International Journal of Cyber Criminology
Vol 5 Issue 1 January - July 2011

Email spam and the CAN-SPAM Act: A qualitative analysis

Szde Yu[1]
The College at Brockport, State University of New York, Brockport, USA

Abstract
Email spam is a rampant activity in cyberspace. This study performed a qualitative forensic analysis on 3,983 spam emails with respect to their content, format, techniques, and their compliance with the CAN-SPAM Act. The findings suggested spammers show little interest in complying with the CAN-SPAM Act, and different purposes of spam determine the format and the techniques spammers use. The rationales behind these spam choices were discussed. Legal and research implications were suggested.

Keywords: Email spam; CAN-SPAM; Technique; Content; Format.

Introduction

Email spam could be one of the most prevalent crimes in the sense that almost every email user probably has received at least a few unsolicited commercial emails at some point of time. The motivation of spam usually involves revenue generation, higher search ranking, promoting products and services, stealing information, and phishing (Hayati & Potdar, 2008). Spam could result in damaging impact on the economy. Some research has claimed spam accounted for nearly 20 billion dollars in lost time and productivity (Yeargain, et al., 2004). Furthermore, McAfee's Carbon Footprint of Email Spam Report (2009) indicated that the energy used to transmit, process, and filter spam amounted to 33 billion kilowatt-hours (kWh), which is equivalent to the electricity used in 2.4 million households or the greenhouse gas emissions as 3.1 million cars using two billion gallons of gasoline.

Despite the effort trying to stop spam, unsolicited commercial emails never seem to stop arriving in our email inboxes. Many email providers, such as Gmail or Hotmail, have diligently developed their spam filters to detect possible spam emails. However, if the filter mistakenly identifies an important message as spam, it could create a problem more than just an annoyance as users might miss an important date or fail to follow up in a communication of great consequence (Weinstein, 2003). Besides, nearly 80% of the energy consumed as mentioned earlier was related to users deleting spam and searching for false positives. Spam filtering accounted for 16% of such energy use, but successful filtering was able to reduce the energy use incurred by spam otherwise. Therefore, accurate spam

[1] Assistant Professor, 161 Albert Brown Building, The College at Brockport, 350 New Campus Drive, Brockport, NY 14420, USA. Email: syu@brockport.edu, szdeyu@gmail.com

filtering saves not only time but also energy and money. Moreover, it should be kept in mind that spam is more than just an economic concern as in the USA it could constitute a criminal offense under the Controlling the Assault of Non-Solicited Pornography and Marketing Act (the CAN-SPAM Act).

This article was intended to perform a qualitative analysis on 3,983 spam emails with respect to their content, format, techniques, and their compliance with the CAN-SPAM Act.

Literature Review

The CAN-SPAM Act

In December 2003, the CAN-SPAM Act was enacted and took effect in January 2004 (Lee, 2005). The Act was enacted in an attempt to regulate interstate commerce by imposing penalties and limitations on sending unsolicited commercial email via the Internet (Yeargain et al., 2004). The Federal Trade Commission (FTC) was authorized to enforce provisions provided in the CAN-SPAM Act, and unsolicited commercial emails that fail to comply with its regulations were declared criminal. The punishment could be a fine up to $16,000 for each separate email in violation of the CAN-SPAM Act (FTC, 2009), or it could be imprisonment (Yeargain et al., 2004). In 2008, the so-called "Spam King" Robert Soloway was convicted under the CAN-SPAM Act for sending fraudulent emails along with two other charges and was sentenced to 47 months in federal prison (Rabinovitch, 2007). The Act also allows states and Internet service providers to file civil lawsuits against spammers (Ford, 2005; Yeargain et al., 2004).

The rationale of the CAN-SPAM Act can be seen as based on the expected utility theory in economics, which posits spammers would only choose to send spam when the expected gains exceed the expected cost (Lee, 2005). Since the Act was aimed to increase the cost to spammers, it was expected spammers would be discouraged. However, this did not seem to be the result. According to Symantec's monthly spam reports, the amount of spam is actually increasing (Symantec, 2010). More than 90% of emails sent in the world now are possibly spam, depending on how spam is defined (Symantec, 2010).

