FORWEB: File Fingerprinting for Automated Network Forensics Investigations

John Haggerty
School of Computing and Mathematical Sciences,
Liverpool John Moores University,
Byrom Street, Liverpool, L3 3AF, UK
+44 151 231 2279
J.Haggerty@ljmu.ac.uk

David Llewellyn-Jones
School of Computing and Mathematical Sciences,
Liverpool John Moores University,
Byrom Street, Liverpool, L3 3AF, UK
+44 151 231 2082
D.Llewellyn-Jones@ljmu.ac.uk

Mark Taylor
School of Computing and Mathematical Sciences,
Liverpool John Moores University,
Byrom Street, Liverpool, L3 3AF, UK
+44 151 231 2215
M.J.Taylor@ljmu.ac.uk

ABSTRACT
A major advantage of information technology is the ease, speed and volume of information that may be shared between hosts. However, this has given rise to concerns over paedophile activity and the spread of malicious digital pictures amongst this community. In network forensic investigations a wealth of information relevant to the investigation will reside within the network itself and on disparate hosts. Current computer forensics tools are designed for the analysis of seized hard drives rather than investigating data within a network. In this paper we present FORWEB, a novel scheme for automated file fingerprinting of malicious pictures resident on Web servers. This approach may be used in forensic investigations to automatically identify repositories of malicious digital pictures on the Internet or to verify the Internet usage of a suspect. A case study and its results demonstrate the applicability of this approach.

Keywords
Computer forensics, file fingerprinting, network investigations.

1. INTRODUCTION
Developments in information technologies provide many exciting opportunities for business and commerce, as well as technical and social challenges. A major advantage of current technologies is the ease, speed and volume of information that may be shared between hosts. However, the subversion of these technologies provides malicious people with many opportunities. For example, within law enforcement this has given rise to concerns over paedophile activity and the spread of malicious digital pictures, in particular indecent images of children.

A major challenge faced by today's forensic examiners is that of network investigations. Whilst much data will reside on the initial suspect's hard drive, a wealth of information will remain within the network itself and on disparate hosts across networks. In many cases, whilst an investigation may begin with just one suspect, it is rare that it will end with only the same suspect [1]. Challenges to network-based investigations include; identifying sources of evidence, hosts residing across geopolitical borders, lack of judicial control, gaining access to evidence and resolving a user's activities to their online activity.

e-Forensics 2008, January 21-23, 2008, Adelaide, Australia.

Current computer forensic tools predominantly focus on analysis of a suspect's hard drive. These tools are not ideally suited for network-based investigations. Network forensics focuses on gathering evidence from the network infrastructure. Current tools enable investigators to undertake activities such as access the wire, collect and store traffic, analyse content and view session data. Many of these tools are designed to provide evidence of a network-based intrusion rather than proactively investigating network hosts. The authors are not aware of any forensics tools that exist for automated searches of the Internet to provide evidence relevant to an investigation of online repositories of indecent images of children or for remotely analysing a host residing on the network.

This paper presents a novel *file fingerprinting* scheme based on signature analysis [2] combined with web robot techniques for the search for evidence of illegal or suspicious digital pictures residing on remote Web servers. Whilst signature analysis and web robots are not new individually, the combination of the two is novel. Once a fingerprint has been discovered further information could then be gained from a manual search of the Web server using the logical file structure of the machine, such as time of file creation, access, modification, etc.

The advantages of the FORWEB (Forensic Web) file fingerprint approach are threefold. First, the speed of analysis within an investigation may be significantly improved by automating the search process. Due to the amount of data and hosts that must be searched within the network environment, manual searches are a time-consuming process. This is not to say that manual inspection may not be required later, but the file fingerprint search can direct the forensic analyst to the relevant areas of the Internet and server based on the reports that the application generates. Second, the FORWEB tool can be used to collect malicious images to be used in further investigations analyzing a suspect's hard drive, if legally able to do so. Third, the tool provides additional evidence beyond the logs residing on the suspect's machine of their Internet usage through

comparison of files on the seized hard drive to those on the Web server.

