Column: Analysis of Digital Traces

Fred Cohen

In part 1 of this series (Cohen, 2011a), Analysis of digital traces is a foundational process by which the examiner, typically using computer software tools, comes to understand and answer basic questions regarding digital traces.

"Input sequences to digital systems produce outputs and state changes as a function of the previous state. To the extent that the state or outputs produce stored and/or captured bit sequences, these form traces of the event sequences that caused them. Thus the definition of a trace may be stated as: "A set of bit sequences produced from the execution of a finite state machine." (FSM)"[1]

Starting with a bag-of-bits

As a fundamental, when handed some set of digital evidence, it is a good working assumption that the examiner doesn't know what it is other than the fact that it is a trace or traces. This is sometimes called a "bag of bits" to indicate that, other than the fact that it is comprised of bits, the examiner really knows nothing more about it.

In cases where the examiner also performed collection, the details of the collection process may also be known, and so forth. The examiner may also rely on statements, paperwork, claims, and all manner of other things to put the bag of bits into context, but at the start of the examination, anything outside of the personal knowledge of the examiner[2] should be treated as speculative and subject to refutation. Analysis is largely about performing computations on the bag of bits and related information to produce analytical products and derived traces. These products are then used to interpret, attribute, reconstruct, present, and otherwise work with the evidence to other examiners, lawyers, triers of fact, etc. But in order to do this, something about the bag of bits must support or refute hypotheses about what it contains.

Redundancy within and between the bag of bits

Redundancy is inherent in human and current computer language, it is fundamental to the notion of syntax and the ability to differentiate legitimate

1 F. Cohen, "Digital Forensic Evidence Examination", 4th ed. 2012. Chapter 5 is used without further citation throughout this column and should be referred to for a more in-depth review of the subject matter.

2 Note that knowledge is not the same as the other elements of the required basis for expertise in US courts; experience, training, skills, and education. Personal knowledge in this case is intended to imply only things the examiner did and saw.

from illegitimate syntax, and without redundancy, reliability[3] cannot be assured. Fortunately, there is a great deal of redundancy in most digital traces. This redundancy comes in two general forms; internal redundancy (within) and external redundancy (between).

Internal redundancy is present within the internal structure of bit sequences within the bag of bits. For example, if the bag of bits contains a sequence of bits produced by a particular global positioning system (GPS) receiver, it might use the GPX format[4] which uses and XML schema[5] and includes the name of the vendor and sequences of points in 4-dimensional space-time. Internal redundancy comes in syntactic requirements of the language and the specific implementation of the device. GPX, "tags" such as "<time>" and "</time>" surround ASCII text indicated in a format "YYYY-MM-DDTHH:mm:ssZ". If content includes a sequence "<time> 2012-05-10T17:35:23Z</time>" an examiner should readily determine it as inconsistent with the internal format of these files, a type C (internal) inconsistency[6], and doubt the reliability of the record. In this case, is that there is no " " (space) between tags and content in the implementation.[7] Thus a header indicating the GPS type combined with the syntax is internally inconsistent.

External redundancy, also called "between" records, relates to external information. For example, we can determine that GPS systems did not exist in 1901 and that therefore, any record indicating a date and time of that era would be inconsistent with the external records. A date indicating "1901-23-49..." would be of the correct format but externally inconsistent, a type D inconsistency, and an examiner should readily doubt its reliability.

Thus, the examiner uses analysis methods to examine traces in light of the redundant nature of such traces to confirm or refute hypotheses about the content in context. In effect, the examiner uses analysis to place content in context and turn the bag of bits into one or more hypothesized meaningful expressions in a syntax associated with mechanisms that produce such sequences. In addition, the examiner uses analysis to exclude hypothesized event sequences and contexts based on type C and D consistency.

Turning the bag of bits into meaningful content in context

The manner in which examiners typically proceed short cuts this, in that they typically start with assumptions and, unless the assumptions are obviously and dramatically violated, continue under them, even in the face of increasing

3 Reliability relates to the extent to which it reflects the reality it purports.

4 See: http://en.wikipedia.org/wiki/GPS_eXchange_Format

5 See: http://www.w3.org/XML/Schema

6 Details of Type C and D in "Digital Forensic Evidence Examination" I.b.i.d.

7 e.g., GPX file produced by a Garmin Oregon 400t hand-held GPS unit.

evidence to the contrary.

For example, using a tool like EnCase™,[8] an examiner might load a "disk image"[9] and start "analysis". EnCase might identify the disk image as containing a region with a Windows™ NTFS file system partition based on the content of the first 512 bytes of the disk image, assuming that region of the image to be a "partition table", and attempt to analyze that region of the disk as if it were such a file system. As long as this process seems to produce sensible results, the examiner will typically ignore all other possibilities, and proceed on that basis. The tool uses designer assumptions to do an analysis, interpret the results of that analysis, and present those interpretations under the set of assumptions provided by the designer and the user, typically doing so implicitly rather than explicitly. The user typically sees only the presentation of interpreted analysis results, and if desired, can drill down into the presentation of interpreted bases in traces for those results.