The CAN-SPAM Act defines spam as unsolicited commercial electronic mail that includes any commercial emails addressed to a recipient with whom the sender has no existing business or personal relationship and not sent with the consent of the recipient, and commercial electronic mail is defined as any electronic mail message the primary purpose of which is commercial advertisement or promotion of products or service (Rogers, 2006). Commercial emails could be considered legitimate unless they violate certain provisions. Emails would be deemed as spam when they were sent in bulk without the recipient's consent and the primary purpose must be commercial in nature (Rogers, 2006), but emails of this nature are not necessarily illegal. The main requirements for a commercial email to be legal include using authentic header information, no deceptive subject lines, identifying the message as an advertisement, providing the real physical location of the business, offering an opt-out choice, and honoring opt-out requests within 10 business days (Federal Trade Commission, 2009). It is also punishable if the emails were knowingly sent through a third party without authorization, or if the emails were sent randomly to an email list that was obtained illegally (e.g. stealing from other online proprietary service or collecting automatedly from websites) (Spammer X, 2004; Yeargain et al., 2004; Ford, 2005). The sender cannot sell or lease the opt-out email addresses and should specify in the subject line any sexually explicit content (Yeargain et al., 2004).

By specifying the rules, the CAN-SPAM Act has been criticized for actually providing guidance for spammers to keep sending more unsolicited commercial emails as long as they follow the rules correctly (Ford, 2005; Yeargain et al., 2004; Weinstein, 2003). Following the rules does not mean spam would become benign. For example, many commercial emails now provide an opt-out option indeed, but some spammers merely use it as indicators for active accounts because now they not only know your email account is in use but also realize you are one of the few who do read their messages, which will very likely to result in more spam messages to come, either from the same or different senders (SpammerX, 2004). Another issue concerning the effectiveness of the CAN-SPAM Act is law enforcement. Because of a preemption clause in the Act, some more rigorous regulations in the state laws may not be enforceable or applicable anymore (Ford, 2005; Maggs, 2006; Yeargain et al., 2004). Also, just like any other cybercrimes, jurisdictional issues are always an obstacle. Even if we assume spammers can be identified, the Act may not apply if they are operating overseas (Maggs, 2006). In his study, Kigerl (2009) found that the CAN-SPAM Act does not have any significant impact on the behavior of spammers. Using an interrupted time series design, Kigerl found the Act has no observable impact on the amount of spam sent and no effect on compliance with the laws.

Spam Techniques

Aside from the law, the bigger concern spammers have probably been regarding how to circumvent the spam filters. As mentioned earlier, while a good spam filter can stop junk emails from coming into our inboxes and thus save time deleting or even reading them, a poor filter can result in too many false positives and become even more problematic (Weinstein, 2003). In light of this, spammers know there must be a way to beat the filter as long as the emails look legitimate. To fool filters, they first need to know how filters recognize spam.

The reliability of spam filters depends on the detection methods they employ. The main methods of spam detection include host-based filtering, rule-cased filtering, statistical algorithms, and white lists (Spammer X, 2004). Host-based filtering is aimed to identify servers that are used to send bulk emails and construct a blacklist for these servers. Rule-based filtering is performed by analyzing each email to determine the likelihood of being spam according to the specified rules. The rules could be defined by the user so as to adapt to the variation of spam. For example, one rule can be set up to analyze the body of the message to detect keywords that indicate pornography. Some statistical algorithms are also used to identify spam. These statistical methods, such as Bayesian filters, typically rely on the previous spam messages to calculate a statistical probability of new emails being spam. Moreover, some more modern forms of algorithms have been developed to adopt user behavior analysis (Stolfo et al., 2010), improve machine learning methods (Lai, 2007; Delany et al., 2004), and replace the currently used Simple Mail Transfer Protocol (SMTP) with Differential Mail Transfer Protocol (DMTP) (Duan et al., 2007) in an attempt to better control spam. In addition to the various statistical methods, some email service providers would compile a so-called whitelist identifying legitimate email sending servers and block emails sent from other email servers unidentified on the list. Another way to train spam filters is through user labeling (i.e., collaborative spam filtering) (Attenberg et al., 2009), where users are allowed to personalize their filters by specifying which message is spam and which one is not.

