A Social Network Discovery Model for Digital Forensics Investigations

A. Karran[1], J. Haggerty[2], D. Lamb[1], M. Taylor[1] and D. Llewellyn-Jones[2]

[1]School of Computing and Mathematical Sciences, Liverpool John Moores University, James Parsons Building, Liverpool L3 3AF
[2]School of Computing, Science and Engineering, University of Salford, Newton Building, Manchester, M5 4WT
e-mail: A.J.Karran@ljmu.ac.uk; J.Haggerty@salford.ac.uk; D.J.Lamb@ljmu.ac.uk; M.J.Taylor@ljmu.ac.uk; D.Llewellyn-Jones@ljmu.ac.uk

Abstract

Our continued reliance on email communications ensures that this type of data remains a major source of evidence during a digital investigation. Due to the many applications and data types, there is no standard email format. Therefore, much current work in the forensic investigation of emails has focused on data extraction. This paper focuses on the investigatory process and posits a model for social network discovery for use in digital investigations. This model is applied to the widely used Enron email corpus to demonstrate its applicability.

Keywords

Digital forensics; social network analysis; visualisation

1. Introduction

With our reliance on email as a communications medium, particularly within the workplace, it is likely to continue to feature as a major resource of evidence during a digital investigation. The identification of both tangible and intangible evidence from large volumes of data is a major challenge for digital investigations involving emails. Social network analysis and visualisation techniques can significantly contribute to evidence discovery and collection by identifying and understanding relationships and data flow between actors and events within the email network.

Digital forensics investigatory models do not currently differentiate between email and any other data. However, intangible evidence such as relationship data is a defining feature of social networks (Wasserman & Faust, 1994) and may provide information pertinent to an investigation. Much work on digital investigations involving email data focus on techniques for the extraction of evidence, for example by focusing on data mining (Wei *et al*, 2008) or clustering algorithms (Bird *et al*, 2006). Recently there has been some focus on process models for investigations that involve email data. These approaches generally provide a theoretical framework or software application (Debbabi *et al*, 2009), which details a process or methodology for the visualisation and extraction of specific email artefacts or features. However, these approaches focus on particular aspects of email data rather than the process itself.

This paper therefore presents a model for social network discovery through email data and is organised as follows. Section 2 presents related work. Section 3 posits the model for social network discovery for investigations involving emails. Section 4 presents a case study and results of applying the approach to the Enron email data set. Finally, we make our conclusions and suggest future work.

2. Related Work

Social network analysis was first formalised as a scientific field by Moreno in the 1930s (Moreno, 1951). It has since become a widely used method of mapping and understanding social structures, adopted in many fields to investigate the underlying structure of interactions and relationships between actors. It can therefore be used to highlight the significance of members within communities. Social network analysis tools and techniques have been used in a wide range of inter-disciplinary studies. For example, this approach has been used to map local communities (Hawe, 2008), highlight voting patterns and regional political affiliations (Faust, 1997), uncloak terrorist networks (Krebs, 2002) and map parts of the Internet in terms of social communities (Cocciolo *et al*, 2007).

Recently, social network analysis has been proposed as an aid for digital investigations. For example, Haggerty *et al* (2009) proposed the Email Extraction Tool (EET) for the extraction and visualisation of email data resident in files on the hard drive. Dellutri *et al* (2009) focus on the identification of social networks through data on smartphones and Web information. This approach aims to reconstruct a user's profile by combining the smartphone's data with social relationships found on the Internet. Wiil *et al* (2010) provide an analysis of the 9/11 hijackers' network and focus on the relationships between actors. This study uses a number of measures associated with social network analysis to identify key nodes. However, these approaches have in common that they focus on the details of extracting and identifying data within specific environments to identify the social network rather than developing the procedures surrounding this activity.

The Enron email data set contains data from about 150 users, predominantly senior management at the company (Cohen, 2010). The data set provides real data that can be used as a test bed in a number of ways. For example, Lin (2010) uses this data set to demonstrate the applicability of their approach in predicting sensitive relationships identified in email communications. Alternatively, Zhou *et al* (2010) use this data set for text analysis which employs a wide variety of statistical techniques to identify value profiles of Enron employees. Alternatively, Collingsworth *et al* (2009) use network analysis of this data set to assess organisational stability. Therefore, this data set provides a means by which the social network discovery model may be tested.

