Forensic Corpora:
A Challenge for Forensic Research

Simson L. Garfinkel

April 10, 2007

Abstract

Research in the field of computer forensics is hobbled by the lack of realistic data. Academics are not developing automated techniques and tools because they lack the raw data necessary to develop and validate algorithms. Investigators that have access to real data operate under legal and practical restraints that prevent the data from being used in research.

To make progress, we must "prime the pump" by collecting or creating forensic corpora that can be used by researchers. We must also pursue targeted technical developments in forensic file formats, knowledge representation, inference techniques, and the presentation of forensic results.

1 Computer Forensics and Today's Forensic Tools

Today's computer forensic research is largely divided according to the kind of data being analyzed, rather than the kind of analysis being performed. There is disk forensics, network forensics, RAM forensics, cell phone and small device forensics, document forensics and software forensics. Research in all of these areas is limited by the inability of experimenters to obtain large datasets that are realistic, varied, and representative of the data from the field. Because they lack data, researchers can't pursue many of the problems faced by today's forensic practitioners.

Today much of the work in the field of computer forensics is focused on visualizing tools, data extraction techniques, and algorithm development. But this work is generally performed on small data sets provided by the experiment. Few algorithms are validated on a wide range of data, and few tools developed by researchers work reliably in the field when they are exposed to data that is not conformant with the test sets. Even more troubling, researchers are missing algorithms and techniques that require massive amounts of information for proper operation.

This paper proposes the creation of large-scale forensic corpora that meet these requirements:

1. **Representative** of data encountered during the course of criminal investigations, civil litigation, and intelligence operations.

2. **Complex**, with intertwined information from many sources. Data objects should have a wide range of sizes. The data should be in many human languages.
3. **Heterogeneous**, generated using a range of computer systems and usage patterns.
4. **Annotated,** so that new algorithms can be readily validated against existing ones.
5. **Available** to researchers operating in an unclassified environment.
6. **Distributed in open file formats**. Tools should be supplied with the corpora to allow for easy manipulation of the data.
7. **Maintained.** Computer systems are constantly changing and evolving. A corpora needs to routinely augmented with new information or else its interest soon becomes largely historical.

Section 2 of this paper discusses existing forensic corpora that are available to researchers. Section 3 discusses what new corpora could enable and argues that today's players will not create a corpora meeting the above criteria without external funding. Section 4 discusses techniques that could be used to create such a corpora.

2 Today's Digital Forensic Corpora

To date, there are only a few large corpora that have been assembled and made available to researchers outside of the organization that did the collection. Some of these cases include:

- **Network forensic data: The DARPA Intrusion Detection Evaluation.** In 1998, 1999 and 2000 the Information Systems Technology Group at MIT Lincoln Laboratory created a test network complete with simulated servers, clients, clerical workers, programmers, and system managers. Baseline traffic was collected. The systems on the network were then "attacked" by simulated hackers. Some of the attacks were well known at the time, while others were developed for the purpose of the evaluation. [5]

  This corpus is available for download over the Internet (#5), contains substantial annotations (#4) and is in tcpdump-formatted files (#6). But the corpus is not representative of today's traffic (#1), it has been criticized for a lack of complexity (#2) and heterogeneity (#3), and it is not being maintained (#7).

- **Disk images: The MIT/Harvard Used Hard Drive Corpus.** Between 1998 and 2006, Garfinkel acquired 1250+ hard drives on the secondary market. These hard drive images have proven invaluable in performing a range of studies, including the development of new forensic techniques [10] and the sanitization practices of computer users [8].

  This corpus is highly representative of drives from the field (#1), complex (#2) and heterogeneous (#3). It has also been made available to researchers on a limited basis (#5) and is distributed in an open file format (#6). But the data is not annotated (#4), and it is not being aggressively maintained due to a lack of funding.

