IP geolocation suspicious email messages

Asmir Butković, Saša Mrdović, Samra Mujačić

Abstract — **As the Internet and electronic mail continue to be utilized by an ever increasing number of users, so does fraudulent and criminal activity via the Internet and email increase. The negative effects of cyber crime activities on the use of the internet for e-business and secure communications increased interest in studying the factors that motivate these criminals, their tactics and what can be done to mitigate their activities. The research in the area of email analysis usually focuses on two areas, email traffic analysis and email content analysis and very little in the area of visual analytics of emails. We have developed a solution to visualize suspicious email messages based on the information provided in the email header (rather than the content of the email). IP mapping tool, called MIPA, uses a Google Map to display the geographic position and integrates InfoDB, WhoIS databases, and the Google Maps API.Thus, the proposed work can be helpful for identifying and investigating suspicious email messages and also assist the investigators to get the information in time to take effective actions to reduce the criminal activities.**

Keywords — **Cybercrime Investigation, Email client, IP geolocation, Maps API**

I. Introduction

Cybercrimes share some similarities with crimes that have existed for centuries before the advent of the cyber space. The major difference is that the internet now provides an electronic platform with the advantages of speed, anonymity and a tool which increases their potential pool of victims. [7] The rapid expansion of computer connectivity has provided opportunities for criminals to exploit security vulnerabilities in the on-line environment. Most detrimental are malicious and exploit codes that interrupt computer operations on a global scale and along with other cyber-crimes threaten e-business.

Email has become one of today's standard means of communication. The large percentage of the total traffic over the internet is the email. Email data is also growing rapidly, creating needs for automated analysis. So, to detect crime a spectrum of techniques should be applied to discover and identify patterns and make predictions. As individuals increase their usage of electronic communication, there has been research into detecting deception in these new forms of communication.[4] The Received lines in email headers contain the list of IP addresses that email has flowed through, as it is passed from one server to the next. Most researchers outlines that IP addresses present in the email headers are the most important tools in fighting email related crimes.

The SMTP protocol specifies that each SMTP relay used to send an email message must add at the beginning of the message's header list a “received” line that contains (at least) information about the SMTP server receiving the message, from where the server received the message, and a timestamp stating when the header was added. These header lines, taken together, provide a trace of the SMTP path used to deliver a message. [10] POP3 (Post Office Protocol) is an email protocol used to retrieve the email messages from the email server to the email client.

Each email can contain more than one “Received” field. This field is typically used for email tracking by reading it from bottom to top. The bottom represents the first mail server that got involved in transporting the message, and the top represents the most recent one, where each received line represents a handoff between machines. Hence, a new received field will be added on the top of the stack for each host received the email and transport it, and to which host the message will be delivered, in addition to the time and date of passing. [6] The email source and ip address location is important as a first line of defence for emailers to know how to track email origin.

IP Geolocation allows us to assign a geographical location to an IP address allowing us to build up a picture of the person behind that IP address. This can have many potential benefits for business, offer much to those in security and other types of application. There are organisations that are responsible for allocating IP address. The Internet Assigned Number Authority(IANA) is responsible for allocating large blocks of IP addresses to the following five Regional Internet Registries(RIR) that serve specific regions in the world: AfriNIC (Africa), APNIC (Asia/Pacific), ARIN (North America), LACNIC (Latin America) and RIPE NCC (Europe, the Middle East and Central Asia). These RIR's then allocate blocks of IP addresses to Internet Service Providers (ISP) who then allocates IP addresses to businesses, organizations and individual consumers.

Using the above information IP addresses can be broken down into graphical locations within few steps but to get a more accurate result than that the user may have to provide additional details to aid the process. [2]

The rest of the paper is organized as follows. In Section II, we describe related work of authors from these areas. Section III explains the problem formulation.