3. A Model for Email Investigation

This section presents a new process model for use in digital forensics investigations of email data. In particular, this model focuses on incorporating social network analysis techniques into digital forensics investigations to elucidate intangible data, i.e. relational information.

Figure 1 illustrates the processes involved in a digital forensics investigation of email data. These processes do not differ much to any other investigation. However, the Analysis stage reflects the need to identify and assess relational information using centrality measures and visualisation techniques. The Investigation column breaks down these processes into further detail of the stages related specifically to email data. The Case Study column illustrates the specific processes used in section 4

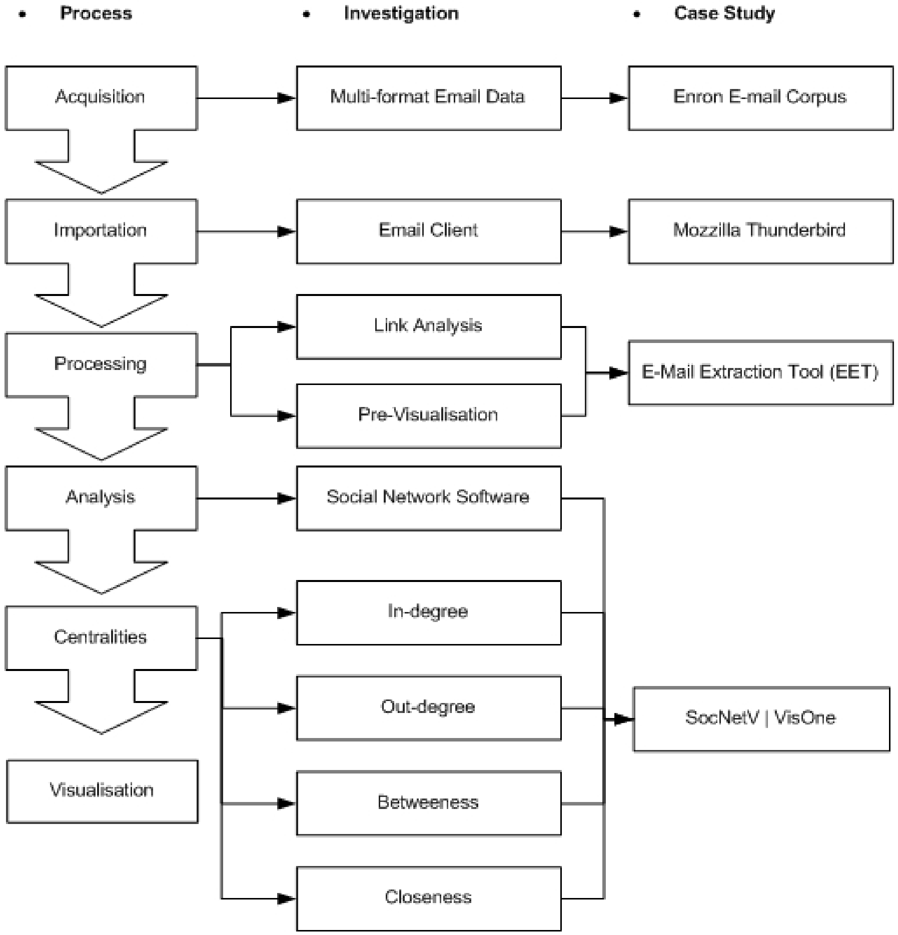


Figure 1: Social Network Analysis of emails methodology.

Process. This column in figure 1 reflects the overall procedures involved in digital forensics investigations. As with any investigation, the data must be acquired in a robust manner, ensuring that the data maintains its integrity. Therefore, when data is imported for analysis, this is done in 'read only' to ensure that data is not modified. As with other investigations, the data has to be processed and analysed. However, with email, the application of analysis techniques such as centrality analysis and visualisation will be applied. These centralities provide alternative views of the social network dynamics (Wiil *et al*, 2010).

