A Framework for the Forensic Investigation of Unstructured Email Relationship Data

John Haggerty, University of Salford, UK

Alexander J. Karran, Liverpool John Moores University, UK

David J. Lamb, Liverpool John Moores University, UK

Mark J. Taylor, Liverpool John Moores University, UK

ABSTRACT

The continued reliance on email communications ensures that it remains a major source of evidence during a digital investigation. Emails comprise both structured and unstructured data. Structured data provides qualitative information to the forensics examiner and is typically viewed through existing tools. Unstructured data is more complex as it comprises information associated with social networks, such as relationships within the network, identification of key actors and power relations, and there are currently no standardised tools for its forensic analysis. This paper posits a framework for the forensic investigation of email data. In particular, it focuses on the triage and analysis of unstructured data to identify key actors and relationships within an email network. This paper demonstrates the applicability of the approach by applying relevant stages of the framework to the Enron email corpus. The paper illustrates the advantage of triaging this data to identify (and discount) actors and potential sources of further evidence. It then applies social network analysis techniques to key actors within the data set. This paper posits that visualisation of unstructured data can greatly aid the examiner in their analysis of evidence discovered during an investigation.

Keywords: *Digital Forensics, Email, Framework, Social Network Analysis, Structured and Unstructured Data*

INTRODUCTION

Email has replaced the letter as the main written communications medium, both in business and personal life. It is estimated that the average employee spends a quarter of their time on email-related tasks and the average number of emails sent by a corporate user per day is 43 (Orloff, 2011). Email can therefore provide the investigator with a wealth of information through both *structured* and *unstructured* data. Emails are structured in that they are formatted according to RFCs 5321 and 5322 (Klensin, 2008a, 2008b). However, they also provide unstructured information through the communications links and contacts that form

DOI: 10.4018/jdcf.2011070101

social networks. Digital forensics software, such as EnCase (Guidance, 2011) or the Forensics Toolkit (FTK) (Access, 2011), are useful for analysing structured data, i.e., the content of the emails in their textual form. However, they do not elucidate the unstructured data, such as relationships within the email network, power relations or network bridges that may be a key concern to a forensics investigation. Social network analysis and visualisation techniques can significantly contribute to evidence discovery and collection by identifying and understanding relationships and data flow between actors. Moreover, it may be used to quickly identify key events of interest within the email social network.

A number of challenges exist to today's digital forensics investigations involving email. As with many forensic investigations, cases routinely involve more than one computer (Richard & Roussev, 2006) and the investigator is unlikely to have access to all computers involved in the email network. Digital forensics investigatory models do not currently differentiate between email and any other data. Current work on digital investigations involving email data focus on techniques for the extraction of evidence, for example, 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, which detail techniques for the visualisation and extraction of specific email artefacts or features. However, this research focuses on particular aspects of email data rather than the wider process. Finally, much of the evidence that is recovered during an investigation may not be analyzed beyond the structured data view. For example, an examiner will manually trawl through the emails 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 may lead the investigator to other relevant sources of evidence, actors related by events or power relationships that are relevant to an investigation.

This paper presents a novel framework for the forensic investigation of unstructured email data. This framework follows traditional digital forensics procedures but incorporates tools and techniques for the triage and analysis of emails. This is achieved by using social network analysis and data visualisation to identify relevant evidence from unstructured data. This paper is organised as follows. In the next section we discuss relevant literature in the areas of social networks and email analysis. Following this, the novel framework is posited. In order to demonstrate the applicability of the approach, relevant stages are applied to the Enron email corpus as a case study. Finally, we make our conclusions and discuss further work.

SOCIAL NETWORK ANALYSIS AND EMAIL INVESTIGATIONS

Investigations involving social networks have raised in prominence due to the reliance of users on their digital communications and the information about a suspect these data sources may yield to the forensic examiner. Computer forensics tools, such as EnCase (Guidance, 2011) and FTK (Access, 2011), are used by examiners to recreate files and data from a suspect's computer. However, these tools are designed to analyse structured data rather than elucidate unstructured data that is present in communications. Moreover, these tools do not provide a visualization of evidence or quantify the importance of actors within a social network.

There are various representations of social networks in the digital world, of which email is just one. Other social network representations include friends on Facebook, followers of a Twitter account or a contacts list within a mobile phone, each with different levels of information that they may provide to the examiner. Our reliance on email within the business environment provides the investigator with an

entropic representation of the suspect's social networks through qualitative and quantitative information derived from both structured and unstructured data.

A social network is an interconnected group or system and the relations, both logical and physical, between the actors. Wasserman and Faust (1994) suggest that the presence of relational information is a critical and defining feature of a social network. There is a tendency to assume that just because actors are linked they must form a cohesive social network. However, this is not necessarily the case and the relationships between network members must be explored to fully understand how these networks function. Therefore, when investigating social networks, we must account for unstructured data as this will help us understand the network dynamics.

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. For example, Granovetter (1973) highlights the strength of weak ties in which these relationships provide new and better opportunities by providing bridges between loosely interconnected groups. Freeman (1978) suggests that analysing centrality measures 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). Lawler and Yoon (1996) focus on commitment and power in exchange relationships. This research attempts to predict how and when people in an exchange become committed to their relationship. Therefore, there is a rich literature from which other disciplines may learn and incorporate into their own research.

