Retrieval and Analysis of Web Search Narratives for Digital Investigations

J.Haggerty[1] and M.J.Taylor[2]

[1]School of Science & Technology, Nottingham Trent University, Clifton Campus,
Nottingham, NG11 8NS
[2]School of Computing & Mathematical Sciences, Liverpool John Moores University,
Liverpool, L3 3AF
e-mail: john.haggerty@ntu.ac.uk; M.J.Taylor@ljmu.ac.uk

Abstract

Our reliance on accessing information on the Web ensures that it provides a wealth of information to a forensics examiner. Current tools for the analysis of a suspect's Web activity return evidence ranging from cached data to URLs visited. However, these tools are not without their limitations, such as textual presentation of results, issues related to private browsing, and links between Web searches and subsequent behaviour. This paper presents a novel approach that visualises search strings in Web browser log files to present a narrative of a suspect's interests, motives and actions over time. The aim of this methodology is to triage and analyse potentially large data sets from a suspect's daily interaction with the Internet and to demonstrate intent during the activity under investigation. In order to demonstrate the applicability of the approach, data from a Web browser history file is parsed for search terms and the results visualised.

Keywords

Digital forensics, Web clients, narrative, data visualization

1. Introduction

The World Wide Web (or Web) and the Internet is pervasive and provides many ways for a user to access the information that they require. The Office of National Statistics (ONS, 2014) suggests that in the UK alone, 73 per cent of the adult population accessed the Internet every day. Moreover, Internet access from mobile devices more than doubled between 2010 and 2013 to 53 per cent. All a user requires to access this information is a Web browser, an application that resides on the host computer and interprets information from Web servers. Information in this software is by default logged, primarily to aid the user experience. However, the reliance on this technology ensures that it provides a wealth of information to a forensics examiner during an investigation regarding a suspect's access and use of the Internet.

The importance of Web log information for an investigation has been known for some time and many forensics tools make use of this data. For example, *Pasco v1.0* (Jones, n.d.) is a forensics application that parses the Internet Explorer *index.dat* file to retrieve evidence of Web sites that a suspect has visited. The data is returned as CSV files providing information such as URLs and times of access. However, the volume of data that such software returns may be considerable with many redundant entries, i.e. entries not related to the event(s) under investigation. Moreover, whilst it extracts Web browser evidence, it does not provide a qualitative analysis of a suspect's activity over time, merely a list of Web sites visited. In order to meet these and other shortcomings, data visualisation tools and techniques can be employed to aid the investigation process and triage evidence returned. As Thomson et al (2013) suggest, data visualization may be used by analysts to alleviate the overheads of interpreting data sets and improve the ability of users to make sense of activity patterns in event logs.

This paper presents a novel approach that visualises search strings in Web browser logs to present a narrative of a suspect's interests, motives and actions over time. In this way, it reduces ambiguity in links between search terms and subsequent Web interaction, as well as providing temporal and textual analysis to demonstrate relational information to the forensics examiner. The aim of this approach is to triage potentially large data sets from a suspect's daily interaction with the Internet to provide further sources of evidence or to support hypotheses during the investigatory process. Moreover, it may be used to prove or disprove the 'Trojan defence', whereby a suspect claims to have inadvertently followed a Web link; Web searches are an active, rather than passive, interaction and may be used to demonstrate intent.

This paper is organised as follows. Section 2 discusses related work in Web log analysis and data visualisation for security and forensics investigations. Section 3 presents an overview of forensic analysis of Web browser logs and posits our approach. Section 4 presents the results of a case study of using the approach on a Web browser history file to demonstrate the applicability of the methodology. Finally, we make our conclusions and discuss further work.

2. Related work

Web log analysis provides a wealth of information to the forensics investigator and the recording of user activity may take place on either the server or the client. Server-side Web logs record information about users that visit the Web site to improve the user experience and their log analysis may focus on content mining, usage mining or structure mining (Hadzic and Hecker, 2011). For example, Hernandez et al (2010) propose a model for the structuring of data to aid the log mining process across formats. Chowdhury et al (2010) propose an approach for mining Web access sequences, mainly focused on online databases. Nithya and Sumathi (2012) propose an approach for pre-processing data prior to analysis to remove usage noise and the recorded presence of Web robots. Whilst these, and other, approaches mine information available from servers, they are limited in their use for forensics investigations as the examiner would have to have prior knowledge of a suspect's use of the Web site.