Investigation. This column reflects the investigation process as it relates to email evidence. The acquisition of email data during an investigation is problematic due to the many email formats available depending on the application used by the suspect. Therefore, importation of the data will depend on the client used, for example, Mozilla Thunderbird stores email data in plain text whilst Microsoft Outlook uses a bespoke format. The data is processed to provide an initial visualisation of links to give the examiner an initial view of the social network. Analysis of the evidence is then conducted using other social network tools. These tools, such as Pajek, SocNetV and VisOne, provide different viewpoints of the network by using statistical measures. These are as follows: *in-degree centrality* indicates a node's receptivity or "popularity" in the network and can be used to identify key network facilitators; *out-degree centrality* indicates the expansiveness of ties within the network that an actor possesses; *betweeness centrality* identifies potential points of information control within the network; *closeness centrality* highlights an actor's ability to interact with other members of the network. Wasserman and Faust (1994) provide a comprehensive description and analysis of these measures.

Case study. This column illustrates how the investigation in this paper has been achieved. The Enron email data set is downloaded and the data converted to Mozilla Thunderbird format due to its plain text format. The initial processing and triage is achieved using the EET tool (Haggerty *et al*, 2009) and provides an initial visualisation of the network. It also provides the platform for easy exportation to other common social network analysis formats, such as Pajek. SocNetV and VisOne tools are used to measure and further analyse the network within this paper.

4. Case Study and Results

This section presents the results of applying the social network discovery model, as detailed in figure 1, to the Enron email data set. This section first places the case study in context by discussing the background to the case. As the data has already been collected, this stage of the model is bypassed, although procedures discussed in (Haggerty *et al*, 2009) would be used. This section presents the link analysis of the data visualised and triaged through EET. It then identifies the social network groups discovered within the data. The data is analysed using social network centrality measures. Finally, the results are discussed.

A. Case Study Background

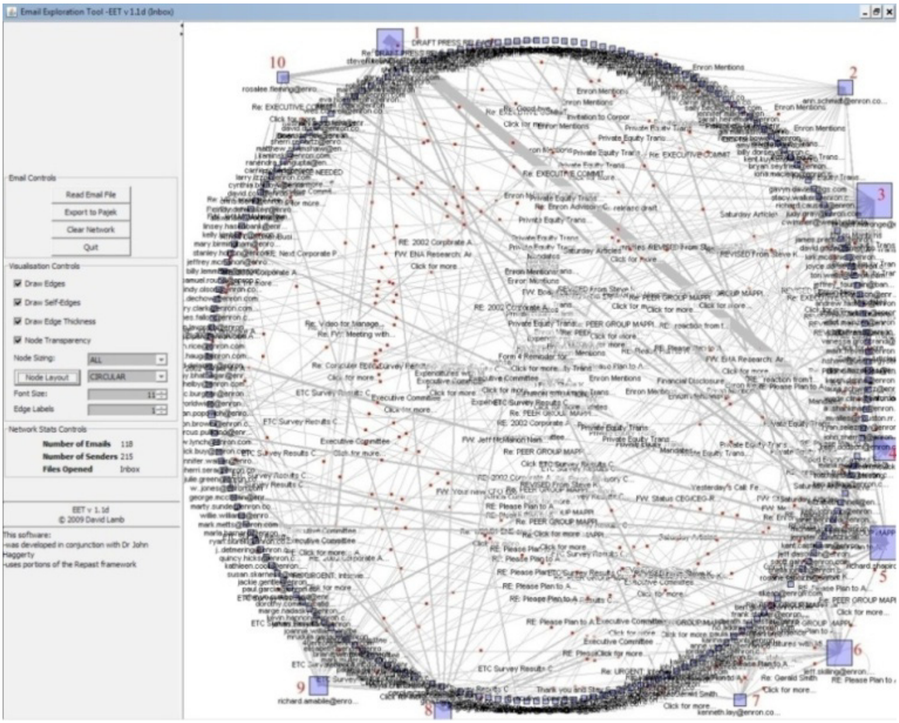
Enron was a large energy company that had expanded from its beginnings in 1985 to employ thousands of workers across 40 countries. The Enron fraud caused shockwaves within the business community when it was revealed in 2001 due to the extent and scale of the case. The fraud resulted in the bankruptcy of the company and dissolution of a large accountancy and audit company. The fraud occurred due to the lack of transparency in the firm's accountancy procedures. The main executives in the company used a series of techniques to perpetrate the fraud, such as accountancy loopholes, employing special purpose entities and poor accountancy practices, in order to hide billions of dollars of debt that the company had accrued.

