CREATING REALISTIC CORPORA FOR SECURITY AND FORENSIC EDUCATION

Kam Woods
School of Information and Library Science
University of North Carolina
Chapel Hill, NC
kamwoods@email.unc.edu

Christopher A. Lee
School of Information and Library Science
University of North Carolina
Chapel Hill, NC
callee@ils.unc.edu

Simson Garfinkel
Graduate School of Operational and Information Sciences
Department of Computer Science
Naval Postgraduate School
Monterey, CA
slgarfin@nps.edu

David Dittrich
Applied Physics Laboratory
University of Washington
Seattle, WA
dittrich@uw.edu

Adam Russell
Graduate School of Operational and Information Sciences
Department of Computer Science
Naval Postgraduate School
Monterey, CA
amrussell@nps.edu

Kris Kearton
NCTAMS LANT DET ROTA
PSC 819 BOX 64
FPO AE 09645
kris.kearton@eu.navy.mil

ABSTRACT

We present work on the design, implementation, distribution, and use of realistic forensic datasets to support digital forensics and security education. We describe in particular the "M57-Patents" scenario, a multi-modal corpus consisting of hard drive images, RAM images, network captures, and images from other devices typically found in forensics investigations such as USB drives and cellphones. Corpus creation has been performed as part of a scripted scenario; subsequently it is less "noisy" than real-world data but retains the complexity necessary to support a wide variety of forensic education activities. Realistic forensic corpora allow direct comparison of approaches and tools across classrooms and institutions, reduce the time required to prepare useful educational materials, and eliminate concerns of exposing students to privacy-sensitive or illegal digital materials. The "M57-Patents" corpus can be freely redistributed without rights-restricted materials, and is available with disk images packaged in both open (Advanced Forensic Format) and commercial (EnCase) formats.

Keywords: Forensics, Corpora, Realistic Data, Education, Security, Tool Validation

1. INTRODUCTION

Digital Forensics combines expertise and methods drawn from computer science, criminology, psychology, and other related fields. Most forensic curricula expect students not only to master existing tools, but also to build an understanding of the strengths and limitations of these tools in their application to real-world data. This understanding fosters the improvement of existing technologies and development of new leading-edge techniques drawing a diverse range of areas including cryptography, machine learning, linguistics, and visualization (Garfinkel 2010).

A fundamental issue in security and forensic education and research is that *real data* is often unsuitable for education purposes due to the presence of information that is confidential. As a result, many of those who teach digital forensics spend a significant amount of their time preparing disk images, packet dumps, memory dumps and other kinds of forensic materials for student use. But the resulting data is often *insufficiently realistic.* A related problem is that many of those who created forensic data for student use, in an attempt to mimic the real-world, inadvertently make the data sets *needlessly complex*. It is exceedingly difficult to create test data sets that are both simple enough for classroom analysis and complex enough to convey external validity.

This project seeks to overcome the paucity of existing constructed realistic corpora that mimic real data without the associated privacy and security concerns. We do this through the creation and distribution of more than 40 digital forensic images, packet dumps, and memory images. These sets are free of privacy-sensitive information, are usable without IRB approval, and are freely redistributable without concern for either privacy rights or copyright.

2. PRIMARY OBJECTIVES

Several key objectives have guided the development of this corpus:

1. *Answer Keys*
   In discussions of this project with educators, the most oft-requested feature is that each digital artifact in the corpus include an "answer key" that explains what information can be found in each artifact, where that information is located, and how the problems should be solved.

2. *Realistic Wear and Depth*
   The digital artifacts are created to contain realistic *wear patterns* and *depth*. From the perspective of the investigator, these systems appear to be normal computers that are used on a regular basis for personal communication, web browsing, application installations, and file creation and transfer.

3. *Realistic Background Data*
   One of the primary difficulties that investigators encounter in real-world cases is distinguishing data that is relevant to the case from the sea of background data. The majority of data sets previously developed for educational purposes contain the scenario data and little else. We address this problem by incorporating realistic background data.

4. *Sharing and Redistribution*
   We intend the majority of the digital artifacts we create to be freely redistributable. Materials that contain commercial or copyrighted data–such as Microsoft Windows executables–are made available in redacted form or distributed as originals to organizations that affirm that they possess the appropriate license (for example, the Microsoft Developer Academic License).

