Attribution of messages to sources in digital forensics cases

Dr. Frederick B. Cohen, Ph.D. California Sciences Institute dr.cohen at all dot net

Abstract

This paper presents advances in and limits on the attribution of messages to sources in digital forensics cases. It overviews current attribution technologies, the limits of those technologies, identifies ways in which attributions are made today and their limits, and discusses available methods for attribution within the legal system and the limits of digital forensic evidence for use in these purposes. It then presents a set of tools for message attribution now in use and identifies how they are being used and the limits of their applicability to forensics. The present and historical situation show a need for improved attribution, and clearly there is a long way to go before a sound scientific basis for attributions of messages to sources will be definitive.

1. Background

This paper is about attributing messages to sources. There has been an increasing interest in this sort of attribution with the increasing volume of messages and their increasing use in legal matters. However, there has not been a similar increase in scientific research or results relating to the methods and tools used for attribution of messages to sources or their limits, reliability, or utility for attribution.

1.1. Overview of digital forensics terms and basics

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. However, traces are not in general unique as to the input sequences that produced them, and the input sequences do not directly demonstrate what non-digital events sequences may have produced them. As a result, additional effort is required to attribute traces to real-world causes. [1]

1.2. Overview of attribution technologies

This challenge has long been an issue in the digital forensics and secure computing arena, and has been the subject of congressional testimony [2] and substantial research efforts. Even such simple things as using timestamps to demonstrate time ordering of events within a single file system, which is required for most notions of causality, has been experimentally found problematic [4], while more complex attributions over networks have been broken down into different levels in order to identify analytical approaches for doing level 1 (closest computer involved), 2 (computer sourcing the datagrams), 3 (individual causing the level 2 computer to act as it did), and 4 (the organization behind the individual) attribution.[3]

Authentication technologies have been used to try to tie actions to actors, but even biometric technologies are often problematic in terms of normal operation, [5] many of them yielding in the range of 2% false acceptance rates for authentication of one individual out of a few hundred candidates when high fidelity information over a long time frame is available.

Usage patterns for attribution have also been problematic, and many authors have tried and failed to achieve substantial success rates using methods like keystroke timing [6], Web click patterns [7], and textual analysis. [8] This messaging is not the only area in which such attribution has proven problematic.

1.3. Limits of attribution with forged records

Attribution experiments to date have largely been done using cooperative participants, and not considered intentional deception. Deception experiments have demonstrated the ability to readily forge essentially all of the methods identified above as well as all other widely published methods that we are aware of. For example, essentially all of the network-based methods have been shown ineffective against low-level technical deceptions [9], and intentional deceptions of this sort have fooled even skilled attackers. [10]

Digital forensic cases involve traces that may or may not be forged, altered, fabricated, or spoliated by either or both sides to the dispute. Because of the enormous number of possible traces and the inherent challenges in finding consistencies and inconsistencies in traces [11] the detection of forgery and attribution in the presence of forgery is even more problematic. While redundant records have been used to detect forgeries [12] and to identify situations in which inconsistencies reminiscent of forgery are not present, [13] the sheer size of the space of traces makes comprehensive or thorough examination of traces infeasible in all but the most trivial cases.[11]
1.4. The need for scientific methods and approaches

Digital traces are highly volatile and easy of forge and spoliate when substantial redundant records are not available. It thus becomes incumbent on the digital forensics community to seek combinations of technical and non-technical means to attribute actions to actors and to create a scientific methodology and approach to doing so. The US Supreme Court has spoken [14] and the National Research Council has concurred. [15] A rigorous scientific approach is needed.

1.5. The standard of proof

As an additional factor, civil matters have a standard of proof of the "preponderance of the evidence", while criminal cases have a higher burden on the prosecution, of "beyond a reasonable doubt".

The "preponderance" standard is, in essence, more so than not, (i.e., >50%). Without sound metrics, claims of exceeding 50% do not have a scientific basis. The jury then must decide without such a basis. In some cases, courts may prohibit information without error rates and likelihood results, because it fails to provide such a basis, and is not held to be more probative than prejudicial. But this is, by no means, a certainty.