The main actors are as follows: Jeffrey Skilling, former President of Enron Corporation and responsible for Enron's introduction of accounting methods that treated anticipated profits as if they were real gains; Kenneth Lay, Enron Chairman and Chief Executive Officer from 1985; Andrew Fastow, Chief Financial Officer who set up a network of off-balance-sheet companies controlled by Enron to hide Enron losses; Jeffrey McMahon and Ben Glisan, former Enron Treasurers Executive; Ken Rice, former Enron executive and President of Enron's broadband service.

The email data used in this paper is from the Enron corporate business network. This data has been released by the Federal Energy Regulatory Committee (FERC) and is sanitised for use (Cohen, 2010). The data consists of a snapshot of the email folders of a large proportion of Enron employees taken in the final year of business and at the time of criminal proceedings for fraud against members of the Enron organisation. This data set comprises approximately 500,000 emails. This case study uses a selection of user email folders to cover the five most important users based on their position within the Enron hierarchy.

B. Link Analysis

Figure 2 illustrates the data set comprising 118 emails and 215 actors visualised in EET. Aggregating the data in this way helps to counter the issue of source-centricity often found in the analysis of email data. Node sizing is used to gain an overall impression of the importance of the main actors in this network, i.e. the larger the node size, the more emails they have sent and/or received. Ten main actors can be identified from this view. These actors have been manually teased out of the main body to give a better impression of how the emails are linked and how information flows around this network group.



☑ Draw Edges
☑ Draw Self-Edges
☑ Draw Edge Thickness
☑ Node Transparency
Node Sizing: ALL
Node Layout: CIRCULAR
Font Size: 11
Edge Labels 1

Number of Emails 118
Number of Senders 215
Files Opened Inbox

EET v 1.1d
© 2009 David Lamb

This software:
-was developed in conjunction with Dr John Haggerty
-uses portions of the Repast framework

Figure 2: Aggregated data set link analysis.

In the link analysis, email headers reveal that Actors 1 and 4 discuss draft press releases about financial disclosure, with Actor 5 receiving overviews and sending them to Actor 6. This communication could be a part of standard corporate operating procedure. The significance becomes apparent only when the nature and content of these emails are considered. Several of the emails used in this aggregated dataset later appeared as exhibits used in evidence. This supports the potential value of triaging through link analysis.

C. Social Network Group Discovery

Figure 3 illustrates the node mapping graph for the discovered social networks, which identifies six distinct social structures. The first network shows a social network with Actor X at its centre. The directed information shows a typical star layout, indicative of a reporting structure. The second structure of significance has Actor Y at its centre. In this network Actor Z can be seen reporting to Actor Y. This is the first appearance of Actor Z in the analysis of the data sets.