The corpus we describe here is accompanied by instructional materials that can be adapted to specific classroom needs and environments. These instructional materials provide ground truth about *who, what, where, when, and how*, regardless of what information may have been lost or is unavailable to the student analyzing the data.

3. REALISTIC FORENSIC CORPORA

Creating realistic forensic corpora that are plausible, internally consistent, and useful in a range of educational contexts is a complex task. As with any attempt to simulate a real-world system, significant planning is required prior to execution of a scenario in order to facilitate the desired outcome. A scenario plan should identify specific education and forensic objectives. For example, if we wish the student or trainees to find a sequence of files that have been transferred from one device to another and subsequently deleted from the original device, these actions must be reflected both in the files themselves (via their final locations, modification and access times, and traces from deletions) and in any logs maintained by the operating system. User activity and any logs or data stores maintained by individual applications (for example, emails sent and received) must be consistent, both in content and any associated metadata.

Scenario sequencing and goals must therefore be carefully planned and transcribed onto calendar dates ahead of time. In more complex scenarios taking place over multi-day periods–in which file systems evolve significantly with use–this is the only way to ensure consistency and limit the introduction of information from non-scenario activity. That is, the scenario calendar must reflect events that take place in *real time*; it is impractical to attempt creation of forensic data that takes place in the future or the past.

In addition to planning the scenario, it is necessary to plan what sort of problems the future forensic student will be asked to solve and how the students will solve the problems. It is also necessary to capture the information required to solve the problem. Significant advance planning is required to adjust the difficulty level of a realistic scenario to meet the needs of a particular class. Introductory classes may focus on learning the operation of tools and on solving discrete problems, while more advanced classes may use exercises designed to simulate the investigative process in more depth. Although it is possible to satisfy multiple needs with a single corpus—and we believe we have done so with M57-Patents—the goal of audience flexibility adds additional complexity, making prior planning even more critical.

Once created, a corpus that is sufficiently realistic can be used for other tasks, such as tool validation and even forensics research. We elaborate on some of the issues involved with existing corpora below, and show how they may be addressed by using a realistic data.

3.1 Training and Education

Most currently available forensic datasets are inappropriate for use in a classroom environment. Drive images with real data acquired from production environments or personal hardware (or purchased from third parties) generally contain private, sensitive, or legally encumbered material. Such images may also contain illegal content. Likewise, data sources from actual forensic investigations can generally not be used in classroom and training environments. Drive images drawn from real environments may further be complicated by the use of security or obfuscation tools–e.g. encryption or steganography–which can impede training when time is limited. Finally, synthesized or collected materials–test datasets, forensic challenges, data constructed by instructor, fake or generated data, and publicly available datasets–present additional issues which may be resolved through the use of realistic corpora.

3.2 Issues with Existing Training Data

There exist a small number of corpora in the form of test data sets and forensic challenges. In our experience these datasets are frequently developed to test a suite of tools rather than as educational aids, and they do not typically represent real-life problems or present specific goals to be accomplished. Realistic corpora can provide specific problems for students to solve while remaining sufficiently complex to exercise available tools. Meanwhile, forensic challenges–including datasets developed by the Honeynet Project1, DFRWS2, and DC33–are often too difficult for students to solve.

Another problem with existing data sets is that the solutions to many of the challenges have already been widely distributed, and as a result answer keys and walkthroughs can be found online. We address this problem by restricting access to our answer keys (through the use of encrypted documents

[1] http://www.honeynet.org/

[2] http://www.dfrws.org/

[3] http://www.dc3.mil/

made available only to instructors).

Finally, there exists a range of public datasets that contain information that seems private but which is not. Examples include Enron emails, YouTube videos, public Facebook profiles, and public chat logs. While these datasets have proven invaluable to researchers for statistical analysis and tool validation, because they are publicly available, well-researched, and frequent subjects of popular media, students may already know what must be found in order to "solve" the associated cases. Realistic datasets can incorporate features common to such datasets (email exchanges, social media interactions) in novel settings that exercise the mechanism of the investigation without the risk of prior knowledge.

