Social Network Visualization for Forensic Investigation of E-mail

J. Haggerty, D. Lamb and M. Taylor

School of Computing and Mathematical Sciences, Liverpool John Moores
University, Byrom Street, Liverpool, L3 3AF
e-mail: {J.Haggerty; D.Lamb; M.J.Taylor}@ljmu.ac.uk

Abstract

E-mail features as a key technology for both the dissemination of information and for social networking. Given the volume of e-mail transmission combined with access opportunities, it is not surprising that e-mails feature heavily during a digital forensics investigation. In these investigations, forensic examiners require an understanding of the social networks to which the suspect belongs for both analyzing the event(s) under investigation and to further exploit potential sources of evidence or other suspects. This paper makes use of visual analytic and social network techniques for digital forensics investigations involving e-mail. We present a novel approach, the *E-mail Extraction Tool* (EET), for automated visualization of client-based e-mail applications and exploring the social networks that these will reveal to the investigator. The case study presented in the paper demonstrates the applicability of the approach to digital forensics investigations.

Keywords

Digital forensics, e-mail, visual analytics.

1. Introduction

The information economy's reliance on computers and networked hosts has seen increases in, and the volume of, data that must be analyzed by forensic examiners. One key technology used by many users is e-mail; for dissemination of information and for social networking. The volume of e-mail transmitted over networks is very large. A study of one Internet Service Provider with only 150,000 customers found that just fewer than 12 million e-mails were delivered per day (Clayton, 2007). Given the volume of e-mail traffic today combined with access opportunities, it is not surprising that e-mail features heavily during an investigation of a suspect's computer(s). In addition, individuals involved in group-related criminal activity such as the dissemination of indecent images of children, organized crime, terrorism and fraud will involve e-mail communications.

A number of challenges exist to today's computer forensic investigations involving e-mail. As with many forensic investigations, cases routinely involve more than a single computer (Richard & Roussev, 2006). In addition, due to our reliance on the medium a large volume of data must be analyzed within tight temporal constraints. Identification and presentation of relevant evidence is a time-consuming process using existing tools and techniques. A major challenge facing law enforcement and national security is accurately and efficiently analyzing this growing volume of evidential data (Chen *et al*, 2004). Finally, much of the evidence that is recovered during an investigation may not be analyzed beyond the data representation or data recovery. For example, an analyst will manually trawl through the e-mails relating to an activity under scrutiny to search for those relevant to the investigation. However, they rarely explore the relevant social relationships and networks that these, and other network communications such as 'chat' sessions, will reveal due to the lack of facility in the tools they have at their disposal. These social networks are potentially great sources of interest as they will lead the investigator to other relevant sources of evidence or actors within a crime.

In order to meet these challenges, two areas of inter-disciplinary research are of direct relevance; visual analytics and social network analysis. *Visual analytics* (VA) is the inter-disciplinary science of analytical reasoning supported by interactive visual interfaces (Jern & Banissi, 2006). VA has direct relevance to computer forensics due to the large datasets that must be analyzed during an investigation. VA provides a number of approaches that allow the analyst to view both tangible data (e.g. event, time of event, etc.) and intangible data (e.g. relationships between actors or events) within a large dataset. *Social network analysis* is a well-established, inter-disciplinary field. This field attempts to visualize and analyze the complex area of social interactions and relationships to identify key actors within those networks, often achieved by some scientific reasoning such as graph theory.

This paper presents a novel approach, the *E-mail Extraction Tool* (EET), for visualizing e-mail accounts and exploring the social networks that these will reveal to the investigator. There are three main advantages that this approach presents to the investigator over existing tools and techniques when analyzing a suspect's e-mails. First, EET is an automated application that mines and visualizes e-mails from the client application files associated with a user resident on the hard drive. As the files are mined by the application, all e-mails including those hidden in replies or forwarded messages, are analyzed. In order to analyze this type of data in existing tools such as FTK, individual messages would have to be read and analyzed, which is a time-consuming process. Second, EET has the advantage over using current visualization tools, such as I2's Analyst's Notebook (I2 Ltd., 2009), which require manual input of data to visualize a social network. This again can be a time-consuming process dependent on the amount of contacts an actor possesses. As will be demonstrated in the case study, the number of individuals within a social network identified by e-mail may be very large. Third, EET provides analysis tools to identify and visualize key actors and the strength of their relationships to others within the social network. Analysis tools beyond just a rudimentary network diagram in applications such as I2 Analyst's Notebook are lacking. As will be demonstrated later in the paper, through its visualization and analysis EET can quickly direct the investigator to relevant e-mails and actors for the investigation, thus reducing the time required to investigate such evidence.