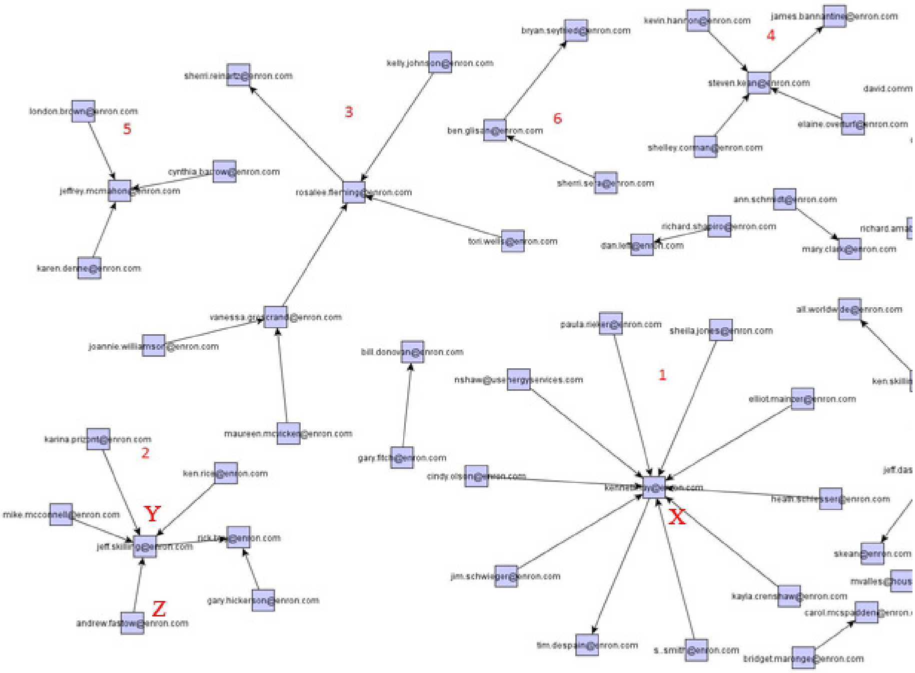
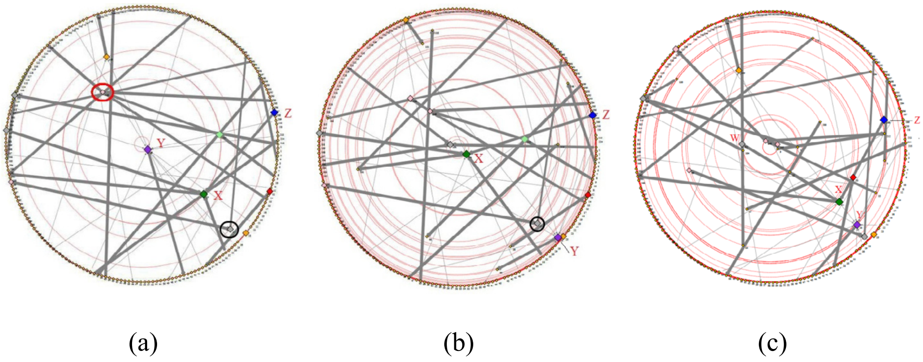


Figure 3: Aggregated data set social network group discovery.

The third social structure features a member of the Enron executive oversight committee and the personal assistants of Actor Y and Actor X. The fourth network features another member of the oversight committee and a public relations executive. The fifth and sixth networks feature two actors of significance in the financial dealings with Enron’s strategic partnership companies. The link analysis and node mapping combine to give a financial context to the analysis, highlighting the major actors within the dataset. If the hypothesis for the process model holds true, the centrality analyses that are considered next should demonstrate this context. That is, it is expected that the actors who are heads of financial matters and those whose position in Enron grants access to this information should prove significant.

D. Centrality Measures

This section details the analysis of the centrality measures used within the case study and presents the quantitative analysis of the email data as a centrality graph, which allows for qualitative interpretation. These circular radial graphs are interpreted by observing the position of nodes (actors); the closer a node is to the centre of the graph, the higher the measure of that node by the centrality measure and therefore of greater interest within a digital investigation. Line thickness indicates the strength of the ties (relationships) with other nodes within the graph. Three centrality measures are used in this study to demonstrate the applicability of the approach to provide the digital examiner with different views of the email network; betweeness centrality, in-degree centrality and out-degree centrality.



(a) (b) (c)

Figure 4: Significant actors identified by (a) betweeness centrality (b) in-degree centrality (c) out-degree centrality.

Figure 4(a) displays the betweeness centrality of the aggregated data. Actors closer to the centre of the graph have a higher betweeness centrality, indicating a greater control over information being sent by email. We can see therefore that the graph indicates that Actor X is most central for communications involving Actor Z. Documents pertaining to the downfall of Enron (Findlaw, 2009) show that collaboration between Actor X, Actor Z and Actor Y sustained the proliferation of the Enron fraud. Within figure 4(a), these actors are highlighted by this quantitative analysis and qualitative information shows that Actor X is central in the network at the time when Enron's financial situation was being called to account. Actor X was able to receive the bulk of network information from both strong and weak ties to other actors within the network. Overall this graph suggests a close-knit social structure able to shape and disseminate financial information to suit the needs of those in prominent roles.