3.3 Tool Validation

Tool validation is an important task in forensics operations and research (Carrier 2005; Beebe 2009). Although other datasets exist for testing tools and providing tool validation, the M57-Patents scenario provides additional datasets that can be correlated across various media, and annotations for verification procedure. A tool could validate itself across traffic and verify that specific traffic was generated by checking images of drives from workstations. A primary advantage of using the M57-Patents corpus is the ability to correlate information from various media sources and to verify that tools are performing the specific functions. Additionally, detailed annotations accompany the corpus. These annotations simplify tool validation, because known attributes are already associated with the datasets.

4. CREATING REALISTIC DATA AND SIMULATING SYSTEM WEAR

Realistic datasets must contain data that now only is consistent with the situation(s) being simulated, but also appears to have been created or manipulated by entities whose personalities, motivations, goals, and modes of interaction are consistent (or can be uncovered) within a particular timeframe. We employed personas—synthetic identities, each with their own backstory, motivation and skills—to research assistants tasked with scenario creation. The personas allowed us to create realistic data and reduced the possibility of accidentally introducing information associated with real identities.

4.1 Scenario Planning, User Roles, and Automation

We created an in-depth "game plan" to help us sequence all scenario events. This plan allowed us to ensure they occur at or near a specific time and allowed us to maintain realism. At the start of each day, the research assistants were given a set of notecards with specific numeric ordering and timing information. This out-of-band communication mechanism provided the assistants with details of which commands to execute, which URLs to visit, and tasks to perform such as sending an email message to another person. Research assistants logged the time that they completed a specific task, and these logs were combined to generate a complete timeline that is included in the corpus teaching materials. The timeline allows teachers to fine-tune in-class exercises and provides a gold standard for identification of activities within the scenario.

Storylines and day-to-day activities were developed following examination of both media accounts and the observation of actually criminal and malicious activities in real-world data. We also based the evidence that we created on the specific types of activity and data storage formats that investigators would uncover during an actual investigation.

Some of the scenario activity was automated via software scripting to provide additional depth to the data contained within the file system and support the illusion of real persons carrying out daily work and personal activities. Specifically, we wrote a program that would automatically generate web traffic according to previously fetched URLs. Careful planning of these scripts was important to ensure each persona remained "in-character" during the whole scenario–*e.g.,* visiting favorite websites repeatedly.

Scenarios with sufficient breadth and depth of planning as well as extensive user activity are valuable in a variety of contexts beyond introductory forensic education. Well-planned and executed scripts produce datasets that embody "ecological validity," can be adapted according to varying instructor needs, and–most fundamentally–reduce the burden on instructors to create their own datasets (a process that is both time-consuming and error-prone).

4.2 Secondary Data Sources

In normal computer crime situations, an incident response team will acquire many types of primary and secondary data in order to fully investigate the situation and report to law enforcement (Eoghan 2004). These data can include bit-identical copies of computer workstations and related computer systems; network packet captures showing suspect communication; central login records from authentication and authorization servers; email spool files; and DHCP lease records. Analysis and correlation of these heterogeneous data sources provides the fundamental basis for a case.

Construction of realistic data corpora allows us to enrich the data that would typically be captured in a real-world investigation with supporting materials that may be used by students to explore details of the scenario background; confirm or refute theories developed about how a particular action or event transpired; or develop experiments structured to test such theories. Supporting data can include network packet captures acquired from a scenario router, memory dumps, and snapshots of critical operating system components such as the Windows Registry. Additionally, while such data are not generally available in a real-world incident response scenario, "live" forensic data such as RAM dumps from running machines provides support for training in techniques that are not yet widespread in professional practice.

5. CORPUS CONTENTS, COLLECTION, AND METADATA

To support realistic computer forensic investigation training, we collected all of the data that would typically be gathered in a real incident response or investigation scenario. Each data component was cryptographically hashed, time stamped, and accompanied by annotations describing relationships within the data and specific criminal actions associated with particular times or data sources. Accidental deviations from the scenario (for example, a missed task) and equipment failures were logged; no attempt was made to artificially insert data into any part of the corpus after the fact.

