Chapter 4

BULK EMAIL FORENSICS

Fred Cohen

Abstract Legal matters related to unsolicited commercial email often involve several hundred thousand messages. Manual examination and interpretation methods are unable to deal with such large volumes of evidence. Furthermore, as the actors gain experience, it is increasingly difficult to show evidence of spoliation and detect intentional evidence construction. This paper presents improved automated techniques for bulk email analysis and presentation to aid in evidence interpretation.

Keywords: Unsolicited commercial email, bulk forensic analysis

1. Introduction

This paper focuses on the examination and interpretation of large email collections as evidence in legal matters. The need for bulk examination methods has become more important because of the high volume of emails involved in unsolicited commercial email (UCE) cases in which one party accuses the other of numerous violations of the law in sending such email.

Current laws typically include statutory damages on the order of $1,000 per email message in cases involving fraudulent email [3, 4, 11]. Some plaintiffs are tempted to acquire and/or produce large volumes of email messages and file suits for millions of dollars. They may configure their environments to accept as many email messages as possible and may involve multiple states in the transmission of email to trigger additional damages on a per state basis. The plaintiffs in some of these cases work together in a loose knit group and use the leverage of high volumes to make the potential risk of litigation very high while driving up defense costs [7]. The plaintiffs acknowledge these techniques and sometimes assert that they are activists seeking to make bulk emailers pay a high price for sending unsolicited email.

Defendants in these cases range across a wide variety of companies. Some appear to be criminal enterprises that violate contracts with multiple marketing firms, lease email platforms from criminal groups to send high volumes of email, regularly flout federal and state laws, steal credit card and other related information used in transactions they facilitate, and when sued, shut down and relocate (to Argentina in at least one case). Other defendants are longstanding advertising firms who – almost without exception – seek to follow the laws regarding advertising, including those related to UCE.

From a technical standpoint, bulk email solicitations involve companies that specialize in different facets of the business. Some create and provide advertising copy and images to their clients or place them on web servers, others send emails to large lists of recipients that they maintain in databases, others handle orders and/or fulfillment, yet others process credit cards and other financial instruments. These companies often subcontract with each other, creating a thriving, competitive market in which entities have intellectual property of different types and enter into arrangements with different customers and vendors. The companies often have exclusive arrangements so that an advertisement will only generate leads to the originator. In many cases, competitors use the resources of other companies (e.g., image servers) without permission, or collect contractually exclusive leads from an inserted advertisement and resell them to their customers.

In the case discussed in this paper [10], the plaintiff asserted that 12,576 email messages were sent by the defendant to the plaintiff in violation of statute [3] and requested damages of about $16 million. The case was eventually ruled in favor of the defendant. Our analysis, while covering both sides, ultimately represents the defendant's perspective more than the plaintiff's perspective. For pedagogic reasons, techniques and results associated with other cases are included without distinguishing them.

2. Challenges

The complexity of Internet business operations complicates the efforts of the plaintiff and defendant. It is often hard to attribute actions to actors, but this is necessary to win a case. Differentiating what came from where, whether images used were actually part of a particular collection, whether a party was making unauthorized use of a competitor's image server, whether emails were in fact from the company whose image server was used, attributing multiple emails to one source when they come from many different addresses and have differing content, and other similar challenges can be daunting. Even the associations of domain ownership, domain names and IP addresses are often complicated by the large numbers of domains, addresses and content, the high rate of change of this information over time, and the lack of timely lookup of relevant information. Furthermore, opponents are not typically cooperative; they obfuscate whenever feasible; sometimes they refuse to answer questions or do not provide documents upon request; they do not retain adequate records or may intentionally destroy records.