This paper is organised as follows. In section 2, we discuss related work. In section 3, the FORWEB approach is posited. Section 4 presents a case study and its results. Finally, we make our conclusions and discuss further work.

2. RELATED WORK

Current practice relies on tools such as *Forensic Toolkit* (*FTK*) [3] and *EnCase* [4] to enable an analyst to investigate data that has been seized from a suspect's computer. These tools are used for storage media analysis of a variety of files and data types in fully integrated environments. For example, *FTK* can perform tasks such as file extraction, make a forensic image of data on storage media, recover deleted files, determine data types and text extraction. *EnCase* is widely used within law enforcement and like *FTK* provides a powerful interface to the hard drive or data source under inspection, for example by providing a file manager that shows extant and deleted files. Whilst these applications provide a robust forensic analysis, they are often time consuming in building a case due to the analyst having to manually read the data, e.g. looking at file contents, recovering deleted files, etc., to determine the relevance of the files to the investigation. These tools support network-based investigations by the analysis of network logs residing on the hard drive rather than the means by which a network may be investigated.

Recent research has recognised the disadvantages of current practice and has therefore proposed alternative approaches. These approaches attempt to not only identify file types, but also known files of a particular type by utilising statistical data derived from file analysis. For example Li *et al.* [5] posit a method based on intrusion detection to identify files of interest. This method models mean and standard deviation information of individual bytes to determine a *fileprint*, or identification of a specific file. This method is dependent on file header data for file categorisation, and therefore requires that the files are not fragmented and for the file system to be intact [6]. As such, Karresand and Shahmehri [7] propose the Oscar method, which determines probable file types from data fragments. This approach, unlike the previous one, aims to identify files based on fragmented data, such as that in RAM, and therefore does not require header information or an extant file system. A disadvantage of this approach is that it uses a more computationally exhaustive statistical measure than Li *et al.* [5] with not much advantage in detection rate, in order to achieve the identification of data fragments.

Previous research into forensic file fingerprinting has focused on the protection of intellectual property and digital rights management (DRM) rather than the creation of fingerprints from malicious files. For example, Schonberg and Kirovski [8] propose a method for multimedia file fingerprinting. This approach focuses on DRM and the creation of fingerprints that may be embedded into a file to ensure the protection of intellectual property and to combat the problem of software piracy. A similar approach proposed in Wong *et al.* [9] aims to protect intellectual property rights. The problem with these approaches is that they do not focus on file fingerprinting for the detection of malicious files, but merely to assert ownership. Therefore, the intellectual property owners must insert data into their files which forms the fingerprint. When searching for malicious files, and in particular indecent images of children, this data is not likely to be embedded in the search space.

The FORWEB fingerprint approach presented in this paper has its basis in signature analysis techniques used in intrusion detection. Signature analysis is widely used in network intrusion detection due to its applicability to searching large data sets within very tight temporal constraints. This method defines a set of rules, or *signatures*, based on the organisational security policy, and applies them to the traffic being analysed. Detection is achieved by comparing an event of interest to a list of known signatures, rather than having to hold information between events [10]. In addition, signature verification is simple and may improve false positive reports. This is achieved by using a single event, for example, a packet or datagram, with which to compare against previously recorded events to determine whether the information represents normal or malicious traffic. Recent work uses techniques such as combining statistical and misuse signatures or machine-learning techniques [11, 12]. Rules (or signatures), once developed, can then be quickly disseminated amongst the application-user community to provide a contemporary defence against these attacks.

3. FORENSIC FILE FINGERPRINTING

This section provides an overview of the FORWEB approach. Fingerprints are created from known malicious digital pictures, after which the FORWEB spider then searches Web servers automatically for evidence of any files of interest. The FORWEB application has evolved from the FORSIGS [2] project for high-speed, large volume analysis of digital pictures residing on the hard drive.