- **The Enron Email Corpus** of messages that were seized by the Federal Energy Regulatory Commission during its investigation of Enron [7]. Although the data is representative of corporate email (#1), complex (#2), and freely available for download (#5), it is not heterogeneous --- all the data is from a single email server. Nor is the data annotated (#4), distributed in a usable file format (#6), or maintained (#7).

Experience with these corpora shows some of the pitfalls inherent in forensic corpora creation. Lincoln's model-based simulation data has been criticized for containing artifacts, but Mahoney and Chan determined that the corpora could be improved by mixing real traffic into the simulation data[15]. The Garfinkel corpus, while legally exploitable by private institutions under the US Supreme Court's ruling in California vs. Greenwood[22], has been barred from use at the Naval Postgraduate School because of concerns that its use on Federal property might violate the Privacy Act and make NPS susceptible to a charge of privacy invasion. The Enron corpus does not include mail headers or attachments, has many invalid email addresses, and has potential moral issues surrounding its use.

Forensic researchers would ideally have at their disposal corpora consisting of packet dumps, email messages, document files, disk images, and so on. These corpora would be current, maintained and annotated. Ideally they could be freely shared between researchers without the burden of classification.

In the absence of such corpora, many researchers rely on data that is generated by self-experimentation. Disk forensic tools are developed using a few file systems from the developer's own computer system. Network forensic systems are based on packets monitored from the developer's own Internet connection. Documents are based on the range of Microsoft Word and Adobe Acrobat files that can found with Google and freely downloaded over the Internet. Spam is provided by the experimenter's on anti-spam filter. Ironically, many of these collections are used for the experimenter's own research but are not made generally available due to privacy concerns!

Because there are few standardized corpora, researchers at different organizations must waste time and money amassing their own low-quality data. And because this data is self-selected, it is very difficult to compare different techniques that are published in the literature.

2.1 Barriers to Corpora Creation

Beyond the DARPA ID evaluation, there has been little interest in the creation of digital forensic corpora explicitly to facilitate research. One reason appears to be concerns about

privacy. Because of privacy concerns and the lack of a sponsor willing to address them up front, many institutions placed roadblocks in the way of would-be researchers.

At universities, the collection of live network traffic, email messages, chat sessions, and other kinds of forensic information has been blocked by Institutional Review Boards (IRBs) under the theory that the collection would necessarily include data from third parties who had not given consent. Even though federal guidelines have very clear exceptions to the need for individual consent in the interest of research that benefits society or when there is no expectation of privacy, many IRBs are overly conservative in their application of these guidelines.

Inside the US government forensic research has been hampered by the application of the Privacy Act, under the theory that the corpora would represent a system of records operated by the federal government containing the names of US citizens who had not be notified. Although this may not be a valid interpretation of the Privacy Act, it has not yet been challenged.

2.2 Industry is Not Leading the Way

Although companies that develop computer forensic tools have their own internal corpora used for validating their tools, they are not making these corpora available to other researchers for competitive reasons. Sadly, the vendors are also passing up the chance to use their proprietary corpora for enabling long-term research into forensic fundamentals. Instead, vendors are pursuing short-term development projects aimed at improving user interfaces, and supporting new kinds of file types and data formats.

For example:

- According to Guidance Software's website, the key advances in EnCase Forensic Version 6 are: support for 400 native file types using Stellent's Outside In technology; text extraction and indexing; support for Unicode; 64-bit support; improved performance when reading Microsoft Outlook PST files; support for FreeBSD and Novell file systems, Apple DMG files, and Gzip compression; and the automatic reading of hard drive serial numbers during acquisition [11].

- According to the Department of the Treasury, the key features in ILook v8 are: identification of new file types; ability to extract part or all of a file system; improved searching and indexing; ability to handle more file systems and file types; ability to crack some kinds of passwords; tools to visualize the layout of files on a hard drive; and the ability to write scripts using Microsoft's ".Net" architecture.[5]

It's not surprising that both Guidance and Treasury are pursuing similar development roadmaps: both are working to satisfy their existing customers—the trained forensic examiners.

