Email Analysis

Client and Web

Moses Schwartz

26 September 2006

CS489 - Digital Forensics

New Mexico Tech

Executive Summary

Analysis of email is especially important not just because email may be used to communicate about things that we might be interested in for an investigation, but because it is a comparatively permanent and public record of those communications.

Data mining techniques are being applied to email analysis -- both large data sets, and smaller ones -- and are being used for applications such as visualizing social networks, classification of messages, and identification of authorship.

Analysis of an individual's email is fairly straightforward, but data mining techniques can still be applied to simplify the task.

Webmail presents an additional difficulty for the forensic investigator, as the email is no longer available on a suspect's hard drive. However, there are typically traces that indicate that a user is using webmail. In extreme cases, webmail providers have been subpoenaed to provide access to email accounts, and have complied.

Analyzing a single email is an interesting task. The email header is the primary source of information, and can be supplemented by server logs. Because email headers can easily be forged, special measures should be taken to determine the validity of email headers.

Table of Contents

Client and Web....1
Executive Summary....2
Table of Contents....3
Introduction....4
Data Mining....4
Analysis of an Individual User's Email....5
Analysis of Webmail....5
Analyzing an Individual Email....5
Future Research....8
References....9

Introduction

Analysis of email is especially important not just because email may be used to communicate about things that we might be interested in for an investigation, but because it is a comparatively permanent and public record of those communications. In the case of a phone call, there is only the record that a call took place; in a spoken conversation, there may be no record at all. Conventional mail can be virtually untraceable, and paper documents are easily destroyed. Email, however, is unique; when a message is sent, the entire message is stored for both the sender and the receiver, and records of the mail being sent are stored on dozens of servers that the message passes through before arriving at its destination. There are a number of ways to analyze email, including: data mining techniques, which may be applied to large or small data sets; straightforward searching of a user's email for certain content; and in-depth analysis of an individual email's lineage.

Data Mining

In general, large bodies of information may be used – data mined – to extract useful information. Email, in particular, is a unique body of information with many aspects that may be mined. During the Federal Energy Regulatory Commission's (FERC) investigation of Enron, FERC publicly released a very large data set of email messages from 158 Enron employees. This data set, dubbed the Enron corpus and now freely available online, contains over 600,000 messages. Obviously, sorting through such a large quantity of emails would be largely unfeasible without some sort of automated process. Such a large data set is the perfect target for data-mining. Social networks can be identified from the flow of email; this information has a number of uses, from increasing efficiency (identifying and removing communication bottlenecks in an organization) to identifying criminal organizations. The content of the email can be mined for such things as author identification, or classification of the emails to simplify the task of analyzing a large body of messages.

The Email Mining Toolkit (EMT) is a data mining tool created by researchers at Columbia University. The toolkit, which is available only with authorization from the authors, is targeted to provide support to law enforcement. There are also many commercial tools to perform data mining on emails, as well as companies that will provide data mining services.

Analysis of an Individual User's Email

The analysis of a single user's email is the most straightforward and easy to understand of the types of analysis mentioned in this paper. Although data mining techniques can (and will) still be applied to an individual user's email, there is likely to be less need – the data set is small enough that it can be feasibly searched manually. Of course, manually searching through a large number of emails is still likely to be distasteful to the majority of investigators. The EMT, discussed above, may be used in this application. There is a wide selection of commercial tools to analyze a single user's email. In cases where email is stored in a text-like format, Grep and Strings are also an effective method of analyzing emails.

Analysis of Webmail

Webmail poses a slight challenge to a forensic investigator because the emails are not conveniently archived and stored on an individual's hard drive. However, webmail will almost always leave traces on the system that accessed the webmail site. Most notable is that the web browser cache may contain messages that were read on the webmail system. However, many browsers are set by default to not cache SSL-encrypted pages (which most webmail sites are), rendering the technique of examining browser cache obsolete. There may, however, still be evidence that the browser accessed the webmail site. Unencrypted pages prior to the login page may be stored in cache, and unless the user is particularly savvy or paranoid, having saved cookies from the site is almost a given. Browser history is also likely to show evidence of the use of webmail. Bookmarks or "Internet Favorites" (in Internet Explorer) are also a good indication that a browser may be being used for webmail.