In addition to the set of data that would typically be collected by an incident response team–and extracted from hardware in a laboratory after the fact–realistic corpora can be augmented with data collected during the execution of the scenario. The data may include disk images, RAM dumps, other device images, and network traffic collected on a day-to-day basis and at the termination of the scenario. Of course, most of this data would *not* be available in a real-world incident response scenario (Brown 2010). We include it to allow for the possibility of *student research projects.* In our experience, many students who attempt original research are overcome by the difficulty of collecting the data that they wish to analyze, and rarely get to the point of doing sophisticated analysis work by the end of a class. By collecting and providing this information, we believe that students interested in doing original research will be more likely to realize their goals.

6. THE "M57-PATENTS" SCENARIO

In the following sections we describe "M57-Patents," a realistic scenario and associated corpus designed for primary use in educational and training exercises. M57-Patents has been designed to closely replicate many of the properties of real-world data.

6.1 Scenario Details

In this scenario, "m57.biz" is a new patent search company that researches patent information for clients. The business of patent search is to generally verify the novelty of a patent before the patent is granted–or to invalidate an existing patent by finding prior art (proof that the idea existed before the patent). At the start of the scenario, the firm has four employees: The CEO and founder Pat McGoo, one IT administrator, and two patent researchers. The firm is planning to hire additional employees as new clients are booked. Since the company is looking to hire additional employees, they have an abundant amount of technology on hand that is not being used.

The role of each employee persona in the scenario was performed by an individual researcher at the Naval Postgraduate School (NPS). Basic activities performed during the scenario included checking and writing email; surfing the Internet; staging and carrying out a variety of malicious and/or "illegal" activities; and using office document creation and other software. Malicious activities appearing in the scenario include but are not limited to theft of company property; proprietary information exfiltration and extortion; use of spyware such as key loggers; and viewing illegal content. (For the purpose of the exercise the "illegal content" are non-copyrighted pictures of common house cats; they are meant to be a simulant of actual illegal content such as child pornography.)

The scenario terminates when police receive information from an individual outside of m57.biz who has purchased a desktop workstation from an advertisement on Craigslist. The purchaser found the aforementioned cat photographs. Investigators are able to trace the machine back to M57. When the police contact the CEO of M57 (Pat), Pat confirms that the hardware has been stolen, and provides a list of additional items stolen from the company inventory. Pat gives consent for the police investigators to search M57 and image all of the company computers, company phones, and removable USB drives. Pat also holds a meeting of his staff and tells them that the police are on their way.

6.2 Personas

The "M57-Patents" scenario includes four main personas representing the employees of the m57.biz company: the CEO (Pat McGoo), the IT administrator (Terry Johnson), and two patent researchers (Jo Smith and Charlie Brown). Unknown to McGoo, several of these individuals are involved in illegal activities including theft, extortion, data exfiltration, and collection and distribution of illegal explicit images.

Several other personas were created outside of the company to simulate real-world interactions. These personas represent friends, acquaintances, clients, and other individuals in contact with the M57-Patents employees. Their involvement included buying company hardware via Craigslist, purchasing exfiltrated patent information from within the company, and normal personal correspondence with the main scenario actors.

6.3 Timeline

The M57-Patents scenario took place within a 17-day period between November 16th and December 11th 2009. Within the scenario, a workday started at 9:00am and ended at 4:00pm. Each day was marked in the timeline by a small number of primary objectives to be completed by research assistants playing the company personas. In addition to these objectives, each persona performed some normal background activity–web browsing, emailing friends and co-workers, patent searches, and writing word-processing documents. Researchers used out-of-band communication to facilitate activity coordination within the lab. As previously mentioned, additional texture was provided through automated web-browsing scripts.

In addition to the scenario, a number of technical procedures were performed each day outside of the scenario. These included verifying that all objectives on the daily activity checklist had in fact been accomplished; confirming that the automation and network capture scripts were running; and making a disk image of each computer.

7. SCENARIO CONSTRUCTION

7.1 Network

The network for M57-Patents consisted of four computers connected to a single switch, which then was connected to a gateway providing a connection to the Internet. Jo required two computers in the scenario. Only one of Jo's computers was on the network at a time; the replacement was made due to Jo's original hardware “failing” a week into the scenario. (In the scenario, the computer does not actually fail, but Jo is told that it fails.) Figure 1 shows the network design used for the scenario.