To bypass IP-based filtering, some spammers have utilized legitimate servers to send their messages. For instance, Yahoo! Groups have been used by spammers to send spam messages because they know it is unlikely for Yahoo servers to be blocked given Yahoo's popularity and reputation. In Appendix 1, there is an actual spam email received from Yahoo! Groups. Although it originated from a genuine Yahoo server and the group really existed, the purpose was to promote a website that sells Asian pornography. Spammers can also switch IP addresses constantly to avoid landing on a blacklist or infect machines to serve as botnets which can be used to send spam without the owner's knowledge (Goodman, et al., 2007). Spammers also have tried to work around identity-based filtering in which Sender Policy Framework (SPF) and DomanKeys Identified Mail (DKIM) were used to verify the email was sent from a real host responsible for sending messages for the particular domain. To beat this, spammers could simply create a new domain for spamming purposes (Spammer X, 2004; Goodman et al., 2007). Spam filters may employ similarity-based matching to comparing known spam messages with new messages so as to recognize spam even when it successfully evades the above mentioned detection methods. In response, spammers would try to insert some random content to make each message look unique (SpammerX, 2004; Goodman et al., 2007). In Appendix 2, there are two messages that contain slightly different subjects, different senders, and seemingly different content, but the hyperlinks embedded in both messages were actually linked to the same website.

Replying on users to report spam is another common detection technique (Taylor, 2006). However, the problem with this method is spam can be defined rather subjectively and spammers could label their own messages as "not spam" to confuse the filters, so if the filters are not personalized they will be susceptible to bad labels (Delany et al., 2004; Attenberg et al., 2009). This puts the anti-spam burden on the users if the machine fails to adapt fast enough. Another technique commonly used by spammers is the attempt to avoid using keywords that suggest spam (SpammerX, 2004). For example, instead of using "viagra" they might intentionally misspell it as "vigara" or they could remove spaces between words (e.g. cheapestviagra) to trick the machine to believe that is a different word. Spammers also have tried to use images instead of text to deceive filters as it is harder for machines to analyze content when the message is hidden in an image even with the optical character recognition (OCR) techniques (Goodman, et al., 2005; Aradhye, et al., 2005; Blanzieri & Bryl, 2009; Dredze, et al., 2007).

Knowing spammers' slyness, email service providers usually combine multiple methods in identifying spam. For example, Gmail allows users to report spam by marking messages they receive as spam, or they can report false positives by marking a message as "not spam". In addition to its spam detection algorithm, Gmail has employed OCR to identify image spam. It also incorporates multiple authentication systems, including SPF and DKIM to verify the sender's true origin, as the sender's information contained in the header often is faked in spam emails (Gmail, 2010; Taylor, 2006).

Spam Content

Although the prevalence tends to change over time (Wall, 2004), the most common types of spam are often found to involve advertising general products, Internet websites or service promotion, financial offers, scams, drugs, and sexual content. (Yeargain et al., 2004; Symantac, 2010; Wall, 2004). Most often spam is used merely for advertising purposes, so the content is similar to legitimate electronic circulars that advertise products,

such as household items, electronic devices, store discounts, or coupons (Wall, 2004). Advertisers choose spam as the means for advertising probably because it costs less than conventional means (Kraut et al., 2005) or simply because the products they intend to sell are not suitable for conventional advertising. Illegal provision of prescription medication, questionable sources of non-prescription drugs, pornography, body enhancement remedies (e.g. penis enlargement or extreme weight loss), and pirated software are among the most advertised products in spam (Wall, 2004; Fogel & Shlivko, 2010; Symantec, 2010; Ting, 2004; Lueg, 2003). Even though most people might cast doubt on the effectiveness of these advertisements, research has suggested some people, albeit relatively few, are indeed purchasing products from these sources (Fogel & Shlivko, 2010; Fogel & Shlivko, 2009). Financial offers which are unlikely to be legitimate (e.g. work at home opportunities, low rate loans, and pyramid schemes) are also common (Symantec, 2010; Lueg, 2003). As another type of spam, scam could be especially harmful because the people who can easily ignore advertisements may find it hard to resist the incentives presented in the fraudulent emails. Scammers usually prey on greed and manage to accomplish their schemes through social engineering, such as asking for personal information first rather than financial information directly (Holt & Graves, 2007; Nhan, et al., 2009).