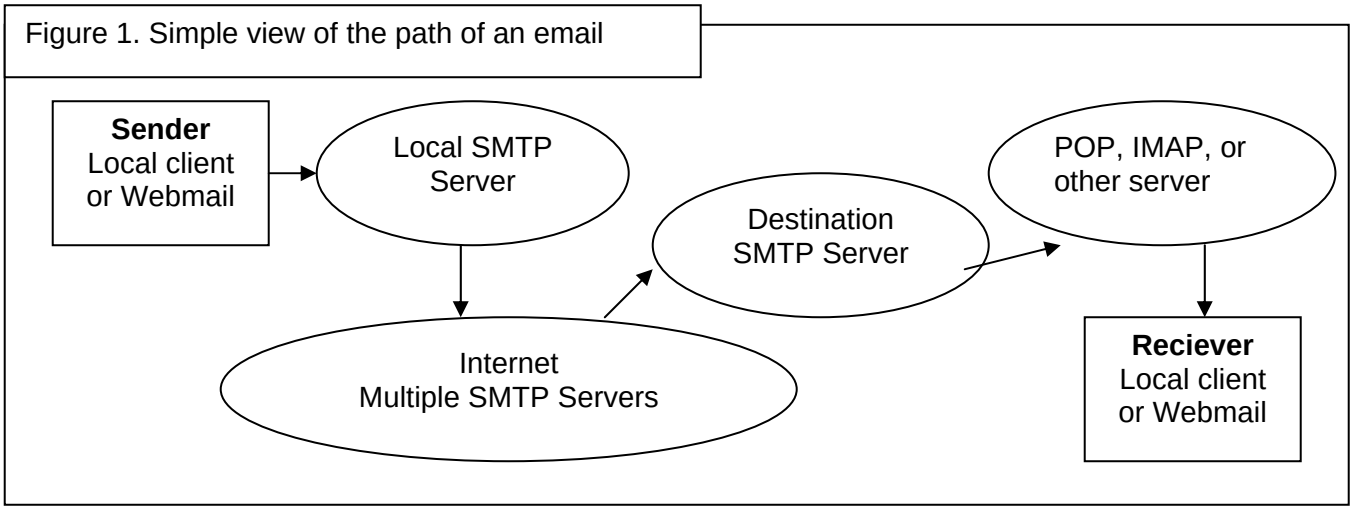
Although the content of the emails may not be immediately available, it is worthwhile to keep in mind that in a criminal case or civil case that goes to court, it may be possible to subpoena the webmail host for access to an account and relevant records.

Analyzing an Individual Email

Although webmail will feature prominently in this section, the analysis of a particular email's lineage is much broader and can be applied to any email. A simple view of the path of an email from a sender to a client is presented in Figure 1. The email originates from the sender, whether from a local email client or a webmail application. When the email is sent, it is first sent to a Simple Mail Transfer Protocol (SMTP) server. That server forwards it to other SMTP servers until it finally reaches the destination server. On reaching its destination, the email is

sent to a Post Office Protocol (POP) server, or any number of similar mail-delivery servers (IMAP is another, and webmail services may use their own servers for this purpose). The receiving client then connects to that server, retrieves the message, and allows the recipient to read it.

Figure 1. Simple view of the path of an email



When the email is sent and when it is received, those respective servers add their own information to the email's header, and most likely log the action. Access to those logs may be required for much analysis, but specifics are outside of the scope of this paper. Considerable information can be gleaned from the header alone.

Suppose Moses, with the address moses@nmt.edu, sends an email from his office on the New Mexico Tech campus to his similarly named friend, with the email address thenewmoses@gmail.com. The subject of this email is "Snakes," and the content "Fish." Below is the entire theoretical email, including all headers.

From: moses@nmt.edu
Subject: Snakes
Date: September 25, 2006 9:35:29 PM MDT
To: thenewmoses@gmail.com
X-Gmail-Received: ca493ed685a8e9ae77165ab2ce345127e5b310b4
Delivered-To: thenewmoses@gmail.com
Received: by 10.90.33.15 with SMTP id g15cs279684agg; Mon, 25 Sep 2006 20:35:32 0700 (PDT)
Received: by 10.35.113.12 with SMTP id q12mr526602pym; Mon, 25 Sep 2006 20:35:32 -0700 (PDT)
Received: from mailhost.nmt.edu (mailhost.NMT.EDU [129.138.4.52]) by mx.gmail.com with ESMTP id 36si2059018nza.2006.09.25.20.35.32; Mon, 25 Sep 2006 20:35:32 -0700 (PDT)