This paper is organized as follows. Section 2 discusses related work in computer forensics, social network analysis and visualization techniques. Section 3 presents an overview of EET. Section 4 presents a case study to demonstrate the use of the EET approach during an e-mail investigation. Finally, we make our conclusions and discuss further work.

2. Related work

E-mail investigations have risen in prominence due to the reliance of users on their e-mail communications and the amount of information about a suspect this data source may yield to the forensic examiner. Computer forensics tools, such as the Forensics Toolkit (FTK) (Access Software, 2009) and EnCase (Guidance Software, 2009), are used by analysts to recreate files and data from a suspect's computer. As such, these tools are useful in recreating the suspect's e-mail messages from their e-mail client or from the mail server. An analyst may view messages, one at a time, to read the communications that the suspect has with others in their social or professional networks. In addition, an analyst may view other vital pieces of evidence, such as network information associated with each e-mail, to build up a picture of their communications. However, these tools do not provide a visualization of evidence or the importance of actors within the social network.

The limitations of current tools and the importance of e-mail within an investigation have led to research interest. For example, (Carenini *et al*, 2005) suggest that one challenge when analyzing e-mail is text mining for hidden e-mails. Hidden e-mails are defined as those that have been quoted in subsequent e-mails but are not themselves present in the e-mail folder in their own right, such as forwarded messages. This issue is of interest within the context of this paper in that forwarded e-mails will indicate to the examiner the wider social groups to which a suspect belongs. However, whilst this approach retrieves the hidden e-mails, it does not provide the analyst with a visualization of the social networks that these may reveal. It is also does not demonstrate the relationships of those actors identified in hidden e-mails to the suspect under investigation.

The interaction and dynamics of social networks has for a long time been of interest in inter-disciplinary research, and in particular, the social sciences and is of relevance to digital forensics. For example, (Freeman, 1978/79) suggests that an understanding of centrality within social networks may provide social science researchers with an understanding of the dynamics of those groups under investigation. In a digital forensics investigation, centrality may provide the examiner with an understanding of the levels of culpability within an act involving a group of people (Haggerty *et al*, 2008**.** Research conducted by (Lawler & Yoon, 1996) focuses on commitment in exchange relationships. This research attempts to predict how and when people in an exchange become committed to their relationship. This has particular resonance in investigating paedophile rings that often require a new member to supply a certain number of original indecent images of children prior to entry to the group. In other criminal activity, such as terrorism and organized crime, new members of a group may have to commit a directed act prior to entry. Therefore, an understanding of group dynamics, centrality and strength of ties will be of great interest to the forensic examiner.

Due to the increased amount of evidence a digital forensics examiner must analyze, visual representations of this information have received attention. For example, (Wang & Daniels, 2005) propose an evidence graph to facilitate the presentation and manipulation of intrusion evidence. Alternatively, (Krishnan *et al,* 2007) posit an approach to visualize large textual datasets and exploit cluster architectures. Tools, such as *Pajek* (de Nooy *et al*, 2005) and *SocNetV* (Kalamaras, 2009) provide social scientists with visualization tools for analyzing social networks. Whilst these tools are useful when looking at social science areas of interest, they are of only a limited use during e-mail investigations as they require manual input of what are large datasets.

Previous work conducted into the visualization of e-mails has not focused on the needs of forensic investigations. (Viegas *et al*, 2004) propose two tools, *PostHistory* and *Social Network Fragments* to visually represent a user's communications. The aim of this research is to provide users with the means to explore personal self-awareness through their data patterns and to remember key events rather than provide a sound forensics-focused approach. This approach differs from EET in that it is user-centric and aims to share a single user's view of their social networks and interactions rather than provide complete strangers with an understanding of an individual's social networks and dynamics. (Kim, 2007) uses e-mail visualization to distinguish between 'spam' and legitimate mail. However, this approach does not identify the activities and relationships of a suspect under investigation.

The next section presents our novel approach for analyzing a suspect's email activity for evidence of social networks and identification of relevant accounts or evidence.