The notion of "reasonable doubt" is determined by juries and judges on a daily basis, but the standard for evidence remains "more probative than prejudicial", and for scientific and technical evidence additional minimum requirements apply. [18]

2. Common mechanisms and their limits

Attribution of messages to sources is problematic for several reasons. Without limit, they include; (1) end-to-end authentication technology is rarely used; (2) many millions of computers are subverted on a regular basis (e.g., worms commonly infect millions of computers); (3) network traffic mechanisms (i.e., NAT, firewalls, etc.) often act to conceal origins and travel patterns; (4) messages and their content are trivially forged at any step of the process; (e.g., by any mail transfer agent) and (5) many parties may have means, motive, and/or opportunity to alter or forge content. Details follow.

2.1. End-to-end authentication technologies

Technologies like digital signatures may reduce the attribution challenges in some cases, but they are not commonly used. When such information is available, it can be checked using mathematical methods, and may be useful in limiting the potential sources of spoliation and fabrication. But they are not uniformly eliminated merely by the presence of this technology. Examples of limits of such systems often include that they; (1) are not not end-to-end, (2) depend on a trustworthy infrastructure and application, depend on control over keys and key distribution, and are not likely to be used by those who don't want to be traced reliably.

2.2. Subverted computers are commonplace

Some of the better known viruses used to create "botnets" have taken control of millions of computers. For example, the "Conficker" worm, averages in the range of 4 million computers under its control, even though ongoing efforts are being undertaken to reduce or eliminate infected computers. [16]

With this many subverted computers, many of which are apparently used to send unsolicited commercial email and twitter "spam", for Internet Relay Chat worms, or leased for other purposes, the IP address or hostname of the source of messages is of only limited value. Many such infestations last for months or years.

2.3. Network traffic mechanisms conceal sourcing

IP addresses were relatively static in allocation up until the late 1990s, but the rate of change of network locations, DNS names and ownerships, and the use of dynamic load balancing, global network diversity, and responses to packet flooding have resulted in far more dynamic assignments of IP addresses.

Even for more permanently and closely controlled computers and addresses, legitimate methods in widespread use subvert attempts to trace messages to sources. For example, (1) proxy servers, (2) gateway computers, (3) cross technology boundaries linking ATM networks with IP networks, (4) network address translation (NAT) gateways, (5) firewalls, (6) large-volume service providers, (7) global load balancers, and all manner of other methods; all limit trace-back.

Mobility has also dramatically changed the nature of networking and address schemes. Many workers work from home, hotels, conference centers, other company facilities, office buildings with addressing mechanisms, copy shops, mobile wireless telephony, coffee shops, malls, Internet cafes, and all manner of other locations. They may send messages from any or all of these locations at different times.

Identity information also widely varies. Individuals having tens, hundreds, or even thousands of different accounts in different systems, all capable of initiating
and terminating messages, each with different modes of identification, authentication, and authorization, many operating as gateways between different sorts of messaging services, and some using anonymizing mechanisms to protect from a wide range of potential leaks, subversions, or other malicious activities.

2.4. Simple forgery is trivial

Forgery of messages is, as a rule, trivial. All someone has to do is use telnet, netcat, or a similar tool to connect to the proper port on a server and use the proper protocol to initiate a message. The message may contain any desired sequence of bytes, and they are, in most cases, no different from the same sequence of bytes sent from elsewhere or other mechanisms.

The mechanisms in widespread use to limit forgery are (1) the placement of reception headers in messages, to provide a trace of sending IP addresses and (2) the use of audit trails in computers en route to record traces of these activities. Forgery of previous path information is feasible within a message, unless audit trails from those systems are available at trial, in which case it requires a forgery of the related audit trails, which may be complex without detection.[13]

The alteration, duplication, spoliation, and loss of messages and logs of messages are also feasible by parties to communications, particularly when they are not encrypted and/or signed. When combined with the use of subverted computers, this process makes the trace-back past those systems far harder, less orderly, and less likely. Subverted systems often have subverted audit mechanisms as well.