3.1 Digital Picture Fingerprint Approach

Figure 1 provides an overview of the FORSIGS application as used for searches of hard drives for malicious digital pictures. This approach forms the basis of the FORWEB scheme.

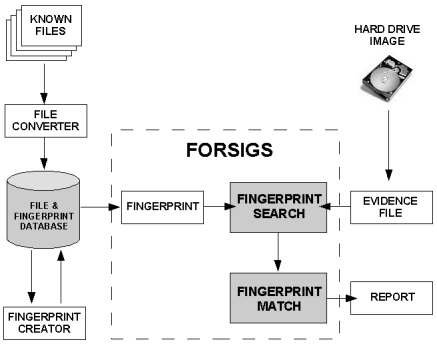

Figure 1. Overview of the FORSIGS application.

Known files collected during an investigation are converted to their hexadecimal form and placed within a file and fingerprint database. Once a file is in the database, a block of that file is extracted through the fingerprint creator. The block is passed to the FORSIGS tool where a number of points form the file fingerprint are used for comparison. The evidence file is acquired from the image of the original hard drive under investigation. The evidence file is searched and if a fingerprint is found, a report of the file(s) location(s), along with any additional information such as embedded EXIF data relating to the file(s), is generated.

FORWEB, as illustrated in figure 2, adapts this scheme to search for and fingerprint files resident on remote Web servers across the Internet. There are a number of applications for the scheme including; identification of repositories of malicious digital pictures on the Internet, verify the Internet usage of a suspect or collect more images for use in hard drive searches if it is legally acceptable to do so.

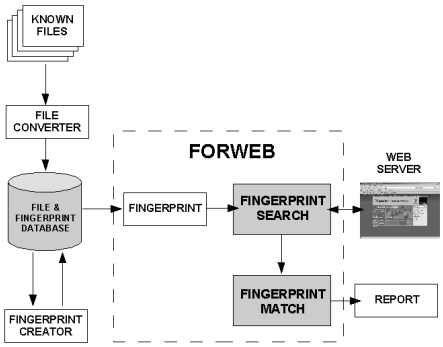

Figure 2. Overview of the FORWEB application.

The fingerprint application reads in the fingerprint block(s) and conducts a search of a Web site, server or multiple sites for known files of interest or other malicious data using standard web-spider techniques. During this process a report is dynamically generated for the analyst detailing the malicious files found along with other pertinent data. Even when performing such an automated scan, the size and quantity of sites means that the process can take some time. By generating the report dynamically at run-time an operator can be made aware of any detected data as soon as it is discovered.

3.2 Digital Fingerprints

Certain file types are more suited to file fingerprinting than others due to their byte value distribution. Figure 3 demonstrates the distribution of byte values for a JPEG file as analysed through the X-Ways [13] computer forensics tool. Due to the compression that this file type requires, there is an even spread of byte values used to create the picture. The ITU T.81 specification describes an efficient coding mechanism for JPEG images based on information-theoretic compression efficiency implying a uniform bit, and therefore byte, distribution [14]. This provides a more robust search for JPEG images as the approach utilises these byte values to create a robust fingerprint.

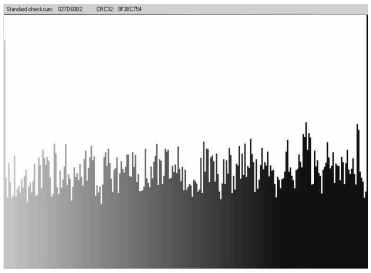

Figure 3. Byte value distribution of a JPEG file.

To create a fingerprint, a known malicious file is passed by the database to the fingerprint creator where a block within the original file is chosen. This may be from anywhere within a file. With many digital pictures being upwards of hundreds of kilobytes in size, the file itself will use many blocks. The first and last blocks of a file are not provided for the fingerprint search. The first block may hold generic but redundant data, such as headers or colour tables of font information, and therefore is less robust for analysis. The last block will include slack space data, and therefore cannot provide a reliable fingerprint. Using either of these two blocks will likely lead to false positives.