3 Applications for Corpora

Today's computer forensic tools were designed to assist law enforcement personnel prepare testimony for use in a court of law. The tools are designed to acquire information from a hard drive, network or device in a forensically sound manner; store that information in some kind of image file or database; allow the operator to browse or search the database for specific kinds of information; and finally, produce a printout that can be handed to the jury.

Most of today's forensic tools are interactive programs designed to be used by a trained forensic examiner. Although these tools are widely used by police departments, the intelligence community, and investigators hired for civil litigation, they have two important limitations: They cannot be used by personnel who have had no forensic training; and they do not readily scale to handle the massive amounts of information collected by today's intelligence operations.

The lack of automation affects both small and large users of forensic tools:

- **Small-scale forensic users**, such as investigators who encounter computers, memory sticks, cell phones, and digital cameras in the field, are not equipped with the tools to assess these devices and determine what hidden information they might contain. Instead, the devices must be sent to specialists; information is not returned for days, weeks, or even months.

  The long delay associated with forensic processing adversely impacts many missions. Indeed, by the time investigators acquire the information, it may no longer be useful.[1]

- **Large-scale forensic users**, such as national intelligence organizations, are being confronted by a flood of captured hard drives resulting from activities in Afghanistan, Iraq and the Global War on Terrorism. These organizations have responded by hiring contractors to run commercial software such a Guidance Software's EnCase [12]. These programs are run to recover individual documents which are then translated into English and entered into document/knowledge management systems such as Harmony [4].

  The use of interactive programs limits surge capacity. Organizations can readily

[1] In 2005 the government of the United Kingdom argued that it needed that nation's counter-terrorism law changed so that suspects could be held in pre-charge detention for 90 dthe UK government in 2005 for extending the term of pre-charge detention for 90 days, instead of the then-current 14 days. One of the reason for this extension, the government argued, was that 14 days was insufficient time for forensic analysis of hard drives captured during the course of terrorism investigations [13].

purchase more computer systems, but they cannot readily hire more trained forensic examiners with the necessary clearances.

Researchers are not pursuing automation because they do not have sufficiently large corpora of forensically interesting data to develop reliable automated algorithms and tools. Instead, much research in both the academic and corporate worlds has emphasized the development of interactive visualization tools. Because they are designed to be operated by a trained individual, tool failures can be more readily tolerated.

3.1 Tantalizing Breakthroughs are just Emerging

Although the preceding paragraphs may give the impression of a bleak research outlook for computer forensics, the reverse is true: there are exciting, tantalizing research findings are just beginning to emerge at conferences and in the research journals. But this work has been hampered by the lack of large-scale standardized corpora that could be used to assist in both the development and in the validation of these techniques.

Two papers presented at the 2006 Digital Forensics Research Workshop demonstrate this point from opposite perspectives:

- Kornblum's paper "Identifying Almost Identical Files Using Context Triggered Piecewise Hashing" showed how a rolling hash algorithm developed for anti-spam work could be adopted to forensic purposes[14]. Typical application include finding and matching altered documents, and determining if a fragment of a file is present on a suspect's hard drive. But Kornblum wasn't able to report on false positives or negatives when the algorithm was run against a standard corpus of Microsoft Word files, because no such corpus exists.

- Garfinkel's paper "Forensic Feature Extraction and Cross-Drive Analysis" presented a new technique for automatically determining the owners of hard drives and for finding hard drives that are used by various social networks [10]. What made this work possible was the possession of a corpus of 750 disk images that Garfinkel had purchased on the secondary market. Even so, that entire corpus is tiny compared to the number of hard drives seized on a regular basis by US intelligence operations. Currently there is no way to know if these techniques will work on a large scale because no larger corpus of hard drive images is available.

4 Research Opportunities for Corpora Creation and Maintenance

We have identified three areas that would benefit from funding: the direct creation of the corpora; the creation of Anonymization tools; and the creation of standardized file formats and tool sets.

