Internet Technologies in Depth. The Technique of Spam Recognition Based on Header Investigating

Dr. Abzetdin Adamov
Chief Information Officer / Head of Computer Engineering Department,
Qafqaz University, Baku, Azerbaijan
aadamov@qu.edu.az

Abstract – E-mail is most effective business and personal communication tool. The popularity, openness and wide availability of this Internet service makes it attractive for advertising of products and services by sending unsolicited e-mails (Spam). The goal of paper is to offer a comprehensive and usable technique to recognize spam that helps to detect and protect users from junk email, fraudulent e-mail threats and viruses. While widespread methods are complex and expensive, proposed technique is based on header investigating without additional tools and hard processing.

Keywords - Internet technologies, e-mail architecture, spam, spam recognition,

I. Internet Message as Communication Tool and Spam

The asynchronous nature of e-mail provides convenience and more effective use of time for communication participants. In contrast to immediate communication means like telephone, email is deferred type of communication. So, instead of immediate reaction, recipients now have the comfort to read, interpret and react on received information later, or do nothing if no action is required [1].

Because of mentioned and other advantages of email communication, the popularity of email as the communication means for business and personal use has risen steadily over the last decade. The following Figure 1. shows rising popularity of the email communication over the last years and some prediction for future.

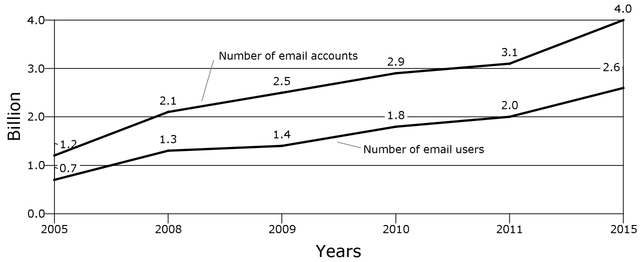

Fig. 1. Email using progress by years

At the same time, because of Internet message concept and architecture, the most important features of which are simplicity, openness, compatibility, standardization, this service is vulnerable for security threats. The most prevalent of them is spam or unsolicited e-mails. The spam continue to evolve becoming more complex and making harder to stop it. Today, spam is not just unwanted e-mail, it's also security problems, viruses, phishing, and other malware. Fortunately, recent years computer professionals and business have intensified the fight against spam and spammers. And yet it is still a serious problem which costs businesses tens of billions and continues to rise from year to year [2]. There are several popular methods for spam detection and prevention like email filtering based on the content of the email, DNS-based blackhole lists (DNSBL), greylisting, spamtraps, etc. [3]. However, most of them require advanced knowledge, special software, or a lot processing time. In contrast to them, technique proposed in this paper does not require any software or special experience. This technique makes it possible to examine e-mail using any e-mail client software, or even with webmail (gmail, yahoo mail). Since this technique is based on email header investigating, it's necessary to observe e-mail architecture, format and meaning of e-mail headers.

II. Email General Structure

As other Internet services the email system based on Internet standards and some dedicated protocols. There are a lot of different email protocols implemented by different email servers, however, some of them are common for all email servers and email clients:

1. Basic email format standard [7] (RFC 5322)
2. Multipurpose Internet Mail Extensions (MIME) standard
3. Simple Message Transfer Protocol (SMTP)
4. Post Office Protocol (POP3) or Internet Message Access Protocol (IMAP)

III. EMAIL PHYSICAL ARCHITECTURE AND PROTOCOLS

The general email architecture consists of two core components and protocols those enable transfer of the electronic text messages between them. The first component is the Email Agent (or email client), which allow users to receive, read, create, and send email. The second component is the Mail Server (or Message Transfer Agent), which is responsible for a message delivery from the source to the destination. As it was mentioned above, there are two key protocols of email system. The SMTP protocol determines the process of message transferring from the source mail server to the destination mail server. The POP3 (or IMAP) protocol defines the process of message retrieval from destination (receiver) mail server to the client's email agent. The software applications developed in accordance to these protocols were named as SMTP and POP3 Internet services, and they actually, form Mail Server itself. The interaction of email components and protocols enable this interaction is show in Figure 2.