Therefore, research in digital forensics has focused on client-side log mining and analysis to identify evidence. For example, Accorsi et al (2011) propose RECIF, an approach that utilises business process logs to identify illegal data flows. Al Mutawa et al (2011) propose an approach for the retrieval and analysis of Facebook Instant Messaging artefacts and identify issues surrounding the recovery of Arabic text. Alternatively, Hai-Cheng Chu et al (2011) focus on the live acquisition of previous

Facebook session evidence utilising the way in which data is written to a computer. Other approaches focus on Web browser logs and in particular the issue of a suspect attempting to disrupt potential evidence about their activities being recorded. For example, Said et al (2011), Ohana and Shashidhar (2013) and Satvat et al (2013) demonstrate the evidence that may be retrieved from a hard drive even when a Web browser is set to private browsing or when portable devices are used by a suspect. However, these approaches often focus on the extraction rather than triage and analysis of potential evidence relevant to the investigation.

The benefits of data visualization and interaction for large data sets have ensured that such approaches have been adopted within the security and forensics domains. Visualization enables an analyst to gain an overview of data during an investigation (Schrenk and Poisel, 2011). For example, Haggerty et al (2014) propose *TagSNet*, an approach for the quantitative and qualitative analysis of email data to enable a forensics examiner to analyze not only actor relationships but also visualize discourse between those actors. Koniaris et al (2013) present visualizations of their results of detecting attackers utilizing SSH vulnerabilities to attack honeypot systems. Giacobe and Xen Su (2011) propose an approach to visually represent security data in a geographical sphere based on IP address rather than physical location. Thomson et al (2013) posit *Pianola*, a system for visualizing and understanding the contents of intrusion detection logs.

Whilst these approaches provide visualisations of evidence, they do not demonstrate intent in committing a crime or taking part in a malicious event(s). The aim of this approach is to triage potentially large data sets from a suspect's daily interaction with the Internet. Moreover, the approach posited in this paper may be used to prove or disprove the 'Trojan defence'. The next section outlines the proposed methodology.

3. Methodology overview

A Web browser is software that enables a host to retrieve data from a Web server that resides on user's device. Web browsers come in several levels of functionality, from text-based applications to software that is able to interpret a variety of information, such as multimedia, images and executable code. The browser market is dominated by four major products; Internet Explorer, Chrome, Firefox and Safari. This software has in common the ability to record information by default about a user's actions when accessing information on the Web, primarily aimed at improving the user's experience. This activity occurs often without the knowledge of the user. They therefore provide a useful resource for a forensics examiner during an investigation as they provide a wealth of information about a suspect's online activity.

The way in which these four applications record, or log, user data differs between software. For example, Internet Explorer uses binary format log files to record a range of information, from cache data to Uniform Resource Locators (URLs) that a user accesses whilst on the Internet. This information is stored in a single file; *index.dat*. The other three applications use a SQLite database format to record information. This information may be both persistent and volatile. For example, Chrome stores historic information in persistent files. However, information on the current session is stored in text format that is replaced when the application is closed and then re-opened.

It should be noted that whilst Web browsers record a wide range of information about a suspect's activity by default, they could circumvent this by utilising the software in 'private browsing' mode. In an attempt to protect users' privacy, many of the main software providers have introduced this feature. The aim of this feature is to allow users to access the Internet without storing data related to the Internet session locally. Such features are useful when accessing the Internet from publicly-used computers, such as in libraries and Internet cafes. However, they may be used by a suspect to disrupt a forensic investigation by reducing the amount of evidence available to the examiner.

The amount of evidence that may be retrieved from private browsing sessions is dependent on the Web browser used by the suspect. For example, as Said et al (2011) and Satvat et al (2013) demonstrate, forensic examiners are able to retrieve data written locally to storage media with varying levels of success. For example, if a suspect were to use a Chrome browser, very little data may be retrieved; if they use Internet Explorer, a wide range of data will have been written to local storage and therefore may be retrieved. This is due to the different schemes used by each application to provide user privacy. Browsers such as Chrome do not write data locally whilst a user accesses the Web whereas Internet Explorer records data and then deletes the data associated with a private browsing session. As demonstrated by Said et al (2011), much of this deleted data may then be retrieved by the forensics examiner using tools and techniques associated with deleted file recovery. Whilst this may not retrieve a coherent list of Web site information that could be viewed in applications such as *Pasco v1.0*, it will result in strings that may be utilised by the methodology outlined in this paper.

