Chapter 20

ANALYSIS OF WEB PROXY LOGS

B. Fei, J. Eloff, M. Olivier and H. Venter

Abstract Network forensics involves capturing, recording and analysing network audit trails. A crucial part of network forensics is to gather evidence at the server level, proxy level and from other sources. A web proxy relays URL requests from clients to a server. Analysing web proxy logs can give unobtrusive insights to the browsing behavior of computer users and provide an overview of the Internet usage in an organization. More importantly, in terms of network forensics, it can aid in detecting anomalous browsing behavior. This paper demonstrates the use of a self-organising map (SOM), a powerful data mining technique, in network forensics. In particular, it focuses on how a SOM can be used to analyse data gathered at the web proxy level.

Keywords: Network forensics, web proxy logs, self-organising map, data analysis, anomalous behavior

1. Introduction

The Internet provides offenders with communication capabilities that did not exist previously. Internet-related crimes are on the rise and Internet abuse by employees is becoming routine.

Network forensics involves capturing, recording and analysing network audit trails to discover the source of security breaches and possible malicious activity [19]. It is becoming increasingly practical to archive network traffic and analyse the data as necessary [6].

A fundamental goal in network forensics is to gather evidence. Evidence can be gathered from various sources depending on the unique nature of the investigation. It can be collected at the server level, proxy level or from several other sources. For example, at the server level, evidence can be obtained from web server logs that record the browsing behavior of site visitors. Furthermore, evidence can be also gathered

from usage data provided by packet sniffers that monitor network traffic coming to a web server.

This paper deals with network forensics—more specifically, the analysis of web proxy data. Analysing data on a web proxy—as opposed to data on a single computer or on multiple computer systems—is significant. Users can delete traces of their Internet behavior from their computer systems. On the other hand, web proxy data pertaining to URL requests made by users is generally accessible only to network administrators and forensic investigators.

Another benefit is that investigators can focus on a single point in the network topology, which saves time that might be crucial to the investigation. For example, they can focus on employees in an organisation who access web sites that promote illegal activities. Since it is not necessary to seize employees' computer systems for evidence recovery, the investigation can be performed without the employees knowing that they are being investigated.

This paper demonstrates how a self-organising map (SOM) [11, 12], a powerful data mining technique, can be used in network forensics to analyse web proxy data. A SOM can reveal interesting patterns, and also serve as a basis for further analysis of the data gathered at the web proxy. More importantly, it can aid in detecting anomalous browsing behavior and in recovering digital evidence.

Sections 2 and 3 provide background information and an overview of SOMs. Section 4 demonstrates the use of a SOM to analyse web proxy data. The final section, Section 5, summarises the conclusions.

2. Background

While computer forensics goes as far back as 1984 [20], network forensics was introduced only in the early 1990s [19]. In general, computer forensics deals with data in a single computer system [22, 26]. Network forensics deals with data that may reside on computers, routers, firewalls, web servers, web proxies and other devices in one or more networks [4].

A crucial part of network forensics is to gather evidence. For example, when an attacker attacks a network, the attack traffic usually goes through a router. As a result, important evidence may be found by examining network logs.

Web proxy logs provide valuable evidence. The purpose of a web proxy is to relay URL requests from clients to a server, receive the responses from the server and send them back to the appropriate clients [17]. The web proxy acts as a gateway between the Internet and browsers on a local network.

Web mining is the extraction of interesting and useful knowledge, as well as implicit information from activities related to the World Wide Web [1]. It is categorised into three main areas: web content mining, web structure mining and web usage mining [16].

Similar to web mining [8, 14–16, 18], analysing web proxy logs can assist forensic investigators (or network administrators) in understanding the browsing behavior of computer users and in providing them with an overview of Internet usage in an organisation. Furthermore, it can assist them in gathering evidence left behind by suspects. Such evidence may involve excessive Internet usage for non-work purposes, or access to web sites promoting pornography and other illegal activities.