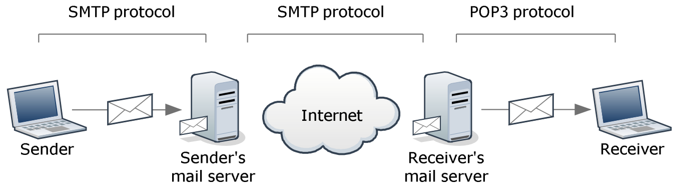

Fig 2.General architecture of email system and protocols

IV. HOW EMAIL DELIVERY WORKS

The email delivery is a whole process of massage transfer from the source to the destination. The Figure 3. shows this process in detail. Let see the process step by step:

1. Using email agent the sender is submitted email for smith@b.com.
2. The SMTP service of the mail server received sender's message resolves the email domain "b.com". To do so the mail server using DNS service (see DNS resolving at [4]) asks the NS server of b.com for the MX record. The MX record specifies the mail server, which is destined to gets all emails with domain name b.com. The name of such a male server is in our example is mail.b.com.
3. Email is routed to the receiver's mail server mail.b.com.
4. The SMTP service of mail.b.com places the email into recipient's mailbox "smith" in the mail store.
5. The recipient checks for email for user smith@b.com using the POP3 service of his email agent. To be able to access to mailbox user has to pass authentication process of the POP3 service.
6. If the authentication module accepts eligibility of the user, the email is downloaded to the user's email agent.

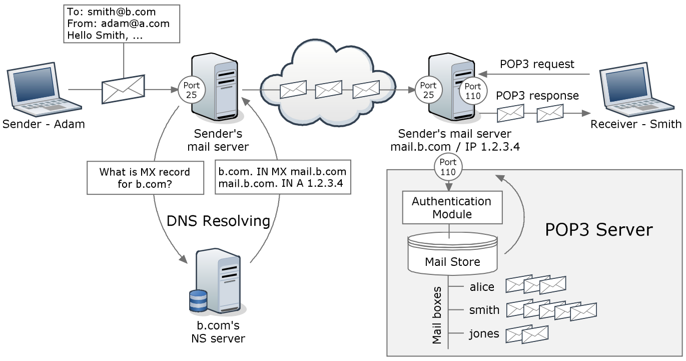

Fig 3. Detailed structure of email delivery

V. THE INTERNET MASSAGE (EMAIL) FORMAT

The first Internet message standard was described by [5] in 1977, which was renewed by [6] in 1982 had been using for almost twenty years. The newest email standard is described in [7] was published in 2008.

According to the last standard the Internet message (or email) consists of an envelope and content (for further more information see [8]). This is illustrated in Figure 4. "a". The envelope, which is part of SMTP protocol, can be viewed as container of message and has information about from whom the message originated (sender) and to whom it is destined (recipient or list of recipients). The existence of sender's information is necessary to be able to send back the error message if the message delivery is failed. The envelope is a temporary container created by source mail server just before passing the message to the destination mail server, as is shown in Figure 4. "b". By the time a message has been delivered to a recipient's mailbox there is no envelope.

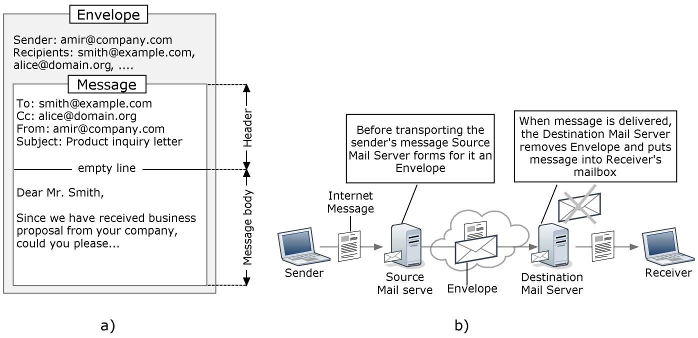
Envelope
Sender: amir@company.com
Recipients: smith@example.com,
alice@domain.org, ...
Message
To: smith@example.com
Cc: alice@domain.org
From: amir@company.com
Subject: Product inquiry letter
empty line
Dear Mr. Smith,
Since we have received business proposal from your company, could you please...
Header
Message body
Before transporting the sender's message Source Mail Server forms for it an Envelope
When message is delivered, the Destination Mail Server removes Envelope and puts message into Receiver's mailbox
Internet Message
Sender
Source Mail serve
Envelope
Destination Mail Server
Receiver
a)
b)