2.5. Means, motive, and opportunity

Many participants in messaging may have the means, motive, and opportunity to forge, alter, spoil, duplicate, or fabricate messages, traffic patterns, content, and other message and infrastructure elements.

The means, as discussed earlier, is simple, and potentially available to anyone with Internet access from anywhere in the world. The opportunity is available to an originator, intermediary, or recipient. (1) An originator can use means, such as a telnet session to port 25 from a library, to originate hard-to-trace messages containing any desired content and forged sourcing information. (2) An intermediary to message traffic, including anyone with suitable access to any such system, can make arbitrary modifications, insertions, removals, or duplications. (3) A receiving computer can do the same, as can a final recipient.

Motive varies widely. For example, recipients or individuals who never received messages may claim that they received them as a basis for legal actions, to enhance claims, to provide alibis, to claim permission, or for similar sorts of assertions. Recipients may also claim that they never received messages to deny responsibility, disclaim assertions, or claim ignorance. Similar motives apply to intermediaries or originators.

Actors with motives include, without limit, friends or enemies of one side or another in a legal matter, parties to the action, innocent third parties through errors and omissions, competitors wishing to shift blame, people seeking to trap other parties into confessions or statements against interest, and people who think they are helping or hurting one party or another.

3. Problems with commonly made claims

The challenge to those seeking to admit messages is to show that the messages are anything but hearsay or that they are admissible as hearsay because of some exception. The challenge for those wishing to prevent admission is to show that it is spoliated, fabricated, incompetent, or hearsay. Attribution augments this in that admitted evidence must show causality. This paper is about attribution and discussion is so limited.

3.1. Claim: Message portions are self-authenticating

A common claim, and a common belief, is that all traces of messages are true, correct, authentic, and caused by their apparent source. Unless challenged, this is likely to be the assumption. Once challenged, hearsay comes into play, and from that point forward, someone has to provide some basis for the acceptance of the traces indicative of the asserted message(s).

It is the responsibility of the offering party to provide testimony that asserts the legitimacy of the offered evidence, and this is normally done by some party asserting that they operate the system that this message arrived at, and that the message is a normal business record. Similarly, a representative of the sending party may be called to testify that it was sent from their system, and that their normal business records so show. To the extent that additional sourcing or delivery information is required, audit trails and authentication mechanisms are leveraged for admission and attribution to sender, path, and recipient.

Once some threshold (50%) has been reached, the burden shifts to the challenger to find inconsistencies showing that normal business records are not normal, accurate, reliable for the purpose, and so forth.
3.2. Claim: Form or style indicates authorship

Claimants may assert common "form" or "style" implies common authorship. For example, the writings of Mark Twain have a seemingly unique style and flair. But there have been many successful forgeries of historic documents that have passed the verbiage test, and many years of effort trying to attribute authorship to form and style with statistical methods have failed to produce results with consistent strong correlations.

Generally, form and style may be used to refute a claim of common authorship in cases when there is no basis to assert intentional deception, but confirmanations are of only limited value. For example, if sentence usage differs significantly based on schooling and word usage varies based on regional differences, indicators of different schooling and word usage would tend to refute claims of authorship, but the fact that I can use the term "youns" and write "needs cleaned" does not mean either that I am from Pittsburgh, where I did grow up, or that I wrote an article that uses those terms, which I now have. At most it makes me one of millions of candidates for authorship.

3.3. Claim: Presence of common sequences

The presence of common sequences within messages tends to lend some amount of credence to common authorship. For example, the use of a seemingly unique phrasing of a thought in two places, one at an earlier time, and another at a later time in a different place, tends to demonstrate that the former was the original creation and the latter a copy. In looking at two documents, one from a few days earlier than the other, if the same phrase appears in both and they have different authorship, it seems hard to support the notion that the latter author created the identical seemingly unique phrasing independent of the former. Such an instance was at issue in a Federal criminal case and it tended to the conclusion that original authorship was from the earlier party, even if this was not a definitive determination of the court. [14]

Tools that look for such sequences are used today for detecting plagiarism in universities, but claiming that the "later" author is guilty is potentially problematic, and the common use of such automated analysis tools is creating potentially enormous liability while making students believe that they are under undue scrutiny. A new market is likely to evolve in the creation of forgery software to take a paper and generate a pseudo-randomly generated variation to escape such tools, and the escalation may continue.