3. The E-mail Extraction Tool

As outlined in sections 1 and 2, the key challenges to computer forensics e-mail investigations are: the number of machines to be investigated, volume of data, evidence identification, relevant social network identification and visual representation of evidence. This section posits the EET tool for visualization of social networks through e-mail account analysis to meet these challenges.

Social network analysis assumes that interpersonal ties between actors are important as they transmit behaviour, attitudes, information, goods and services (de Nooy *et al*, 2005). Within a digital forensics investigation, this is reflected in both tangible evidence (e.g. events, times/dates, etc.) and intangible evidence (e.g. relationships, strength of ties, etc.). As (Viegas *et al*, 2004) posit, e-mail and instant messaging capture some of the most meaningful social interactions that people have online every day with people participating in a variety of social groups. Therefore, these will provide a wealth of information about a suspect to the forensics examiner.

Often, people will reflect their social groups in their e-mail accounts by organizing individual messages into folders. For example, in a work-based e-mail account folders may be organized by the various roles they perform within the organization, administration information and personal or professional contacts. Within a home-user account folders may be organized by family, friends, hobbies/interests or social activities. Therefore, folder organization will provide the forensics examiner with a rudimentary indication of the suspect's view of their wider social networks. However, of further interest during an investigation will be:

- How are these social networks organized?
- Who are the key actors within these networks?
- What is the role of the suspect within these networks?
- What are the strong ties within these networks (i.e. who are the core of actors with whom the suspect corresponds regularly)?
- Are there specific individuals that bridge disparate clusters of actors?

The forensic examiner will be faced with analyzing both tangible and intangible evidence during the course of an investigation. The approach for visualizing intangible data posited in this paper may not provide evidence *per se*, but it provides a powerful analysis technique particularly where groups of suspects are involved, such as paedophile rings, terrorist networks or organized crime. This approach highlights roles and actors within these networks, their level of involvement, their centrality to the investigation and provides clues for the examiner as to other potential sources of evidence or co-conspirators.

The EET approach sets out to discover two dimensions of e-mail patterns; the social networks to which a suspect belongs and the strength of ties between actors within that network. This approach differs from current approaches in that it does not focus on textual context of individual e-mails or uncovering hidden e-mail fragments. It should be noted that this view of the network is a suspect-centric snapshot, i.e. as we are analyzing the suspect's computer, the social networks will be from the suspect's point of view. A suspect-centric visualization differs from traditional social network approaches in which the entire network is assumed and visualized. Due to the nature of digital forensics investigations, the examiner is often provided with only a partial view of the wider network. Therefore, widely used actor centrality measures (for further reading, see (Freeman, 1978/79)) used in social network theory, e.g. out-degree, betweenness or closeness, are less applicable as the analyst will rarely have an overview of the whole network to which the suspect belongs. However, strength of ties, indicated by quantitative, events-based analysis of key actors' communications, can provide an understanding of the wider social networks and the power or centrality within those networks. In order to achieve this, the social patterns derived from the FROM, TO, CC and SUBJECT data in both messages sent to and received by the suspect are used. In addition, by mining the e-mail application's associated files for this events-based information, EET allows for the hidden e-mail problem identified in (Carenini *et al*, 2005) and is taken into account in the visualization of the social network. As will be demonstrated by the case study, examiners are able to analyze social network dynamics without having to examine the content of the e-mails themselves. However, EET provides the facility to read individual e-mails, should the examiner require.

The software has, at its most basic level, three main areas of functionality; file reading and processing, graph organization and graphical output (via internal visualization, and optionally, a Pajek-compatible file format). These functional points are covered in more detail in the next section, whereas here, the software operation is broken down into its component parts. This helps to illustrate both how EET works and the flexibility of the design to allow adaptation for domain-specific applications.

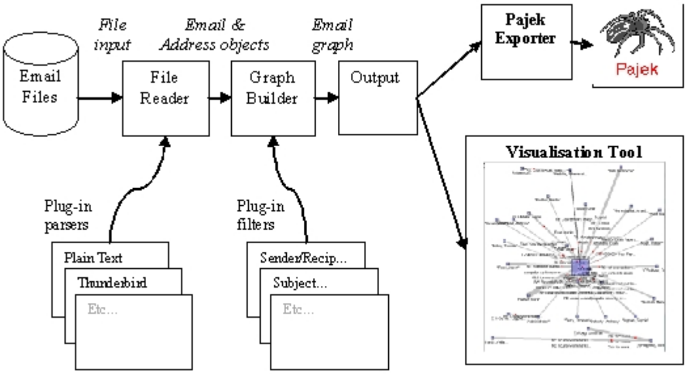


Figure 1: Overview of EET.