Considerable work has been done in the area of web usage mining [1, 2, 5, 10, 24]. In general, web usage mining involves three phases: data pre-processing, pattern discovery and pattern analysis. Web usage mining seeks to reveal knowledge hidden in web server log files, including statistical information about site visitors, and the preferences, characteristics and navigational behavior of computer users.

The self-organising map (SOM) approach used in this paper is fairly similar to web usage mining. SOMs have been used by researchers in a wide variety of fields [3, 7, 21, 25], but they have rarely been used in digital forensics.

SOMs have been used to cluster web pages according to user navigation behavior [23] and to organise web documents based on their content [13]. Our work uses a SOM to cluster, visualise and analyse web usage data gathered at a web proxy. Multi-dimensional data in web proxy logs is transformed by a SOM to two-dimensional data, which can be visualised and analysed more efficiently by forensic investigators.

3. Self-Organising Maps

The self-organising map (SOM) [11, 12] is a neural network model that has been widely used to cluster and visualise high-dimensional data. Clustering attempts to group data with similar characteristics [27]. Visualisation is the process of mapping complex data to a graphical representation to provide qualitative notions of its properties.

A SOM is used to map high-dimensional data onto a low-dimensional (typically two-dimensional) space. It consists of two layers of units (neurons), the input layer and the output layer. Each unit in the input layer, which represents an input signal, is fully connected with units in the output layer. The output layer forms a two-dimensional grid of units, where each unit represents a unit of the final structure.

A SOM employs unsupervised competitive learning, in other words, the learning process is entirely data driven and units in the output layer compete with one another. The learning process involves two steps: identifying the winning unit and updating unit weights. When an input pattern is presented to the input layer, the winning unit in the output layer is the one whose weights are closest to the input pattern in terms of the Euclidian distance [9]. After the winning unit is determined, the weights of that unit and its neighboring units are adjusted. The learning process continues until the SOM produces acceptable results or a pre-set limit is reached on the number of iterations.

The effect of the learning process is to cluster similar patterns. An additional step (determination of cluster boundaries) is required to visualise clusters. This is done by calculating the unified distance matrix [9]. The size of a cluster is the number of units allocated to the cluster.

One of the advantages of a SOM is its ability to manifest possible correlations between different dimensions of input data in component maps [28]. Each component map displays the spread of values in a particular dimension. Correlations are revealed by comparing component maps.

A SOM is also ideal for association and classification. Association seeks to identify correlations in data. Classification maps a data item into one of several predetermined classes. Network forensic investigations typically involve the analysis of enormous amounts of data. The ability of a SOM to support the clustering, correlation, association, classification and visualisation of multi-dimensional data make it an attractive tool for network forensics.

4. Analysing Web Proxy Data

We used a SOM to analyse web proxy logs for twenty computer users in an organisation. The proxy logs, which were generated by a Squid Proxy [29] over a period of one month, contained data pertaining to 374,620 HTTP requests. A Squid Proxy log has the following format:

```
time:etime:client:log-tag:size:request:url:userid:hierarchy
```

Our approach to analysing web proxy data in support of network forensics is fairly similar to web usage mining [1, 2, 5, 10, 24]. The following subsections discuss the three main phases: data pre-processing, pattern discovery and pattern analysis.

4.1 Data Pre-Processing

Data pre-processing is concerned with data cleaning and data transformation. The goal of data cleaning is to remove irrelevant information.

Data transformation, on the other hand, converts raw data items into structured information.

Eliminating irrelevant data simplifies SOM learning and reduces processing time. Only certain data fields are required to analyse the browsing behavior of computer users and Internet usage in the organisation. For example, the `client` and `userid` fields in the web proxy logs refer to the same person; therefore, only one field is required and the other is classified as irrelevant data. The following fields were deleted after the data cleaning process: `etime`, `client`, `log-tag` and `request object`.

Next, certain data transformations were performed on the web proxy logs. The `timestamp` of each request was converted into a more readable format, namely, `1121182139` was transformed to `2005/07/12 15:28 Thursday`. Furthermore, strings (e.g., `day`, `request`, `url`, `userid` and `content type`) in the proxy logs were converted to numerical values to speed up processing.

