Chapter 6

EXPLORING BIG HAYSTACKS

Data Mining and Knowledge Management

Mark Pollitt and Anthony Whitledge

Abstract The proliferation of computer-generated evidence in court proceedings during the last fifteen years has given rise to the new science of digital forensics and a new breed of law enforcement officials, "computer forensic examiners," who apply the rules of evidence, investigative methods and sophisticated technical skills to analyze digital data for use in court proceedings. This paper explores the technical challenges facing the law enforcement community and discusses the application of data mining and knowledge management techniques to cope with the increasingly massive data sets involved in digital forensic investigations.

Keywords: Digital forensic process, data mining, knowledge management

1. Introduction

The term "forensics" means "relating to, used in, or appropriate for courts of law or for public discussion or argumentation" [7]. In the judicial system, forensics refers to a branch of science, e.g., forensic pathology, devoted to providing data and conclusions as evidence in judicial proceedings. The goal of forensic procedures and protocols is to meet federal and state rules governing the admissibility and use of evidence in courts of law. For example, the Federal Rule of Evidence 901 [15] requires that information be "authenticated" to the court before it may be admitted into evidence. All forensic sciences have rigorous requirements designed to ensure that a trail exists backward from the scientific test to the original sample, and to prove that the scientific conclusion is about the sample in question.

Digital forensics has been defined as: "...the application of science and engineering to the legal problem of digital evidence" [12]. Digi-

tal forensics, then, is the science of collecting, preserving, examining, analyzing and presenting relevant digital evidence for use in judicial proceedings.

In the early days—around 1990—digital forensics focused almost exclusively on copying information from computers. With few tools available, examiners sometimes took days or weeks to retrieve pertinent information from a 5 MB or 10 MB hard drive. Usually, the examiner printed the information for the investigator, who could then use it in the same manner as any other piece of evidence. Most computer programs produced paper reports as their output, and the printouts did not seem to differ from other documents created by more traditional means.

During the past fifteen years, the availability of inexpensive computers and the explosive growth of the Internet have significantly altered how people communicate and record their transactions and interactions. The old paper-centric society has been transformed to a computer- or network-centric culture. Society has adapted the old ways of doing business to the new technologies and, at the same time, has found many new ways of utilizing them. Two enablers of this revolution are the dramatic increase in storage capacity and the equally dramatic decrease in the per-unit cost of storage.

Inexpensive storage and new technologies make it possible to create vast amounts of information quickly and easily, and eliminate the cost-based need to delete much of this information. Most investigations and legal cases focus on only a fraction of the information contained in a data set. Finding relevant information in a data set collected from an individual or business has become a search for the proverbial needle in a haystack. This paper discusses the inability of current forensic methodologies to deal effectively with the increasing volumes of data collected by investigators. In particular, it examines two approaches—data mining and knowledge management—that might provide the needed paradigm shift.

2. Data Volume

The FBI has reported dramatic increases in the number of cases involving digital evidence and the volume of digital evidence: from 2,084 cases and 17 terabytes of data in FY 1999 to 9,593 cases and 776 terabytes of data in FY 2004 [6] (see Figure 1). Data for FY 2005 indicate further increases in the number of cases and the volume of evidence.

This trend is not an FBI anomaly and is, in fact, representative of the entire digital forensics community. Researchers at the University of California, Berkeley have studied the amount of digital information

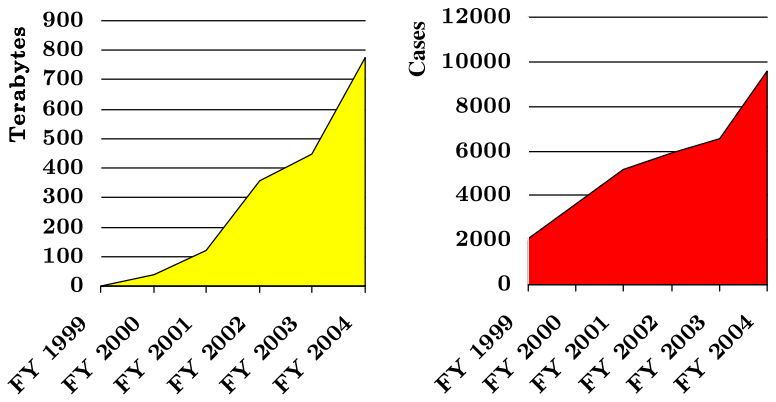

Figure 1. FBI-reported data and case volumes for FY 1999–2004 [6].

created annually [9]. By their estimates, the entire planet produced between 2 and 3 exabytes (10^{18} bytes) in 1999 and about 5 exabytes in 2002, corresponding to a growth of 30% per year. To put this volume of information in perspective, the Berkeley researchers equated the 5 exabytes to the equivalent of 37,000 Libraries of Congress, or more than 800 MB of data per living person. Incredibly, as much as 92% percent of the new information is stored on magnetic media, primarily hard disks [9]. Another source [16] has estimated that this volume of information represents "all the words ever spoken by human beings." This trend is likely to continue for the foreseeable future.