Current tools for social network analysis often require users to write scripts which are interpreted by the software to present the network visually. Applications such as *Pajek* (Vlado, 2011) or *SocNetV* (Kalamaras, 2011) visualise network information via the connection of vertices through arcs and edges. Once produced, the network and the relationships between actors may be explored. These applications provide various tools for measuring networks. For example, *SocNetV* provides a number of measures and layouts for analysing actors' power and centrality within the network based on graph theory. In addition, weighting may be applied to network edges to represent strength of ties. The use of weighting (for example by value or frequency) provides a deeper analysis of network dynamics (Perer & Schneiderman, 2009). Due to scalability issues in these visualisations where large graphs, such as those identified through email, soon become counter-intuitive, other approaches to network visualisations have been proposed. For example, Lee *et al.* (2006) propose *TreePlus*, a tree layout approach to explore social networks. These trees can be expanded or reduced to provide a more intuitive view of larger networks. Falkowski *et al.* (2006) suggest an approach for analysis of subgroup evolution in social networks. This approach uses a number of views to facilitate analysis, and displays the network in a graph but organised along a temporal plain. Hu and Gong (2010) present a visualisation of individuals' spatial-temporal social networks through three-dimensional graphs. Snasel *et al.* (2009) propose the use of Galois lattices for social network analysis in a small-scale network comprising 18 actors over 14 social events.

Despite issues of scalability in current tools, social network analysis has been proposed as an area that would benefit digital investigations. For example, Wang and Daniels (2008) suggest a graph-based approach to aid the investigation of network attacks. These approaches may also be used for data held in email applications. For example, Haggerty *et al.* (2009) proposed the Email Extraction Tool (EET) for the automated 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 these actors. This study uses a number of measures associated with social network analysis to identify key nodes. Wei *et al.* (2008) suggest a data mining approach to detect email spam. Whilst this approach does not visualise data, it does use techniques associated with this type of approach. Kim (2007) takes this further by incorporating visualisation of data to distinguish between 'spam' and legitimate emails. However, this approach does not identify the activities and relationships of a suspect under investigation. Henseler (2010) suggests an approach for filtering large email collections during an investigation based on statistical and visualisation techniques. This method uses the Enron email corpus as the basis for its results. Hadjidj *et al.* (2009) propose a forensic email analysis framework based on an automated tool which utilises statistical and machine-learning techniques.

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 robust procedures for the analysis of unstructured data. The next section proposes such a framework.

INVESTIGATION FRAMEWORK

This section presents a framework for use in the investigation of email data. In particular, this approach focuses on incorporating social network analysis techniques into digital forensics investigations to elucidate structured and unstructured data. In this way, key pieces of information and evidence sources beyond the computer which holds the data may be identified.

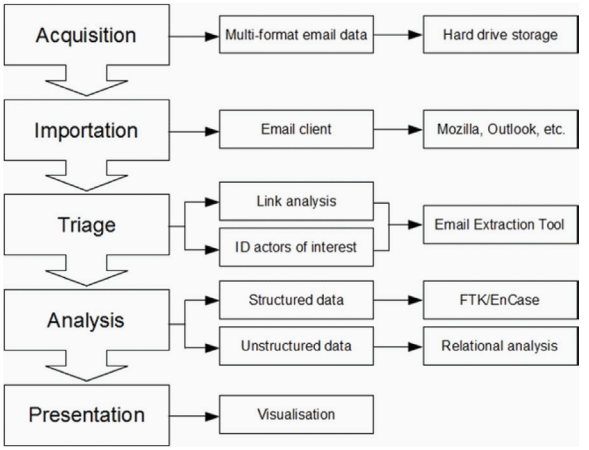
Social network analysis assumes that interpersonal ties between actors are important as they transmit behaviour, attitudes, information, goods and services (de Nooy et al., 2005). Within a digital forensics investigation, this is reflected in the unstructured data. As Viegas et al. (2004) posit, email 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.

Often, people will reflect their social groups in their organization of their email accounts by moving individual messages from their inbox into folders. For example, in a work-based email 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 and 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 includes:

- 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 roles do the other actors play within the network?
- What are the strong ties within these networks (i.e., who are the core of actors with whom the suspect regularly corresponds)?
- What is the relationship between the suspect and other actors in the network?
- Are there specific individuals that bridge disparate clusters of actors?
- Can other potential sources of evidence be identified through the email flow?

As discussed earlier, the forensic examiner will be faced with analyzing both structured and unstructured data during the course of an investigation. The approach for visualizing unstructured 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

Figure 1. Overview of the framework for the investigation of email data



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 suspects.

Figure 1 illustrates the framework for the investigation of email data. These processes do not differ much from any other investigation. However, the Triage and Analysis stages reflect the need to identify and assess relational information from the unstructured data. This is achieved through the use of software specifically aimed at this type of analysis.

As with any investigation, the data must be acquired in a robust manner, ensuring that the evidence maintains its integrity. Therefore, data is imported from the hard drive storage located on an image of the original hard drive. This is done as 'read only' to ensure that the data to be analysed is not modified. Email client files are located in software-specific directories. For example, Mozilla Thunderbird stores email data in text format in the following directories dependent on the operating system: C:\Documents and Settings\[UserName]\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) (Haggerty et al., 2009).

Data is imported from the email client directories as discussed above. Email, especially in the Windows environment, can be problematic to the investigator due to the number of formats and conventions available. Mozilla Thunderbird emails are the less problematic as they are stored in text format so there is less work to do in parsing the information. Email clients such as Microsoft Outlook are more problematic as the data is stored in a bespoke format. However, if emails are in the more complex Outlook PST file format, they can be converted to the Thunderbird, i.e., text, format using tools such as *libPST*.

Data is triaged using the Email Extraction Tool (EET) (Haggerty et al., 2009). This software automatically reads the data in the operating system location for the email client to perform a number of functions. The EET approach sets out to provide an initial view of two dimensions of email patterns; the social networks to which a suspect belongs (link analysis) and the strength of ties between actors within that network (identify actors of interest to the investigation team). This method differs from current approaches in that it does not focus on the textual context of individual emails. 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 network links will be from the suspect's point of view. An initial view of the suspect's strength of ties, indicated by quantitative, events-based analysis of key actors' communications, can provide an initial 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, and CC data in both messages sent to and received by the suspect are used. In addition, by mining the email applications 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. In effect, messages in email chains are treated as individual emails and explored for social network information. The Triage stage enables the investigator to gain an initial impression of the social networks evident in the suspect's emails and to identify key actors or further sources of evidence. In addition, it processes the data for output into other social network analysis tools, such as *Pajek*.