4.2 Pattern Discovery

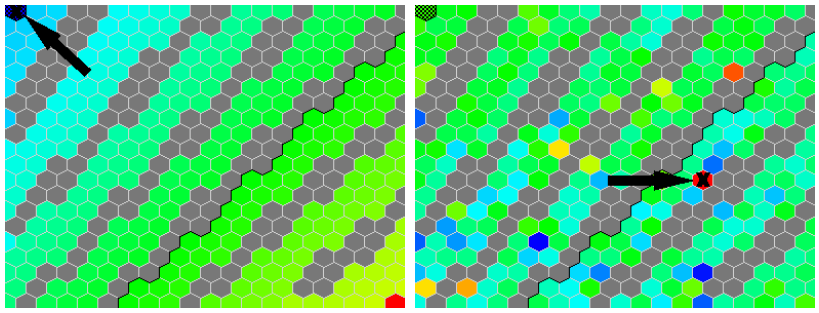
Pattern discovery draws on algorithms used in data mining, machine learning and pattern recognition to detect interesting patterns. These patterns can be further analysed during the pattern analysis phase to gain better insights into the data and to aid in evidence recovery. The SOM data mining technique was used for pattern discovery in our study.

Pattern discovery using a SOM occurs after the data pre-processing phase. Note that the SOM's learning process terminates when the SOM produces acceptable results or a pre-set limit is reached on the number of iterations.

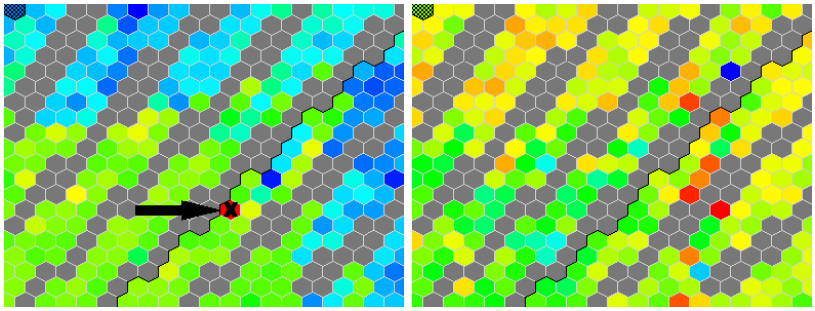
4.3 Pattern Analysis

Two-dimensional maps, which are displayed as hexagonal grids in Figure 1, are generated after the SOM completes its learning process. These "component maps" reveal variations in the values of components across the map. Each component map visualises the spread of values in a particular dimension. In the images in Figure 1, the color blue indicates low values, red indicates high values and the other colors (e.g., green and yellow) represent intermediate values. The color grey indicates that no data is mapped to that particular unit (or hexagonal grid). Color versions of the images are available at `mo.co.za/bib.htm`.

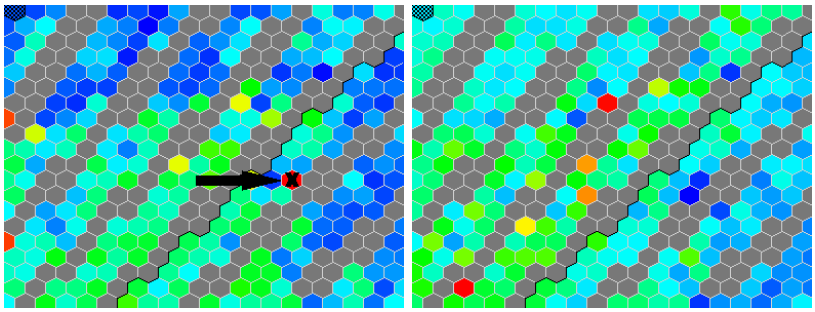
The two-dimensional component maps are a powerful visualisation aid. An application has been implemented to support the analysis of component maps by forensic investigators. Each unit in the map contains information about HTTP requests. A unit is selected by clicking on it,



(a) Component map (time); (b) Component map (day).