7.2 Workstations and Devices

Workstations used in the M57-Patents scenario were prepared as clean environments. First, we purged the hard drives with a single pass of NULL characters over the entire hard drive of each machine. From this clean state, a single partition was created onto which the operating system was installed from original installation media. All of the hard disk images were formatted with NTFS. Once installed, the systems were updated via Windows Update.

Five other devices were used in the scenario and subsequently imaged: four USB drives and one cell phone. In the scenario, one of the USB drives is Jo's personal storage device, while the remaining three were are “work” drives belonging to M57. Control of at least one drive changed during the period of the full scenario. The cellphone was likewise used for personal purposes by one of the employees and plays a part in at least one of the criminal activities.

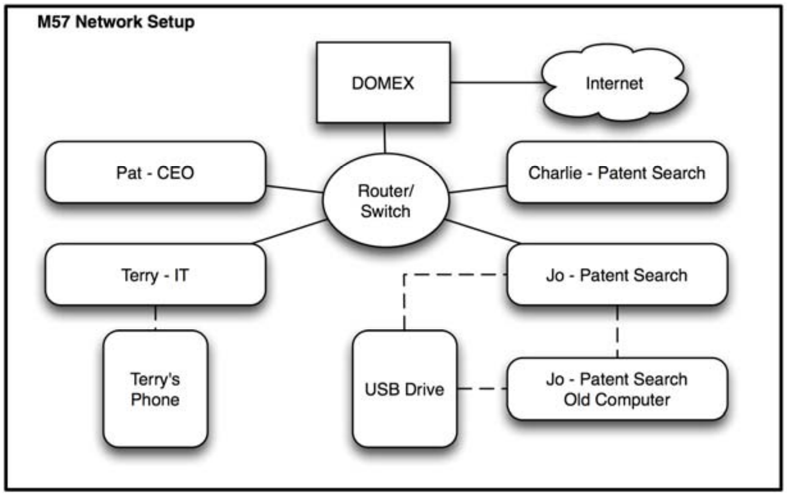

Figure 1: M57-Patents network setup, isolated through server (DOMEX). By isolating the network in this manner, it was relatively easily to tap all network traffic (at DOMEX) and to consistently route the scenario's email traffic.

7.3 Disk Images, Memory, and Network Captures

The workstation hard drive for each persona was imaged at the end of every workday (excluding weekends and holidays) using the *aimage* disk imager. At the end of the scenario the hard drives were imaged again. The disk images are stored in the Advanced Forensic Format (AFF) from which raw disk images can readily be extracted (Garfinkel et al. 2009a; Garfinkel 2009b).

RAM contents of each workstation were also captured daily, except for weekends and holidays. The contents of RAM were extracted using both *win32dd* and *mdd*. We provide both versions for download.

Four USB devices and one cell phone used during this time were imaged once at the end of the scenario. The USB devices are stored in AFF as well as RAW format. The cell phone contents were imaged via the SIM card. This method was feasible since–at the beginning of the scenario–the phone's settings were altered to store all of the non-multimedia data to the SIM card.

A network tap was placed on the gateway's interface using *tcpdump*. Data was collected every day the scenario was in operation, including weekends and any holiday that occurred during the scenario. The *tcpdump* script produced a daily *tcpdump* file with dmp file extension. The network capture dumps are currently available as single-day downloads as well as within a package containing capture data for every day of the scenario.

8. DISTRIBUTION CONSTRAINTS

The M57-Patents scenario is intended for free, public distribution. Because of this, a fundamental goal of the design and implementation was to remove instances of copyrighted material and personally identifiable data.

This section addresses copyright issues, scrubbing private information, answer key distribution, and the generation of simulated objectionable material.

8.1 Redacting Real-World Information

Separation of the scenario environment from the real world is difficult; in complex scenarios some real-world information inevitably seeps into the corpus either through user error, improperly configured services, or simply as a consequence of unforeseen issues inherent to the environment. As an example, in the M57-Patents scenario the workstations were connected to an isolated network using a local outgoing mail server (for emails between employee personas) that stamped each email with header information identifying the domain as *nps.edu*.