Asmir Butković, Police Support Agency of Bosnia and Herzegovina, Sarajevo, Bosnia and Herzegovina, asmir.butkovic@psa.gov.ba

Saša Mrdović, University of Sarajevo/Faculty of Electrical Engineering, Sarajevo, Bosnia and Herzegovina, sasa.mrdovic@etf.unsa.ba

Samra Mujačić, University of Tuzla, Faculty of electrical engineering, Tuzla, Bosnia and Herzegovina, samra.mujacic@untz.ba

The next section describes the proposed approach, presents our software solution and the implementation details of each component. Section V concludes the paper with future work.

II. Related Work

In this [1] paper, authors presented IP mapping tool called MIPA, that mapping IP addresses to geographic coordinates based on information retrieved from RIPE and InfoDB database using Google Maps API functions. The results indicate that this solution can be used effectively in the process of cybercrime investigation and of great help to law enforcement agencies.

In paper [2] authors have provided an overview of IP Geolocation applications and methodologies both traditional and those that attempt to push the envelope. The methodologies presented here vary both in their complexity and accuracy; as such, and they don't claim any one method as the ideal solution. The optimal approach is therefore highly sensitive to the type of application being developed.

This paper [4] proposes to apply Association Rule Mining for Suspected Email Detection.(Emails about Criminal activities). Deception theory suggests that deceptive writing is characterized by reduced frequency of first person pronouns and exclusive words and elevated frequency of negative emotion words and action verbs . They apply this model of deception to the set of Email dataset, then applied Apriori algorithm to generate the rules.

In this [6] paper, authors evaluated the performance of several machine learning-based classifiers and compared their performance in filtering email spam based on email header information. These classifiers are: C4.5 Decision Tree (DT), Support Vector Machine (SVM), Multilayer Perception (MP), Nave Bays (NB), Bayesian Network (BN), and Random Forest (RF). They adopted header-based email spam filtering by including additional header information features that found to be of a great importance to improve the performance of this technique.

III. Problem Formulation

Though email was originally developed for sending simple text messages, it has become more robust in the last few years and preferred form of communication. The ease, speed and relative anonymity of email has made it a powerful tool for criminals and one possible source of data from which potential problem can be detected.

The question is simple, how to determine from where an email message has originated ?

With the recent surge in cyber attacks, there is a growing demand for effective security analytics tools. Though, there are advanced data collection techniques in the form of honeypots and malware collectors, the value of data are only as useful as the analysis technique used.

One of the primary drawbacks of current security analytic tools is the lack of visualization controls to effectively analyze the data. [8] Simple Mail Transport Protocol (SMTP) is the standard transport protocol used today to send mail across the internet. Most email systems that send mail over the Internet use SMTP to send messages from one server to another; the messages can then be retrieved with an email client using either POP3 or IMAP.

The header section contains a lot of information relating to the mail. Typically an email passes through at least four relays before reaching the recipient’s inbox. During the SMTP transactions the hostname and IP addresses of the relays participating in transmission of the email are recorded in the “Received:” lines in the email header, so it contains the IP addresses of all servers involved in routing email from one point to another. These lines when read from bottom to top provide the path traversed by an email from the sender to the recipient. With this addresses a user can track an email origin and possibly the identity of the sender. An example line might look like the following:

Received: from mail3.mcsignup.com ([173.231.184.51]) by COL0-MC2-F47.Col0.hotmail.com with Microsoft SMTPSVC(6.0.3790.4900); Wed, 25 Sep 2013 09:59:42 -0700

This example line says that the machine COL0-MC2-F47.Col0.hotmail.com received the message from the mail3.mcsignup.com. As we have said, the key line is the first one internal to the recipient's organization, which gives the IP address of the outside machine that delivered the message across the internet to the recipient's organization. Identifying and processing this line in clients, extract the IP addresses of the sender is one of the goals of this paper.