Once a block has been loaded for comparison, sixteen points of reference in both blocks are compared to see if they are the same. There are two reasons why a sixteen-point reference scheme is used in the file fingerprint approach. First, the sixteen-point reference is the standard for physical fingerprint comparisons in England and Wales [15]. It is noted that other countries have different standards and the application can be altered to reflect the needs of different legal systems. Second, to defeat the possible attack of a suspect altering byte values, a small number of comparison bytes are used rather than comparing the whole block. If using the entire block for comparison, only one byte would need to be changed to defeat the approach. However, as only sixteen points of reference are used in the fingerprint approach and the points of reference can be randomised when creating the fingerprint, a malicious person would have to alter a large number of bytes within the entire file.

Figure 4 illustrates the digital fingerprint comparison process. The evidence file and fingerprints are loaded

into the FORWEB application as illustrated in figure 2. Data in the evidence file is read block-by-block and compared to the fingerprints of files of interest. If all sixteen points in the block are matched to that of a fingerprint, a match is reported.

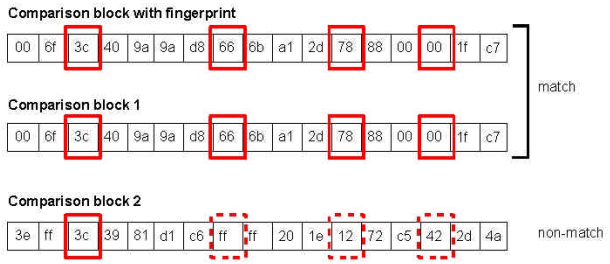

Figure 4. Digital fingerprint matching.

The probability that all sixteen points within a block will match that of the fingerprint is remote. A single byte can take any one of $2^8 = 256$ distinct values. The probability that a single fingerprint byte will match an arbitrary byte within a file with an even distribution of independent values will therefore be 1/256. The probability that all sixteen points will present a perfect match is therefore $(1/256)^{16}$ or approximately 2.29×10^{-1532}.

3.3 Web-spider approach

The FORWEB application utilises a standard approach to Web searching based on Web robot (or Web spider) techniques, and has been built on top of the W3C's own *webbot* Web robot [16]. To initiate a search an initial starting page is chosen, for example, the main home page of an image gallery or photo-sharing Website. The FORWEB application downloads the HTML content for the page, detecting embedded links during the download process. The destinations of these links are then downloaded in parallel and the process repeated, so that the robot automatically traverses across the site – and any sites linked to it – as shown in Figure 5. Any images linked by the HTML, or appearing as inline images on the page, are also downloaded during this process and then tested against the fingerprints using the process described in the preceding section.

Web robots use a well-tested procedure that is commonly utilised by search engines to build up their search databases. Moreover, the W3C's *webbot* tool incorporates various techniques to ensure efficiency and a high throughput of data, such as allowing concurrent parallel connections, caching and HTTP/1.1 pipelining. Indeed, it is used as the W3C's primary tool for HTTP performance measurements [17].

A number of refinements can be used to allow the FORWEB robot to negotiate Web pages in a variety of ways. For example, by specifying a search prefix (or regular expression) the robot can be instructed to search only a single site, a set of sites, or a subset of a site, which allows for more focussed investigations.

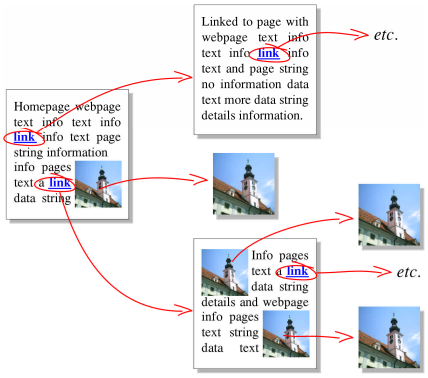

Figure 5. The Web-spider searches by following links.