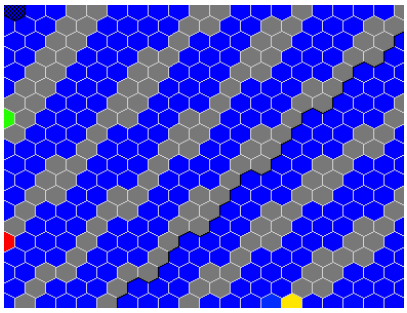
(c) Component map (request); (d) Component map (content type).



(e) Component map (URL); (f) Component map (userid).

Figure 1. Component maps generated from web proxy data. (Color images are available at `mo.co.za/bib.htm`).

which causes it to be highlighted with a small checker board (see top left corner of Figure 1(b)). Information about the unit is then presented in a dialog box at the bottom of the screen (see Figure 2).



(g) Component map (size).

Figure 1 (continued). Component maps generated from web proxy data. (Color images are available at `mo.co.za/bib.htm`).

4.4 Analysis of Component Maps

Figure 1(a) reveals variations in the time periods that computer users made HTTP requests. The map has three portions: blue (top left), green and red (bottom right). The blue portion denotes the 12 am to 9 am time period; green denotes 9 am to 3 pm, and red, 3 pm to 12 am. Upon viewing the map, it is immediately obvious that the green portion is significantly larger than the others, implying that Internet usage mostly occurred from 9 am to 3 pm.

Figure 1(b) presents variations in the specific days that HTTP requests were made. HTTP requests occurred mainly in the middle of the week: Tuesday, Wednesday and Thursday (green), as opposed to Friday, Saturday and Sunday (blue and red).

Figure 1(c) reveals variations in HTTP requests. POST (green) and GET (blue) were common requests. Usually, the GET method is by far the most common request for a specific URL; Figure 1(c) clearly shows that is, in fact, the case.

Figure 1(d) presents variations in the type of data returned by HTTP requests. The map shows that the content is predominantly images, applications and text (green and yellow), with text being the most common.

Figure 1(e) reveals variations in the URLs of the requests. The requests involved 4,263 distinct domain names, which were replaced with numerical values. Most of the requests involved domains that were mapped to lower numerical values (blue and green). The most popular domain was `www.google.co.za` (represented by 1), which was accessed 43,952 times (approximately 10% of requests).

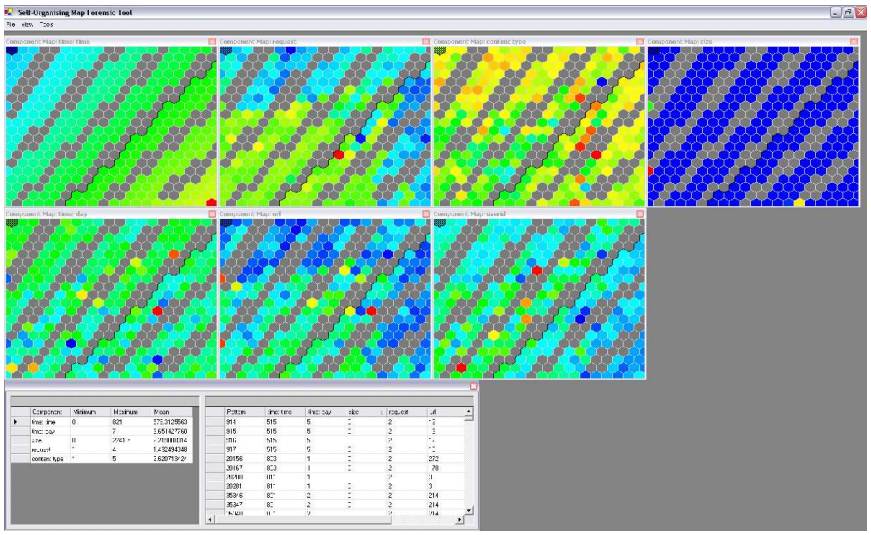


Figure 2. Screenshot of network forensics application.

Figure 1(f) presents variations with regard to users who made the HTTP requests. In particular, the map provides information about user browsing behavior and Internet usage.