Such common sequences can be sought across messages with search methodologies, but the automatic identification of all common sequences is too complex because the set of all sequences is so large. [11] Therefore, such methods require some sort of initial identification of syntactic elements that may be used to reduce the size of the problem. For example, line searches, sentence searches, header searches, URL searches, and so forth have been used for this purpose.

The presence of common sequences does not imply common origination, even if it could be established that it implies common original authorship. The problem is that copying in the digital realm is so easy and copies are not, in general, differentiable from originals in the purely digital realm. This is in fact the very basis for the use of forensic imaging on a bit-by-bit level; the equivalence of a bit regardless of the media of storage.

Authorship is different than origination of a message. Just because the words of Mark Twain are contained within a message does not mean that Mark Twain sent the message. Claims that the presence of a URL, a sequence of words, an address, a phone number, or any other sequence in a message, is not a sound basis for attribution of origination or transmission path. The methods and motives identified earlier clearly show many alternative explanations.

However, the presence of such sequences does imply some sort of access or other mechanism by which the same sequences came to appear in multiple messages or by which they came to be in a message when they were previously in a different form. For example, if a seemingly unique sequence of terms, a complex and original formula, a substantial set of numerical values associated with a process, or other similar thing exists in a message and there was only one party that previously had knowledge of that information, a case for attribution to the party with the unique knowledge as the start of the causal chain may be made.

Many people also use personal knowledge of common history to communicate in ways that seem potentially attributional. For example, in the drug trade, the credit card theft trade, and in other similar small and relatively closed groups, language is it might be asserted that a particular usage implies membership in a particular group. This may be problematic because there are many cases in which investigators have penetrated such groups by learning the language and successfully worked on undercover cases using it.
Similarly, confidence workers often learn the language of their victims and use it to become more trusted.

As a technical matter, it is feasible to both identify different specialized dialects and word usages, and to create messages using these dialects and usages to personalize messages. Identification requires only a syntax tree, and generation can be done with translation methods. These sorts of methods have been used in intelligence operations and undercover work for many years, and this limits their attribution utility.

3.4. Claim: Similarity of a group of messages

There is now significant use of software to assert plagiarism in cases when student papers appear to be similar to other previous papers. Unfortunately, the studies and metrics associated with these techniques are far from definitive, and these methods commonly produce false positives that get adjudicated on a case-by-case basis from within the institutions that apply them. For legal purposes, such measures may be problematic because of a lack of sufficient scientific basis for their introduction as evidence, but many copyright infringement cases are settled based on the impressions of a judge or jury as to similarity.

A major problem with the use of similarity in grouping messages for attribution is that there are often so many parameters available that many links may be formed. For example, a tool described later finds the largest similar groups of characteristics and features for a collection of messages. For a corpus of 4053 messages, 7531 different groupings were found, and there is no basis that we have been able to identify for why one grouping should be favored over others in terms of independently asserting similarity. A study of some set of messages for making a more definitive determination might be instructive, but how it would extend to a different set of messages is unclear.

3.5. Claim: Similar timing or other physics bases

Another asserted claim is that information related to timing, addresses, or names attribute to sources. As in the general case for similarity, there is no generic basis for such comparisons being more valid than other comparisons unless based on external properties of the digital system that provide a reason to assert such an attribution. Timing is an instructive example.

Because of the nature of computational complexity, physical distance, processing speeds, concurrency limits, disk speed, or other similar physical limitations of devices and mechanisms in use; timing results may be determined with a high degree of certainty to be inconsistent with the physics of digital information. [1] In these cases, attribution can sometimes be ruled out, with appropriate caveats about the nature of the systems in place. A similar approach can be used when location information is available and can be authoritatively asserted.