The personas in the M57-Patents scenario were performed by students and researchers acting out events in a pre-defined timeline. Although the prescribed events and business behavior was detailed, the actors may have accidentally engaged in activities outside of this detailed scenario. For example, at least one researcher inadvertently logged into his personal email system via a web browser. A process was put in place for any team member who introduced this type of information into the scenario to create a detailed report of the occurrence–time, site visited, and other information that would help scrub the information. At the conclusion of the scenario we scanned for a number of identifiers including the usernames and email addresses of all of the researchers. When these were found, we excised the TCP streams from the network captures and examined the hard drives to determine if the information had been recorded (it was not).

8.2 Simulated Objectionable Material

A significant asset of the M57-Patents corpus is the ability to simulate objectionable material–such as child pornography–without exposing users to actual illegal content. M57-Patents simulates such material by using images and videos of cats. The simulated material consists of 43 images and four movie files. This source material appears at various resolutions and is present in several pieces of media obfuscated with various methods. In addition, the simulated contraband is distributed in a hash database called the "Monterey Kitty" hash set. This hash set can be used with existing commercial utilities such as EnCase and FTK to automatically locate objectionable material (Guidance Software 2010; Access Data 2010).

Because a goal of the scenario design was to keep all contained information free of commercial and otherwise license-restricted media assets, the images and videos for this set were created from scratch by the researchers.

9. DISTRIBUTION AND ACCESS

9.1 Annotation, Sharing, and Publication

The M57-Patents scenario corpus provides numerous benefits to forensics research—and especially student research. It allows for the M57-Patents data to be published and publicly shared in a variety of forms, since it does not contain private or legally sensitive information. Published research using the M57-Patents corpus can be validated and reproduced, because the data is freely available. Because the data is already collected, students can spend their time developing new forensic approaches, rather than collecting data. Finally, dataset annotations distributed along with the disk images simplify familiarization with the corpus, development of classroom materials, and identifying and extracting data relevant to specific actions within the scenario.

9.2 Distribution

The M57-Patents corpus is currently available for download from the main corpus portal at digitalcorpora.org.[4] Individual workstation images (in AFF and RAW formats), RAM dumps (captured both by *mdd* and *win32dd*), and network captures can be downloaded directly from the site via a calendar link map. Because these materials are relatively large (more than 400GB for the full corpus), we have provided a peer-to-peer option for acquisition and sharing of the data between researchers and educators via BitTorrent files with permaseeds at iBiblio at the University of North Carolina, Chapel Hill. This facility allows us to create customized “views” into the raw corpus that can be downloaded as single packages depending on the needs of the organization or individual. We provide torrents for each set of workstation drive images captured during the scenario, the full set of RAM dumps, the full set of network packet captures, a “police evidence” torrent consisting of only those materials that would typically be collected during incident response, and a torrent linking the entire corpus.

9.3 Annotations, Timeline, and Answer Keys

In addition to the drive images and other raw M57-Patents data, a set of annotations, answer keys, and a full scenario timeline are available to provide background support for the scenario, detail the planning and execution of each criminal action, and provide a master reference for the events during each scenario day. The annotations include some materials to enhance the realism of the scenario and frame the process of the investigation. These include four detective reports prior to and including seizure and imaging of the M57 hardware; a search warrant and affidavit (modeled after real warrants issued in the state of California), and an informal report which can be distributed to students describing the employees of the M57 company and layout of the company’s IT infrastructure.