Analysis extends the findings of the triaged data to examine both the structured and unstructured data. The structured data refers to the email textual information, for example, header information and the email contents. Existing tools, such as FTK and EnCase provide the facility to read this data in a forensically sound manner. Depending on the quantity of emails recovered, this may be a time-consuming process. However, investigators can use the results of the triage stage to identify only those emails relevant to the investigation, thereby reducing this time considerably. Unstructured data analysis refers to the examination of the emails for relational information. The use of social network analysis tools, such as *Pajek* or *SocNetV*, enable the examiner to quantify the relationships identified in the email data by using statistical and mathematical techniques to provide different viewpoints of the network. Quantifying the relationships enables the investigator to identify the role that actors within the network fulfil in relation to the event(s) under investigation, particularly to identify culpability (Haggerty et al., 2008) or potential sources of further evidence.

Unstructured data may be analysed in a number of ways using existing social network analysis techniques to quantify relational data. Tools, such as *SocNetV*, apply graph theory to the network data and enable the investigator to measure and visualise the network in a variety of ways. A combination of various measures provides an overview of the dynamics of the network from different viewpoints and an appreciation of the actors' or groups' roles therein. *Out-degree centrality* measures the expansiveness, or number, of actors that a particular actor possesses or accesses. It therefore measures which actors provide the most potential connections within the network. *In-degree centrality* is the inverse of out-degree centrality to indicate a node's receptivity or "popularity" in the network and can be used to identify key network facilitators. *Betweeness centrality* identifies potential points of control within the network. Those that act as chokepoints (are the most between others) may be centres of power, control and influence because they can choose to retain or share information with other actors in the network. Betweeness recognises that communication flow within a given network often does not rely on adjacent actors but moves along geodesics. Because they are focal points of communications within the group and subgroups, certain actors facilitate contact and communications within the network and can therefore be seen as major channels of influence. For further details on quantitative analysis of social networks and the details of these measures, see Freeman (1978) and Wasserman and Faust (1994).

The final stage of the framework is the presentation of evidence. The visualisations of data produced in the previous two stages are a key part of the presentation as they can be used to elucidate any wider results obtained during the investigation. These visualisations will be supported by evidence obtained through structured data analysis.

CASE STUDY AND RESULTS

This section demonstrates the applicability of the framework, as detailed in the previous section, by applying it to the Enron email corpus. The aim of this case study is not to provide a full analysis of this data set, but to illustrate the process of investigating unstructured email data. The Enron email corpus contains data from approximately 500,000 emails, the accounts of which were predominantly senior management at the company. This data has been released by the Federal Energy Regulatory Committee (FERC) and is sanitised for use in research (Cohen, 2009). The data set provides real data from a large-scale fraud that can be used for research 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 email analysis framework may be tested.

Background to the Enron Fraud

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. For a detailed report of the investigation, see Powers et al. (2002).

The main actors in the Enron fraud are as follows:

- Jeffrey Skilling: former president of Enron Corporation; Chief Operating Officer from 1999; responsible for Enron's move into energy wholesale and Internet trading; introduced accounting methods that treated anticipated profits as if they were real gains. He was convicted of multiple federal felony charges (conspiracy, securities fraud, false statement and insider trading) in 2006.
- Kenneth Lay: Enron Chairman and Chief Executive Officer (CEO) except for the months of Skilling's tenure from 1985. He was found guilty of ten counts of securities fraud but died before sentencing.
- Andrew Fastow: Chief Financial Officer; set up a network of off-balance-sheet companies controlled by Enron to hide Enron losses. He received a six-year sentence on charges of conspiracy, insider trading, money laundering, and wire and securities fraud.
- Jeffrey McMahon and Ben Glisan: former Enron treasurers executives; chiefly responsible for brokering the relationships and financial dealing between Enron and its strategic partnerships.
- Ken Rice: former Enron executive and president of Enron's broadband service, found guilty of over 15 counts of securities fraud.
- Sherron Watkins: considered as a whistleblower of the Enron fraud. She wrote a concerned email to Kenneth Lay warning of financial statements irregularities.

The Enron fraud ran over a number of years before being discovered in 2002. Due to the complexity of the case, it took another four years until conviction of the main protagonists. The timeline of key events are shown in Table 1.

Table 1. Timeline of key events in the Enron fraud

Year	Key events
1999	Board of Directors allows Andrew Fastow to run private business partnerships to purchase poorly performing Enron assets. Andrew Fastow creates LJM and Chewco for this purpose. These are later found to have been created to disguise Enron debt and inflate profits.
2000	Enron implicated as a major culprit in the California energy shortage.
2001	February - Jeffrey Skilling named CEO of Enron. August – Jeffrey Skilling resigns as CEO of Enron, stating he needed more time with his family. Kenneth Lay returns from retirement to become CEO of Enron. Enron reports losses of $618 million. Sherron Watkins expresses concerns over Enron business practices. Enron publicly discloses an open inquiry by the Securities and Exchange Commission into the limited partnerships between Enron and the companies Chewco and LJM owned by Andrew Fastow. Andrew Fastow is ousted from his position at Enron. Enron releases information that it has overstated profits by an amount of $586 million dollars. Enron files for bankruptcy protection.
2002	Investigations into Enron's financial dealings begin. Kenneth Lay resigns as CEO of Enron stating too much of his time would be taken up with investigations. An Enron board report concludes that Enron executives intentionally manipulated company profits. Andrew Fastow arrested on charges of fraud and money laundering.
2003	Eleven Enron executives indicted. Ben Glisan pleads guilty to fraud.
2004	Andrew Fastow pleads guilty to two counts of conspiracy and receives a lesser sentence for turning state's evidence. Jeffrey Skilling and Kenneth Lay indicted on counts of wire fraud, securities fraud, insider trading and conspiracy.
2006	Kenneth Lay and Jeffrey Skilling convicted of multiple counts of fraud and conspiracy. Both men receive heavy jail sentences.