Large volumes render the detailed examination of each email sent by an individual much too time consuming for the legal calendar to sustain. It is common for a few CD-ROMs of new evidence to be proffered within a few days of an expert report deadline, or a day or two before a deposition involving the individual identified as knowledgeable about the content. Evidence also commonly includes content that, upon inspection, leads to additional sources of evidence that have to be identified and sought. From a tactical standpoint, this evidence is sometimes provided in an obscure form and as a small part of a large collection of other content, perhaps as a scanned printout of an extraction of a log file included in tens of thousands of pages of other material.

These and other challenges point to the need for tools that can automate many aspects of analysis while supporting interpretation by the expert in a timely and accurate manner. Furthermore, it is important to be able to apply and modify these tools as new information appears.

3. Tools and Techniques

The most common tools used for analysis are small programs involving Perl scripts, shell scripts and Unix commands such as `grep` and `awk`.

3.1 Application of Common Tools

Using common tools presents certain tradeoffs. Writing or modifying scripts on short notice can lead to difficulty in verifying their operation. Off-by-one errors, misses and makes are commonplace [1]. For example, if a directory contains a set of files corresponding to what is purported to be one email per file and the goal is to find the number of files containing some critical content element, a typical script might be:

```
grep "critical content element" * | wc
```

Two problems with this script are: (i) multiple instances of the string on different lines in one file cause a miscount of the number of files containing the content; and (ii) the occurrence of more than one instance on a single line cause an undercount of the number of instances in the collection. With thousands of emails, a count of 7,543 that is off by one is hardly substantial, unless the email left out is unique in some manner. But offering the wrong count may produce a challenge from the other side and may degrade the standing of the expert and the quality of the report. Several approaches are available to deal with these sorts of errors. The most important step is to clearly define the objective of the analytical process and to properly report the results.

3.2 Issues of Legal Definition

In one case [10], a key issue was the number of applicable emails. The plaintiff asserted that there were 12,576 "emails," but the evidence provided contained 1,421 "actual emails," i.e., sequences of bytes of the proper format from the proffered file corresponding to what a user would consider an email [8]. The legal definition in this case counted email messages once for each recipient, leading to multiple counts of a single email. Even so, this definition did not clarify how the 1,421 actual email messages became the 12,576 emails asserted by the plaintiff.

3.3 Date and Time

Another issue was the relevant dates for the suit [10]. Because of statutes of limitations, effective dates of laws and legal filing dates, authoritative dates and times of events can be very important. Dates and times in emails depend on several factors. The content of proffered emails may not be trustworthy; dates and times stamped by computers may differ from real-world dates and times; and because time passes as email is in transit, an email sent before a deadline can arrive after it.

An anchor email was used in the case to rehabilitate dates and times [10]. This leveraged the fact that the plaintiff's emails were handled by the vendor Postini, which put date and time stamps on the emails in transit. While the collection of emails may have been forgeries, the assumption that they were not led to the use of the Postini date and time stamps as anchors. Independent contemporaneous emails were used to independently validate Postini date and time stamps. These emails, which had known date and time characteristics, were exchanged between systems under the control of the experts and went through the same Postini servers during the period in question. This reconciliation of date and time information excluded all but 242 of the emails in the case.

There are clearly other date and time issues related to email messages. One of the bases for legal claims stems from damages due to the reduction in available bandwidth, storage, CPU time or other resources. Evidence of damage must be in tangible form and, unless detailed records

Table 1. Extracted email arrival and delay times.

Arrival Time	**Delay Time**
06/27/02 07:33 AM	+0000-00-00 00:00:02
06/27/02 07:53 AM	+0000-00-00 00:00:06
06/27/02 09:11 AM	+0000-00-00 00:00:04
06/27/02 11:55 AM	–0000-00-00 00:00:03
06/27/02 02:41 PM	+0000-00-01 21:24:25
06/27/02 06:23 PM	+0000-00-01 13:06:42
06/27/02 08:12 PM	+0000-00-01 20:16:02
06/27/02 08:24 PM	+0000-00-01 13:09:01
06/27/02 09:12 PM	+0000-00-02 01:12:32

are kept, this is hard to show. One way to demonstrate damage is to analyze `Received:` headers of emails that show arrival times at servers [6]. The fundamental task is to demonstrate a correlation between these times and email volumes.