Figure 1(g) reveals variations in the size of data returned by HTTP requests. At least 90% of the map is blue, corresponding to HTTP requests that returned data less than 1,000 KB in size. However, three units have a color other than blue; the corresponding HTTP requests were for data ranging from 1,000 KB to 7,000 KB.

Figure 2 presents a screenshot of an implemented application that presents all seven component maps to a forensic investigator. By comparing the maps, the investigator can review Internet usage and analyse the browsing behavior of users. The browsing behavior includes HTTP request times and URLs, the type of data requested from specific URLs, and the size of data requested from specific URLs,

Detecting anomalous browsing behavior is important in many network forensic investigations. Ideally, this is accomplished by examining the irregular portions of component maps—regions where a specific color has fewer occurrences. For example, in Figure 1(a), the irregular portion in the blue zone is marked with a cross (indicated with an arrow). It indicates HTTP requests made during an unusual time period (between 12 am and 6 am). However, an investigation of the URLs, and the type of data requested revealed no suspicious activity.

In Figure 1(b), the irregular portion is the red zone marked with a cross (indicated with an arrow), which corresponds to HTTP requests that were made on Saturday. The use of this map in conjunction with the other maps reveals possible correlations between the various dimensions. It appears a correlation exists between the red portions in Figures 1(b) and 1(e) (indicated with arrows). Upon investigating the URLs and the type of data requested, it was found that suspicious activities had in fact occurred. First, Figure 1(d) indicates that the majority of the requests were for images. Second, Figure 1(e) shows very few red regions, indicating that URLs represented by red were visited rarely. On examining the original proxy logs, it was observed that a particular user visited several adult web sites on a Saturday and the contents retrieved by the user were mainly images.

In Figure 1(c), the irregular portion of the map is again the red zone marked with a cross (indicated with an arrow), which corresponds to the CONNECT method. The CONNECT method is normally used to tunnel a connection through an HTTP proxy. By investigating the URLs and the type of data requested, it was found that a user was conducting Internet banking, which was not deemed to be an unauthorised activity.

Although certain activities deviate significantly from others, they may not necessarily be unauthorised or illegal. For example, in the example above, one anomalous incident involved Internet banking while the other involved visits to adult web sites. Therefore, when anomalous activity is suspected, it is necessary to conduct a detailed examination of the original web proxy logs.

5. Conclusions

Self-organising maps (SOMs) can be used in network forensics to analyse web proxy data with a view to investigating browsing patterns of computer users and detecting anomalous behavior. The component maps resulting from two-dimensional mappings of data produced by a SOM offer a powerful framework for visualising and analysing the large volumes in data contained in web proxy logs. By comparing different component maps, a forensic investigator can rapidly obtain an overview of Internet usage and an understanding of the browsing patterns of computer users, including anomalous behavior. Only when anomalous behavior is indicated, is it necessary for the investigator to conduct a detailed analysis of the web proxy logs. This can contribute to an increase in the quality of forensic investigations and a reduction in the amount of effort, especially in network forensics, which involves the collection and analysis of large quantities of data.

References

[1] A. Abraham and V. Ramos, Web usage mining using artificial ant colony clustering and linear genetic programming, *Proceedings of the IEEE Congress on Evolutionary Computation*, vol. 2, pp. 1384-1391, 2003.

[2] B. Berendt, Web usage mining, site semantics and the support of navigation, *Proceedings of the Workshop on Web Mining for E-Commerce: Challenges and Opportunities*, 2000.

[3] J. Brittle and C. Boldyreff, Self-organizing maps applied in visualising large software collections, *Proceedings of the Second International Workshop on Visualising Software for Understanding and Analysis*, 2003.

[4] M. Caloyannides, *Privacy Protection and Computer Forensics*, Artech House, Boston, Massachusetts, 2004.

[5] R. Cooley, B. Mobasher and J. Srivastava, Data preparation for mining World Wide Web browsing patterns, *Knowledge and Information Systems*, vol. 1(1), pp. 5-32, 1999.

[6] V. Corey, C. Peterman, S. Shearin, M. Greenberg and J. van Bokkelen, Network forensics analysis, *IEEE Internet Computing*, vol. 6(6), pp. 60-66, 2002.