A choice of depth-first or breadth-first search strategies can be used, depending on the nature of an investigation. For example, if the aim of an investigation is to exhaustively search a particular site to locate any of the images in the signature database, then a "shallow" breadth-first approach (i.e. all hyperlinks on a page scanned first) might be appropriate. Alternatively, if the investigation is less focussed, or there is a desire to locate as many of the images in the signature database as may be found on a given Website, then a "deep" depth-first search approach (i.e. first hyperlink from homepage, then the first link on that page and so forth) would be more appropriate.

4. CASE STUDY AND RESULTS

The FORWEB program is tested by generating a number of fingerprints from JPEG image files and then searching various Web sites where we knew the images would appear to establish whether a fingerprint match would succeed. The performance of the tool is also tested under various conditions and with an increasing number of fingerprints.

The results for the tests on two sites are shown in Table 1. The table shows eight tests performed on the photo sharing Web site Flickr (tests 1-8). A photo set containing 101 images was set up, and the application instructed to search the set to establish any fingerprint matches. The results from a further eight tests (tests 9-16) are then shown for a search of the ACSF conference Web site that is held on servers at the University where the tests were carried out. This site also contained a collection of 101 images to test against the fingerprints in the database. The main differences between the two tests were the layout of the Web sites, and the nature of the network conditions under which they were performed. The Flickr tests represent a more realistic scenario, whilst the ACSF tests reduce the impact of the network on the results and provide a better reflection of the underlying performance.

The tests were conducted on the University network where the local area network is 100BASE-T 100Mbps Ethernet. This is connected to the wider Internet via the Net North West 1000 Mbps connection to the SuperJANET5 UK education and research backbone. The computer running the tests was a 3.4GHz Windows PC with 1GB of RAM.

In every case all images with fingerprints present in the database were correctly identified. The output from test 10 can be seen in Figure 6, where the additional EXIF data generated by the program can also be seen.

```
Loaded 1 fingerprints
Fingerprint 0 found starting at block 11

----------------------
File: http://www.cms.livjm.ac.uk/acsf2/ACSF2007%20001.jpg
Fingerprint matched: 0

JPEG EXIF data follows
File name    :
http://www.cms.livjm.ac.uk/acsf2/ACSF2007%20001.jpg
File size    : 179320 bytes
File date    : Mon, 16 Jul 2007 14
Camera make  : Canon
Camera model : Canon PowerShot A300
Date/Time    : 2007:07:12 11:37:18
Resolution   : 2048 x 1536
Flash used   : Yes (auto, red eye reduction mode)
Focal length :  5.0mm  (35mm equivalent: 109mm)
Digital Zoom : 3.200x
CCD width    : 1.65mm
Exposure time: 0.017 s  (1/60)
Aperture     : f/3.6
Whitebalance : Auto
Metering Mode: matrix
----------------------


Accessed 222 documents in 12.11 seconds (18.33 requests pr
sec)
  Did a GET on 222 document(s) and downloaded 94M bytes of
document bodies (8160640.0 bytes/sec)
  Did a HEAD on 0 document(s) with a total of 0K bytes
```

Figure 6. Digital fingerprint matching.

Table 1 shows the results of applying the approach to the two Web sites. The GET column represents the actual number of files downloaded from the Web server. The HEAD column represents the files that were checked but not downloaded (this would usually occur for files with a MIME type that the robot is not interested in, such as PDF or CSS files).

Table 1. FORWEB results applied to two separate sites.

Tests	Fingerprints	Get	Head	Time (s)	Requests/s	Bytes/s
1	0	408	1170	747.28	2.11	15816.5
2	1	752	1170	748.45	2.57	35575.6
3	10	752	1170	729.24	2.64	36513.1
4	2000	752	1170	746.22	2.58	35682.1
5	4000	752	1170	744.20	2.58	35788.8
6	6000	752	1170	766.48	2.51	34738.7
7	8000	752	1170	746.61	2.57	35663.1
8	10000	752	1170	746.69	2.57	35659.7
9	0	110	0	11.56	9.51	8509917.
10	1	222	0	12.11	18.33	8160640.
11	10	222	0	12.33	18.01	8015670.
12	2000	222	0	14.28	15.55	6919486.
13	4000	222	0	34.06	6.52	2901012.
14	6000	222	0	45.63	4.87	2165856.
15	8000	222	0	58.41	3.80	1691901.
16	10000	222	0	69.20	3.21	1427932.