Built on the literature, this study was intended to qualitatively examine spam messages in an attempt to see how the content of spam is related to the techniques used by spammers, and how individually they violate the CAN-SPAM Act.

Methods

The current study analyzed 3,983 email messages that were identified as spam in five separate Gmail accounts. These email addresses were created purposefully for attracting spam. They were intentionally used for online registration on various websites and forums. They were also frequently left on some websites so that they would be easily found by automated software used by spammers to harvest email addresses. All five accounts have been in use since the year of 2003, but only the spam emails received from May 2010 to August 2010 were kept and analyzed. False positives were removed from the analysis.

The analysis was conducted on a daily basis, which means each message was analyzed either on the same day it arrived or the day after. The analysis consisted of three steps. First, the content and format of each message were examined. As suggested in the literature, spam can be serving multiple purposes. In this paper, according to the content, spam emails were categorized as follows:

1. Scam/fraud
    a. Lottery/award
    b. Business proposals
    c. Requests for favors
    d. Benefits/charity
    e. Financial offers (e.g. loans or credit cards)
    f. Phishing
2. Advertising/promotion
    a. Pornography
    b. Sex-related drugs/products
    c. Other drugs
    d. General products/commercial websites (nonsexual)
    e. Computer software

    f. Educational programs
    g. Other
3. Unclear purposes

In terms of format, spam emails were classified as follows:

1. Plain text
    a. Contact information
    b. No contact information
2. HTML
    a. Hyperlink
    b. No hyperlink
3. Image spam
    a. Redirected: by clicking on the image the user is redirected to another website
    b. Not redirected
4. Hybrid (Image + HTML)
5. No content

Next, the second step of the analysis was aimed to examine the techniques employed in each spam email. The techniques identified included the following:

1. Sender IP origin: a forensic tool eMailTraker Pro was used
2. Third party server: this was examined by comparing the domain name of the email address to the domain name of the sender server
3. Inserting random text: random text in the body or in the subject line could deceive spam filters to believe it is not bulk email
4. Using images instead of text: image spam could be used to avoid keyword search
5. Misspelling: spammers could intentionally misspell sensitive words

Finally, the legality of each email was also determined in accordance with the CAN-SPAM Act. The Act specifies several requirements for commercial emails to be legit and the current study tested the following rules insomuch as possible:

1. Opt-out choice: the recipient should be provided an option to opt out.
2. Non-deceptive subject: the subject line needs to clearly indicate its advertising nature and sexually explicit content needs to be marked as such
3. Honest header: the sender's information should not be deceiving
4. Physical address: the Act requires a valid physical address of the business
5. SPF: Gmail uses SPF validation to ensure the email was not sent via a third party machine without permission

Results

Content

Table 1 summarizes the content of the spam emails analyzed in this paper. Of the 3,983 emails, 909 of them were categorized as scam. The ruses commonly employed in this type of spam included claiming the recipient is the winner of a lottery, proposing a business partnership that promises financial income, asking for assistance in transferring funds, calling for donation in the name of charity, offering loans or credit cards without background check, and phishing. Phishing usually involved imposture of a legitimate website, a bank, or a government agency, inducing people to reveal personal information, such as bank accounts, user name and password, or social security number. Companies like

Ebay, Facebook, and PayPal appeared to be the impostors' favorite choices. They often intend to trick people to enter their username and password to those websites on a phishing site. Once they steal your login credentials, they could have access to your personal information (e.g. Facebook), financial information (e.g. Ebay), and even funds that have been deposited (e.g. PayPal). Considering how many users these websites have, there is a fair chance to trick the recipient to think the message is indeed from Facebook if he or she happens to be a Facebook user. Luckily, spam filters seem to be able to distinguish those fake ones for the most part.