However, the fact that timing and other similar things may be used within a physics methodology to rule out an attribution, does not mean it can be used to definitively rule such an attribution in. Claims of timing may be couched in consistency or inconsistency terms, and generally will require an anchor event [1] or some other similar mechanism to make definitive, unless the event is based on a claim by the other side. As circumstantial evidence, such assertions may have a cumulative effect and help form a positive attribution, but on their own, they are problematic at best.

3.6. Other sorts of claims and their problems

There may be any number of other claims, such as similarity or other patterns, common IP addresses, common from and to headings, common server version numbers, and server reception sequence numbers. Generally speaking, unless there is a basis in information physics, statements of the parties, or external facts, they cannot be used to rule an attribution in or out. But they may demonstrate consistency or inconsistency, and may have a cumulative effect to the extent that there are a substantial number of them and they are all consistent with the asserted attribution.

Clearly, all of these sorts of claims are problematic. In effect, people making such claims without a sound scientific basis are putting up straw men, perhaps with the notion that those who challenge their claims will be unable to unravel them.

4. How to attribute with higher surety

In many cases, more certainty is desired from an attribution, if only because it is almost certain that the opposition will claim that the attribution is not based on anything other than hearsay and speculation. There are two general classes of approaches that have been identified for higher surety attribution of traces of digital messages; (1) consistency and inconsistency, and (2) using legal process to gain additional records.

4.1. Checking for consistency and inconsistency

The fundamental technical approach to attribution and validating or invalidating traces as to their utility in
issues; (1) apply the legal process to other sources of

There are two basic approaches to dealing with these

across jurisdictional boundaries to avoid legal process.

anonymity to their users, and some of them operate

example, anonymizing services offer some level of

and third parties are disinterested or uncooperative. For

the case for attribution becomes stronger and stronger,

chain from the origination point is successfully made,

problems. Additional records are sought on a step-by

spoliation, fabrication, and to find other similar

validate or invalidate attributions, to show consistency

Consistency analysis is done on these records to

authenticate those, and seeking related, records.

Records should be produced in digital form and

added to the collection of traces available for analysis.

Consistency analysis is done on these records to

validate or invalidate attributions, to show consistency or inconsistency with other records and identify spoliation, fabrication, and to find other similar problems. Additional records are sought on a step-by-step basis to trace the messages back to their origins.

As the collection of these records grows and a causal chain from the origination point is successfully made, the case for attribution becomes stronger and stronger, until it becomes very difficult to refute. A complete path and history is rarely available, but enough traces are often identifiable to allow a strong case to be made.

4.3. Uncooperative parties and unavailable records

In many cases, parties to litigation are uncooperative and third parties are disinterested or uncooperative. For example, anonymizing services offer some level of anonymity to their users, and some of them operate across jurisdictional boundaries to avoid legal process. There are two basic approaches to dealing with these issues; (1) apply the legal process to other sources of

information, and (2) use an incomplete causal chain to exceed the threshold required for the standard of proof.

The use of legal process can lend substantial support to the attribution effort, and anonymizing services have been compelled to generate, keep, and produce records for legal cases. But across jurisdictional boundaries, this is less effective, and for cases where illegal entries are involved and no such records are normally produced, this is sometimes problematic.

When a continuous path cannot be shown, it is often feasible to show previous steps in a path, even when no trace of intermediate steps is available. In most cases, a missing step in a message path is not problematic if the steps on either side of it can be shown to coincide. For example, a message going from A to B to C to D to E that can be shown to have passed from A to B and D to E, and where records from B and D are independent of the parties to the matter at hand, may be adequately probative and carry sufficient weight to provide an adequate attribution, assuming there aren't inconsistent records identified that put the attribution into question.

5. Automation for analysis

As message volume increases, legal process time limits mandate automated analysis.