The methodology posited by this paper aims to use search terms recorded in Web browser logs or elsewhere in local storage to provide a qualitative narrative of a suspect's activity and interests over time. This approach aims to overcome issues with tools such as *Pasco v1.0*. For example, if private browsing is used by the suspect, little evidence will be recorded that could be retrieved. However, as discussed above, strings containing search terms may be stored elsewhere. Search terms provide a qualitative view of a suspect's activity, motives or intentions. With the wide range of devices that a suspect has at their disposal, a search may be performed on one computer but the access of the Web server made from another. The analysis of search terms can be used to reduce the ambiguity of Web access and subsequent actions by a suspect. For example, a suspect may search for, "possible sentence for stabbing a person randomly on a bus", which may result in subsequent interactions with legal Web sites, perhaps not indicating the link between the search and the resulting behaviour. Finally, search terms may be used to provide a temporal and textual analysis to demonstrate the *intention* to engage in criminal or malicious activity.

As discussed above, Web logs come in a variety of formats and this affects both the location of information that is recorded and the techniques to retrieve the data. For example, Chrome utilises a number of files in SQL format that record a user's Web activity, such as ‘History’ and ‘Archived History’ whilst Internet Explorer uses the *index.dat* file to record a wide range of data and Web preferences (for a detailed discussion on *index.dat* file format and artefacts, see Jones (2003)). Information that may be retrieved ranges from URLs visited by the user to auto-complete information. The Chrome databases are located in `%systemdir%\Users\%username%\AppData\Local\Google\Chrome` (Windows 7) and `/home/$USER/.config/google-chrome/` (Linux) folders. The *index.dat* file is located in `%systemdir%\Users\%username%\AppData\Local\Microsoft\Windows\History\History.IE5\` (Windows 7) and it may also store historic information in daily, weekly, and monthly history logs.

As illustrated in figure 1, the methodology posited in this paper has three main areas of functionality: evidence retrieval, Web log mining and analysis, and data visualisation.

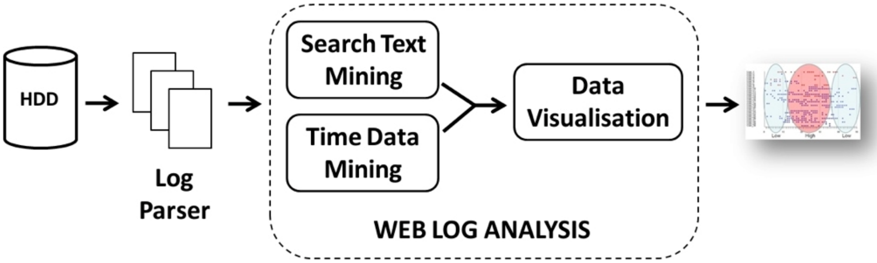

Figure 1: Overview of the Web log mining methodology.

As with any investigation, the data must be acquired in a robust manner, ensuring that the evidence maintains its integrity. Therefore, an image of the original hard disk drive (HDD) or storage media is made. The Web log files are accessed on this image and are pre-processed to provide a textual format for further analysis through text mining. The log parsers depend on the Web browser that is used by the suspect. For example, *SQLite3* can be used to view the Web logs of Chrome or Chromium browsers whereas the *Pasco v1.0* utility (Jones, n.d.) may be used to access Internet Explorer *index.dat* files. Information is written to local storage in binary format, thus the log parsers provides a means to read the data. The log parser returns information from the Web browser file format into text providing information such as Web sites visited and time of access. Amongst these results are records of search engine accesses and search terms used.

The pre-processed log files are passed into the program for textual and temporal analysis prior to visualisation. Textual analysis accesses URL descriptors associated with search strings to extract the search term used by the suspect. The Search Text mining function searches the URLs returned by the Log parser for strings associated with searches, such as “`?q=`” and “`#q=`” and strips the search terms from the URLs. The advantage of using this short search text rather than a search engine URL is that it can be used to identify searches made in a range of Web-based databases, such as streaming multimedia repositories and Web-based map sites. The data is also searched for dates and times of access. In this way, the forensics examiner is able to not only gain an overview of what the person searched for, but also the times that they made those searches. This can then be related back to the investigation to place the suspect's activity in context of the event(s) under investigation.