Table 1. Content of Spam

Type of Scam	Sub-categories	No. of Emails (Total: 3983)	%
Scam/Fraud			
	Lottery/Award	251	6.30
	Business proposal	108	2.71
	Requests for favors	39	0.98
	Benefits/Charity	108	2.71
	Financial offers	225	5.65
	Phishing	178	4.47
	Total	909	22.82
Advertising/Promotion			
	Pornography	1188	29.83
	Sex-related drugs/products	995	24.98
	Other drugs	376	9.44
	General products/websites	247	6.20
	Computer software	49	1.23
	Educational programs	100	2.51
	Other	30	0.75
	Total	2985	74.94
Unclear purposes		89	2.23

Most of the scam emails were designed to solicit money from the recipients. In order to accomplish that, they almost always promise a large amount of money in return. For example, in one of the emails, as a reward for helping an African government transfer 20 billion dollars to a bank in the UK, the commission was promised to be 10% of the total amount. In another email, it claims that the recipient's email address was chosen to win 1,000,000,000 pounds sterling in a lottery. Only after the alleged winner responded did they ask you to pay advancement fees in order to begin the process. The amount of the advancement fees can vary from hundreds to tens of thousands or more. Obviously, scammers are hoping greed will cloud a rational person's judgment enough to think it is worth the risk. It would be even better if the recipient is not so rational to begin with.

Although scam is rather prevalent in spam, the majority of spam emails were meant for advertising (see Table 1). More than half of such spam emails contained sexual content. They either involved pornography or tried to promote some drugs or products that allegedly could enhance sexual performance. The message in Appendix 3 is merely one of the numerous examples. Online pharmacies also constituted a sizeable proportion of spam.

In this study, 9.44% of the spam was promoting drugs, other than sex-related drugs. For prescription drugs, they usually offered a price lower than regular pharmacies but it is doubtful where the drugs were obtained. Besides drugs, many small businesses use spam as cheap advertisement for promoting products ranging from household items to jewelry. Some advertisements were associated with some sort of educational programs that offer a degree. It is not yet clear whether these programs are legitimate or they are actually a form of scam, but a simple online search suggested the legitimacy of these programs largely was questionable at best. Low priced computer software was featured in 49 emails in this study. It is reasonable to assume the software on sale is not genuine. Some of them actually clearly indicated the software was "cracked". There were 89 spam emails that did not manifest a clear purpose. For example, many of them had no content. This could be a result of spammers' trial and error.

Format

In general, there were three basic formats identified in the spam emails collected from this paper (see Table 2). The message could be written in plain text or HyperText Markup Language (HTML). Some messages contained only images. Plain text emails are usually an unpopular choice for spammers because this format does not allow for many tricks to be used, unless the spammers resort to steganography (e.g. semagrams or open codes) or cryptography (EC-Council, 2010), but it is not sensible for spammers to encode their messages if the recipients cannot decipher them. In contrast, HTML renders more leeway for spammers to be creative. To say the least, in HTML format, spammers can embed links, track whether the link has been clicked on, insert images, and disguise the text or links to make them look innocent.

Table 2. Format of Spam

Format	Features	No. of Emails (Total: 3983)	%
Plain Text			
	Contact Info.	216	5.42
	No Contact Info.	82	2.06
	Total	298	7.48
HTML			
	Hyperlink	1881	47.22
	No Hyperlink	5	0.13
	Total	1886	47.35
Image Spam			
	Redirected	228	5.72
	Not Redirected	50	1.26
	Total	278	6.98
Hybrid (Image+HTML)		1485	37.28
No Content		36	0.90

In the current study, only 298 messages were written in plain text. Among them, 216 emails offered some form of contact information in the text. The contact information

provided in plain text emails was usually a web address or an email address. Occasionally, a phone number or instant messenger information would be provided, but this was rare. Most spam emails utilized HTML and almost all of them had at least one hyperlink embedded, which can redirect the recipient by one simple click. Technically, image spam could be seen as a form of HTML format, but in this paper it was separated on purpose to distinguish text-only spam and image-only spam. It is important to note, however, image-only spam does not mean there is no text. Rather, the text in image spam is hidden in the image files. Thus, the machine could be misled to think that is a photo, while the recipient can still clearly see the text message. The emails shown in Appendix 1, 2, and 3 are all an example of spam in HTML format. Appendix 5 is an example of plain text spam.