Data volume, however, is not the only issue. If it were, simply adding computing resources would solve the problem. Rather, the increase in volume is indicative of increasingly massive data sets that hide small amounts of information relevant to an investigation. The goal of a digital forensic examination is to develop "information of probative value." The determinants of success in this process are knowing what information to look for, discovering where to find it, and then recognizing it when it is encountered. However, existing methods, especially those that rely on the actions of a single forensic examiner or investigator, do not scale well and, therefore, do not adapt to large data sets.

3. Digital Forensic Process

Digital forensics has evolved over two decades. While the processes of investigation, examination and analysis of digital evidence have not been studied extensively, several papers address the general area [1, 3, 8, 10, 13]. While varying in their focus, all the papers recognize digital

forensics as part of a larger investigative process and one that contains definable functions. This view is consistent with the observations of the authors of this paper, each of whom supervised a national digital forensics program for a major U.S. law enforcement agency.

Over time, the digital forensic process has become somewhat standardized with slight variations based on organization, jurisdiction, purpose and available resources. The process has several components.

- *Media Acquisition:* Obtaining evidence by duplication or seizure.
- *Preservation/Duplication:* Ensuring the integrity of the evidence by a chain of custody and duplication of the original evidence.
- *Documentation:* Documenting the physical media, partitions, file systems and data.
- *Data Reduction:* Using known hash values and other techniques to eliminate data without probative value.
- *Data Selection:* Attempting to identify and extract probative information, e.g., using string searches, header examination, reconstruction of file system structures and files, content examination and the selection of data for further analysis.
- *Examination and Analysis:* Discovering and recording the technical provenance of data and its investigative context.
- *Reporting:* Creating written, oral and electronic products to communicate the results of the examination.

While digital forensic tools have become much more robust, the underlying process has changed little since the 1990s. As we discuss below, two characteristics of the digital forensic process appear to be limiting and both of them involve the need for human judgment.

The classic tasking from investigators to digital forensic examiners is: *"Tell me what is important that will help solve my case."* This tasking makes two assumptions: the examiner knows what is important in the case, and the examiner can identify what is important. Both assumptions place substantial reliance on the examiner's ability to locate high value information and to recognize its significance. Success in this endeavor often boils down to the examiner's or investigator's ability to distill the volume of digital evidence into a manageable data set and to recognize important information when it is encountered. We would suggest that, at best, this process is becoming very difficult and, at worst, it is a gamble.

Thorp [14] has observed that IT-enabled change is creating dilemmas for management. So much information is being delivered by technology that many people now feel that they are drowning in information or are being forced to work with the wrong types of information. Thorp also emphasizes that while IT investment is producing more data, it is not providing the means to translate it into information and knowledge that have business value. This is a very apt description of the state of digital forensics: digital forensics has the ability to manage data, not information and knowledge.

4. Analyzing Large Data Sets

Current digital forensic practices are hardware-centric, focusing almost exclusively on the collection, extraction and presentation of data to investigators. What many forensic practitioners call "analysis," is really part of the extraction process. The conversion of email messages to text-searchable files, the piecing together of data fragments from slack space and other areas of the disk, and the recovery of deleted files are processes that provide raw data to investigators. They are not analytical processes that help investigators understand the data or point them to the most important pieces of information.

Meanwhile, the increasing sophistication of operating systems and storage devices, and the massive volumes of stored data often result in forensic products that are too large and too sophisticated for their forensic examiner customers to understand or use effectively. Subject matter experts—agents and detectives—generally do not have strong computer backgrounds or skills. They may have little familiarity with data storage concepts, file system layouts, metadata and the relationships between the stored data and its content and meaning. Investigators also do not have the skills or equipment to handle the volumes of data being collected and provided to them: data volumes in the 500 GB to 1,000 GB range are increasingly common in fraud cases, and data volumes are on the rise in all types of cases.

On the other hand, the trend in law enforcement agencies is to move digital forensics into a laboratory model that imposes a rigorous scientific approach. This also separates forensic examiners from the consumers of their work product: agents, investigators and detectives who are responsible for cases. The result is a corps of forensics examiners who may not understand the subject matter of cases and are unable to provide adequate analytical assistance to consumers.

It is increasingly apparent that the task of analyzing case data comes after the forensic examiner's work ends and before the agent's investiga-

tive work begins. Forensic examiners generally do not have the analytical skills or the subject matter expertise to assist in data analysis. Indeed, neither forensic examiners nor investigators see substantive data analysis as part of their job descriptions. Given the complexity of analyzing terabytes of data and their background and training, it is unlikely that agents and investigators will master this vital part of the process at any time in the near future.