The initial URL prior to parsing is illustrated below. The search string inputted by the user is inserted into the URL within a search engine. The format of the recorded search is; <search engine URL> <query> <additional parameters>. As illustrated below, the search is indicated by the "`?q=`" string and the words used in the query are separated by "+". In the case below, the search term is repeated indicating an OR query. The additional parameters include information on the Web browser being used to access the site as well as input encoding. The text mining function extracts the search term used by the suspect by extracting the information following the "`?q=`" string. This therefore produces a string, "`abc+xyz`", from the example below.

```
https://www.google.co.uk/search?q=abc+xyz&oq=abc+xyz&aqs=
chrome..69i57j0l5.5985j0j8&sourceid=chrome&espv=210&es_sm
=122&ie=UTF-8
```

Data is passed to the visualisation function for temporal (timeline) and text visualisation. The temporal visualisation in the case study shows the forensics examiner the date of last access for the search term. It is recognised that users will repeat search terms to recall Web sites rather than bookmark their results but this information is not recorded in Chrome. Visualisation of the text is in two forms; whole search terms and individual words. The whole term visualisation places the search in context and shows the forensics examiner the intention of the suspect; it provides contextual narrative. This is visualised as a 'network of searches', similar to a mind map and relationships between searches made on the same day are highlighted by clicking the network node. Visualisation of individual words demonstrates common themes in the overall search narrative of a suspect by quantifying those words.

This section has provided an overview to the search term analysis approach that provides both textual and temporal visualisations of a suspect's online activities. In the next section, we demonstrate the applicability of the proposed approach by applying it to a Chrome history file.

4. Case study and results

The History file of a Chrome browser is accessed to show user activity over a period of ten days. Log parsing outlined above is achieved using *sqlite3*. As discussed in the previous section, Chrome records a wealth of information about a users' online activities and stores this data in a number of database tables, such as downloads, URLs, keyword search terms, and meta(-data). Data is exported to text format from the URL table within the database to be passed to the application discussed above for temporal and text mining. Information recorded in the URL table includes the URL itself, number of visits and time of last visit. Therefore, in the temporal analysis, only the last date that the user visited a Web site is recorded and this is extracted by the temporal mining function.

Over the ten-day period, a total of 149 URLs were accessed, of which 41 contained search terms used in Google searches. Figure 2 provides a temporal view of the searches extracted from the Chrome History database. As discussed above, Chrome only records the last access (and number of times the URL was visited) and therefore, only one entry per search term is available. However, certain patterns of activity are clearly discernible. For example, searches around the 'Heartbleed' bug are followed by searches related to OpenSSL itself. Searches related to wireless security are followed by searches for downloads and tutorials related to exploitation of such networks. The temporal analysis therefore suggests a search narrative associated with malicious activity related to the security of networks. The extent to which the suspect has engaged in such activity could be ascertained by further analysis of all URLs visited, for example using *SQLite3*, and searches of the hard drive for software associated with security exploitation.

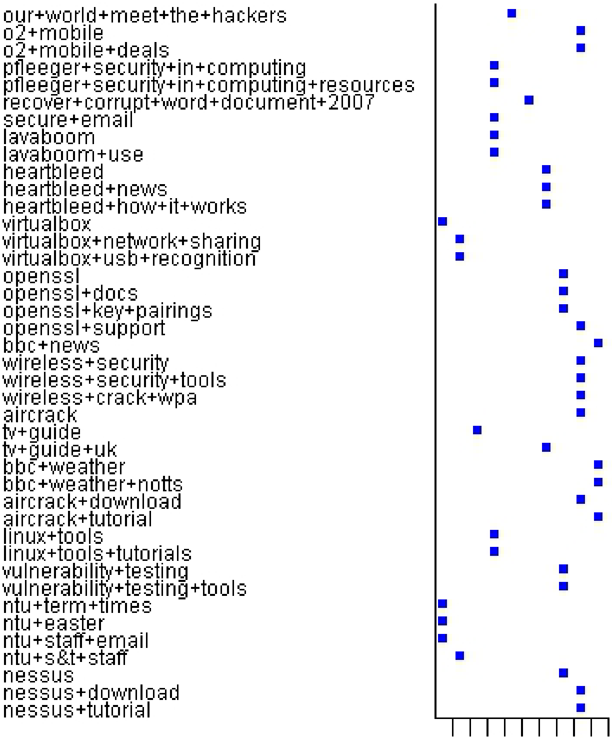

Figure 2: Temporal view of Web searches.