The analysis of `Received:` headers is complicated by the use of multiple time zones and time differentials between computer date and time settings. The approach used in [10] was to recast all dates and times in Universal Coordinated Time (UTC) and then examine time differences from hop to hop, where each "hop" corresponded to the `Received:` time stamp of a computer in the processing sequence. This involved: (i) parsing all the `Received:` headers; (ii) normalizing times to UTC; (iii) determining the distance (in hops) from final arrival point for each header; (iv) correlating paths through the email system so that comparable paths are compared with each other and not with other paths; (v) identifying time differentials by hops for common paths as a function of time; and (vi) relating these time differences to email volumes.

Despite the analysis, some puzzling outcomes were encountered. Some emails traveling along certain paths were delayed by days whereas other emails of similar size and content and sent along the same paths either earlier or later arrived within seconds. No crashes or other disruptions during the same time frames occurred to explain the anomalies, and they remain unexplained to this day. An inverse relationship between volumes of emails and delivery times can lead to the conclusion that these emails actually improved system performance. But this is ridiculous because correlation is not causality.

Table 1 presents email arrival times and delays. Each row corresponds to a different email message and the arrival times are sequenced in chronological order. The time delay is the interval between the first

Table 2. Number of emails arriving at different hops by date.

Date	Hop 1	Hop 2	Hop 3	Hop 4	Hop 5
10/01/03	4	4	3	2	0
10/02/03	9	9	9	9	0
10/03/03	8	8	8	8	0
10/04/03	6	6	6	6	0
10/05/03	11	11	10	10	0
10/06/03	11	9	8	7	0
10/07/03	23	20	19	18	1
10/08/03	11	11	11	11	0
10/09/03	12	9	6	6	0

arrival at the plaintiff's servers and the final internal delivery. Note that the emails with delivery times in excess of one day (non-zero `yyyy-mm-dd` value for the delay time) arrived before and were delivered after those processed and delivered within seconds of arrival. The emails were unexceptional in size, makeup and content. This appears to refute claims that emails were delayed by high volume.

Table 2 shows the number of arrivals at different hops in the plaintiff's infrastructure on different days (Hop 1 is the final hop). While some emails could arrive just before midnight and be delivered early the next morning or could be at different distances from their final destinations, in this case, none fit this pattern. All the emails had at least three internal hops before delivery and their times were consistently within a few seconds. Detailed examination showed that some of the excess emails had long delays and others were duplicates generated by the plaintiff.

3.4 Deliverability of Emails

Emails asserted must be "deliverable" in that there must be a user who can actually receive them. Some plaintiffs configure systems to accept any and all SMTP sequences, causing them to receive misdirected emails, emails to nonexistent users or emails to cancelled accounts. This is problematic because it may constitute interception of private communications, which is illegal in some jurisdictions and may be in violation of policies and/or contracts.

Common legal interpretation is that such actions invite the emails and, therefore, cannot be the basis for claims associated with undesired email transmission. SMTP refuses emails to recipients that do not exist without allowing the server to enter a state where data (header or body) can be received. Any receipt for non-existing users may constitute an invitation. In [10], there were 133 invited emails that could not have been delivered, leaving only 109 actual emails to be considered. In other cases, tens of thousands of emails have been similarly excluded, and courts have ruled that the activity was designed to generate law suits and was not the intent of the statures.

Demonstrating this fact involves discovering user identities. This is done from lists of user names in password files, server logs, configuration files associated with remote access servers, and document requests. In [10], RAS server logs, password files and other discovery led to this information, which had to be correlated with the emails in time to determine when the identities were valid.