A notable exception is the Internal Revenue Service's Criminal Investigation Division (IRS-CID), which selects its digital forensic examiners from its cadre of experienced, senior agents. Its "Computer Investigative Specialists" (CISs) are thus both subject matter experts and forensic experts who often work closely with case agents to help with data analysis and "translate" the technical issues that affect the quality of the evidence. Although this works well for the IRS, it may not be adaptable to other law enforcement agencies. The IRS-CID is a specialized unit with jurisdiction over relatively few crimes; it is, therefore, easier for a CIS to be a subject matter expert on IRS casework than it would for a forensic examiner in a larger agency like the FBI. Additionally, many federal law enforcement agencies and police departments may not have the funding to devote investigative resources to digital forensics.

The remainder of this paper discusses strategies for dealing with the gap between forensic examination and investigative analysis in large cases. From the earliest days, law enforcement has adapted tools and techniques used for other purposes, e.g., systems administration and network analysis, to the digital forensic process. This paper proposes the next step: applying concepts and analytic tools designed for business data to the forensic process.

5. Data Mining

Data mining entails the discovery of meaningful patterns, rules or summaries in data (see, e.g., [2]). Data mining techniques cover both exploration and analysis. They may be automated or semi-automated, and are capable of dealing with large volumes of data.

Data mining is all about patterns and relationships between data elements. The use of data mining by businesses to understand and use and transactional and other data is well established. Criminal investigators face the same challenges: to understand and use the data collected during investigations. Specifically, investigators must locate and understand the information that is relevant to their cases. It is equally important to find relevant information that the investigators did not know existed.

Investigators typically start with some knowledge and suspicions, and work to uncover facts that support their hypotheses. However, the massive volumes of case data make this an increasingly difficult task. Furthermore, patterns, connections and facts hidden in the data may not be discoverable without the assistance of analytical tools.

Current techniques for dealing with digital data range from merely printing out documents for subsequent analysis to the application of sophisticated data mining techniques. The techniques and tools used invariably depend on the importance of the case and the investigative dollars allocated to it. Thus, in smaller cases, an investigator is usually left to work with whatever hardware and tools are locally available. Large cases, or cases that attract media attention, are generally given greater resources and access to technically sophisticated analysts.

However, in most investigations, the analysis of digital data is limited to the manual examination of data files, string searches of data sets, and the use of databases to accumulate and sort pertinent information. While some forensic tools have scripting and programming capabilities, their use is normally limited to forensics experts, not the general investigative population. Consequently, automated tools are rarely, if ever, used to locate and analyze relevant evidence.

There are several reasons why investigators have limited experience with sophisticated analytical tools and data mining techniques and why these tools and techniques are scarcely used in investigations. In most law enforcement agencies, investigators and management have little or no understanding of the benefits data mining could bring to investigations. Moreover, agencies typically do not have funds for procuring tools and training investigators on the effective use of data mining and automated analytical techniques. Often, the cost and time required to convert digital evidence into a common format for analysis is a significant factor in the decision not to engage modern analytical techniques.

However, attempts to use advanced analytical tools in investigations have been successful, although not as successful as might be expected. The lead time to develop such tools dictates that generic approaches be adopted and such approaches may be too general to assist with specific cases. In other instances, investigators and developers expect too much from a tool. For example, visualization tools, e.g., the commercially-available Analysts Notebook, are effective at elucidating connections between related elements in a data set, but they do not help with data reduction or provide other types of analysis that might establish otherwise undiscoverable relationships.

As might be expected, the most successful results have come from cases where custom applications are tailored to a specific investigation

or data set. However, such applications are expensive and not adaptable to other types of cases and other data sets, although they do demonstrate their power and utility in investigations. It is expensive to convert collected case data to a common format needed for data mining and to develop the custom tools to mine the data effectively. Such an effort requires a cross-functional understanding of data mining operations and investigations. Yet, few in law enforcement understand data mining or the role it might play in investigations and even fewer data miners understand the goals, methods and requirements of criminal investigations.

In the authors' opinion, data mining tools can provide a means to bridge the gap between the forensic examiner's output and the investigator's quest for relevant information. It would be most useful, indeed, for law enforcement to have a set of tools designed to assist agents and analysts in developing patterns from data that may indicate criminal conduct, identifying individual items of relevant information from a large data set, and reducing data volume by eliminating redundancy and filtering irrelevant and unimportant information.

6. Knowledge Management

Knowledge management is "[a] discipline that promotes an integrated approach to identifying, capturing, evaluating, retrieving and sharing all of an enterprise's information assets. These assets include databases, documents, policies and procedures, and previously uncaptured tacit expertise and experience in individual workers" [11].