Figure 3 illustrates two network views which can be used by the forensics examiner to explore relationships between search terms used by a suspect. Search terms are arranged around the user to show their association. In addition, the examiner can select a node to show other search terms used on the same day. In this way, temporal information can be discerned and potentially follow the intentions of the suspect during a particular time period. In the figures below, the search term of interest is selected and the related searches are highlighted in red. In the figure on the left, the term “wireless+security+terms” is selected and on the same day, “o2+mobile”, “o2+mobile+deals”, “openssl+support”, “wireless+crack+wpa”, “aircrack+ download”, and “nessus+download” are terms also searched. In the figure on the right, “openssl+docs” is selected and other searches were made on the same day for “openssl”, “openssl+key+pairings”, “vulnerability+testing+tools”, and “nessus”. In this way, we are able to determine that the suspect has been actively searching for network security issues and tools that may be used to exploit systems, depending on the context of the investigation.

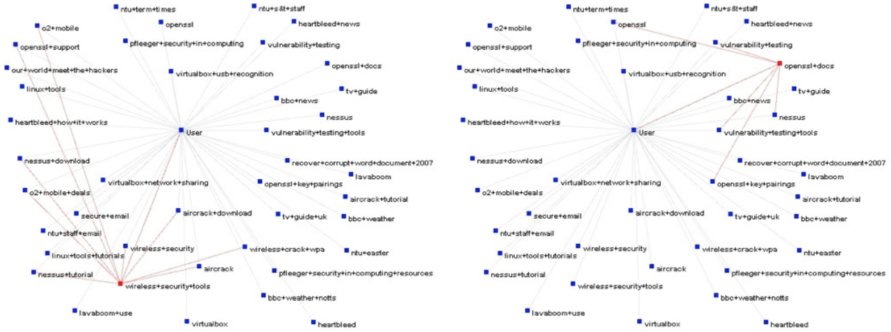

Figure 3: Network view of Web searches.

Figure 4 illustrates the key themes across the data set and the search narrative of the suspect. The visualisation in this figure is a tag cloud of all words whereby the font is sized by frequency of occurrence. Clearly discernible in this view is the prominence of “openssl” as a search term and its relationship with “heartbleed” could be explored using the visualisations above. In addition, the suspect also makes searches related to “security” and “tools”. If this were an investigation into suspected computer and network misuse, these terms would identify intention and suggest forensic analysis of the hard drive for tools associated with such activity.

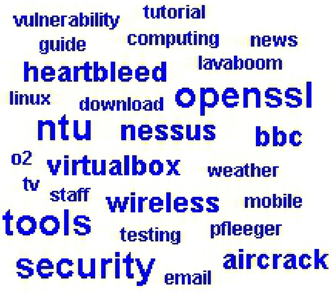

Figure 4: Overview of Web search words.

As demonstrated by the case study, search narratives provide a wealth of information to the forensics examiner. The methodology posited above may be used to triage potentially large data sets from a suspect’s daily interaction with the Internet to provide further sources of evidence or to support hypotheses during the investigatory process. Moreover, as Web searches are active rather than passive in that a user must enter the terms of the search, it can be used to identify intention during an event(s) of interest and potentially prove or disprove a Trojan defence.

5. Conclusions and further work

Our reliance on accessing information on the Web ensures that it provides a wealth of information to a forensics examiner during an investigation regarding a suspect’s access and use of the Internet. Current tools for Web browser log analysis may result in considerable evidence with many redundant entries and do not provide a qualitative analysis of a suspect’s activity over time, merely a list of Web sites visited in textual form.

Data visualisation may be used forensics examiners to alleviate the overheads of interpreting evidence and to make sense of patterns in system logs. Therefore, this paper has presented a methodology to visualise search strings in Web browser logs to present a narrative of a suspect’s interests, motives and actions over time. The aim of this approach is to triage potentially large data sets from a suspect’s daily interaction with the Internet to identify Web search narratives. As Web searches are an active, rather than passive, interaction they may be used to demonstrate intent in criminal or malicious activity. In order to demonstrate the applicability of the approach, data from a Chrome history file is parsed for Web search terms and the results visualised. The case study demonstrates that visualisation of the temporal and relational information of search narratives can aid a digital investigation. Further work aims to extend the proposed approach to larger data sets, explore further visualisation techniques, and identify other Web interactions such as database queries.