Fig 4. Email format and envelope concept

There is no inherent relationship between recipients' addresses in the envelope and the addresses in the header section (such as To, Cc, Bcc), however according to [8] (RFC 5321) appropriate header fields can be used to form recipient(s) list. That can be imagined like the postal mail with the destination address on the envelope, at the same time it may have address on the top of the message within envelope, which does not make sense for delivery. It is why, sometimes recipient receives message even if he can't find his address within the recipients' list.

VI. Email Header Investigating and Spam Recognition

The content of email includes header fields and message body. The meaning of the header fields is to provide receiver's email agent with descriptive information about message, such as sender, receiver, date, subject, etc. The header block contains several textual lines each of which presents syntax: "header title: value" (look at Figure 4. "a"). The body separated from header fields by empty line, contains textual information the sender is sending to the recipient. The primary header fields specified by [7] (RFC 5322) are shown in Table 1.

TABLE I
Internet Message Header Fields

Header	Description
From:	The name and email address of the message originator
Date:	The local date and time when the message was written or sent
Message-ID:	Machine readable unique identifier generated by mail server; designated to prevent multiple delivery, and to use as reference in In-Reply-To
In-Reply-To:	Used for reply messages only, and contain Message-ID of the original message(s), creating relational tree of messages
To:	Email address(es) of the primary recipient(s)
Cc:	Email address(es) of the secondary recipient(s). Generally, used to indicate recipients whose don't have immediate relation to the matter, however should be informed
Bcc:	Same as Cc, but hidden from recipients. SMTP removes this header field before delivering of the message
Subject:	Textual human readable summary of message
Content Type:	MIME type of the message content, designed for email agent to display message properly
Received:	Contain information about all mail servers that were involved in the message delivery
References:	Like In-Reply-To uses Message-ID(s), but designed to identify a thread of correspondence
Keywords:	Keywords specified by sender
Reply-To:	Email address should be used when recipient replies to message
Return-Path:	This header indicates the email address of message's sender. The value of this header has to be same as "From" address of the SMTP Envelope
Delivered-To:	The email address of recipient
Sender:	Actual sender of the message (generally, used address listed in the From)

The level of importance of each header field in message formation is different. For example, any internet message must include *From:* and *Date:* fields, and should include *Message-ID:* and *In-Reply-To:*. The rest of fields are optional or are managed automatically by mail servers. The one of the most important headers *Received:* is deserved to be reviewed in more detailed way. This header significantly simplifies the fight against spam and spammers. When we receive unsolicited bulk email, our email agent program normally shows only the standard *To:, From:, Subject:*, and *Date:* headers, as for any other email. At the same time, the *From:* address may appear to be from someone we well know, or from some organization whose name we respect or trust. In reality these spoofed messages do not originate from the address that appears in the *From:* header. To see the real address message was sent from, it is necessary to control *Received:* filed, which tells us the route the message took when it was sent to us.