The software comprises the following components, as illustrated in figure 1:

- The e-mail files – repository of investigation files, exported from the e-mail clients;
- The file readers – used to process varying file formats, with suitable filters/parsers;
- The graph builders – build the investigation network from the e-mail and address objects with suitable filters / construction algorithms.
- The output processors – convert the built graph into a useful format;
- Visualisation Tool – displays the resulting e-mail graph within the application,
- Pajek Exporter – exports into a Pajek-compatible file with labelled vertices and edges contained within an `arcslist`.

This section has provided a brief overview of EET. In the next section, a case study is presented to highlight the novelty and the benefits of using EET during an investigation involving e-mail analysis.

4. Case study

To demonstrate the applicability of EET to computer forensics investigations, this section presents a case study. The section begins with an overview of the Mozilla Thunderbird e-mail client before presenting examples of using the application and the results that these return.

The recovered e-mails in this case study are from a suspect using the Mozilla Thunderbird e-mail client. E-mail, especially in the Windows environment, can be problematic to the investigator due to the number of formats and conventions available. Mozilla Thunderbird e-mails are the less problematic as they are stored in text format so the application has less work to do in parsing the information. However, EET has been tested with other formats. If e-mails are in the more complex Outlook PST file format, they can be converted to the Thunderbird format using *libPST*. The Thunderbird convention is illustrated in figure 2 below. The folder storing the e-mails can be found at `C:\Documents and Settings\[User Name]\ApplicationData\Thunderbird\Profiles\` (XP), `C:\users\[User Name]\ AppData\Roaming\Thunderbird\Profiles\` (Vista), `~/.thunderbird/xxxxxxxx. default/` (Linux) and `~/Library/Thunderbird/Profiles/xxxxxxxx.default/` (Mac OS X) (Mozilla, 2009). It contains two types of files: a mail summary file indexing the e-mails and an *mbox* format file containing the e-mails themselves. EET mines the *mbox* files for relevant information and outputs the results. As discussed in the previous section, the suspect will often organize their folders by social group. Therefore, we can utilise this to view either individual folders (i.e. individual social groups) or all folders together (i.e. all social groups) to provide varying levels of analysis into the social networks.

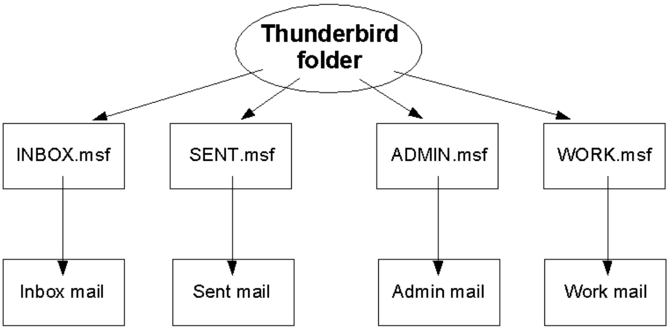
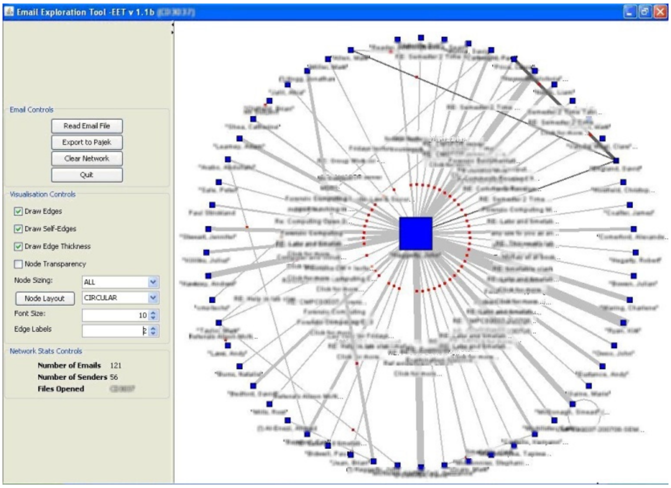


Figure 2: Illustration of the Thunderbird e-mail architecture.