Davenport and Prusak [5] describe the relationships between three key elements involved in knowledge management: data, information and knowledge. Data is defined as "a set of discrete, objective facts about events." This accurately describes the portion of the digital forensic process that focuses on files, file systems and structures. A hard drive is thus interpreted as a massive number of data elements.

Davenport and Prusak describe information as "data that makes a difference" [5]. Data becomes information when the recipient recognizes value in the data. Information becomes knowledge when one or more of four processes occur: comparisons, consequences, connections and conversations. The conversions of data to information and information to knowledge are fundamental to knowledge management, and involve the application of human intellect [5].

Thorp [14] describes knowledge management in terms of a "benefits realization approach." He proposes two techniques: modeling and value assessment. The former technique uses a "results chain" method, where outcomes are modeled against assumptions, initiatives and contri-

butions. The latter incorporates four criteria to evaluate the conversions: alignment, benefits, integration and capability/efficiency.

Another key aspect of knowledge management is the conversion of tacit information to implicit information. In the forensic and investigative context, this is clearly a significant issue, as it frames the inquiry and either contributes to or detracts from the outcome [4]. Currently, the conversion of tacit information between examiners/investigators and attorneys is done on an *ad hoc* basis, without any formal structure.

Knowledge management can significantly enhance the practice of digital forensics. Three areas for future research are modeling; data, information and knowledge valuation; and developing effective strategies and methodologies for conversion of tacit information to explicit knowledge. Failure to develop effective mechanisms will likely impede the effectiveness and utility of digital evidence.

7. Conclusions

Criminal investigators at all levels have to deal with increasingly complex cases with massive amounts of digital evidence. Meanwhile, the gap between digital forensics experts, who collect, secure and present digital evidence, and investigators who use digital evidence, is widening. The outputs of forensic examiners are often too large and too complex for investigators to use effectively. The law enforcement community needs assistance in managing large data sets. Also, it needs analytical tools that can work across the disparate data types in investigative data sets. These tools should be easy to use and should produce results that the investigators can understand and explain in court.

Data mining tools and knowledge management principles can (and should) be developed to help address these issues. In particular, research should be directed at developing knowledge management strategies specific to law enforcement that will operate within the specific context of criminal investigations. Furthermore, research should focus on forensic applications of data mining tools and developing analytic tools for forensic examiners and criminal investigators.

References

[1] N. Beebe and J. Clark, A hierarchical, objectives-based framework for the digital investigation process, *Proceedings of the Digital Forensics Research Workshop*, 2004.

[2] M. Berry and G. Linoff, *Data Mining Techniques for Marketing, Sales and Customer Support*, John Wiley, New York, 1997.

[3] B. Carrier, An event-based digital forensic investigation framework, presented at the *Digital Forensics Research Workshop*, 2004.

[4] B. Crowley, Tacit knowledge and quality assurance: Bridging the theory-practice divide, in *Knowledge Management for the Information Professional*, K. Srikantaiah, M. Koenig and T. Srikantaiah (Eds.), Information Today, Medford, New Jersey, 2000.

[5] T. Davenport and L. Prusak, *Working Knowledge: How Organizations Manage What They Know*, Harvard Business School Press, Boston, Massachusetts, 1998.

[6] A. DiClemente, Digital forensics: Current status and future directions, presented at the *First IFIP WG 11.9 International Conference on Digital Forensics*, 2005.

[7] Farlex, Inc., The Free Dictionary (www.thefreedictionary.com).

[8] G. Hama and M. Pollitt, Data reduction – Refining the sieve, presented at the *Second International Conference on Computer Evidence* (www.digitalevidencepro.com/Resources/Sieve1.pdf), 1996.

[9] P. Lyman and H. Varian, How Much Information 2003? (www.sims.berkeley.edu/how-much-info-2003), 2003.

[10] M. Pollitt, A framework for digital forensic science, presented at the *Digital Forensics Research Workshop*, 2004.

[11] The Provider's Edge, LLC., Knowledge Management Basics (www.providersedge.com/kma/km_overview_km_basics.htm), 2003.

[12] A. Sammes and B. Jenkinson, *Forensic Computing: A Practitioner's Guide*, Springer-Verlag, New York, 2000.

[13] P. Stephenson, Modeling of post-incident root cause analysis, *International Journal of Digital Evidence*, vol. 2(2), 2003.

[14] J. Thorp, *The Information Paradox: Realizing the Business Benefits of Information Technology*, McGraw-Hill, Toronto, Canada, 1999.

[15] U.S. Government, *Federal Rules of Evidence* (judiciary.house.gov/media/pdfs/printers/108th/evid2004.pdf), 2004.

[16] R. Williams, Data Powers of Ten (www.davedoyle.com/help/data.html), 2005.