5.1. The current volumes of messages in cases

Legal cases are starting to involve larger volumes of messages. For example, three pending cases in which attribution of messages to sources are key involve (1) thousands of usenet messages from a collection of more than 100,000 such messages; (2) tens of thousands of asserted email messages claimed to have been extracted from a collection of more than a million such messages; and (3) about 100,000 asserted messages from a collection purported to contain more than 100 million messages. Large-scale SMS, Twitter, IRC, News, and other similar sorts of events are starting to appear and will likely expand over time. Clearly, automated tools are required in order to examine such collections in time frames of months. Such tools must also pay careful attention to time and space complexity issues and efficiencies, or they will take too long to meet the needs of the legal system.

5.2. Criteria for tool and methodology acceptance

Reliability information on tools and methodologies, while not strictly required for all legal processes, is a sound basis for challenge, and tools are considered more reliable by the courts if they have been the
subject of significant testing, scientific methodology, and peer review. [14][15][18]

5.3. A toolset for facilitating analysis

To facilitate working on high volume cases involving messages, tools have been developed, tested, and applied, and complexity analysis, methods and results presented in legal matters. The toolset described here is customized to electronic mail, Usenet messages, and similar formats. Processing is equivalent for most messaging schemes, but syntax for initial processing may differ based on the formats of messages.

Tools need to be applied by people with suitable expertise and in a methodology that is appropriate to getting answers suitable to legal needs. In particular, and as an example, just because methods like string searches can be "perfect" in the sense of accurately detecting a regular expression within a larger sequence, doesn't mean that the people using those tools search for the correct expressions, that the things found are suitable to the next step in an automated process, or that results are meaningful to the matter at hand. As a result, on a stepwise basis, and in the aggregate, redundant tool usage and results checking must be undertaken to verify actual results in actual cases. Tools include, without limit:

**Hypermall:** produces (linear time) a set of Extracts, one file per message, named sequentially, containing a forensically sound message subsequence according to the message format specification. It also provides a "separator", "headers", and "body" file with similar names and sequence numbers, each in their own directory. This has been implemented as a modification of the hypermail.pl script that has been widely available over the Internet for many years. The "Header" files are implemented as a sequence of derived traces with multi-line headers joined into a single "entry" (line) for ease of use and compatibility with common search tools. It works for all message collections with proper message separator syntax.

**Firstwords:** produces (linear time) a directory containing a file for each header type in which the message identifier, location (entry number) within the Extract, and content of each header is placed. This greatly facilitates header analysis and searching. It is implemented with a simple parser that searches for the header name separator and appends the header to a file of same the name as the header. The ":" separator is used in twitter, usenet, and mailbox formats.

**Received-Analysis:** parses (linear time) "Received:" headers and produces date and time cross-references, fields within headers, such as By, For, From, and With, normalized times to UTC and formats for easy sorting, produces time sequence analysis with time deltas for time stamps within each header, and produces a time zone analysis that associates time zones with machines over time and feeds into a time zone shift analysis that figures out when computers change time zones and lists the times and tracking information before and after the time zone changes. These headers commonly appear in usenet postings and email messages, but are less useful in message formats with central servers instead of store and forward mechanisms.

**Histograms:** produces (linear time) histograms of each "hop" in all messages for analysis of subsequences of delivery processes. This is done by separating the 1st, 2nd, and so forth "Received:" header entries gathered earlier, into groups by header number, and producing various counts, such as the number of nth headers present on a given day. These can then be sorted by date with hop counts for each date put in a table that allows easy viewing of the gross progress of messages through processing steps. Again, this is for store-and-forward mechanisms only.

**MD5-matches:** produces (linear time) lists of hashes of Extracts and identified sets of identical hashes for separators, headers, and bodies. This is then used to feed into other matching algorithms to find matched headers with different bodies or separators, matched bodies with different headers or separators, and other similar groupings indicative of duplication, spoliation, forgery, or fabrication. This is useful for all formats.

**Correlator:** produces lists of occurrences of "A in B" that coincide with occurrences of "C in D", where A and C are regular expressions and B and D are header files or other similarly formatted files. This allows analysis functions like searching for all messages containing a range of IP addresses that also have a particular string within a "Message-ID:" header. This is useful for all formats, but is predominantly an analytical tool and not an automated mechanism for creating derived traces for analysis.