Figure 3 below shows the visualization of a single folder as viewed in EET with actors organized in a circular display with the suspect in the centre. For reasons of confidentiality, identifying features such as names and subject lines have been disguised. The control panel, which is collapsible, is divided into three key areas: e-mail controls, visualization controls and network statistics controls. E-mail controls allow the examiner to read in e-mail files, clear the network from the screen and quit the application. In addition, there is the functionality to export files to a Pajek format to make use of analyzing files in this or other existing social network analysis tools. The visualization controls provide the examiner the facility to manipulate the output once the network has been painted to screen by drawing edges and self edges (i.e. e-mails where the sender has copied themself in to the message they have sent), draw edge thickness, toggle node transparency, change node sizing according to the analysis, change network layout, change font size and change the number of edge labels.

The icons in the network are organized by size in that the more traffic associated with an e-mail (both in the e-mail header and in forwarded or reply messages), the larger the icon. The strength of ties through volume of e-mail is also reflected by the edge thickness. Node sizing can be changed to reflect five options in measuring the traffic: all e-mails, received, recipients, senders and sent. The network can be initially displayed as circular, clever circular and random. Once displayed, individual nodes can be moved around the screen to aid analysis and clicking on an individual node will highlight the flow of traffic between actors. The edges represent the relationships between actors within the network. The facility to change the number of edge labels was seen as a key development in the visualization. The edge labels are the subject line of individual e-mails and can be clicked to display the actual e-mail(s) of interest. In a busy network, a large number of edge labels detrimentally affect the visualization and therefore the user has a choice as to the number that they wish to view, or to switch this function off. Finally, the network statistics controls provides information about the network to the examiner and includes the number of e-mails, number of senders and the folder (file) that is open in the network view.



Email Exploration Tool -EET v 1.1b [redacted]

Visualisation Controls

☑ Draw Edges

☑ Draw Self-Edges

☑ Draw Edge Thickness

☐ Node Transparency

Node Sizing: ALL

Node Layout: CIRCULAR

Font Size: 10

Edge Labels: 2

Network Stats Controls

Number of Emails 121

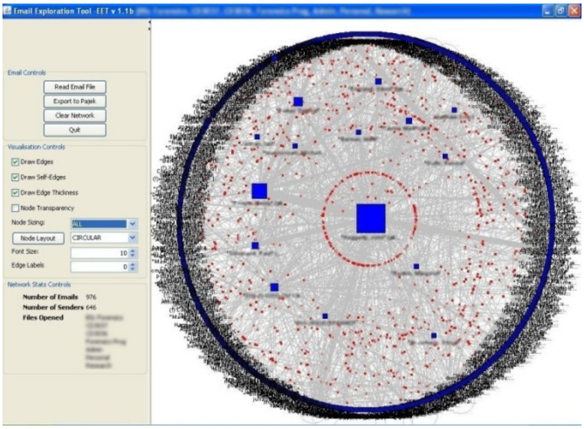
Number of Senders 56

Files Opened [redacted]

Figure 3: EET interface displaying a single e-mail folder.

The number of actors identified by an e-mail network analysis will vary depending on the environment and the individual under investigation. For example, in the case of corporate e-mail it is likely that the suspect will have a large number of contacts, identifying both explicit and implicit social networks. The social networks identified will be both those associated with particular roles that the individual is undertaking and possibly some personal contacts. In an investigation involving a user's home e-mail account, the networks identified will far more based on personal, social contacts. Therefore, the investigator must take into account the environment from where the evidence is seized.

Figure 4 illustrates an entire corporate-based e-mail account. The figure demonstrates the complexity of social network analysis within this environment and the many actors that will be identified. In total, 976 e-mails involving 646 actors are visualized. To recreate this many actors in a manual tool such as I2 Analyst's Notebook would take a significant time: by mining the e-mail files themselves, this process takes seconds. The original visualization is in circular format. However, EET identifies a number of key actors as having strong relationships or of importance within the social network. The suspect is placed in the centre of the diagram, and as can be readily seen by icon size, as would be expected they are actively engaged in the network. In order to identify further actors of interest to the investigation, actors with large icons associated are dragged into the central space. As can be seen by edge thickness, these actors have strong links to the suspect and are in regular contact, accounting for a large proportion of e-mail communications. Further qualitative analysis would be required by the investigator to identify the precise nature of the relationships between the suspect and these actors. This could be achieved by clicking on the edge labels to see the associated e-mail subject lines or through reading the e-mails themselves through the EET interface. However, the visualization quickly identifies 14 key actors that will be of interest to the investigation out of the total 646 actors. These may be either potential witnesses or their computers a source of evidence.