Next problem is determining the geographic location of this IP address or sender's suspicious email message.IP Geolocation is the process of obtaining the geographical location of an individual or party starting out with nothing more than an IP address. WHOIS provides publicly available information that allows one to query a remote WHOIS database for registration information of a domain name. Generally, a WHOIS record contains a full name, address, telephone number and email address of the Internet Service Provider (ISP). A WHOIS search accepts IP address as an input for querying. In this way, it forms a relationship between the owner of an IP address and its ISP. There are a lot of online services which provides mappings between an IP address and a geographical location of its Internet Service Provider (ISP) (including country and city). It can be efective to locate email messages from it IP address. [11]

Google Maps API is ideal candidate for this task, due to its simple integration and it is free for commercial use providing that the site on which it is being used is publicly accessible and does not charge for access.

We have available MIPA software tool that has integrated all of the functionality necessary to determine the geographic location of various IP addresses.

IV. Problem Solution

Since the focus of this paper is on the analysis and processing of suspicious email messages we have developed an additional module for MIPA application that represents some kind of email client. The MIPA software tool has already had a well-realized solution for IP address geolocation approach to solving this problem is to focus on finding the origin of email messages and their visualization. The functionality of this new part of the MIPA system is described below.

The proposed system first extracts useful features from the email header. The POP3 protocol is used to retrieve the email messages from the mail server to the our application (as email client program). Using the POP3 protocol email messages are carried out from the mail server and passed over through the network to the recipient's mailbox (our program). These processes are done through client/server command dialogue. The POP3 client/server communication procedure is achieved through a set of steps which can be divided into three main states (AUTHENTICATION state, TRANSACTION state and UPDATE state). [9] The POP3 client/server procedure starts when the email client wants to retrieve it from the mail server. We use hotmail mail account for our demonstration here since hotmail is one of the most popular free email service.

To configure an account in our email client we need to have the following details: the username, the password, the incoming server (pop3.live.com) and the port(995). The sequence is almost the same for accounts of other email services (e.g. for Gmail incoming server:pop.gmail.com and port:995). The procedure starts from our application by creating a new TCP connection between the our application and the email server on port 995.

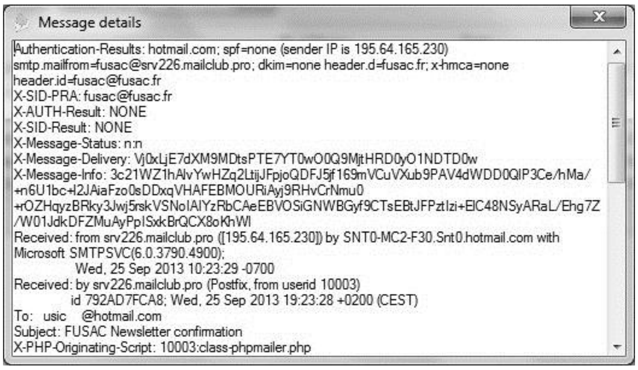
Once the TCP connection is created, the client waits for an acknowledgment from the Email server that contains +OK greeting message which informs the client that the Email server is ready to start commands dialogue between them. In „AUTHENTICATION state", the client must identify himself to the POP3 Email server by using USER and PASS command which contains the username and the password of email account. If the username and the password is correct, which is a positive case, the server must respond using +OK command, or otherwise the server will respond with -ERR command. Once the USER and PASS commands are performed with positive responds from the POP3 server, the client has an access to an appropriate maildrop.

The POP3 client now has entered into the "TRANSACTION State" and can start commands dialogue process. The first command performed from the POP3 client in this state is STAT command. The information sent from the POP3 server contains the number of Email messages in the maildrop and the message size of the maildrop. These information, in our example, are shown in Fig. 1, on the ListBox control with a Label „Email server responses".

The next command performed from the POP3 client is TOP [msg] [lines] command for only retrieving part of the message. The following syntax: *TOP [msg] [0]* command, is used to extract the headers of the email for each email message in the maildrop. (e.g. TOP 1 0 return the headers for message 1)

After that, we extract four features from the email header, namely „From", „Subject", „ Received From" and „Date" for each email message in the maildrop, and add to the first list on the form. To find the actual location from which the email originates we pick the "Received From" IP that is at the bottom of the list on the header view.