3.5 Detecting Duplicates and Near-Duplicates

Identifying the cause of duplication is necessary for a plaintiff to assert the authenticity of records. Also, it enables a defendant to assert that records are spoliated. In [10], eleven actual email messages were duplicates that were somehow produced by the plaintiff's processing. In other cases, many thousands of duplicates have been identified.

Duplicates appear in many forms. The most obvious is an exact copy of an email sequence including headers, body and separator; in this case, a byte-by-byte comparison yields an exact match. The analysis can be performed by computing a hash value for every sequence, sorting the values to identify duplicates and verifying the duplicates via byte-by-byte comparison.

In other cases, only parts of email sequences are identical, such as delivery information, message identifiers, Message-ID fields, dates and times, and the rest of the headers and bodies. These duplicates are problematic for the plaintiff because they may indicate evidence spoliation. Examples of observed matches indicative of spoliation include:

- Identical sequences except for the `From` separator in `mbox` files [8].
- Identical sequences except that they indicate additional `Received:` headers.
- Identical sequences except for different date and time stamps on otherwise identical `Received:` headers.
- Emails with identical `From` separators but different headers and bodies.
- Emails with identical headers by different content.
- Emails with indicators of cut and paste operations used by web browsers supposedly sent by automated email mechanisms.

- Emails containing content indicative of being processed by receivers such as the systematic addition of content to email bodies of different formats from different sources.
- Emails with identifiers in headers showing different sourcing than other emails supposedly from identical sources.

These and other indicators of spoliation have been seen in actual cases. However, detection is problematic in large volumes of email because human interpretation is often required to make determinations about legitimacy.

The general detection approach is to create matching software that performs imperfect matches of portions of evidence. While matching one item to a large set of other items is straightforward, matching each of n items to each other item requires $O(n^2)$ time. Other methods trade off space for time: using hash values requires $O(n)$ time for hash generation followed by $O(log\ n)$ time for sorting, resulting in $O(n\ log\ n)$ time. This may have to be done for each of a large number of different types of matches. For example, if the removal of each line in the headers is to be considered, this comes to the average length of headers multiplied by the previous time. If altered header lines are to be considered, then this has to be broken down further.

A more comprehensive approach looks at evidence in terms of sequences of words or other symbols starting with length one and going up to some maximum sequence length. Each sequence can then be given a unique number and all components (emails, headers or other subsequences) with equal numbers are determined to match to the specified level of similarity. The analyst then has to interpret the meaning of these matches. Unfortunately, this approach leads to very large numbers of matches, and the analyst must again find a way to explore only subsets of the matches in order to keep time and costs down.

One of the most effective approaches is for a human to perform rapid comparisons of sequences of components. Similar components can then be compared in more detail and any obvious mismatches excluded.

3.6 Grouping Extracts for Comparative Analysis

Another approach for detecting anomalies in analysis and interpretation is to create an error model and look for identified error types. One class of error types and method for grouping emails is by the structure of headers. While header lines are largely unstructured, they normally begin with a sequence of characters followed by a colon (:) and continue on lines that start with whitespace [6]. A simple parser can add the lines starting with whitespace to the previous lines starting with a header identifier to allow line-by-line parsing, which makes parsing and analysis easier to program.

The extraction of email sequences from an `mbox` file produces a set of sequences ("extracts"). Extraction of header lines from extracts and identification by extract, line in extract and header identifier allows a wide range of analytical techniques to be applied with relative ease. Using disjunctions, conjunctions and other similar operations allows easy analysis such as the detection of all emails with no `cc:` field containing a particular domain name. The analyst can then use these operations on emails to find similarities and differences relevant to the case at hand.

In [10], it was important to identify emails containing particular IP address ranges in the `Received:` headers as recorded by the plaintiff's computers. Extracting this data is non-trivial because `Received:` lines have a non-standard format. However, once a parser is designed for the particular header lines of the mail transfer agent (MTA) software in use, the IP addresses can be associated with extracts, lines in extracts and the `by` portion of `Received:` headers to include only the desired extracts. Claims regarding these extracts could then be analyzed by customizing other analysis program snippets for the specific claims.