It is clear from these results that the main overhead for the robot is the time taken downloading the files themselves. Consequently there is a significant difference between time taken to traverse the Flickr pages as compared to the ACSF pages, even though the latter contained nearly four times the quantity of data. This is partly due to the nature of the respective sites: the Flickr site comprises many small individual downloads, compared to the fewer larger files on the ACSF site. However, the main difference is in the nature of the network: the ACSF site could be accessed across the LAN without needing to make connections across the wider Internet.

A more detailed analysis of these results provides an idea of the overhead in checking the fingerprints themselves, as compared to the downloading of the data. Figure 7 plots the time taken to undertake each of the searches against the number of fingerprints in the database. Note that every JPEG image discovered on the site must be checked using the FORSIGS checking algorithm against every signature in the database. The results for the ACSF site – where network bandwidth impacted least – suggest a steady, linearly increasing time as the number of fingerprints increases. Nonetheless, the overhead is still incredibly low even for a very large number of fingerprints. We anticipate a police investigation would involve checking against a large database of fingerprints, it is clear from the results that fast testing can be maintained. Moreover, we claim it remains significantly faster than the performance of a human operative.

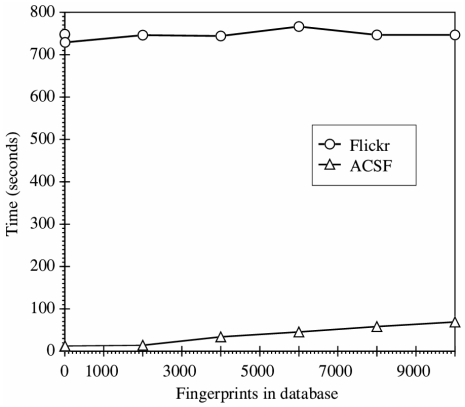

Figure 7. Search times for Flickr and ACSF sites.

By comparison, the increasing trend of the ACSF tests is not reflected in the Flickr results, which remain fairly consistent even as the fingerprint database size increases significantly. We posit that this – again – is due to the network overhead. Network transfer is largely able to continue during the fingerprint checking process, and the overhead of these checks is dwarfed by the time needed to download the data itself. Even though a human operator may be able to refine the search compared to the brute-force nature of the robot, it is worth noting that much of this overhead is a necessary aspect of any search for picture files.

Whilst it is difficult to make direct comparisons between the speed of a human operator and the FORWEB Web robot, we nonetheless believe the results demonstrate the viability of FORWEB as an investigative tool. Moreover, the results highlight the ability of FORWEB to test very large numbers of fingerprints at great speed.

5. CONCLUSIONS AND FURTHER WORK

The FORWEB tool has been developed to aid in investigations that require the discovery of digital pictures on Web servers and the World Wide Web. It takes a database of picture fingerprints generated using the FORSIGS forensics tool and negotiates a Web site or collection of sites to identify any relevant images.

Through the use of a number of test case studies on real-world Web sites we are able to demonstrate the viability of the process, and provide an indication of the time needed to search photo album and file sharing sites.

Although we consider the tool already to have potentially useful application, further testing will be required to establish how well it compares to a human operator. In future work, we therefore aim to undertake further tests. At present, the robot is able to navigate sites based on a number of criteria. However, we believe there is great scope for the development of intelligent navigation techniques that are not aimed at exhaustive searches, but which rather use reasoning to undertake a more focussed approach.

Finally, we also aim to consider the kind of procedure that law enforcement agencies might need to adopt if they were to use such a tool. In particular, the question of admissibility of evidence gained in an environment where content could be changed from one second to the next is particularly problematic. It may be appropriate to require searches be performed from multiple separate independent Internet connections simultaneously. We also intend to examine the effect of resizing and re-sampling or otherwise altering images.