4.1 Direct Corpora Creation

Government funds could pay for the direct creation and maintenance of forensic corpora. There are four approaches to creating data sets that are both forensically sound and exploitable. With the exception of the first, all of these techniques would themselves require research advances:

- **Legally acquire material** on the secondary market that can be used for research. The US Supreme Court has ruled that there are no privacy rights in refuse [2]. We believe that this legal doctrine can be extended to the sale of computer equipment on the secondary market.

  We believe that the Privacy Act issues in such a corpus, including those issues identified by the NPS attorneys, can be overcome. For example, other attorneys might conclude that the information on the hard drives are not actually "records" at all under the terms of the Act because the information is not used by the government to make a determination regarding the named individuals. Alternatively, the information might be classified as "statistical records" and therefore exempt from the Act's notification requirement. Yet another option would be to restrict an acquisition project to used equipment purchased outside of the United States.

- **Anonymization tools** could be used to create an anonymized corpora from actual data.

- **Hiring consenting adults** to have their computers monitored, and thereby create corpora. This approach was used by the CALLFRIEND speech corpus[3]. Such data would need to be scrubbed to eliminate private data from non-consenting parties and could be performed automatically.

- **Using improved models** to create simulated data.

4.2 Anonymization Tools

There has been considerable work done on anonymization of network information. But most work to date has limited itself to log files and network flows; little if any work has attempted to create full-content anonymized data that could be used as inputs for network forensics research.

For example, Pang, Allman, Paxson and Lee developed a tool for anonymizing network packet traces.[17] Version 0.1, the only version ever released, limits itself to anonymizing IP and TCP *headers* for the purpose of protecting the privacy of an organization's network topology; it was assumed by its authors that the content of network traffic would not be released. The Network Security and Architecture Laboratory at the Georgia Institute of Technology has developed a packet-level anonymization system that maps IP addresses, anonymizes IP, UDP, TCP and ICMP headers, and has protocol-level

sanitizers for DNS, Samba, and IRC. [16] But the system is only designed to anonymized attack traffic—it doesn't have the facilities to anonymize normal traffic.

One of the most promising approaches that we have identified to date is the Framework for Log Anonymization and Information Management (FLAIM), under development by the Log Anonymization and Information Management Working Group at the National Center for Supercomputer Applications. FLAIM is a framework that allows multiple anonymization policies to process a variety of data sources. FLAIM's current goal is to support the anonymization of logfiles so that network information can be shared for both research and operational response[19]. We believe that it may be possible to extend FLAIM to packet files.

With such technology, consenting adults could be hired to produce a forensic corpus. The resulting data could then be automatically scrubbed using an X-out policy that would eliminate information from individuals who had not given consent.[2]

4.3 The Need for Standardized File Formats and Tools

Yet another problem facing corpora creation is the lack of standardized file formats to represent the corpora and their annotations, and the lack of tools to develop, maintain, and distribute these corpora.

For example, with the exception of the W3C's Extended Log File Format for web server logs, there is no standard for log files. Nor is there a for disk images, although the file format used by EnCase is widely used.

Garfinkel has developed a general purpose Advanced Forensics Format for disk images and other kinds of forensic information[9]. Support for AFF has been added to the popular open source Sleuth Kit [21]. Nevertheless, AFF is not widely used.

References

[1] Byers, Simon, "Scalable Exploitation of, and Responses to, Information Leakage Through Hidden Data in Published Documents," AT&T Research, 2003.

[2] California v. Greenwood, 486 U.S. 35 (1988). http://supreme.justia.com/us/486/35/case.html

[3] Canavana, Alexandra, and George Zipperlen, *CALLFRIEND American English-Non-Southern Dialect*, Linguistic Data Consortium, Philadephia, 1996.

[2] For example, if the consenting adult replies to a message and includes the original message in their reply, the scrubber would X-out the quoted original message but not the response.