Even in cases involving more than 100,000 emails, the separation of email extracts by headers and the analysis of each header can be completed in a day or two. This tends to yield a great deal of information about emails. Simple sorting of headers rapidly yields information about similarities and differences. Types of information that can be detected include:

- Headers that are misspelled or otherwise differ from normal expectations.
- Associations of emails to other emails based on unique header fields or other header content.
- Sequencing information about the infrastructures involved in email transport.
- Details of the protocols, MTAs, hardware and software involved.
- Attribution information associated with unique identifiers.
- Groups of emails apparently sent from, through or by the same or similar MTAs, systems and mechanisms.

The following is a tree depiction of an email handling process (with details intentionally obfuscated). Such a tree can be generated by analyzing `Received:` headers and may include a variety of subfields depending on analyst needs.

```
0 325802 B.net
    1 325090 mail.R.com
        2 325090 mail.R.com
            3 215585 mail.H.com
                4 232   mail.R.com
                4 24    other.H.com
            3 109301 other.H.com
                4 5     mail.R.com ...
```

In the tree above, almost all the emails (325,090 of 325,802) arriving at B.net arrived through mail.R.com, all of which came again through mail.R.com, most of which came through mail.H.com with most of the remainder coming through other.H.com. In this case, a close relationship exists between B.net, R.com and H.com. Interestingly, some emails originally arriving at mail.R.com go through mail.H.com, and back to mail.R.com before being delivered to B.net. Depending on where the various servers are located, this looping between providers may be evidence of intentional forwarding of emails through multiple jurisdictions to add damages to the legal action.

Given the definitions used by the plaintiff's expert in [10], the total number of emails that could be the issue came to only 175 out of the original asserted claim of 12,576. These included 98 actual emails combined with the 34 unique active recipient addresses that were potential recipients of the actual emails. This analysis alone reduced the potential damages from more than $10 million to less than $200,000. But, as we discuss below, this was not the end of the issue.

3.7 Signups, Invitations and Other Causes

Another substantial limit on UCE cases is that laws tend to exonerate defendants who fulfill user requests or email users with a pre-existing relationship. A user who requests information on a web site may intentionally or inadvertently agree to terms and conditions granting the right to send or cause to be sent email that would otherwise be categorized as UCE.

When large volumes of emails are involved, it may become problematic for the plaintiff to prove that the emails were not solicited or that the addressees requested cessation of the emails. The plaintiff presumably has to show that the emails in question were unsolicited, which requires the presentation of legal documents signed by the individual recipients or proffering evidence associated with things like customer complaints to make the case. For hundreds of thousands of emails sent to hundreds or thousands of recipients, this would involve an enormous quantity of paperwork and would cause disturbance to numerous customers.

In practice, high volume UCE cases usually involve a small number of individuals who act on their own or assert their role as email service providers to sue defendants repeatedly. They try various methods to enable them to assert large numbers of emails, such as taking over the accounts of previous users, allowing emails directed to other recipients to be sent to them, making copies of emails sent to users and resending the duplicates to themselves, and forwarding emails through multiple jurisdictions to create additional penalties. These tricks may work in cases that are poorly defended, and some plaintiffs have won high-valued default judgments against defendants. One such defendant lost suits totaling more than one billion dollars; but the owners closed their business and left the country years ago (after apparently committing fraudulent activities). It seems unlikely that the judgments will ever yield real compensation. When plaintiffs create the conditions associated with high volumes of UCE, defendants are often able to show that the emails were invited. For example, when email identified as "spam" is sent, resent or forwarded, this may constitute an invitation and the forwarder may be liable for the violation.