Acquisition and Importation

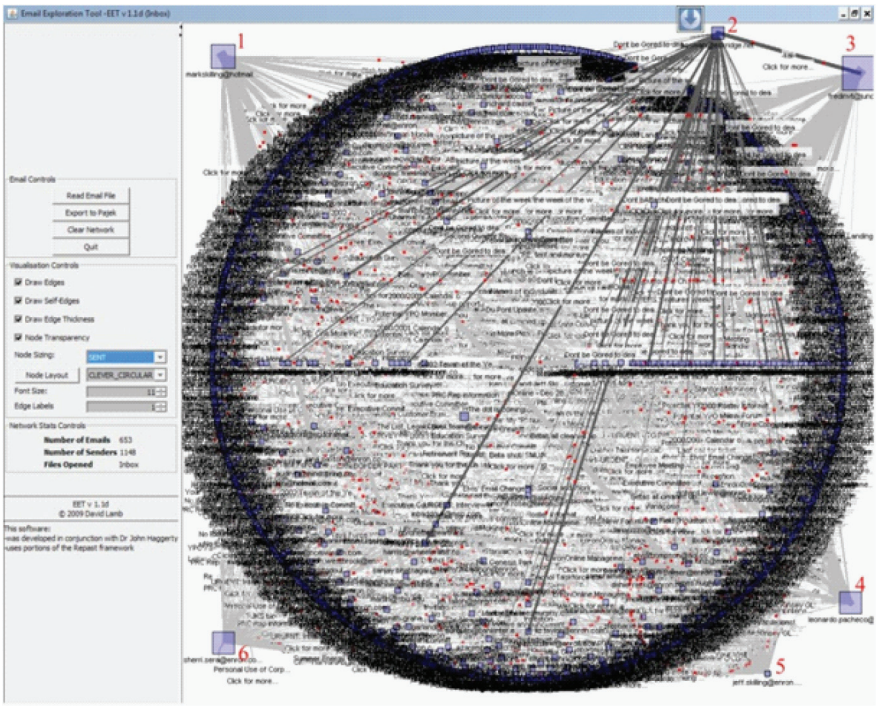
In a digital investigation, images will be taken of all hard drives and the data will be analysed in a robust manner following well described procedures. These hard drive images will contain the email files and the email client preferred by the user. The Enron corpus is available both in mbox and PST formats online. Due to the quantity of email data available in the Enron corpus, this paper will concentrate on two of the main actors in the fraud, Jeffrey Skilling and Andrew Fastow, to demonstrate the applicability of the approach. This reflects many digital forensics investigations where one or two suspects will form the starting point for an examiner. Moreover, it will apply their data to the two stages in the framework related to the analysis of unstructured data; triage and relational analysis.

Triage

As discussed above, the Triage stage aims to identify important links between the suspect and other actors in the network or to discount actors from the investigation. In this way, the number of actors and emails requiring analysis may be reduced considerably. The EET tool has a number of network views to aid the examiner in this task and it is outside the scope of this paper to discuss them here. However, more details of the software, including the different network views that it affords to aid analysis are detailed in Haggerty et al. (2009).

Figure 2 illustrates Jeffrey Skilling's "discussion threads" emails organised using EET's 'Clever Circular' layout. This layout organises actors in a circle, with actors with stronger links in the network laid out horizontally in the

Figure 2. Jeffrey Skilling Inbox "discussion threads" view by senders



centre of the circle. Node sizing is by sender. Edge labels comprise the subject line of the email(s) sent between the actors. This particular network contains 653 emails with 1148 actors. The data is filtered to select only those emails that contained in-depth discussions between Jeffrey Skilling and members of Enron corporate body and discussions with others external to the company.

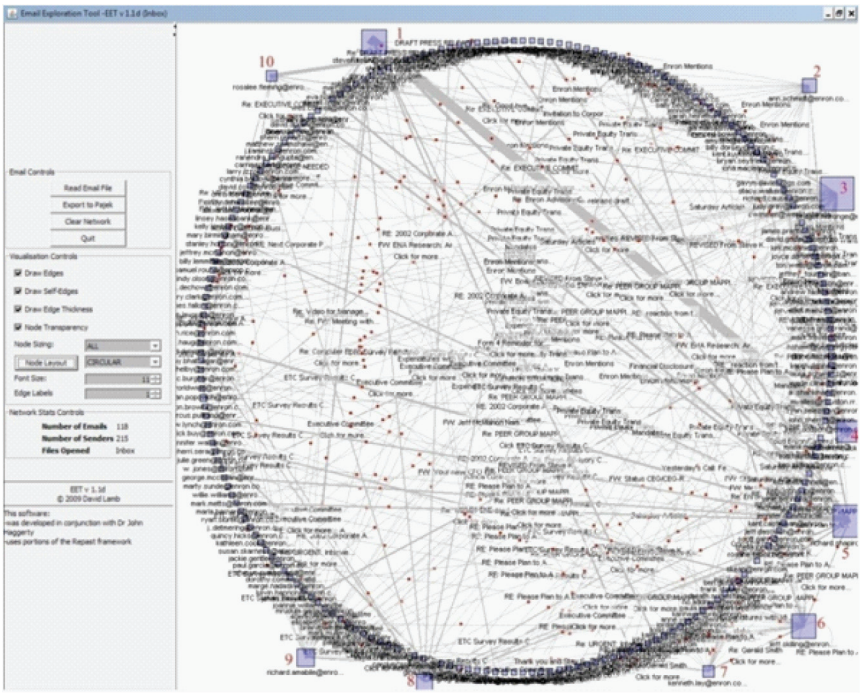
In Figure 2, we see Jeffrey Skilling (actor 5) receding into the noise of the network. Five actors appear significant in this view, one of whom (actor 6) may be identified as a potential source of further evidence due to her prominence in this network view and as her company role as Skilling's Personal Assistant (PA). The PA will naturally be prominent in the social network featuring the main actor as the position naturally bridges many network groups within the organisation.