The system also has the option to retrieve the email message from the maildrop by sending RETR command with the message's number which requires to be retrieved to the client's mailbox. An example is shown in Fig. 2.



```
Authentication-Results: hotmail.com; spf=none (sender IP is 195.64.165.230)
smtp.mailfrom=fusac@srv226.mailclub.pro; dkim=none header.d=fusac.fr; x-hmca=none
header.id=fusac@fusac.fr
X-SID-PRA: fusac@fusac.fr
X-AUTH-Result: NONE
X-SID-Result: NONE
X-Message-Status: n:n
X-Message-Delivery: Vj0xLjE7dXM9MDtsPTE7YT0wO0Q9MjtHRD0yO1NDTD0w
X-Message-Info: 3c21WZ1hAlvYwHZq2LtijJFpjoQDFJ5jf169mVCuVXub9PAV4dWDD0QIP3Ce/hMa/
+n6U1bc+l2JAiaFzo0sDDxqVHAFEBMOURiAyj9RHvCrNmu0
+rOZHqyzBRky3Jwj5rskVSNoIAIYzRbCAeEBVOSiGNWBGyf9CTsEBtJFPztIzi+ElC48NSyARaL/Ehg7Z
/W01JdkDFZMuAyPpISxkBrQCX8oKhWI
Received: from srv226.mailclub.pro ([195.64.165.230]) by SNT0-MC2-F30.Snt0.hotmail.com with
Microsoft SMTPSVC(6.0.3790.4900);
          Wed, 25 Sep 2013 10:23:29 -0700
Received: by srv226.mailclub.pro (Postfix, from userid 10003)
         id 792AD7FCA8; Wed, 25 Sep 2013 19:23:28 +0200 (CEST)
To:  usic   @hotmail.com
Subject: FUSAC Newsletter confirmation
X-PHP-Originating-Script: 10003:class-phpmailer.php
```

Fig. 2 Example of a message source

The next step is to look up the IP address. In this work, we use the functionality of MIPA tool previously developed and described in paper[1], to do just that.

Matching latitude/longitude coordinates found during this process associate each email address with geographical information and add to the second list on the form (Fig. 1). Geographical information includes country, region, postal code, city, and latitude/longitude coordinates.

Latitude/longitude coordinates are necessary to display the geographical locations on the Google map during visualization.

This system also offers additional features such as mapping individual IP address to geographic location (Fig. 3), mapping the multiple IP addresses at the same time and calculate the distance between them (Fig. 4).

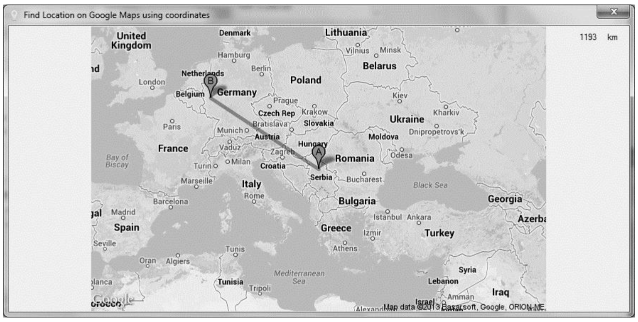


Fig. 3 Example of mapping individual IP address

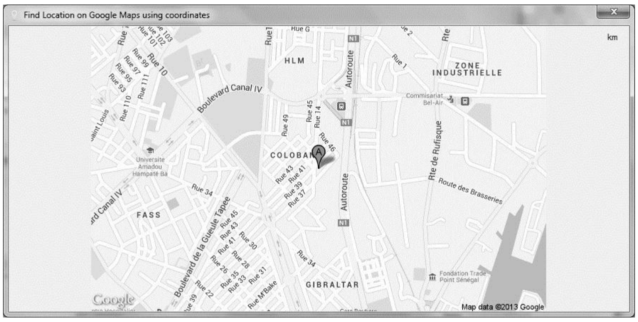


Fig. 4 Example of mapping the multiple IP addresses

V. Conclusion

It is an indisputable role that email has in e-business as well as its importance as a means of communication between individuals and organizations. However, its ease of use, speed and relative anonymity produced that has become a favorite means of communication between criminals and committed a number of different crimes.