**Match-correlator:** produces a (n^2 time and space) set of files containing the number and percentage of lines that are identical between each pair of files from a collection. This is particularly useful in finding both perfect and imperfect matches between message sequences, such as duplicated messages forwarded for multiple receptions and comparisons between message...
collections from different steps along a path from place to place. Given, for example, that N lines of header are present and K lines of header are used for "Received:" headers, messages with less than N-K differences (> (N-K)/N similarities), the messages must be related, because, at a minimum, they have an identical reception. This is applicable predominantly to small sequences from messages because of the time and space complexity, which makes times very long for large sequences, such as bodies of large messages.

**Received-tree:** produces a (n log(n) time) tree structure indicative of the paths by which messages reached their final destination. This is particularly useful in identifying anomalous deliveries, identifying structure in delivery infrastructure, and producing collections indicative of common sourcing and delivery paths. This has been shown in [20] along with other related techniques. It is useful for all store-and-forward messaging systems.

**N-tuples:** produces a (n^2 time and space) general purpose grouping analysis of the largest sets of groups of messages and characteristics and/or features so that related sets of characteristics, features, and messages can be examined. This is, in general, problematic in that the number of different symbol sets from which to choose N-tuples of symbols is very large. [21]

A restricted version of this analysis, called "greatest common factor" (GCF), seeks the largest group of sets of factors (i.e., characteristics and features parsed from messages) common to sets of messages. Factors may include presence or absence of particular headers, locations and sequencing of headers, IP addresses and address ranges, software names and versions, date and time stamps, etc. This restricted analysis has proven helpful in grouping large collections of messages into subsets, but it requires that the analyst identify and parse for what constitutes meaningful factors, and results must be examined for sufficient semantic basis.

### 5.4. How the tools help identify inconsistencies

The fundamental use of tools such as these are to (1) help identify consistencies or inconsistencies, and (2) support investigative processes. The identification of inconsistencies and consistencies generally involves internal trace consistency (type C) and consistency with external "events" (type D). [11]

Type C inconsistencies in messages include things like (1) messages with identical identification numbers or similar tags that have different sourcing or content, (2) identical headers with different bodies, (3) multiple messages with identical "unique" identifiers, (4) supposedly identical sources that travel in excess of realistic travel rates for the transport technology, (5) overly or underly consistent delay times in the apparent processing of messages, (6) ordering errors and differences in header sequences, (7) commonalities in messages with different identified sources and travel paths, (8) self-indicating integrity flaws like digital signature mismatches, (9) large numbers of messages with highly correlated content and headers that are supposed to be independent, and (10) travel patterns inconsistent with normal processing.

Type D inconsistencies depend on specific claims and external factors, but often include things like (1) time zones inconsistent with asserted locations, (2) claims of delays inconsistent with timings and volumes, (3) claims of commonalities inconsistent with uncommon elements, or (4) claimed attributions not more consistent with traces than alternatives.

Identifying these sorts of inconsistencies is easier in high volume with these sorts of tools. In some cases, tools identify the inconsistencies directly, but they may also provide derived traces or arrangements of traces that allow more rapidly human cognitive detection.

More examples of consistencies and inconsistencies are identified in [1], but in general, the list of all such consistencies and inconsistencies is enormous [1][11]. Failure to find inconsistencies does not imply that they don't exist, but finding them is revealing because, in science, refutation is generally more powerful than confirmation. [19]

### 5.5. Complexity issues with tools

The time and space requirements for these analytical functions is particularly important because, as the number of messages grows, algorithms with even n^2 complexity become problematic. For example, 10^5 messages at n^2 produces 10^10 steps, which may take hours to days on even a parallel processor, and may take terabytes of space, depending on the particular sequences that have to be stored.

Another complexity issue arises when claims are made about the use of tools and those claims are inconsistent with the complexity known for the class of problems being addressed. Time and space limits of real systems produce breakpoints where any given tool will fail and a secondary approach will need to be used.