Another significant actor identified in this view (actor 1) represents a member of Skilling's family, a fellow Enron employee. The edge labels suggest that these emails represent personal jokes and other topics not relevant to the investigation. This actor stands out because of the sheer volume of traffic, as the emails are sent to many other actors within the network as well as to Skilling.

Actors 2, 3 and 4, are highlighted as significant on this view. Actors 2 and 3 represent members of networks that are external to the Enron corporate structure. Email headers suggest that these are exclusive "insider" networks comprising relationships formed outside of the Enron structure. These relate to fraternity membership and relationships formed by network members at college.

The final significant actor (actor 4) represents another network bridge. This actor (Leonardo Pacheco) bridges the relationship and networks between the current CEO, Jeffrey Skilling, and former CEO, Kenneth Lay. Historical documents detailing the downfall

Figure 3. Andrew Fastow's Enron email network



of Enron show that this actor was a confidante of Kenneth Lay and disseminator of financial information around the greater Enron company network. Given this information and the actor's significant position in the network it would be natural to assume some knowledge of the fraudulent activities of other members. However, although this actor received higher than usual bonuses from Enron, he was not implicated in the Enron fraud.

Figure 3 illustrates Andrew Fastow's email networks consisting of 188 emails with 215 unique actors. This data comprises four different sources within the Enron data set as many of his contacts were external to the company and therefore not included in the email corpus.

Ten main actors can be identified from this view. These actors have been teased out of the main body to give a better impression of how the e-mails are linked and of how information flows around this network group. Node sizing ALL, i.e., both senders and receivers, was used to gain an overall impression of the importance of the main actors within this network. From this view we can see some familiar actors, such as Kenneth Lay, Jeffrey Skilling, PAs Rosalee Fleming and Sheri Serra, and some new actors, Ken Rice, Jeffrey McMahon and Ben Glisan. The email headers suggest that this set of actors and the context of the data set relate to finance. Indeed, all of the actors except the PAs have corporate positions congruous to this.

Actors 1 and 4 discuss potential draft press releases concerning financial disclosure, with actor 5 receiving overviews and sending them on to actor 6. At face value, these links present nothing of real significance, as this sort of communication could be assumed to be a part of standard corporate operating procedure. It is only in the nature and content of these email communications that their importance becomes apparent. A significant portion of the email communications that were used in this aggregated data set later appear as being entered as exhibits used in evidence by the Department of Justice.

This section has demonstrated the relevance of the Triage stage within the framework by elucidating unstructured data, i.e., relationships and sizing nodes by prominence in the network, on a small sample of the available data. In this way, key potential sources of evidence and significant actors have been identified, whilst others, such as Skilling's familial relation can be discounted. The next section will use relational analysis on these networks to further explore the unstructured data to provide evidence to the examiner.

Relational Analysis

This section details the analysis of the centrality measures to present a quantitative analysis of the email data as a centrality graph, which allows for qualitative interpretation. These circular radial graphs, created by exporting data from EET to SocNetV, 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.

Figure 4 illustrates the betweeness centrality of actors in the Jeffrey Skilling data set presented above. Given the source-centric nature of the data, it is to be expected that Skilling (actor A) would be placed at the centre of the graph. His PA Sheri Serra (actor B), would also appear more central than other actors given the nature of the actor's relationship with Skilling.

Figure 5 illustrates the in-degree view of the Skilling network. Again, Skilling (actor A) is the central actor receiving a lot of information from the network. An email alias of Skilling (actor C) also appears in this graph and shows a higher level of in-degree centrality than many other actors. This could indicate a separate private store of emails which may be used for archival purposes or as a means of disseminating information from these discussions outside of the Enron corporate structure. The next most prominent actor in this graph is Skilling's PA (actor B). James Hughes (actor D) is another actor identified by using this measure of centrality. This actor's position in the network displays a high level of access to information from across the network, highlighting this actor as an information broker. This interpretation is borne out by documents from the Enron trial which show that this actor was a key member in enabling the flow of fraudulent information around and beyond the Enron corporate network. Former CEO Kenneth Lay (actor E) is also highlighted using this measure. This actor's position in the graph indicates only a marginal position of importance within Skilling's network for this data set.

Figure 6 displays a graph for out-degree centrality of the Skilling network. This graph shows the main actors from the in-degree centrality graph retreating into the background of the network. As discussed, these actors received the most information from the network. Skilling (actor A) and his PA (actor B) are still prominent in this graph as information recipients but this role is marginal at best. Instead a large number of other network actors come to the fore with none standing out as significant due to the sheer number of actors. An interesting feature of this graph is the number of connections between these actors and Skilling. Similar to the in-degree graph, Skilling has exceptional connectivity within the network.

