a number of “tags” whose values contain authenticating data for the message being sent. In the example above, the tags used are:

v= This tag defines the version of this specification that applies to the signature record.
a= The algorithm used to generate the signature (plain-text; REQUIRED). It supports "rsa-sha1" and "rsa-sha256", Signers usually sign using "rsa-sha256".
c= It is the canonicalization algorithm 1.e. the method by which the headers and content are prepared for presentation to the signing algorithm.
d= It is the domain name of the signing domain.
h= It is a colon-separated list of header field names that identify the header fields presented to the signing algorithm.
q= It specifies the query method used to retrieve the public key which by default is dns.
s= It is the selector used in the public key.
bh= The signature data or public key, encoded as a Base64 string.

The example of DomainKeys signature is given below. DomainKeys signature has been signed with SHA1 algorithm.

DomainKeys-Signature: a= rsa-sha1; q=dns; c=simple; s=s512;
d=alice.com; b=XXX……………………………==;

When an e-mail message is signed with DomainKeys, it will include a number of “tags” whose values contain authenticating data for the message being sent. In the example above, the tags used are:

a= It is the encryption algorithm used to generate the signature which by default is "rsa-sha1".
q= It specifies the query method used to retrieve the public key which by default is dns.
c= It is the canonicalization algorithm 1.e. the method by which the headers and content are prepared for presentation to the signing algorithm.
s= It is the selector used in the public key.
d= It is the domain name of the signing domain.
b= The signature data or public key, encoded as a Base64 string.

Date header represents the date e-mail was composed and submitted for delivery. However, this filed can also be spoofed (Banday 2010a) as has been done in this sample e-mail message. It can be easily noticed by comparing its value in Para 23 with the dates in the *Received* header fields.

Message-Id is the message Identification attached to the e-mail message. Every e-mail has a unique message ID that helps the administrators to locate the e-mail in server log. Usually every sending server uses its own custom algorithm to generate this unique number and append domain name to this to make it unique on the internet. This ID can also help to identify the domain of the sender but it can also be forged to confuse the investigators.

The first *Received* header field representing the trace information contains the IP address of the machine used to send the e-mail message. On tracking this IP address several cases as explained below are possible:

i. The IP address in the *Received* header field maps to direct connection having a static IP address. In this case, this address is the address of the sender's computer. However, if the IP address is dynamic then the logs of the proxy or SMTP server need to be obtained for continuing the e-mail tracking.

ii. The IP address contained in the *Received* header corresponds to some proxy server. In this case, proxy server's log must be obtained to track the sender. Open proxy server may raise some issues for the investigators because they do not maintain a strict log of activities. In case SSL is used to log on to *HTTP* based e-mail server, proxy can not be an issue because IP address of the client shall be recorded. Corporate proxy servers may not be strictly time synchronized as they may be using Network Time Protocol (*NTP)* and thus may impede the investigation. *ISP* proxy servers usually maintain a strict and time synchronized log (use *STIME* protocol) and have a clear devised policy to cooperate with the investigators.

iii. The tracked IP address maps to some tunnelling server. In this case, tracking source of e-mail will be difficult because tunnelling may be done in different ways and some are not logged.

iv. The IP address in the *Received* header field maps to SMTP server. In this case, the SMTP server log must be obtained. IP address may map to SMTP server belonging to ISP, or some corporate or an open relay. In all cases, logs stored must be obtained. If the logs are strictly time synchronized, then the sender can be tracked easily. ISP and corporate SMTP servers can provide further details about the particular user such as his contact details and credit card number.

v. The IP address contained in the *Received* field resolves to Annomizers or re-mailers. In this case, investigators must obtain logs and original e-mail message from the anonymous SMTP or HTTP servers. Further, in case the anonymity is a paid service, user account details must also be obtained.

It is also possible to add one or more false *Received* headers in the data field of the message with an intention to freeze the investigation. Investigators must pay careful attention to all fields of the *Received* headers with respect to each other especially in terms of delivery methods and date & time. If the delivery methods vary or the time & date differ considerably, then false headers can be easily identified. Otherwise, the investigation shall have to investigate all IP addresses and request logs from all servers. It may be very difficult to track a sender from the IP address if the sender has tampered IP address at packet level (Ehrenkranz et al. 2009). Once the source of the e-mail message under investigation has been determined or some one is strongly suspected for being the source, his or her computer, e-mail client software, web browser, etc. are investigated for traces of evidence.

5. CONCLUSION

E-mail forensics, a specialized kind of network forensics deals with the investigation of content of e-mail messages to identify the actual sender, recipient, data and time when it was sent, etc. It also involves investigation of source and destination systems and intermediate devices used for its delivery. The header analysis is carried out on the content of e-mail message to determine its legitimacy. Spoofing, unauthorized networks open mail relays, annomizers or re-mailers, open proxy, SSH tunnel or port-redirector, botnets and untraceable Internet connections are common approaches by which senders lie to recipients about their true identities. Various software tools which essentially perform automated header analysis and network device inspections have been developed to assist speedy investigations. These also include features for abuse reporting, support for multiple mailbox formats, e-mail classification, etc. It has been found that header fields of e-mail message that can directly reveal the identity of sender can be forged unless compatible security protocols are used at both sending and receiving servers. However, first received header of the message contains the original IP address of the computer used to send the e-mail message, which can be tracked to identify the sender. In case IP address contained in the first received field maps to some Internet resource, sender can be tracked by identifying its identity from the logs maintained by servers or various network devices.