From a technical standpoint, showing that protocols would not or could not have sent the emails unless the plaintiff acted to enable them may be adequate. Plaintiffs have an obligation to mitigate damages as well. For example, configuring email servers to allow all emails (not just to known users) to be sent or forwarding to a "dummy" account are clear indications of inviting emails. Showing the technical basis for such claims typically means examining log files, configurations and ancillary information, testing configurations in reconstructions with data from the case, and showing that the configurations produce the results at issue. In some cases, only screenshot images are available that are purported to depict the configuration of a product. Lacking version information and other relevant material, the analyst must ultimately make assumptions and draw conclusions based on the assumptions. But the evidence can often help. For example, in one instance, a configuration screen clearly indicated that the MTA was configured to forward emails identified as "spam" to a third party under a different email address. This party intentionally and knowingly sent the emails in question to the plaintiff, making him potentially liable.

When emails to users are copied, potential liability arises based on contracts with users or privacy regulations. To the extent that plaintiffs do not retain such information or fail to produce it, preservation has been held to be required as soon as a plaintiff is aware of the potential for legal action [2]. While there is a duty for UCE mailers in the United States to retain information on removal requests and not send additional emails, signups are typically held as proprietary information by vendors involved in different aspects of the business. For example, an advertiser almost certainly would not have information about the individuals who receive its advertisements and can only process removals by providing them to the solicitor. Contracts between advertisers and UCE mailers typically provide for the timely removal of users from mailing lists and include requirements for following applicable laws and regulations. This presumably limits the liability of advertisers, but laws vary on how this liability may apply to entities who order the insertion of advertisements.

Other causes may be asserted for individual emails. In one instance, an email was shown to have been received a second time six months after it was originally delivered. This makes the case for examining system logs and information on system failures, crashes, reboots, break-ins and other events that might cause delays. The example mentioned above appears to be the result of a restoration from old backups after a crash.

3.8 Compliance with Internet RFCs

In many cases involving high volumes of UCE, plaintiffs have claimed that the headers were false or misleading based on their compliance or non-compliance with Internet RFCs [6]. As of this time, courts have not ruled that RFCs constitute legal contracts or are enforceable in a legal sense. Nevertheless, claims are made with regard to RFCs in many cases and expert witnesses are called upon to testify with regard to RFCs, their interpretations and the extent to which they may have been violated.

- **HELO Lines:** One of the most common assertions made by plaintiffs in these cases is that a "HELO" indicates a fraudulent source. The HELO exchange is used in the initiation of an SMTP (RFC 821 [6]) exchange in which the sending computer is supposed to send HELO followed by a string. The HELO information is sometimes recorded in a `Received` line associated with that hop in the delivery process. RFC 2821, the updated version of SMTP, uses "EHLO" instead of HELO to initiate its processing, indicating to the receiving server that the RFC 2821 protocol applies. RFC 2821 indicates that in cases when the HELO protocol is used, RFC 821 must be used to process email.

  Most of the emails we have seen in bulk email cases conform with RFC 821 instead of RFC 2821. RFC 821 specifies domain names such as `localhost`, but it does not assert any requirement of authenticity. Email recipients never see the HELO lines sent to SMTP servers unless they examine log files associated with the email. Moreover, the recording of HELO information is neither mandatory nor is it intended for users. The normal presentation of an email does not include the area that contains HELO information. Many commonly used email clients have versions that send the name of the receiving (instead of the sending) computer in the HELO line apparently because the authors misread the RFCs.

Filtering based on the HELO information is sometimes used to prevent emails from known undesirable source domains. Some MTAs check the IP address against the domain name using a DNS number-to-name lookup and place a warning in the header to notify spam filters of a mismatch. However it is extremely common for DNS names and IP addresses to not match in a number-to-name lookup for several reasons, including when (i) large numbers of domain names are associated with a single IP address; (ii) proxy servers are used for delivering email; (iii) email delivery services are used for delivering email; (iv) servers are named incorrectly during configuration; (v) default server names are not updated during a configuration; and (vi) emails are sent from mobile locations (e.g., coffee shop, bookstore or hotel room). Dynamic DNS introduces additional complications and multiple answers to name-to-address lookups are not compensated for in many reverse lookup approaches.