Considering the different ways to approach solving this problem, we came to the conclusion that the analysis the origin of email message or the IP address of the sender's e-mail message could help a lot in identifying the persons responsible for these criminal activities. Also, given the fact that the IP geolocation can offer much in security, we tried to combine these two approaches into a single hybrid solution. The result is an improvement of the MIPA software tool that integrate different technologies, data sources and methods of analysis of data in a single product that can be a valuable addition to the arsenal of tools that Cybercrime investigators can use.

As one of the possible ways of expansion of our research we are considering the option of storing a large number of email messages of different email accounts in the database for the general analysis and study of the eventual connection of suspicious email messages sent to different email addresses.

References

[1] A.Butkovic, F.Orucevic,A.Tanovic, “Using Whois Based Geolocation and Google Maps API for support cybercrime investigations”, WSEAS International Conference on Circuits, Systems, Communications, Computers and Applications (CSCCA '13), pp.194-201, June 2013.

[2] Taylor, J Devlin, K Curran, "Bringing location to IP Addresses with IP Geolocation", Journal of Emerging Technologies in Web Intelligence, pp. 273-277, Aug 2012

[3] J. Goodman, “"IP addresses in email clients", Microsoft Research, Redmond, 2004

[4] S. Appavu, M. Pandian "Association Rule Mining for Suspicious Email Detection A Data Mining Approach ", Intelligence and Security Informatics, pp. 316-323, May 2007

[5] S.Nizamani, N. Memon, U.Kock, P. Karampelas, “Modeling Suspicious Email Detection Using Enhanced Feature Selection” , International Journal of Modeling and Optimization, Vol.2, August 2012

[6] O.Jarrah, I. Khater, B. Duwairi, “Identifying Potentially Useful Email Header Features for Email Spam Filtering”, January 2012

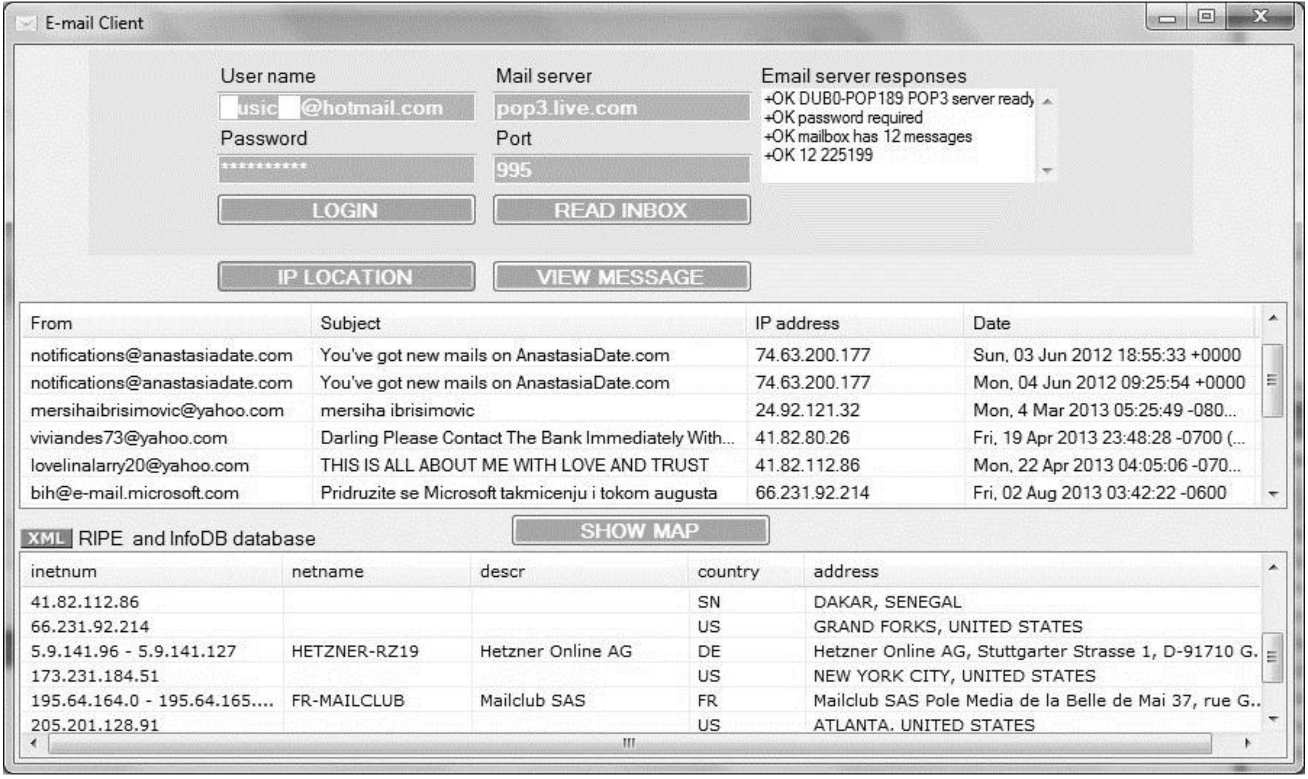
[7] O. B. Longe, V. Mbarika, M. Kourouma, F. Wada, R. Isabalija, “Seeing Beyond the Surface, Understanding and Tracking Fraudulent Cyber Activities”, pp. 124-135, December 2009