Email Exploration Tool -EET v 1.1b

Email Controls

Visualisation Controls

☑ Draw Edges
☑ Draw Self-Edges
☑ Draw Edge Thickness
□ Node Transparency

Node Sizing: ALL

CIRCULAR

Font Size: 10

Edge Labels 0

Network Stats Controls

Number of Emails 976
Number of Senders 646
Files Opened

Figure 4: A suspect's corporate e-mail account.

In the case of the account above, the suspect provides some preliminary social analysis to the investigator in the way in which they organize their e-mail folder.

Many of the folders are organized based on the various roles that they perform within the organization. However, the suspect also uses this account as their primary e-mail address and therefore accesses it for both business and home use. The suspect has a folder called 'Personal' where all personal e-mails are stored. Therefore, the investigator is able to gain a better understanding of the suspect's social life beyond work and identify key personal, and therefore closer, contacts that would be of interest to the investigation.

Figure 5 illustrates the suspect's 'Personal' folder and has been annotated for clarity. In order to enhance the visualization, the EET control bar has been contracted to provide a larger visual analysis space and the figure shows senders only. The investigator can see quickly the dynamics of the suspect's personal social networks. The suspect is located at A. In general, there are two distinct subnets. The first appears to contain stronger ties to the suspect with fewer contacts but thicker lines indicating volume of traffic. The strength of ties suggests that the actors in this subnet would be closer to the suspect and would be likely to hold information pertinent to the investigation or may contain co-conspirators. These would be key actors for further investigation. The second subnet is indicated by a large volume of traffic sent from just two actors, C, to a large distribution list. These two actors send a large volume of traffic to two inter-connected subnets. The large e-mail lists and the few replies that these receive identify these actors as disseminators of information. However, there are a number of bridges between the two disseminators and the suspect, indicated by B. These bridges have a powerful relationship with the suspect as they have the choice to forward (or not) information from the disseminators and would also be of interest to the investigation as they could hold possible evidence due to this role.

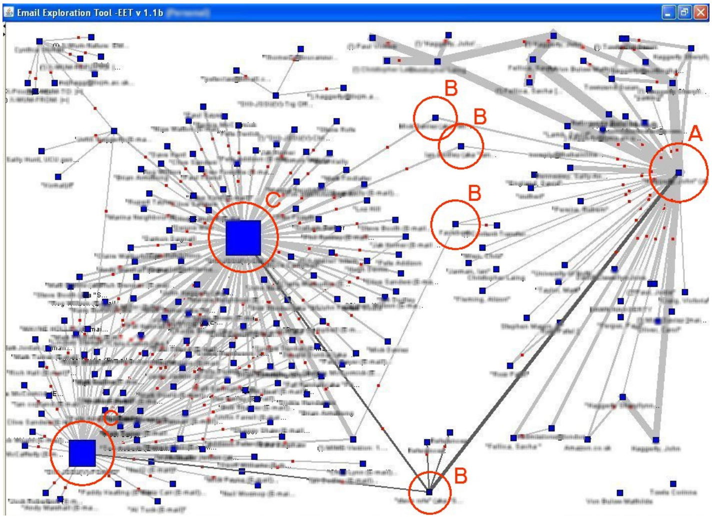


Figure 5: The suspect's 'Personal' e-mail folder.

As we can see by the case study, the investigator can be directed to potential sources of evidence through the automated network visualizations and can gain a better understanding of the social networks to which they belong. This type of analysis has particular use in the investigation of criminal activities that involve groups, such as paedophile rings, terrorist cells or organized crime.

5. Conclusions and further work

E-mail is the principal medium today for the dissemination of information and social networking. It is therefore not surprising that e-mail communications will feature heavily in most digital forensics investigations, whether examining computer-based crime or not. E-mails provide the examiner with a wealth of information regarding a suspect's activities and contacts. This information may be in the form of tangible evidence, such as time and date of a computer-related event, or intangible evidence, such as the relationships that a suspect has with others in his social networks.