Appendix 4 shows an example of image-only spam. Most image spam would redirect the recipient to another website when the image is clicked on. When the image is large, it is very likely for a person to trigger the redirection by accident. In the current study, image-only spam (6.98%) was not as common as HTML text-only spam (47.35%) or spam in a hybrid format (37.28%). In the hybrid format, images and HTML text could both be found and they were always accompanied by hyperlinks that would redirect the recipient to other websites with one simple click. Image-only spam is less popular probably because many email service providers, such as Gmail, now would automatically disable the images when showing the message. This is true for non-spam emails as well, because this would allow the message to be loaded faster, and the recipient can decide if he or she wants the images to show by their own discretion.

The spam emails collected in this paper were not all in English. It should be noted that the spam received in each person's inbox has much to do with the person's online habit and language preferences. A multilingual person apparently would have a better chance to receive spam in multiple languages. Table 3 presents the languages of spam in this paper. English appeared to be the most common language in spam. Although this was due to the email user's language preferences, it could also in part be accounted for by the fact that most spam emails were sent from the United States (Symantec, 2010). Besides English, more than 30% of spam was in Chinese (including Simplified and Traditional Chinese), followed by Japanese. Together these three languages accounted for more than 95% of the spam emails analyzed in this paper.

Table 3 Language of Spam

Language	No. of Emails	%
English	2323	58.32
Traditional Chinese	743	18.65
Simplified Chinese	509	12.78
Japanese	228	5.72
Spanish	47	1.18
Korean	33	0.83
Russian	20	0.50
Italian	3	-
German	2	-
Unidentified	75	1.88
Total	3983	100

Techniques

Presumably, the techniques employed by spammers in the current study were not quite effective in bypassing the spam filter guarding Gmail accounts, since these spam emails had been categorized as spam already. False positives were excluded from analysis, as mentioned. False negatives did occur 11 times during the research period and they were reported as spam by the user immediately. Nonetheless, it did not seem like those 11 spam emails really defeated the filtering algorithms, because even though the message could pass filtering in one account, a seemingly identical message was usually found caught in other accounts. There was no evidence suggesting the false negatives were technically or fundamentally different.

As indicated in the literature, one popular technique used by spammers is to send bulk emails from a foreign server. This could help avoid detection since the server could be out of the jurisdiction. Table 4 shows the origin of spam in this paper. About 40% of spam came from North America, followed by Asia (28.04%) and Western Europe (18.33%). The top five countries were as follows: USA (33.09%), Taiwan (12.55%), China (7.98%), UK (7.43%), and Japan (5.05%).

Table 4. Origin of Spam

Origin	No. of Emails	%
North America	1608	40.37
South America	81	2.03
Western Europe	730	18.33
Eastern Europe (including Russia)	188	4.72
Africa	209	5.25
Asia	1117	28.04
Unidentified	50	1.26

Another spam technique is utilizing a third party server to distribute spam. This could help avoid detection, and this technique could also help bypass spam filters if the server being used is unlikely to be on the blacklist. To the extent possible, this paper looked up and compared the sender's email domain name with the sender's IP address. In 1,639 spam emails, it could not be determined whether the email was sent through a third party server. However, 2,021 messages were found to be sent from a server whose domain owner was different from the host of the email domain. For instance, in one email, the email domain appeared to be 'unimstores.com', which belongs to GoDaddy.com (an American domain registrar company), but the sender's server, according to the IP address, belongs to an Internet service company in Petersburg, Russia. It should be noted that even if it was a third party server, it does not necessarily mean it was illegal if the third party knowingly provided service. Nevertheless, the inconsistency between who claimed to be the sender and who actually sent the email usually indicates an attempt for deception, especially when it involves a foreign server.

Next, some spammers would insert random text in order to let each message appear unique even though it is indeed sent in bulk. For example, in Appendix 3, at the end of the subject line 'lxcy' was inserted. In another practically identical email 'xk87' was inserted at the end of the subject line. Sometimes the random code was added outside the

HTML tabs, so it was not visible unless the recipient examined the source code. In the current paper, this technique was found in 804 (20.19%) messages.