Figure 4(b) illustrates actor in-degree centrality with those having more incoming emails (i.e. information) placed closer to the centre of the disc. In this graph the former treasurer shares the highest in-degree measure with actor X. An interesting observation in this graph is the position of the replacement treasurer marked with a circle. The FERC Enron investigation timeline supports such an inference indicating that the position of treasurer within Enron was in flux and changing hands. They therefore had not had time to have made as many connections within the email network as the previous treasurer. From a digital investigation viewpoint, this may suggest that they will not be able to provide as much evidence.

Figure 4(c) illustrates the out-degree centrality for actors within the data set, with those having more outgoing connections being closer to the centre. In this graph the personal assistants for Actor X and Actor Y are the most prominent, indicating the volume of information they send into the network on behalf of their superiors. These would be actors not necessarily under investigation but would be able to provide a wealth of information to the investigatory team.

The next actor of significance is Actor W, the head of the Enron broadband division. This is the first time his presence is evident in the centrality measures. This signifies that this actor was in a position, both within Enron and the network, to disseminate information widely across the network. This information took the form of financial "overviews" of the market position of Enron broadband. It was later reported that this was code for fraudulent disclosure of financial information (Murphy, 2006). Actor W was seen as a *protégé* of Actor Y. Actor W's position in the graph of Figure 4(c) indicates a relationship in which he reports directly to Actor Y, who then disseminated the fraudulent reports to the rest of the network.

E. Discussion

The social network discovery model illustrated in figure 1 provides a framework by which digital forensics investigators may analyse email data and explore intangible evidence contained therein. Link analysis visualisation provides a rudimentary visualisation of the email data and can be used to effectively triage the potentially large data sets to identify just key actors within the network. This can then be used to inform the social network group discovery, whereby relevant subnets within the data may be assessed. Finally, the application of social network analysis tools and techniques to measure the network not only provides different viewpoints of the network, but also quantifies an actor's role, and therefore potentially their culpability, in an event or set of events. The results obtained from applying the model to the Enron data concur with those established during the FERC investigation, indicating that the techniques used provide a valid indication of real world activity. Moreover, the case suggests that the most significant results can be gained by aggregating key actor data into a single data set, given the source-centric nature of email data. This aggregation provides the additional benefit of enabling the investigator to recreate the role of actors whose data may have been excised or obfuscated.

When combined with supporting evidence, the qualitative and quantitative analysis of this social network provides some indicative and interesting conclusions. The executive oversight committee of Enron took a stronger stance on the financial dealings of the company and its strategic partnership companies. The previous financial controller was an unwilling participant in fraudulent activities. His replacement was promoted and co-opted on the basis of tractability and willingness to commit fraudulent acts. The holder of this role (Actor W) became a prominent actor within the management reporting structure of the network. This position allowed the doctoring of information that flowed around the network so as to benefit the prominent actors. Actor Z, the architect of the strategic partnership fraud methodology, is seen in these graphs but with a low profile. The information they received from the network appears to come from two more prominent sources – Actors X and Y – indicating strong ties to these actors. These actors and relational information are identified by applying the model posited in this paper.

5. Conclusions and Future Work

Our reliance on email communications, particularly within the workplace, ensures that this type of data will continue to feature during digital forensics investigations. Email data provides not only evidence of the flow of information through a network, but also an indication of actor relationships. This relational information may not only identify potential sources of evidence in cases involving many actors, but also when measured, it may provide a quantified assessment of a suspect's culpability in an event or set of events.

Recent work in digital forensics email analysis has focused on the identification and extraction of specific artefacts due to a lack of standardisation across these applications. Little work beyond those proposing specific tools has been conducted into proposing general frameworks for the wider digital investigation of email data. Therefore, this paper has posited a social network discovery model for email investigations. This model reflects the processes required to triage, visualise, analyse and quantify both tangible and intangible information contained within the network. The model has been applied to the Enron email data set to demonstrate its applicability.