This paper has introduced the *E-Mail Extraction Tool* (EET), an application for the automated analysis and visualization of e-mail files to identify the social networks to which a suspect belongs. This application mines the files associated with a mail client and presents a visual representation of individual folders or the entire account. This approach has a number of advantages over existing approaches. As the files are mined by the application, all e-mails including those hidden in replies or forwarded messages, are analyzed. In order to analyze these files in existing tools such as FTK, individual messages would have to be read and analyzed, which is a time consuming process. In addition, EET has the advantage over using current visualization tools that require manual input of data to visualize a social network. Finally, EET provides analysis tools to identify and visualize key actors and the strength of their relationships to others within the social network which are lacking in existing tools. Further work aims to enhance the visualization and analysis functions of EET.

6. References

Access Data (2009), *FTK Forensic Tool Kit*, http://www.accessdata.com, accessed 2009.

Carenini, G., Ng, R., Zhou, X. & Zwart, E. (2005), "Discovery and Regeneration of Hidden Emails", *ACM Symposium on Applied Computing*, pp. 503-510.

Chen, H., Chung, W., Xu, J.L., Wang, G., Qin, Y. & Chau, M. (2004), "Crime Data Mining: A General Framework and Some Examples", *Computer*, Apr 2004, pp. 50-56.

Clayton, R. (2007), "Email Traffic: A Quantative Analysis", *Proceedings of the 4th Conference on Email and Anti-Spam CEAS 2007*, Mountain View, CA, 2 – 3 Aug 2007.

de Nooy, W., Mrvar, A. & Batagelj, V. (2005), *Exploratory Social Network Analysis with Pajek*, Cambridge University Press, UK.

Freeman, L.C. (1978/79), "Centrality in Social Networks", *Social Networks*, vol.1, pp. 215 – 239.

Guidance Software (2009), *Encase*, http://www.guidancesoftware.com, accessed 2009.

Haggerty, J., Taylor, M. & Gresty, D. (2008), "Determining Culpability in Investigations of Malicious E-Mail Dissemination within the Organisation", *Proceedings of the 3rd Annual Workshop on Digital Forensics and Incident Analysis (WDFIA 2008)*, Malaga, Spain, 9 Oct 2008.

I2 Ltd. (2009), *I2 Analyst's Notebook 7 Product Overview*, http://www.i2.co.uk/products/analysts_notebook/, accessed 2009.

Jern, M., Banissi, E., Andreinko, G., Muller, W. & Keim, D. (2006), "European Research Forum Panel Session Envisioning Research Challenges in Visual Analytics", *Proceedings of the 10th International Conference on Information Visualisation (IV'06)*, London, UK, 5 – 7 Jul 2006, pp. 5-8.

Kalamaras, D.B. (2009), SocNetV, http://socnetv.sourceforge.net/, accessed 2009.

Kim, U. (2007), "Analysis of Personal Email Networks using Spectral Decomposition", *International Journal of Computer Science and Network Security*, vol. 7 (4), Apr 2007, pp. 185 – 188.

Krishnan, M., Bohn, S., Cowley, W., Crow, V. & Nieplocha, J. (2007), "Scalable Visual Analytics of Massive Textual Datasets", *Parallel and Distributed Processing Symposium*, Long Beach, CA, 26 – 30 Mar 2007, pp. 1 – 10.

Lawler, E.J. & Yoon, J. (1996), "Commitment in Exchange Relations: Test of a Theory of Relational Cohesion", *American Sociological Review*, vol. 61 (1), pp. 89 – 108.

Mozilla (2009), http://www.mozilla.org/support/thunderbird/faq, accessed 2009.

Richard III, G.G. & Roussev, V. (2006), "Next-Generation Digital Forensics", *Communications of the ACM*, vol. 49 (2), Feb., 2006.

Viegas, F.B., Boyd, D., Nguyen, D.H., Potter, J. & Donath, J. (2004), "Digital Artifacts for Remembering and Storytelling: PostHistory and Social Network Fragments", *Proceedings of the 37th Hawaii International Conference on System Sciences*, Hawaii, USA, 5 – 8 Jan 2004.

Wang, W. & Daniels, T.E. (2005), "Building Evidence Graphs for Network Forensics Analysis", *Proceedings of Advances in Computer Security and Applications Conference (ACSAC 2005)*, Tucson, AZ, 5 – 9 Dec 2005.