6. REFERENCES

[1] Tendler, S., "1,200 Arrested in British Paedophile Raids", The Times Online, 18 Dec 2002, available from www.timesonline.co.uk/tol/news/uk/article803233.ene, downloaded 18 Jul 2007.

[2] Haggerty, J. & Taylor, M., "FORSIGS: Forensic Signature Analysis of the Hard Drive for Multimedia File Fingerprints", in IFIP International Federation for Information Processing, Vol. 232, *New Approaches for Security, Privacy and Trust in Complex Environments*, Venter, H., Eloff, M., Labuschagne, L., Eloff, J. & von Solms, R. (eds.), (Boston, Springer), 2007, pp. 1-12.

[3] The Forensics Toolkit, available from http://www.accessdata.com, accessed August 2007.

[4] Guidance Software Encase, available from http://www.guidancesoftware.com, accessed August 2007.

[5] Li, W. J., Wang, K., Stolfo, S. & Herxog, B., "Fileprints: Identifying File Types by n-gram Analysis", *Proceedings of the 6th IEEE Systems, Man and Cybernetics Assurance Workshop*, West Point, NY, USA, June, 2005.

[6] Karresand, M. & Shahmehri, N., "Oscar – File Type Identification of Binary Data in Disk Clusters and RAM Pages", *Proceedings of IFIP SEC 2006*, Karlstadt, Sweden, 22 – 24 May, 2006.

[7] Karresand, M. & Shahmehri, N., "File Type Identification of Data Fragments by their Binary Structure", *Proceedings of the 2006 IEEE Workshop on Information Assurance*, US Military Academy, West Point, NY, 21-23 June, 2006.

[8] Schonberg, D. & Kirovski, D., "Fingerprinting and Forensic Analysis of Multimedia", *Proceedings of MM '04*, New York, NY, USA, 10-16 October, 2004, pp. 788-795.

[9] Wong, J.L., Kirovski, D. & Potonjak, M., "Computational Forensic Techniques for Intellectual Property Protection", *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 23, no. 6, June 2004, pp. 987-994.

[10] Krugel, C. & Toth, T., "Distributed Pattern Detection for Intrusion Detection", *Proceedings of Network and Distributed System Security Symposium 2002*, San Diego, CA, USA, 2002.

[11] Haggerty, J., Shi, Q. & Merabti, M., "Early Detection and Prevention of Denial-of-Service Attacks: A Novel Mechanism with Propagated Traced-Back Attack Blocking", *IEEE Journal on Selected Areas in Communications*, vol. 23, no. 10, October 2005, pp. 1994-2002.

[12] Liang, Z. & Sekar, R., "Fast and Automated Generation of Attack Signatures: A Basis for Building Self-Protecting Servers", *Proceedings of Computer and Communications Security 2005*, Alexandria, VA, USA, 7-11 Nov 2005.

[13] X-Ways Forensics, available from http://www.x-ways.net/, accessed August 2007.

[14] ITU/CCITT, "Information Technology – Digital Compression and Coding of Continuous-Tone Still Images – Requirements and Guidelines T.81", September 1992.

[15] Evett, I.W. & Williams, R.L., "A Review of the Sixteen Point Fingerprint Standard in England and Wales", *Fingerprint Whorld*, vol. 21, no. 82, October, 1995, and also the *Journal of Forensic Identification*, vol. 46, no. 1, January/February, 1996.

[16] Nielsen, H. F., "Webbot - the Libwww Robot," W3C, available from http://www.w3.org/Robot, downloaded 5 August 2007.

[17] Nielsen, H. F., Gettys, J., Baird-Smith, A., Prud'hommeaux, E., Lie, H. W., and Lilley, C., "Network performance effects of HTTP/1.1, CSS1, and PNG," *Computer Communication Review*, vol. 27(4), pp. 155-66, October 1997.