Other techniques identified included using image instead of text and using misspelling to obfuscate sensitive words. As discussed, image-only spam was seen in 278 emails and a mix of text and images could be found in 1,485 messages (see Table 2). As for purposeful misspelling, this technique was only analyzed in English spam. Of 2,328 spam emails in English, 539 (23.15%) were found to contain purposeful misspelling. In Appendix 3, that email contained several misspellings in the subject line as well as in the content. First, 'inche', 'Guaranteeed', and 'Incredib1e' all appear to be typos. Second, spaces were removed in several key phrases, such as 'yourPenis', 'Rock-HardErections', and 'NaturalPenisEnhancement'. It can be a combination of both, such as 'Intense0rgasns'. These errors may not impede a real person's understanding of the message being conveyed, but spammers obviously hope they could deceive spam filters. This technique was most likely to be employed in spam aimed to promote sex-related drugs or products.

CAN-SPAM Violations

The spam emails were further analyzed according to the legal regulations. First, it is required for commercial emails to provide an option for recipients to opt out and once they opt out, their email address should be removed from the mailing list within 10 days. In this study, only 268 spam emails offered an opt-out choice in the message. Even if they did, it would be difficult to know if that actually works. Second, the subject line should not be deceptive. It needs to clearly indicate its advertising nature. In this aspect, the majority of the spam emails actually met the requirement. Approximately 75.62% of the spam emails did not bother to hide their true intention. Although it was not always clear what they were trying to advertise, it was reasonable to assume that an average person should be able to realize that was a commercial email by reading the subject line. The major exception is when the email was meant for scam. Of course, no one would indicate this is a scam in the subject line. In this sense, they are always deceptive. If scam is the purpose, the message often appeared to be more personal than commercial. For example, in Appendix 5, the subject reads "Greeting in the name of God", and the context without a close look does seem like normal correspondence. Some fraudulent emails would directly declare you a winner of an award in the subject, knowing that most people would be intrigued by the possibility of monetary gains. In sum, while commercial advertisements generally do not bother to be deceptive on the subject line, fraudulent emails might be in disguise from the subject line to their ulterior intent.

Another legal requirement is to mark messages that contain sexually explicit content. Most sex-related spam would in the subject line make it obvious that it was about sexual performance or pornography, but there was never a clear warning about such adult content. In other words, an innocent person who does not know what 'CheapestViagra' or 'BetterEjacu1ation' means might still unwillingly stumble across sexual explicit content. On the other hand, those subjects could already be sexually explicit enough, which makes warning seem pointless.

Moreover, as many spammers intend to conceal their identities, they are prone to forge the header information associated with the spam emails. The sender's email address can be easily faked. A DNS validation was performed to test if the email server was active. It should be noted DNS validation can only test the validity of the server, but it does not guarantee the particular email address is valid. Even so, 1,076 emails (27.01%) failed the

DNS validation, which follows the email address that appeared as the sender did not exist because the server was not operational or even not existent. In this study, the analysis was done within 48 hours after the message arrived, so the likelihood that the server went down right after the message was sent is slim.

Besides fake email addresses, 557 emails (13.98%) certainly failed to meet the honest header requirement for other reasons. A number of them used 'me' as the sender's name; a few of them used the same address for both the sender and the recipient; some even had an 'unknown sender'. A further examination revealed that 2,018 spam emails (50.67%) were not honest about the recipient's email address as the one showing in the header was not the one that actually received that message. Spammers often Bcc: the recipients so that they do not need to reveal who is actually receiving the message and they can hide the fact that the message was sent in bulk, which results in the deceptive header information. Overall, dishonest header information was found in at least 2,119 emails.

In addition to honest header information, a physical address is also stipulated in the CAN-SPAM Act. In this paper, only 183 emails (4.59%) provided a physical address in the message, and 166 of them were a PO Box. The ones that did provide a physical address were mostly promoting general products unrelated to sex or drugs.

The Act also prohibits spam from being sent via a server without permission. In this regard, Gmail employs SPF validation. SPF records specify which servers are designated for sending emails for each domain. There were four SPF validation results observed in Gmail. A message can pass the SPF test or fail it. Very often in spam emails, the SPF result showed neutral, which means the server is neither permitted nor denied by the SPF records. In some rare cases, the validation could not be completed possibly due to non-response from the server. In this study, 701 emails passed, 297 failed, 2,984 were neutral, and 1 was incomplete. According to this result, the SPF validation employed by Gmail does not seem very decisive. It is not clear why most cases resulted in a neutral decision. One possible explanation is the incomprehensiveness of the SPF records. A neutral decision alone is not sufficient for declaring a message spam.