- **False Sender Identities:** Another common claim by plaintiffs is that the use of a fictitious name or email address in a sender identity (e.g., `From:` line) is deceptive. Some plaintiffs have claimed that the use of an email address not containing the name of the sender is fraudulent because it misleads the recipient into believing that the sender is someone he is not. While this might appear to be a cogent argument, Internet systems often use fictitious names and pseudonyms, including names like `accounting` in RFC 821 and a wide variety of other sender names in emails from almost any company that can be identified.

  In most of the cases where this claim has been made, the plaintiff also uses false names as do the plaintiff's providers and customers, making the claim that much more problematic. However, the issue is not all that clear in law. There is a real possibility that some court will eventually rule differently, making pseudonyms and anonymized names problematic as well.

  Experts called to testify about Internet conventions and other common usage may examine the use of naming by the plaintiff and defendant, their ISPs, other providers, supply chain entities and government agencies, including the court of competent jurisdiction. We have found that it is a cogent argument to demonstrate that the court making the ruling does the very thing the plaintiff claims to be fraudulent. While some may decide to give an opinion about a fictitious name being misleading, this is problematic. Unless the digital forensic expert is also an expert in linguistics, he risks having his credibility destroyed along with the rest of his testimony.

Forged sender identities may be identified by an expert so long as there is a basis for showing that the user identity was used without the permission of the real person. The potential for forgery already exists, for example, when senders claim to be the recipient or use the identity of a different individual. A recent case [5] involved claims of more than 2,000 Usenet postings using a U.S. Chess Federation board member's name to discredit him and gain his board position. Credibly tracing these to the sources then becomes the issue.

- **False Received Headers:** Emails are sometimes sent with forged `Received:` headers to mislead recipients who attempt to trace email. These are problematic in individual email cases when forgers use realistic sequences. But such forgeries are not as trivial as they may appear, especially in volume cases because of timing and consistency problems with forgeries and the use of legal means to obtain records from other sites. Usually such forgeries involve a common intermediary associated with many other reception sequences, which are easily detected when presented in a tree format as shown in Section 3.6.

3.9 Inconsistencies

The examination of subject lines for deceptive content typically requires a linguistics expert. However, technical analysis has shown inconsistency in claims in high volume email cases. Claims typically require the explicit identification of specific statutory violations associated with each asserted email. Since many `Subject:` lines may be identical or nearly identical, analysis for consistency may reveal weaknesses in the plaintiff's claims. In a recent case involving more than 10,000 emails, approximately 30% of the claims about `Subject:` headers were inconsistently made, and the plaintiff lost a summary judgment. Inconsistencies in claims also goes to other issues and should be examined in high volume cases using automated techniques.

3.10 Assessment of Damages

Experts may also be asked to assess damages under trespass laws. Damages in such cases involve physical damage, deprivation, conversion or lost value or rights. In high volume email cases, only deprivation typically applies, and only to the extent that the plaintiff can prove quantified, time-framed, tangible, unmitigatable damage caused by uninvited messages by the defendant [9]. To date, plaintiffs in high volume email cases have largely failed to produce such proof.

3.11 Tracing Emails

To demonstrate causality, it is often necessary to trace emails to their origins. There are three common approaches for determining causality. One is to go from the destination back to the source step by step using subpoenas and gather evidence along the way. This approach has proved to be successful when applied properly.

The second approach takes a shortcut to the origin. In [10], the plaintiff responded deceptively to a UCE by providing a false lead to the seller. When the seller responded to the lead, plaintiff accused the seller of causing UCE to be sent and used this action to conduct a trace from the seller onward. This strategy might have worked if the plaintiff had not lied about the response, which brought up the counterclaim of unclean hands and whether the forward trace had yielded only a single sending chain. In [10], this evidence helped clear the defendant because, as it turned out, a fraudulent intermediary had violated exclusive lead generation contracts by selling the generated leads to many advertising agencies and was, thus, not acting on behalf of the defendant. Problems with this approach are (i) it may not produce a unique sender because of lead sharing; (ii) the entity sending the advertisement may not be the entity who "benefits" from it; (iii) care must be taken to ensure that the process is properly recorded; and (iv) just because one email produces this behavior does not mean that others will produce the same result.