Future work aims to develop the model further by highlighting relevant actors in email social networks in an automated way and testing the methodology by using data for case studies from other domains.

6. References

Bird, C., Gourley, A., Devanbu, P., Gertz, M., and Swaminathan, A. (2006), "Mining e-mail social networks", Proceedings of the 2006 international Workshop on Mining Software Repositories, Shanghai, China, 2006, pp. 137-143.

Cocciolo, A. Chae, H. S., Natriello, G. (2007), "Using social network analysis to highlight an emerging online community of practice", Technical Report available at http://edlab.tc.columbia.edu/files/cscl_final_1.pdf (accessed 28 February, 2011).

Cohen, W.W., 2010. "Enron Email Dataset", http://www.cs.cmu.edu/~enron/, 21 August, 2010 (accessed 26 February, 2011).

Collingsworth, B., Menezes, R., Martins, P. (2009), "Assessing Organizational Stability via Network Analysis", Proceedings of IEEE Symposium on Computational Intelligence for Financial Engineering, Nashville, USA, 2009, pp. 43-50.

Debbabi, M. Hadjidj, R. Lounis, H. Iqbal, F. Szporer, A. Benredjem, D. (2009), "Towards an integrated e-mail forensic analysis framework", Digital Investigation, Volume 5, Issues 3-4, March 2009, pp.124-137.

Dellutri, F., Laura, L., Ottaviani, V., Italiano, G.F. (2009), "Extracting Social Networks from Seized Smartphones and Web Data", Proceedings of the 1st International Workshop on Information Forensics and Security, London, UK, 2009, pp. 101-105.

Faust, K. (1997), "Centrality in Affiliation Networks", Social Networks, Volume 19, pp. 157-191.

Findlaw, http://fl1.findlaw.com/news.findlaw.com/wsj/docs/enron/sicreport/chapter2.pdf (accessed 28 February, 2011).

Haggerty, J., Lamb, D. & Taylor, M. (2009), "Social Network Visualization for Forensic Investigation of E-Mail", Proceedings of the 4th Annual Workshop on Digital Forensics and Incident Analysis, Athens, Greece, 2009, pp. 81-92.

Hawe, P., Ghali, L. (2008), "Use of social network analysis to map the social relationships of staff and teachers at school", Health Education Research, Volume 23, 2008, pp. 62-9.

Krebs, VE. (2002), "Uncloaking Terrorist Networks", First Monday, Volume 7, Number 4, April 2002.

Lin, H. (2010), "Predicting Sensitive Relationships from Email Corpus", Proceedings of the 4th International Conference on Genetic and Evolutionary Computing, Shenzhen, China, 2010, pp. 264-267.

Moreno, J. L. (1951), Sociometry, Experimental Method and the Science of Society: An Approach to a New Political Orientation, Beacon House, Beacon, New York.

Murphy, K. (2006), "No Love Lost at the Enron Trial", Business Week, February 15, 2006, http://www.businessweek.com/bwdaily/dnflash/feb2006/nf20060215_9015_db017.htm (accessed 28 February, 2011).

Wasserman, S., Faust, K. (1994), Social Network Analysis: Methods and Applications, Cambridge University Press, Cambridge.

Wei, C., Sprague, A., Warner, G., Skjellum, A. (2008), "Mining spam e-mail to identify common origins for forensic application", Proceedings of the 2008 ACM Symposium on Applied Computing, Fortaleza, Brazil, 2008, pp.1433-1437.

Wiil, U.K., Gniadek, J., Memon, N. (2010), "Measuring Link Importance in Terrorist Networks", Proceedings of the 2010 International Conference on Social Networks Analysis and Mining, Odense, Denmark, 2010, pp. 225-232.

Zhou, Y., Fleischmann, K.R., Wallace, W.A. (2010), "Automatic Text Analysis of Values in the Enron Email Dataset: Clustering a Social Network Using the Value Patterns of Actors", Proceedings of the 43rd Hawaii International Conference of System Sciences, Hawaii, USA, 2010, pp. 1-10.