Now we will try to understand how to find original source of the suspicion email through investigating the email header. To do so, firstly we need to be able to see the full email header. Generally, all email client programs (even webmail services like Gmail, Yahoo, etc.) have appropriate function to display full header of any message in your inbox. Let see the header of message I have received recently is shown in Figure 5.

```
1.   Delivered-To: my.address@gmail.com
2.   Return-Path:
     <SRS0=M78ycc=RT=p3slh174.shr.phx3.secureserver.net=
     lindaadleen2@qafqaz.edu.az>
3.   Received: ........................

4.   Received: by 10.220.162.197 with SMTP id w5cs344529vcx;
       Sun, 17 Oct 2010 05:24:20 -0700 (PDT)
5.   Received: from bosmailscan05.eigbox.net ([10.20.15.5])
       by bosmailout03.eigbox.net with esmtp (Exim) id
     1P7SHj-0007rH-Qy
       for www.adamov@gmail.com; Sun, 17 Oct 2010 08:24:19 -
     0400
6.   Received: from p3slh174.shr.phx3.secureserver.net
     (localhost.localdomain [127.0.0.1])
       by p3slh174.shr.phx3.secureserver.net
     (8.12.11.20060308/8.12.11) with ESMTP id o9HCOF7n030063
       for <aict2011@qafqaz.edu.az>; Sun, 17 Oct 2010
     05:24:15 -0700
7.   Received: (from lindaadleen2@localhost)
       by p3slh174.shr.phx3.secureserver.net
     (8.12.11.20060308/8.12.11/Submit) id o9HCOEvK030054;
     Sun, 17 Oct 2010 05:24:14 -0700 Date: Sun, 17 Oct 2010
     05:24:14 -0700
8.   Message-Id:
     <201010171224.o9HCOEvK030054@p3slh174.shr.phx3.
     secureserver.net>
9.   To: aict2011@qafqaz.edu.az
10.  Subject: xxxxxxxxxxxxxxxxx!!!!!
11.  From: vangelis@mail.ru
```

Fig. 5. Email header investigation to find the original source of spoofed message

The header has been slightly modified by removing most eleven *Receive:* fields. The *Receive:* headers appear in reverse order. So, the first *Receive:* header from bottom (see line 7) presents the original source of the message. The line “from lindaadleen2@localhost” shows information about computer the message was sent from. Probably, spammer uses SMTP service installed on his computer in order to send bulk mail. The next line shows the name of the first mail server involved in delivery process “p3slh174.shr.phx3.secureserver.net”, the exact date and time of receiving, and unique id assigned by server to this message. The id is unique for particular mail server and can be used for tracking of the message. The two headers *To:* (see line 9) indicates to whom the message is sent and *Delivered-To:* (see line 1) indicates by who it is received, are supposed to be same. Furthermore, other two headers *From:* (see line 11) and *Return-Path:* (see line 2) are also supposed to be same. The fact that they are not same testifies the spam nature of the message.

CONCLUSION

The increasing popularity of e-mail-based communication without significant change in architecture makes this tool vulnerable to many styles of attack. In order to enhance the reliability of email, it is crucial to be able to verify addresses in *From:* and *To:* headers. The verification method based on headers investigation makes it possible to distinguish wanted email from spam (junk, bulk, unsolicited) email with quite high level of accuracy.

REFERENCES

[1] Value-Added Services for Next Generation Networks, Thierry Van de Velde, Auerbach Publication, 2008

[2] Ferris Research: Cost of Spam, http://www.ferris.com/research-library/industry-statistics/

[3] Shawn Hernan; James R. Cutler; David Harris (1997-11-25). "I-005c: E-Mail Spamming countermeasures: Detection and prevention of E-Mail spamming". Computer Incident Advisory Capability Information Bulletins. United States Department of Energy. Retrieved 2007-01-06.

[4] Abzetdin Adamov, Neglected point of Internet performance. How to choose the right DNS service, http://aadamov.wordpress.com/

[5] RFC 733, Standard for the format of ARPA network text messages, 21 November 1977, http://www.ietf.org/rfc/rfc0733.txt

[6] RFC 822, Standard for the format of ARPA internet text messages, August 13, 1982, http://www.ietf.org/rfc/rfc0822.txt

[7] RFC 5322, Internet Message Format, October 2008, http://tools.ietf.org/html/rfc5322

[8] RFC 5321, Simple Mail Transfer Protocol, October 2008, http://tools.ietf.org/html/rfc5321)