The third approach is to use information in the bodies of emails. This typically involves the assertion that a URL contained within an email is used by a defendant in their business to track or display advertisements. If this is relied upon by the defendant, then the theory is that it should be an adequate record to show that the defendant caused the email to be sent. Problems with this approach include (i) competitors can and regularly do use "image servers" of others in their businesses so that other companies pay for the space, artwork and bandwidth while they gain the financial advantages; (ii) a malicious actor could provide the information for the purpose of damaging the defendant's reputation;

and (iii) someone else could use the URLs for any purpose, including for falsifying the records to create a legal action. Other sorts of commonalities have similar problems, but some success may be gained by using this information in conjunction with the first approach.

4. Conclusions

Making a case against bulk email senders involves most of the same elements one would use in any legal case involving digital evidence. However, challenges to digital evidence in the larger sense [1] must be met in order to make a case against a competent defense. Key factors that differentiate bulk email cases from other matters are: (i) the evidence must be explored using automation and any automated techniques used must meet legal standards; (ii) contemporaneous records should be properly identified, collected and preserved to obtain the evidence necessary to prove the case; and (iii) increased care should be taken because small mistakes tend to get amplified by volume. Poorly constructed cases, exaggerated claims, spoliated evidence and large volumes of invited emails are likely to be detected by a competent defense, especially in cases involving large monetary claims. In the case [10] discussed in this paper, the defendant won a summary judgment. While digital evidence played a substantial role in the decision, as always, the evidence and analysis are applicable in the context of the specific case. Nevertheless, the techniques may be applied to a variety of bulk email cases.

References

[1] F. Cohen, *Challenges to Digital Forensic Evidence*, ASP Press, Livermore, California, 2008.

[2] C. Crowley and S. Harris (Eds.), The Sedona Conference Glossary: E-Discovery and Digital Information Management, The Sedona Conference, Sedona, Arizona, 2007.

[3] Government of California, Article 18: Restrictions on Unsolicited Commercial E-Mail Advertisers, *West's Annotated California Codes (Business and Professions Code)*, §17500 to §18999.99, pp. 101–117, 2008.

[4] Government of Maryland, Definitions, Commercial Electronic Mail (Subtitle 30), *Michie's Annotated Code of the Public Laws of Maryland (Commercial Law)*, pp. 476–477, 2005.

[5] D. McClain, Member of U.S. Chess Federation's board is asked to resign in dispute over an election, *New York Times*, January 15, 2008.

[6] J. Postel, RFC 821: Simple Mail Transfer Protocol, Information Sciences Institute, University of Southern California, Marina del Rey, California (tools.ietf.org/html/rfc821), 1982.

[7] SpamLinks, Anti-spam laws (spamlinks.net/legal-laws.htm).

[8] Sun Microsystems, `mbox`, Manual pages from `/var/qmail` (version 1.01), Santa Clara, California (www.qmail.org/qmail-manual-html /man5/mbox.html).

[9] Supreme Court of California, Intel Corporation v. Hamidi, *West's Pacific Reporter (Third Series)*, vol. 71, pp. 296–332, 2003.

[10] U.S. District Court (Northern District of California), ASIS Internet Services v. Optin Global, Inc., Case No. C-05-5124 JCS, December 17, 2008.

[11] U.S. Government, Controlling the Assault of Non-Solicited Pornography and Marketing Act, Public Law 108–187, 108th Congress, *United States Statutes at Large*, vol. 117(3), pp. 2699–2719, 2004.