6. References

Accorsi, R., Wonnemann, C. and Stocker, T. (2011), "Towards Forensic Data Flow Analysis of Business Process Logs", *Proceedings of the Sixth International Conference on IT Security Incident Management and IT Forensics*, Stuttgart, Germany, 2011, pp. 3 - 20.

Al Mutawa, N., Al Awadhi, I., Baggili, I. and Marrington, A. (2011), "Forensic artifacts of Facebook's instant messaging service", *Proceedings of the International Conference on Internet Technology and Secured Transactions*, Abu Dhabi, United Arab Emirates, 2011, pp. 771 - 776.

Chowdhury Farhan Ahmed, Syed Khairuzzaman Tanbeer and Byeong-Soo Jeong (2010), "Mining High Utility Web Access Sequences in Dynamic Web Log Data", *Proceedings of the 11th ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing*, London, UK, 2010, pp. 76 - 81.

Giacobe, N.A. and Sen Xu (2011), "Geovisual analytics for cyber security: Adopting the GeoViz Toolkit", *Proceedings of the IEEE Conference on Visual Analytics Science and Technology*, Providence, RI, USA, 2011, pp. 315 - 316.

Hadzic, F. and Hecker, M. (2011), "Alternative Approach to Tree-Structured Web Log Representation and Mining", *Proceedings of the IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology*, Lyon, France, 2011, pp. 235 -242.

Haggerty, J., Haggerty, S. and Taylor, M. (2014), "Forensic Triage of Email Network Narratives through Visualisation", *Journal of Information Management and Computer Security*, forthcoming.

Hernandez, P., Garrigos, I. and Mazon, J-N (2010), "Modeling Web logs to enhance the analysis of Web usage data", *Proceedings of Workshops on Database and Expert Systems Applications*, Toulouse, France, 2011, pp. 297 - 301.

Koniaris, I., Papadimitriou, G. and Nicopolitidis, P. (2013), "Analysis and Visualization of SSH Attacks Using Honeypots", *Proceedings of EuroCon*, Zagreb, Croatia, 2013, pp. 65 - 72.

Jones, K.J. (2003), "Forensic Analysis of Internet Explorer Activity Files", Technical Report, available from http://www.foundstone.com/us/pdf/wp_index_dat.pdf, accessed 26 Mar 2014.

Jones, K.J. (n.d.), "Pasco v1.0", available from http://www.mcafee.com/uk/downloads/free-tools/pasco.aspx, accessed 26 March 2014.

Nithya, P. and Sumathi, P. (2012), "Novel Pre-Processing Technique for Web Log Mining by Removing Global Noise and Web Robots", *Proceedings of the National Conference on Computing and Communication Systems*, West Bengal, India, 2012, pp. 1 - 5.

Office of National Statistics (ONS) (2013), "Internet Access - Households and Individuals, 2013", Government Report, http://www.ons.gov.uk/ons/rel/rdit2/internet-access---households-and-individuals/2013/stb-ia-2013.html, accessed 26 March 2014.

Ohana, D.J. and Shashidhar, N. (2013), "Do Private and Portable Web Browsers Leave

Incriminating Evidence?", *Proceedings of IEEE Security and Privacy Workshops*, San Fransisco, CA, USA, 2013, pp. 135 - 142.

Said, H., Al Mutawa, N., Al Awadhi, I. and Guimaraes, M. (2011), "Forensic Analysis of Private Browsing Artifacts", *Proceedings of the International Conference on Innovations in Information Technology*, Abu Dhabi, United Arab Emirates, 2011, pp. 197 - 202.

Satvat, K., Forshaw, M., Hao, F. and Toreini, E. (2013), "On the Privacy of Private Browsing – A Forensic Approach", *Proceedings of the International Workshop on Data Privacy Management*, London, UK, 2013.

Schrenk, G. and Poisel, R. (2011), "A Discussion of Visualization Techniques for the Analysis of Digital Evidence", *Proceedings of the Sixth International Conference on Availability, Reliability and Security*, Vienna, Austria, 2013, pp. 758 - 763.

Thomson, A., Graham, M. and Kennedy, J. (2013), "Pianola - Visualization of Multiva-riate Time-Series Security Event Data", *Proceedings of the 17th International Conference on Information Visualisation*, London, UK, 2013, pp. 123 -131.