[4] http://digitalcorpora.org/corpora/scenarios/m57-patents-scenario

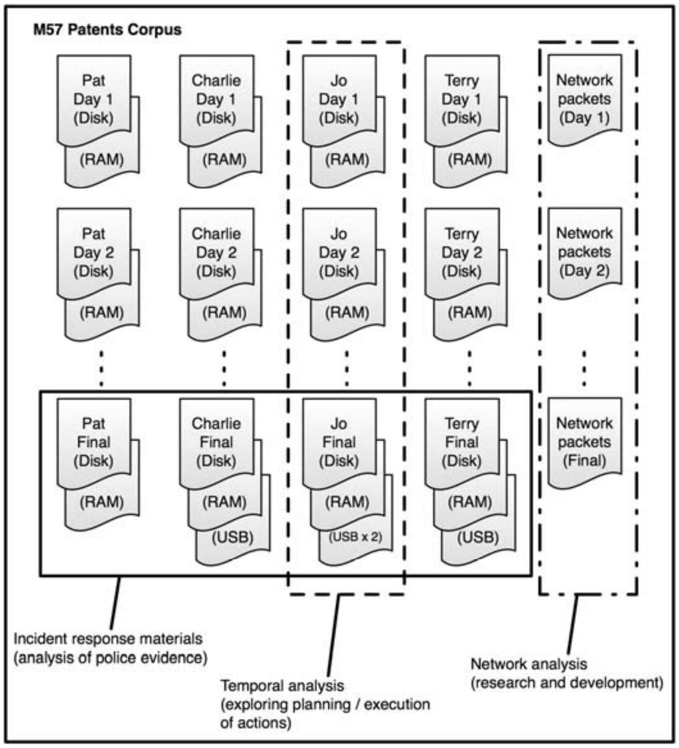

Figure 2: Overview of M57-Patents materials extracted during execution of the scenario. Various educational objectives and levels of analysis can be supported with "slices" of the scenario materials.

The full scenario timeline details (by day and time) any criminal acts carried out by employees including theft, exfiltration of data, extortion, and possession of illegal digital materials. Individual reports are available for each of these activities that further elaborate on the process, in particular providing paths within the disk images to relevant files, messages, software installations, and deleted content. Contents of the employee email accounts are provided as separate text files. Finally, a collection of the simulated illegal images is provided along with MD5, SHA1, and SHA256 hash tables to support various educational exercises.

The distribution of the answer keys is a primary concern for every forensic educator. The M57-Patents answer key is available for download only in encrypted form. The passphrase can be distributed to known professional and academic educators on request. Additionally, the educator must demonstrate that he or she is an educator, professor, or some other individual involved in the teaching of forensics material. This process will not prevent every student from obtaining an answer key (with sufficient effort), but it does introduce a reasonable barrier against cheating.

9.4 Copyright Issues

A clear concern in distributing the drive images is, “Can we legally distribute drive images that contain copyrighted files?” In particular, the M57-Patents data sets contain binaries from Microsoft Windows XP and Microsoft Windows Vista operating systems. For this corpus, Microsoft executables and libraries were disabled in the publicly available images by altering data at the start of the bitstream. While these images cannot be mounted as live workstations, this form of redaction has little to no effect on common methods of investigation using commercial or open source tools. For the end-user who has a MSDNAA license, non-redacted images can be provided upon receipt of the license. During production of the corpus, researchers were likewise careful to avoid downloading rights-restricted digital media such as music, videos, photos, or commercial software. For example, instead of using Microsoft Office, the fictional M57 company uses Open Office.

Later we plan to distribute a tool that can replace the redacted data in the disk images, allowing us to minimize the size of data that must be archived.

10. LESSONS AND FUTURE WORK

The lessons learned from creating the M57-Patents corpus–particularly in terms of handling accidental pollution of the dataset with personal identifiable information, legally encumbered data, and other sensitive materials–are informing our data creation methodologies for additional corpora. Future work will build on these lessons and will be documented to guide other researchers and educators who wish to create their own datasets.

One seemingly simple but deeply important lesson of this work concerns the day-to-day recording of scenario activity and any deviations from the planned timeline. These records provide a ground truth for the finalized timeline and reduce the likelihood of mismatches between scenario answer keys and what is actually found in the data. Understanding that human error and hardware failures are both likely in extended scenarios allows us to build a degree of flexibility into the initial scenario and plan for minor redactions (which we can reliably perform) rather than extensive manipulation of the data after the fact (a process that is error-prone and may further contaminate the data).

In addition to these issues, we are examining ways to enrich realistic corpora with additional activities and related records, including in-scenario communications with third parties or partner organizations, more complex and nuanced personas engaging in more of the kinds of everyday activities performed by real people (e.g. use of social media services), and more finely-grained records of run-time data from scenario host systems (e.g. process listings, network connection logs, and changes to Registry key settings). While some of this data is available in existing corpora, systematizing the process of its collection and organization will streamline the creation of educational materials and allow instructors to focus more efficiently on areas of interest within the data.