An example of a misinterpretation based on analytical assumptions presented to an examiner by EnCase[10] was the presentation of a date and time indicating writing a document in the middle of the Atlantic ocean when in fact it could not have been produced there.[11] In this particular case, erroneous interpretation and representation was the result of a shift in time zones between daylight savings and standard times between the date used by the examiner and present at the beginning of the records under examination and the dates associated with the specific file under examination. In the same case, automated analysis also ignored the second of pairs of date and time stamps within files where there were differences between those dates and times indicative of different time bases in different systems.

All current tools that perform automated analysis, interpretation, and presentation, produce these sorts of results, and it is the job of the modern examiner to understand this. In particular, it is important for the examiner to understand the specifics of the analytical process, examine the results of analysis against the original traces and methods used, and recognize inconsistencies leading to false interpretation and presentation. Just because these sorts of faulty assumptions and mechanisms are present in these tools, doesn't make the results invalid. It does, however, put the onus on the examiner to understand the limits of their tools.

8 This is one of the most popular and commonly used tools in digital forensics today and is produced by Guidance Software.

9 Typically a representation of the bit sequence found on a disk drive or partition within a disk drive.

10 There is no intent to disparage this product as opposed to others, it is only a popular example.

11 United States v. Bayly, et. al., United States District Court, Southern District of Texas, case no. Cr. No. H-03-363.

Analytical methods

There are a relatively small number of well understood, published, and peer reviewed analytical methods used in digital forensics today. The generally fall into a set of areas outlined here, and differ between structured (i.e., following specific rules for syntax and typically produced by fully automated mechanisms based on digital data) and unstructured (i.e., the result of codification of naturally occurring phenomena into digital representations, such a digital photographs or sound recordings) content.

Feature and characteristic detection and analysis
Based on assumptions and hypotheses regarding the bag of bits, and subject to refutation at any time, traces are parsed into syntactic structures and the particular elements within those structures. This is a finitely recursive process of identifying a context (i.e., characteristic), identifying content (i.e., features) within that context, and then treating the content as context for further feature and characteristic detection and analysis. For structured content, characteristics like the document type and its syntax form the context for identifying features like combinations of words used within it and
types of spelling errors, if any. In the unstructured content arena, characteristics like the arrangement of pixels in a two dimensional grid contained within a graphical image are treated as context for extracting and analyzing features, such as areas that look like eyes, tables, or grass.

Recursively, sentences and may be analyzed for language, syntax, spelling, sentence structure, word usage, and so forth. And eyes in a picture may be analyzed as for presence within a face, number and placement, eye color, and so forth. The resulting recursive structures may be further analyzed for consistency with internal or external records, such as whether any people have 5 eyes, or when capitalization is normally used.

Symbol set identification
Part and parcel of the analysis process is the assumption and validation of symbol sets. For example, XML is generally composed of ASCII character sets, excluding select byte codes and forcing other byte codes (e.g., the code for "<") to be used only in specific ways and in specific places. Identifying symbol sets is vital to parsing and to differentiating internal and external consistencies.

Structured and unstructured content are generated from and analyzed to produce symbolic representations. The symbol sets of representations act to define and restrict the analytical framework, and inconsistencies with the analytical framework above base rates are strong indicators of an error in assumptions or hypotheses of the analysis process.

Trace typing
Based on symbol set identification, trace typing is done to identify the specific type of the trace. Typically, this can exist at many levels, such as determining

that content is consistent with ASCII text, in a line-oriented format with fields separated by commas, containing fixed and variable length fields, etc. This can be used to hypothesize about the mechanisms associated with the trace, for example, if the trace is typed to a particular version of a particular device. This may then be used to perform other analysis under the assumptions regarding the operation of the mechanisms known to produce these types of traces.

Parsers, search methods, and related mechanisms

Search is one of the mainstays of digital forensic analysis. In its essence, search looks for patterns within bit sequences. Well known and longstanding methods for computerized search have been studied over many years and they are applied to look for exact sequence matches and regular expressions. Other sorts of search are far more rare, but in the broad sense, parsers may also be used for search. In this case, finite state machines (FSMs) are run against sequences of bits to identify symbol structures within the syntax assumed for parsing. They typically produce parse trees that are then analyzed further to identify content of interest, or elements are placed in databases for subsequent searching and analysis.

Normalization and derived traces

Rather than trying to specify all ways in which the same content may be expressed, normalization is used to translate traces into derived traces that reflect a standardized form of the content. For example, all ASCII coded characters may be mapped into lower case characters so that searches may proceed regardless of the case of the lettering. Similarly, "Jim", "James", "Jimmy", "Jimbo", and "J.Jones@JamesJones.Com" might be mapped into "James" as normalization and placed into a derived trace so that searches for the named individual will find all of those forms. Time and dates may all be translated into YYYY-MM-DDTHH:mm:ss.dddd format, while multiple spaces, tabs or other whitespace separators may be translated into a single space. The list goes on and depends on notions of equivalence or similarity in syntax and semantics.