These graphs help to interpret Skilling's position as a key actor in the Enron network. The Triage stage suggests that this actor has a high degree of control over information flowing out of this network and being disseminated publicly. Furthermore, from the centrality measures, this actor shows up as key in both the receipt and dissemination of fraudulent documents within the Enron network. Evidence for this can be seen from his links with James Hughes and Kenneth Lay and subsequent documentary evidence (USDOJ, 2006). Both

Figure 4. Jeffrey Skilling betweeness centrality

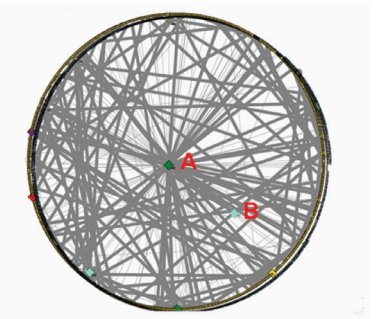
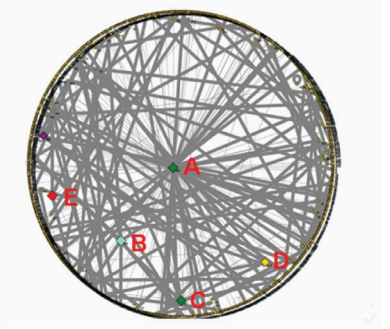


Figure 5. Jeffrey Skilling in-degree centrality

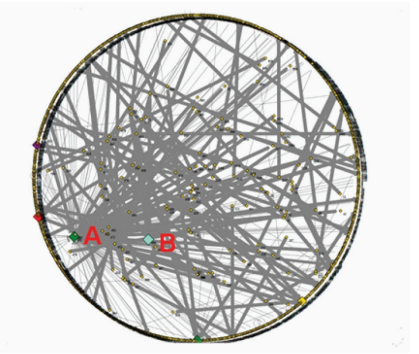


Skilling and Lay are prominent in the data set, whilst Andrew Fastow's presence is barely detectable despite documents from the time of the Enron collapse showing that he was one of the main actors in the Enron fraud.

Figure 7 illustrates the betweeness centrality measurement of Andrew Fastow's data set. From the graph, the financially related data involving Fastow (actor H) indicate that Kenneth Lay (actor E) is the most central actor. This was at the time when Enron's financial situation was being called to account. Another actor of significance, Steve Kean (actor F), was Chief Financial Officer of Enron. The next actor of significance in this graph highlighted with a red circle (actor G) is Rosalee Fleming, Lay's PA. A PA's presence can be largely considered as an agent of the actor they represent. What is significant about this actor here, with importance to both the framework and the larger context of the case study, is that PAs tend to bridge the social networks of those they represent and will be potential sources of further evidence.

Skilling (actor A) is highlighted as a significant actor within the network. He is displayed here with strong ties to Fastow who remains a peripheral figure in this graph, with little ability to influence the flow of information within the network. Also of significance is Ben Glisan (actor I), who was chief treasurer for Enron at the time of the collapse and who was found guilty of wire and securities fraud. His position in this data set displays his ability to influence and disseminate information around the network. Of equal significance to Skilling is Vanessa Groscrand (actor J). This actor is a member of the Enron advisory committee which provides some oversight of the financial transactions undertaken by the company.

Figure 6. Jeffrey Skilling out-degree centrality



The in-degree measurement of this data set in Figure 8 displays Jeffrey McMahon as possessing the highest in-degree centrality measure (actor K). McMahon was the former treasurer (superseded by Glisan) of Enron and responsible for brokering the early partnerships between Enron and Fastow's three companies, which accounts for his position of prominence within this graph. McMahon was an unwilling participant in the fraudulent activities of other actors and was thus replaced with Glisan (actor I). This would account for Glisan's position of prominence.

The actor possessing a next highest in-degree centrality in this graph is Skilling (actor A). His position in the graph indicates that he is a central figure for communications directly involving Fastow and the financial dealings between Enron and the subsidiary companies. As in previous graphs, the PAs of Lay and Skilling appear as significant (actors L). Also, Fastow (actor H) appears to be of little significance. He possesses only one strong tie to another actor (Skilling) within the network and receives all information updates from this source.

Figure 9 illustrates the out-degree centrality for actors within the data set. In this graph the PAs (actors L) for Lay and Skilling are the most prominent. This is an indicator of how much information they send into the network on behalf of their respective superiors. The next actor of significance is Ken Rice (actor M) and it is the first time his presence is seen in all of the centrality measurements thus far. As the head of the Enron broadband division, he was indicted on 37 counts of fraudulent activities. Email topics within this dataset included financial "overviews" of the position of Enron broadband within the market, which was later reported as code for fraudulent disclosure of financial information.

The next actor of significance is Steve Kean (actor F) who was a senior vice president of Enron in charge of public affairs and also a member of the Enron advisory council. His presence in this position indicates that he is a prime sender of information with the ability to monitor much of the information that flowed across this network (as indicated by a position of prominence in the betweeness graph, see Figure 7). Another actor of prominence is Sheri Serra (actor B), Skilling's PA, during the time period of this data set. Of equal out-degree centrality in this graph are Skilling (actor A) and Vanessa Groscrand (actor J). The increased prominence of members of the Enron executive oversight committee, such as Groscrand, is of special significance in these graphs. This increase indicates that the oversight committee was clearly beginning to take a more proactive approach to overseeing the actions and financial transactions of some of the key members of the network.