[8] A. Muallem, S. Shetty, S. K. Hargrove , “Visualizing geolocation of spam email”, pp. 63-68, April 2013

[9] H. Bazar, R. Sureswaran, O. Abouabdalla, “A new approach to enhance e-mail performance through pop3 protocol”, International Conference on Network Applications, Protocols and Services, November 2008

[10] B. Leiba, J. Ossher, V.T. Rajan, R. Segal, M. Wegman,"SMTP Path Analysis", CEAS 2005

[11] G. Zhang, "An analysis for anonymity and unlinkability for a VoIP conversation", Springer 2010



User name: [redacted]usic[redacted]@hotmail.com

Mail server: pop3.live.com

Password: **********

Port: 995

Email server responses:

```
+OK DUB0-POP189 POP3 server ready
+OK password required
+OK mailbox has 12 messages
+OK 12 225199
```

From	Subject	IP address	Date
notifications@anastasiadate.com	You've got new mails on AnastasiaDate.com	74.63.200.177	Sun, 03 Jun 2012 18:55:33 +0000
notifications@anastasiadate.com	You've got new mails on AnastasiaDate.com	74.63.200.177	Mon, 04 Jun 2012 09:25:54 +0000
mersihaibrisimovic@yahoo.com	mersiha ibrisimovic	24.92.121.32	Mon, 4 Mar 2013 05:25:49 -080...
viviandes73@yahoo.com	Darling Please Contact The Bank Immediately With...	41.82.80.26	Fri, 19 Apr 2013 23:48:28 -0700 (...
lovelinalarry20@yahoo.com	THIS IS ALL ABOUT ME WITH LOVE AND TRUST	41.82.112.86	Mon, 22 Apr 2013 04:05:06 -070...
bih@e-mail.microsoft.com	Pridruzite se Microsoft takmicenju i tokom augusta	66.231.92.214	Fri, 02 Aug 2013 03:42:22 -0600

RIPE and InfoDB database

inetnum	netname	descr	country	address
41.82.112.86			SN	DAKAR, SENEGAL
66.231.92.214			US	GRAND FORKS, UNITED STATES
5.9.141.96 - 5.9.141.127	HETZNER-RZ19	Hetzner Online AG	DE	Hetzner Online AG, Stuttgarter Strasse 1, D-91710 G.
173.231.184.51			US	NEW YORK CITY, UNITED STATES
195.64.164.0 - 195.64.165....	FR-MAILCLUB	Mailclub SAS	FR	Mailclub SAS Pole Media de la Belle de Mai 37, rue G..
205.201.128.91			US	ATLANTA, UNITED STATES

Fig. 1 MIPA Email client interface