[4] CNN.COM, “Studies: Al Qaeda both complex and dull,” February 17, 2006. http://www.cnn.com/2006/WORLD/meast/02/16/jihad.study/

[5] Cunningham, Robert K., Richard P. Lippmann, David J. Fried, Simson L. Garfinkel, Isaac Graf, Kris R. Kendall, Seth E. Webster, Dan Wyschogrod, and Marc A. Zissman, "Evaluating Intrusion Detection Systems without Attacking your Friends: The 1998 DARPA Intrusion Detection Evaluation," in Proceedings ID'99, Third Conference and Workshop on Intrusion Detection and Response, San Diego, CA: SANS Institute, 1999. http://www.ll.mit.edu/IST/ideval/pubs/1999/Evaluating_IDs_DARPA_1998.pdf

[6] Department of the Treasury, “About ILookv8”, http://www.ilook-forensics.org/iLookv8.html

[7] Enron Email Dataset. http://www.cs.cmu.edu/~enron/

[8] Garfinkel, S. and Abhi Shelat, "Remembrance of Data Passed: A Study of Disk Sanitization Practices," IEEE Security and Privacy, January/February 2003. http://www.computer.org/security/garfinkel.pdf

[9] Garfinkel, S., "AFF: A New Format for Storing Hard Drive Images," *Communications of the ACM*, February 2006. http://www.simson.net/clips/2003.15_972.FinalPaper.pdf

[10] Garfinkel, S., "Forensic Feature Extraction and Cross-Drive Analysis," The 6th Annual Digital Forensic Research Workshop Lafayette, Indiana, August 14-16, 2006.

[11] Guidance Software, “EnCase Forensic Version 6—Why You Should Upgrade,” http://www.guidancesoftware.com/products/ef_transition.aspx.

[12] Guidance Software, EnCase. http://www.guidancesoftware.com/

[13] House of Lords, House of Commons, Joint Commtitee on Human Rights, *Counter-Terrorism Policy and Human Rights: Terrorism Bill and Related Matters*, Third Report of Session 2005-06, HL Paper 75-I, HC 561-i. Official URL: http://www.publications.parliament.uk/pa/jt200506/jtselect/jtrights/278/27802.htm . Report also available at: http://www.statewatch.org/news/2006/feb/jt-hr-cttee-terrorism-05.pdf

[14] Kornblum, J. “Identifying Almost Identical Files Using Context Triggered Piecewise Hashing,” DFRWS 2006. http://www.dfrws.org/2006/proceedings/12-Kornblum.pdf

[15] Mahoney, Matthew V., and Philip K. Chan, “An Analysis of the 1999 DARPA/Lincoln Laboratory Evaluation Data for Network Anomaly Detection,” TR CS-2003-02, Computer Science Department, Florida Institute of Technology, 2003.

[16] Network Security and Architecture Laboratory, Georgia Institute of Technology. http://www.ece.gatech.edu/research/labs/nsa/honeynet/tools/pcap-anon.shtml

[17] Pang, Ruoming, Mark Allman, Vern Paxson, and Jason Lee, “The Devil and Packet Trace Anonymization,” *Computer Communication Review*, January 2006. http://www.icir.org/enterprise-tracing/devil-ccr-jan06.pdf.
[18] Slade, Robert. *Software Forensics*, McGraw Hill Professional, 2004
[19] Slagell, A., Lakkaraju, K., and Luo, K., "FLAIM: A Multi-Level Anonymization Framework for Computer and Network Logs," 20th USENIX Large Installation System Administration Conference (LISA '06), Washington, D.C., Dec., 2006
[20] TalkBank, http://www.talkbank.org/
[21] The Sleuth Kit, http://www.sleuthkit.org/sleuthkit/desc.php
[22] US Supreme Court, California vs. Greenwood, 486 US 35, argued January 11, 1988; opinion issued May 16, 1988.