Similarity analysis and related methods

Similarity analysis is based on some definition of relationships between traces. The relationship is codified in a metric which is then measured between different traces. The result of applying the metric is then used to establish similarity relative to that metric. For example, two email messages may be similar in size if they contain the same number of bits. Multiple relationship metrics may be applied to establish a set of factors that are similar between sets of bit sequences, so that groups of traces are identified as similar or dissimilar to a level with respect to the defined relationship metric.

Time sequencing, travel patterns, and related methods

Analysis of time, movement, and event sequencing is particularly interesting in digital forensics because of the desire to establish what happened when and the

availability of a very rich set of records relating time at varying precision and accuracy. While timestamps may record time and date to the second or millisecond, the basis for those times relative to events at issue are somewhat more dubious. For example, an accurate record of the execution of a program to the nearest second is commonly available, but the process of execution may have lasted for a period of minutes, hours, or days. Understanding what the timestamp actually reflects in terms of that execution may not be provided by the timestamp. Most analysis today simply sorts by time and providers the ordered list of identified records, but this is often misleading in terms of the actual event sequence or relevance. Time sequences are often used to establish travel patterns, such as the use of sequences of credit card transactions at different retail outlets being used to establish that the person using the credit card went from place to place or was or was not capable of being at a particular place at a particular time. But analysis is not attribution.

Anchor events

Anchor events are events external to the traces that can act to tie down traces to externalities. For example, if a message contains bit sequences that are typically associated external systems, events in those external systems may be used to anchor the events asserted to be related to the records reflected in the traces. Traces produced by electronic mail processes typically include sequences bits that include "Received:" headers reflecting timestamps added by mail transfer agents in the path from origination to destination. By finding records of other messages passing through the same external MTAs in the same time frame, and when those records' timestamps are independently determined reliable (e.g., by the examiner having operated the systems that allow timestamps to be validated as reliable), those anchor events provide external context that can be used in analysis.

Building sieves and counting things

Many examinations involve producing counts of various things. For example, a count of how many times a particular telephone number appeared in a log of calls made by a suspect might be relevant to establishing that a relationship existed between two parties or their phone numbers. Many other things are counted in analysis, and this is an area where computers are particularly useful and reliable, if properly applied. In order to count things, computers typically sieve in or out the things of interest or non-interest, leaving the sieved portion of traces to be counted. For example, to find the number of times two phone numbers communicated to each other when the individuals associated with those phone numbers were known to be in different cities, a sieve might be produced to extract relevant phone records and the results counted. Note that such a sieve is not typically available off-hand, and that the examiner is typically called upon to build such a sieve. Once build, many examiners share the details of their methods with others and thus build up a library of partial solutions to analytical problems that they reuse or alter for another purpose

over time.

Presentation and human cognitive analysis

The human visual cortex and brain is far better at rapidly detecting certain classes of patterns than computers. As a result, one of the most common analytical techniques is to produce a graphical image reflective of a set of traces relative to a context and have the examiner identify things of interest to the matter at hand. An example of this is in the analysis of graphical depictions of patterns of communications between groups, where people very quickly identify "key players" once the data is presented in an amenable manner. Similarly, when experts examine things like email headers, they rapidly detect things that "just don't look right", and can often explain them once seen. After this has been done a number of times, there is a tendency for someone to come up with automation to perform such analysis, and the automation of the analysis area largely grows by turning human cognitive methods into automated programs to perform the same or similar functions without the dependency on human judgment, and with repeatability and scalability that far exceeds what people can do.

Traceability to original traces.

A final critical factor in analysis is that analytical results are normally traceable directly to the specific traces associated with those results. Thus, unlike programs that merely sort times, a forensic analysis of times associated with traces will ultimately have to be able to be shown to relate the sorted times to the traces used to producing those times. Thus derived traces need to link back to their origins, normalization requires association with the original traces that were normalized, and so forth.

A final comment

This description of analysis and its methods is not comprehensive, but it may be a reasonable starting point. To the extent that many things are missed in this description, other works attempt to be more comprehensive.[1] But this is a growing and evolving field, and more is better when it comes to identifying methods that have been applied, studied, tested, and published. As always, we welcome your expansion of the art and science and our lists of elements of those.

In our ongoing efforts to define and detail the science and art of digital forensics, standard terminology and common understandings have been found to be an important and largely unfulfilled need.[12] But findings also indicate that by starting to use common words we produce common understandings and consensus around the issues of the emerging science. By describing the field as

12 F. Cohen, "Update on the State of the Science of Digital Evidence Examination", Conference on Digital Forensics, Security, and the Law, 2012

a whole, and in this short piece the elements of analysis, we hope to bring about a unified language and understanding of the field that will help the emerging science to form and the practitioners of the art to communicate and operate as scientists.

But consensus does not come from me telling you what to think or how to say it. It comes from increasing numbers of members of the field adopting common definitions, terminology, and methodology, applying it themselves, and demanding it of others. This is up to you as my readers to decide. As always, feedback helps, and we welcome it. Add your voice to the consensus by responding to this editorial with your views.