[7] G. Deboeck, Financial applications of self-organising maps, *Neural Network World*, vol. 8(2), pp. 213-241, 1998.

[8] M. Eirinaki and M. Vazirgiannis, Web mining for web personalization, *ACM Transactions on Internet Technology*, vol. 3(1), pp. 1-27, 2003.

[9] A. Engelbrecht, *Computational Intelligence: An Introduction*, Wiley, Chichester, United Kingdom, 2002.

[10] M. Géry and H. Haddad, Evaluation of web usage mining approaches for users' next request prediction, *Proceedings of the Fifth ACM International Workshop on Web Information and Data Management*, pp. 74-81, 2003.

[11] T. Kohonen, The self-organizing map, *Proceedings of the IEEE*, vol. 78(9), pp. 1464-1480, 1990.

[12] T. Kohonen, *Self-Organizing Maps*, Springer, Berlin-Heidelberg, Germany, 2001.

[13] T. Kohonen, S. Kaski, K. Lagus, J. Salojarvi, J. Honkela, V. Paatero and A. Saarela, Self organization of a massive document collection, *IEEE Transactions on Neural Networks*, vol. 11(3), pp. 574-585, 2000.

[14] P. Kolari and A. Joshi, Web mining: Research and practice, *IEEE Computing in Science and Engineering*, vol. 6(4), pp. 49-53, 2004.

[15] R. Kosala and H. Blockeel, Web mining research: A survey, *SIGKDD Explorations*, vol. 2(1), pp. 1-15, 2000.

[16] Y. Li, X. Chen and B. Yang, Research on web-mining-based intelligent search engines, *Proceedings of the International Conference on Machine Learning and Cybernetics*, 2002.

[17] C. Maltzahn and K. Richardson, Performance issues of enterprise-level web proxies, *Proceedings of the ACM Sigmetrics International Conference on Measurement and Modeling of Computer Systems*, pp. 13-23, 1997.

[18] B. Mobasher, N. Jain, E. Han and J. Srivastava, Web Mining: Pattern Discovery from World Wide Web Transactions, Technical Report TR96-050, Department of Computer Science and Engineering, University of Minnesota, Minneapolis, Minnesota, 1996.

[19] S. Mukkamala and A. Sung, Identifying significant features for network forensic analysis using artificial techniques, *International Journal of Digital Evidence*, vol. 1(4), 2003.

[20] M. Noblett, M. Pollitt and L. Presley, Recovering and examining computer forensic evidence, *Forensic Science Communications*, vol. 2(4), 2000.

[21] U. Payer, P. Teufl and M. Lamberger, Traffic classification using self-organizing maps, *Proceedings of the Fifth International Networking Conference*, pp. 11-18, 2005.

[22] M. Reith, C. Carr and G. Gunsch, An examination of digital forensic models, *International Journal of Digital Evidence*, vol. 1(3), 2002.

[23] K. Smith and A. Ng, Web page clustering using a self-organizing map of user navigation patterns, *Decision Support Systems*, vol. 35(2), pp. 245-256, 2003.

[24] J. Srivastava, R. Cooley, M. Deshpande and P. Tan, Web usage mining: Discovery and applications of usage patterns from web data, *SIGKDD Explorations*, vol. 1(2), pp. 12-23, 2000.

[25] S. Tangsripairoj and M. Samadzadeh, Application of self-organizing maps to software repositories in reuse-based software development, *Proceedings of the International Conference on Software Engineering Research and Practice*, vol. 2, pp. 741-747, 2004.

[26] J. Vacca, *Computer Forensics: Computer Crime Scene Investigation*, Charles River Media, Hingham, Massachusetts, 2002.

[27] J. Vesanto, Using SOM in Data Mining, Licentiate Thesis, Helsinki University of Technology, Helsinki, Finland, 2000.

[28] J. Vesanto, Data Exploration Process Based on the Self-Organizing Map, Doctoral Thesis, Helsinki University of Technology, Helsinki, Finland, 2002.

[29] D. Wessels, Squid Web Proxy Cache (www.squid-cache.org).