11. CONCLUSION

Realistic corpora provide an effective means to improve forensics education. Through careful design and implementation, corpora such as M57-Patents include data with sufficient depth and complexity to support a wide variety of classroom activities without the “noise”, legal encumbrances, and privacy issues associated with real-world datasets. The mechanisms we have described here produce controlled environments that are designed to feel organic rather than contrived; can be quickly assessed with the existing timelines and answer keys; and support sharing and discussion among forensics educators.

Efficient compression and packaging of the corpus simplifies distribution and reduces storage overhead for instructors. The set of “police evidence” materials associated with M57-Patents–those materials that would be captured by an incident response team–is just over 40GB in size. Most of the scenario tasks can be investigated using just this data. Daily disk images provide further mechanisms for temporal analysis and evolution of the file systems, and the corpus includes a plethora of data that can be analyzed using memory and network analysis tools.

Realistic corpora such as M57-Patents can be used for multiple purposes at a variety of complexity and difficulty levels–in undergraduate classrooms and lab, for training exercises, and to support further research and development of digital forensics tools and techniques.

ACKNOWLEDGEMENTS

This work has been supported at the Naval Postgraduate School and at UNC-CH by “Creating Realistic Forensic Corpora for Undergraduate Education and Research” (NSF Award DUE-0919593). We would also like to recognize the research team responsible for production and recording of the M57-Patents scenario: Adam Russell, Henry Gilbert, Stephen Lavelle, Kristian Kearton, Nate Phillips, John Judd, and Joshua Gross.

REFERENCES

Access Data. 'Forensic Toolkit (FTK) Computer Forensics Software.' http://accessdata.com/products/forensic-investigation/ftk. Accessed Feb 19, 2010.

Beebe, Nicole. "Digital forensics research: the good, the bad, and the unaddressed." Fifth Annual IFIP WG 11.9 International Conference on Digital Forensics. 2009.

Brown, Christopher. "Computer Evidence: Collection and Preservation." Charles River Media. Boston, MA. 2010.

Carrier, Brian. "File System Forensic Analysis." Addison-Wesley. Upper Saddle River, NJ. 2005.

Casey, Eoghan. "Digital Evidence and Computer Crime: Forensic Science, Computers, and the Internet (2nd ed.)." Elsevier Academic Press. Amsterdam, Netherlands. 2004.

Cohen, M.I. "PyFlag – An advanced network forensic framework." Proceedings of the 2008 Digital Forensics Research Workshop (DFRWS). http://www.pyflag.net. 2008.

Cohen, M.I., S. Garfinkel and B. Schatz. "Extending the Advanced Forensic Format to Accommodate Multiple Data Sources, Logical Evidence, Arbitrary Information, and Forensic Workflow." DFRWS 2009(a).

Digital Forensics Association. 'Formal education: college education in digital forensics.' http://www.digitalforensicsassociation.org/formal-education/. 2010. Accessed February 19, 2011.

Garfinkel, S. "Digital Forensics Research: The Next 10 Years." Proceedings of the 2010 Digital Forensics Research Workshop (DFRWS). 2010.

Garfinkel, S. "Providing Cryptographic Security and Evidentiary Chain-of-Custody with the Advanced Forensic Format, Library, and Tools." The International Journal of Digital Crime and Forensics, Volume 1, Issue 1, January-March 2009(b).

Garfinkel, S., P. Farrell, V. Roussev and D. Dinolt. "Bringing Science to Digital Forensics with Standardized Forensic Corpora." Proceedings of the 2009 Digital Forensics Research Workshop (DFRWS) 2009(c).

Guidance Software, Inc. 'EnCase Forensic.' http://www.guidancesoftware.com/forensic.htm. Accessed Feb 19, 2010.

Kirschenbaum, M. G., R. Ovenden and G. Redwine. "Digital Forensics and Born-Digital Content in Cultural Heritage Collections. Council on Library and Information Resources. Washington, D.C. December 2010.