6. REFERENCES

Allma, E., Callas, J., Delan, M., Libbey, M., Fenton J. & Thomas, M. (2007). DomainKeys Identified Mail (DKIM). Internet Engineering Task Force (IETF), RFC 4871.

Banday MT, et al., (2010a) “Analyzing Internet e-mail date-spoofing”, Digital Investigation (2010), doi:10.1016/j.diin.2010.11.001.

Banday, M.T., Qadri, J.A. (2010b). “A Study of E-mail Security Protocols,” eBritian, ISSN: 1755-9200, British Institute of Technology and E-commerce, UK, Issue 5, Summer 2010, pp. 55-60. Available online at:

http://www.bite.ac.uk/ebritain/ebritain_summer_10.pdf.

Banday, M.T., Qadri, J.A., Shah, N.A. (2009). "Study of Botnets and Their Threats to Internet Security,". Sprouts: Working Papers on Information Systems, 9(24). http://sprouts.aisnet.org/9-24.

Berthold, O., Federrath, K¨opsell, H. S. (2000), “Web MIXes: A system for anonymous and unobservable Internet access”, In Proc. of Designing Privacy Enhancing Technologies:Workshop on Design Issues in Anonymity and Unobservability, July 2000.

Crocker, D. (2009), “Internet Mail Architecture”, RFC 5598, July 2009. http://tools.ietf.org/pdf/rfc5598.pdf, (25-Mar-2011).

Cherry, S.M. (2001), "Remailers Elude E-mail Surveillance", IEEE Spectrum, 38 (11), p.69 2001, 10.1109/MSPEC.2001.963268.

Dusi, M., Gringoli, F. Salgarelli, L. (2008), "A Preliminary Look at the Privacy of SSH Tunnels," in Proceedings of the 17th IEEE International Conference on Computer Communications and Networks (ICCCN 2008), (St. Thomas, U.S. Virgin Islands), Aug. 2008.

Ehrenkranz, T. and Li, J. On the state of IP spoofing defense. In Proceedings of ACM Trans. Internet Techn.. 2009.

Emmanuel S Pilli, R C Joshi and Rajdeep Niyogi, (2010), “A Generic Framework for Network Forensics”, International Journal of Computer Applications 1(11), February 2010, pp.1–6.

Hastings, N. E, McLean, P. A., (1996), “TCP/IP spoofing fundamentals”, In Proceedings of the IEEE 15th Annual International Phoenix Conference; 1996. pp. 218-224.

Marwan A. Z., (2004), “Tracing E-mail Headers”, Proceedings of Australian Computer, Network & Information Forensics Conference on 25th November, School of Computer and Information Science, Edith Cowan University Western Australia 2004, pp. 16-30.

Natarajan, M., Reddy, S., Allam, Moore, L. A. (2009), Tools and Techniques for Network Forensics, International Journal of Network Security and its Applications, 1(1), April 2009, pp. 14-25. Available online at: http://airccse.org/journal/nsa/0409s2.pdf.

Landsiedel, O., Niedermayer, H., Wehrle, K. (2005), An Infrastructure for Anonymous Internet Services, In IWI2005, Chiba/Tokyo, Japan, May 2005.

Radvanovsky, B. (2006), “Analyzing spoofed e-mail header”, Journal of Digital Forensic Practice, 1(3), 2006, pp. 231-243.

Resnick, P. editor, (2001), “Internet message format”, Internet Engineering Task Force (IETF); 2001. RFC 2822.

Shue, C. A., Gupta, M., Lubia, J. J., Kong, C. H., Yuksel, (2009), “A. Spamology: A study of spam origins”, In the 6th Conference on Email and Anti-Spam (CEAS) (2009).

Shunman, W., Ran, T., Yue, W., Ji, Z., (2003), “WLAN and its security problems”, In 4th International Conference on Parallel and Distributed Computing, Applications and Technologies (PDCAT2003), 2003, pp. 241–244.

Vivek S. P., Wang, L., Park, K., Pang, R., Peterson, L. (2004), “The dark side of the Web: an open proxy's view”, SIGCOMM Comput. Commun. Rev. 34, 1 (January 2004), 57-62. DOI=10.1145/972374.972385. Available online at: http://doi.acm.org/10.1145/972374.972385.

Wong, M., Schlitt, W. (2006). Sender Policy Framework (SPF) for Authorizing Use of Domains in E-MAIL, version 1. Internet Engineering Task Force (IETF), RFC 4408.