When combined with supporting evidence, the quantitative analysis of this social

Figure 7. Andrew Fastow betweeness centrality

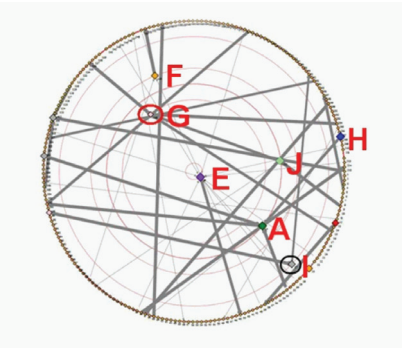
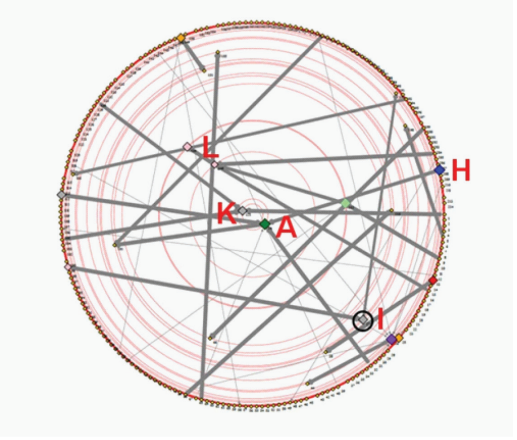


Figure 8. Andrew Fastow in-degree centrality

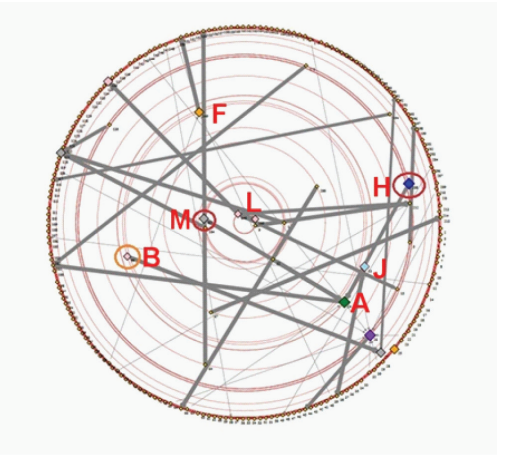


network provides some indicative and interesting conclusions. The executive oversight committee of Enron took a stronger stance on the financial dealings of the company and its strategic partnership companies. The previous financial controller, Jeffrey McMahon, was an unwilling participant in fraudulent activities. His replacement, Ben Glisan, was promoted and co-opted on the basis of tractability and willingness to commit fraudulent acts. Glisan 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. Andrew Fastow, 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 – Kenneth Lay and Jeffrey Skilling – indicating strong ties to these actors. These actors and relational information are identified by applying the framework posited in this paper.

CONCLUSION

Our reliance on digital communications, of which email is a subset, ensures that this type of data will remain prominent in a digital forensics investigation. Emails comprise both

Figure 9. Andrew Fastow out-degree centrality



structured and unstructured data. Structured data provides qualitative information to the forensics examiner and is typically viewed through tools such as EnCase and FTK. Unstructured data is more complex as it comprises information about a social network, such as relationships, identification of key actors and power relations, and there are currently no standardised tools for its forensic analysis. In addition, visualisation of this data can greatly aid the examiner in their understanding of the evidence.

This paper presents a framework for the analysis of unstructured (and structured) data and applies it to a case study to demonstrate its applicability. Link analysis visualisation provides a rudimentary analysis of the email data and can be used to effectively triage the potentially large data sets to identify key actors (and discount others) within the network. This can then be used to inform the deeper analysis of unstructured data, whereby relevant actors may be assessed. The application of social network analysis tools and techniques to measure the network not only provide 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 framework to the Enron data concur with those established during the FERC investigation, suggesting that the techniques used in this analysis provide a valid indication of real world activity. Moreover, the case study suggests that 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, such as Andrew Fastow, whose data may have been excised or obfuscated.

Further work aims to develop the framework and appropriate tools for digital investigations. For example, there is a temporal element to unstructured data that would further enhance the understanding of social network information by identifying key nodes of influence and the development of the network over time. In addition, work aims to incorporate network narrative analysis tools. In this way, the investigator may visualise not only the network relationships, but the structured data's influence on network behaviour.

REFERENCES

Access Data. (2011). *FTK forensic tool kit.* Retrieved from http://www.accessdata.com

Bird, C., Gourley, A., Devanbu, P., Gertz, M., & Swaminathan, A. (2006). Mining e-mail social networks. In *Proceedings of the International Workshop on Mining Software Repositories* (pp. 137-143).

Carenini, G., Ng, R., Zhou, X., & Zwart, E. (2005). Discovery and regeneration of hidden emails. In *Proceedings of the ACM Symposium on Applied Computing* (pp. 503-510).

Cohen, W. W. (2009). *Enron email dataset.* Retrieved from http://www.cs.cmu.edu/~enron/

Collingsworth, B., Menezes, R., & Martins, P. (2009). Assessing organizational stability via network analysis. In *Proceedings of the IEEE Symposium on Computational Intelligence for Financial Engineering* (pp. 43-50).

de Nooy, W., Mrvar, A., & Batagelj, V. (2005). *Exploratory social network analysis with Pajek*. Cambridge, UK: Cambridge University Press.

Dellutri, F., Laura, L., Ottaviani, V., & Italiano, G. F. (2009). Extracting social networks from seized smartphones and web data. In *Proceedings of the 1st International Workshop on Information Forensics and Security* (pp. 101-105).

Falkowski, T., Bartelheimer, J., & Spiliopoulou, M. (2006). Mining and visualizing the evolution of subgroups in social networks. In *Proceedings of the International Conference on Web Intelligence* (pp. 52-58).

Freeman, L. C. (1978/79). Centrality in social networks. *Social Networks*, *1*, 215–239. doi:10.1016/0378-8733(78)90021-7

Granovetter, M. S. (1973). The strength of weak ties. *American Journal of Sociology*, *78*(6), 1360–1380. doi:10.1086/225469

Guidance Software. (2011). *Encase.* Retrieved from http://www.guidancesoftware.com

Hadjidj, R., Debbabi, M., Lounis, H., Iqbal, F., Szporer, A., & Benredjem, D. (2009). Towards an integrated e-mail forensic analysis framework. *Digital Investigation*, *5*(3-4), 124–137. doi:10.1016/j.diin.2009.01.004

Haggerty, J., Lamb, D., & Taylor, M. (2009). Social network visualization for forensic investigation of e-mail. In *Proceedings of the 4th Annual Workshop on Digital Forensics and Incident Analysis* (pp. 81-92).