Received: from localhost (localhost.localdomain [127.0.0.1]) by localhost.localdomain (Postfix) with ESMTP id 09FF4436164 for <thenewmoses@gmail.com>; Mon, 25 Sep 2006 21:35:32 -0600 (MDT)
Received: from mailhost.nmt.edu ([127.0.0.1]) by localhost (mailhost.nmt.edu [127.0.0.1]) (amavisd-new, port 10024) with ESMTP id 11225-05 for <thenewmoses@gmail.com>; Mon, 25 Sep 2006 21:35:30 -0600 (MDT)
Received: from [192.168.1.2] (cs-fitch017.nmt.edu [129.138.21.110]) by mailhost.nmt.edu (Postfix) with ESMTP id 6FD4B436030 for <thenewmoses@gmail.com>; Mon, 25 Sep 2006 21:35:30 -0600 (MDT)
Return-Path: <moses@nmt.edu>
Received-Spf: pass (gmail.com: best guess record for domain of moses@nmt.edu designates 129.138.4.52 as permitted sender)
Mime-Version: 1.0 (Apple Message framework v752.2)
Content-Transfer-Encoding: 7bit
Message-Id: <77E313EF-271F-4AD0-A8D3-81263BF7B083@nmt.edu>
Content-Type: text/plain; charset=US-ASCII; format=flowed
X-Mailer: Apple Mail (2.752.2)
X-Virus-Scanned: by amavisd-new-2.3.1 (20050509) (RHEL AS) at nmt.edu

Fish

From this header, we can determine a considerable amount about the sender. The `Date:` field is set by the sender's email client, and can easily be forged. However, by the time the email has been delivered, a number of `Received:` fields have been added, showing exactly the time that the receiving servers got the message. In each of those `Received:` fields there is also an SMTP ID – this number is "a unique identification assigned by each intermediate relay or gateway server." The SMTP ID is changed often (every day or more), with the result that we can determine approximately when a server received the message from this ID alone. Similar to the SMTP ID is the Authentic Message-ID String, which is in the `Message-Id:` field. This unique ID makes it possible to identify this particular message in logs on any server that handles the message and logs the event. Also note that the X-Mailer: field indicates what email client the sender is using, which also gives us clues about the sender's operating system. Perhaps most useful for identifying the sender – certainly more useful than the `From:` line, which is easily forged – is the final `Recieved:` field, which tells us the sender's IP address. Not only does it tell us the IP address of the sender (129.138.21.110), but also the subdomain (cs-fitch017), and the sender's address on his own LAN. With a little bit of background information, we could easily determine from the subdomain that the email originated in room 017 of Fitch Hall, and that this room is being used by the computer science department.

The headers from emails sent through webmail do not have as much personally identifiable information, but many of the fields are more reliable (that is, it's harder to forge a header when

using a webmail system), and most webmail systems include the IP address of the sender in their headers. This IP address can be obfuscated or hidden through the use of open proxies or, worse, anonymous proxies, but this is unfortunately beyond the scope of this five-page paper. Forgery of headers was alluded to in our sample analysis several tiems, but was not touched on in any great detail, nor will it be anywhere in this paper. There are numerous ways to forge email headers, all of which a forensic investigator must be aware of.

Future Research

As email analysis techniques develop, especially on the data mining front, we can expect to see new and more effective systems for the extraction of data from email sets, as well as the classification of email content. There has been research into natural language processing on the Enron corpus, and even incremental developments may prove useful to forensic investigators. Social network analysis, using email analysis to determine social relationships, also seems like an interesting area.

References

Akin, Thomas, "WebMail Forensics," Presented 2003 at BlackHat.

de Vel, O., "Mining E-mail Content for Author Identification Forensics"
O. de Vel
SIGMOD Record, Vol. 30, No. 4, December 2001

Farwell, L.W., "Email Forensics," Presentation.

Klimt, B., Yang Y. "Introducing the Enron Corpus," Language Technology Institute, Carnegie Mellon University, Pittsburgh, PA. http://www.ceas.cc/papers-2004/168.pdf

Krebs, Valdis. "Data Mining Email to Discover Social Networks and Emergent Communities." 2003. http://www.orgnet.com/email.html

Savatore, S., "Email Mining Toolkit Supporting Law Enforcement Forensic Analyses," Columbia University, New York, NY