Finally, while complexity analysis may identify known limits on best case performance, many tools
5.6. Examples from recent cases

As a standard process in large-volume cases, the author now runs a sequence of processes using these tools. They have revealed inconsistencies indicative of, without limit, (1) fabrication of the collections (e.g., mailbox files not in the proper order for the manner in which they would normally be assembled by a mail transfer agent), (2) fabrication errors (e.g., identical headers with different bodies, headers too similar to represent different messages, supposedly unique message-IDs in multiple messages), and (3) groupings of messages by similarity (e.g., subsets of messages from collections with identical sequencing of header fields at identical locations, subsets of messages with identical body lengths and a high percentage of identical content at identical locations within those bodies, subsets of messages containing identical URLs, groups of messages with identical class C network addresses indicated in headers, groups of messages with similar geographic locations).

The utility of such results can only come from their consistency or inconsistency with the issues in the legal matter. For example, if a fabricated collection is claimed to be original writing, consistency with fabrication tends to defeat the claim of originality. A claim of identical authorship or origin is inconsistent with different sourcing or delivery processes, and the assertion of identical mechanisms from specific groups of messages is consistent with identical locations of content in sequence within a message.

Of course some of these consistencies or inconsistencies are more probative than others, and the courts have to decide whether and to what extent they go to admissibility and/or weight. The presence of an inconsistency that indicates fabrication, for example, would likely go to the issue of admissibility and weight, while similarity measures, which may indicate that some things are similar while others are dissimilar, are more likely go to weight, and the explanations and other evidence supporting and refuting those claims by the various parties will then be considered.

What these tools do not do, is provide the basis in corpa for specific claims as to weight, or provide definitive results as to what specifically caused the specific sequences at issue that come to be. At best, they can help to limit the envelope of event sequences that could have produced the traces at issue, or act to allow comparisons between corpa.

There is a clear lack of adequate basis in experience and supporting studies to support claims about the extent to which similarity measures such as those identified here imply attribution as to authorship or origin. The presence of such tools does not change this. But such tools do provide the means for making systematic comparisons between different corpa.

In one case, an examiner identified a complex header sequence associated with a party and immediately identified the sequence in 64 of 200,000+ messages. Tools revealed that sequences and content of headers were almost identical. This supported extending the attribution to the additional 64 messages, which when examined, showed an independent linkage back to the same party. These 64 messages were also part of an independently identified set asserted attributable to the behavior of the party. This consistency provides a basis for linking attributions and increasing their combined weight. While the tools used to perform the analysis could cause the apparent consistencies, as in all such use of tools by forensic examiners, the results were independently verified by the examiner and made available to the other side, thus eliminating the analytical tools as the source of the consistency.

6. Summary and conclusions

Attribution is problematic in the case of messages, more than in the case of files or other similar traces, because of the need to track causes to effects across infrastructure as it is handled by many systems and mechanisms. The need for attribution is increasingly involving message traffic, and the methods in use for local forensic analysis are inadequate to the needs for messaging-based attribution.

As the size of these traces in these cases grows, the time and space for analysis grows, and the certainty of results become harder to verify. Analysis of collections of hundreds of thousands of messages is feasible for the sorts of consistency analysis currently done, but this is not likely to be true for millions of messages or for more complex sorts of analysis anticipated to detect more complex sorts of forgeries and provide attribution when far more sources of information are involved.

Legal process must be mixed with technical process to produce the sets of traces required in order to
attribute across infrastructure and properly associate a source and travel path with messages as received. Automated tools are needed to improve the likelihood of detecting consistencies and inconsistencies of traces and between traces and events in order to make attribution reliable and well suited for the legal system.

As the surety requirement of the standard of proof increase, so does the need for stronger attribution methods. With the increasing use of digital forensic evidence in the form of messages, the pressure to create, validate, test, and measure methods and tools will increase.

This paper is the beginning of the work toward a scientific methodology for attribution, but it is far from the end. A great deal of experimental work will be needed to validate more and more such methods and to work toward a more definitive set of tests for forensically sound attribution of messages.

7. References