Haggerty, J., Taylor, M., & Gresty, D. (2008). Determining culpability in investigations of malicious email dissemination within the organisation. In *Proceedings of the 3rd Annual Workshop on Digital Forensics and Incident Analysis* (pp. 12-20).

Henseler, H. (2010). Network-based filtering for large email collections in e-discovery. *Artificial Intelligence and Law*, *18*(4), 413–430. doi:10.1007/s10506-010-9099-3

Hu, B., & Gong, J. (2010). Modeling individual-based social network with spatial-temporal information. In *Proceedings of the International Conference on Management and Service Science* (pp. 1-4).

Kalamaras, D. B. (2011). *SocNetV.* Retrieved from http://socnetv.sourceforge.net

Kim, U. (2007). Analysis of personal email networks using spectral decomposition. *International Journal of Computer Science and Network Security*, *7*(4), 185–188.

Klensin, J. (2008a). *RFC 5321: Simple mail transfer protocol.* Retrieved from http://tools.ietf.org/html/rfc5321

Klensin, J. (2008b). *RFC 5322: Internet message format.* Retrieved from http://tools.ietf.org/html/rfc5322

Lawler, E. J., & Yoon, J. (1996). Commitment in exchange relations: Test of a theory of relational cohesion. *American Sociological Review*, *61*, 89–108. doi:10.2307/2096408

Lee, B., Parr, C. S., Plaisant, C., Bederson, B. B., Veksler, V. D., Gray, W. D., & Kotfila, C. (2006). TreePlus: Interactive exploration of networks with enhanced tree layouts. *IEEE Transactions on Visualization and Computer Graphics*, *12*(6), 1414–1426. doi:10.1109/TVCG.2006.106

Lin, H. (2010). Predicting sensitive relationships from email corpus. In *Proceedings of the 4th International Conference on Genetic and Evolutionary Computing* (pp. 264-267).

Orloff, J. (2011). *35 interesting statistics about email.* Retrieved from http://www.theemailadmin.com/2011/05/35-interesting-statistics-about-email/

Perer, A., & Schneiderman, B. (2009). Integrating statistics and visualization for exploratory power: From long-term case studies to design guidelines. *IEEE Computer Graphics and Applications*, 39–51. doi:10.1109/MCG.2009.44

Powers, W. C., Jr., Troubh, R. S., & Winokur, H. S., Jr. (2002). *Report of investigation by the special investigative committee of the Board of Directors of Enron Corp.* Retrieved from http://news.findlaw.com/wsj/docs/enron/sicreport/index.html

Richard, G. G. III, & Roussev, V. (2006). Next-generation digital forensics. *Communications of the ACM*, *49*(2), 76–80. doi:10.1145/1113034.1113074

Snasel, V., Horak, Z., Kocibova, J., & Abraham, A. (2009). Reducing social network dimensions using matrix factorization methods. In *Proceedings of the Conference on Advances in Social Network Analysis and Mining* (pp. 348-351).

United States Department of Justice (UDOJ). (2006). *Kenneth L. Lay and Jeffrey K. Skilling Jury trial – Government exhibits.* Retrieved from http://www.justice.gov/enron/exhibit/04-18/index.htm

Viegas, F. B., Boyd, D., Nguyen, D. H., Potter, J., & Donath, J. (2004). Digital artifacts for remembering and storytelling: PostHistory and social network fragments. In *Proceedings of the 37th Hawaii International Conference on System Sciences* (pp. 1-10).

Vlado, A. (2011). *Pajek.* Retrieved from http://vlado.fmf.uni-lj.si/pub/networks/pajek/default.htm

Wang, W., & Daniels, T. E. (2008). A graph based approach toward network forensic analysis. *ACM Transactions on Information and System Security*, *12*(1), 401–433. doi:10.1145/1410234.1410238

Wasserman, S., & Faust, K. (1994). *Social network analysis: Methods and applications*. Cambridge, UK: Cambridge University Press.

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

Wiil, U. K., Gniadek, J., & Memon, N. (2010). Measuring link importance in terrorist networks. In *Proceedings of the International Conference on Social Networks Analysis and Mining* (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. In *Proceedings of the 43rd Hawaii International Conference of System Sciences* (pp. 1-10).

John Haggerty is Lecturer in Information Systems Security at the University of Salford. After some years working in the 'real world', he returned to academia in 1999 to undertake a MSc and then PhD in Network Security at Liverpool John Moores University, which was completed in 2004. His main research interests include network security, digital forensics, social network analysis and mobile computing. He has published approximately 30 peer-reviewed papers in journals, conferences and books in these areas. In addition to these publications, he has developed a novel commercial application for digital forensics investigations involving multimedia images which is currently undergoing tests with police high-tech crime units.

Alexander J. Karran studied for his BSc and MSc degrees in the School of Computing and Mathematical Sciences at Liverpool John Moores University. He is currently a PhD student within the School of Psychology at Liverpool John Moores University undertaking a project that combines both psychology and computing.

David J. Lamb is a post-doctoral researcher in the School of Computing and Mathematical Sciences at Liverpool John Moores University working in Software Engineering for Large-Scale Systems, and Computer Security. He is currently an active member of the school's Applied Research Computing Group and contributes to the EU ANIKETOS Computer Security project. He is a member of the Institute of Engineering and Technology, and has 12 years' experience developing software systems for both industry and academia. This includes a wide variety of systems from small-scale mobile clients through desktop office systems, to web-based enterprise systems. He has developed for clients in both the UK public and private sectors; recent clients include a social housing trust and the National Health Service in the North-West of England.

Mark J. Taylor is a senior lecturer in computing at Liverpool John Moores University. He is a Chartered IT Professional, a Chartered Engineer and a Chartered Scientist.