Eventually, of the 3,983 spam emails analyzed in this study, only 108 emails did not have apparent violation of the CAN-SPAM Act. If a PO Box is not deemed an acceptable physical address, then only five emails met all the regulations of the Act.

Discussion

In this article, 3,983 spam emails were analyzed. The majority of them were aimed at advertising or promotion. Sex appeared to be a dominant theme in this type of spam. This is consistent with prior studies (Fogel & Shlivco, 2009). While prior studies rarely examine the relation between the content and techniques, this study found sex-related spam also manifested most 'deviance' in that various techniques were commonly applied to avoid spam detection. For example, random text and intentional misspelling were constantly used in spam with sexual content. Emails of this kind almost inevitably violated the CAN-SPAM Act by failing to offer an opt-out choice and/or a physical address. They never showed warning about their sexual content while the subject line itself often was sexually explicit, which is a violation of the Act (Yeargain et al., 2004). They were also very likely to present deceptive header information, such as fake sender's or fake recipient's email address. Although HTML was their favorite format of choice, image spam was not usually seen in spam with sexual content. This was true no matter where the spam originated. Even when images were incorporated, the images used in sex-related spam were mostly

not sexually explicit themselves. It almost seemed like spammers tend to be cautious about the photos they use even though they are promoting sexual content.

As most spam involves advertising (Yeargain et al., 2004; Symantac, 2010; Wall, 2004), the present study took a closer look into the format of this type of spam. Image spam was most frequently seen in spam aimed at promoting other products or websites. The images always were accompanied by hyperlinks that would lead to the promoted websites directly. A combination of image and HTML was a popular format in spam except for a certain type of scam. Other studies rarely address how the purpose of scam can affect the format (Edelson, 2003; Kumaraguru et al., 2007), but the present study found as the scheme varies, the way scammers organized their messages also varies. For example, phishing emails relied heavily on HTML. In an attempt to lure the recipient to a phishing site, the message often appeared to be colorful and eye-catching so that it would be tempting enough to induce mouse-clicking (Dhamija et al., 2006). The obvious flaw in this type of scheme usually can be found in the web address. The phishing site's web address is unlikely to be the same as the genuine one. Therefore spammers tend to use colorful images or text, allowed by HTML, to disguise their links. On the other hand, when a fake business proposal (e.g. advance fee scam) was intended, the message would seem like a formal letter in plain text so that it would look more personal rather than commercial. Although it looks personal, it is usually not really personalized as they do not really know your personal information until you provide them (Chang, 2008).

In the present study, the spam techniques mostly used by scammers were sending the message from a foreign server and using image spam. Intentional misspelling was nonexistent probably because this would only make them look less professional and hence reduce credibility, which is not good for a scam. Actually, scam should be seen beyond merely a spam problem. Social engineering usually plays a more important role in scam than technology does (Holt & Graves, 2007; Nhan, et al., 2009). In this sense, scammers could be actually more traceable than normal spammers, because they need to stay active in communication with potential victims.

When the sender was soliciting a reply (e.g. lottery winning scheme), some sort of contact information would be provided (usually an email address), but there would be no opt-out choice or physical address. Most spammers do not expect a response as their primary goal was only to get the message out. In such cases, fake information was very common and the header information was more likely to be dishonest, and spammers often did not even bother to hide the violation of the CAN-SPAM Act, which is consistent with the previous findings (Kigerl, 2009). However, if a particular website was being promoted, it was not uncommon to see the sender's email address share the same domain name as the website, which suggests the authenticity of the sender's email address. It is because it is unnecessary to fake the sender's email address, for the website is more traceable than the email address. This type of email usually had no problem passing SPF validation. Deception was less seen when the website was not porn-related or drug-related. The purpose of spam determined how much the spammers attempted to legitimize their messages, even though technically they all failed to meet all requirements. When the purpose was more deviant (e.g. sex or drugs), there was practically no regard for the legal requirements.

In previous studies of spam, language was rarely taken into consideration. In contrast, there were ten languages identified in the spam emails in the present study. The majority of them were in English and a substantial portion was in Chinese. Except for the

linguistics inherent in different languages, there was no obvious distinction found in spam of different languages in terms of the purpose. In all languages, sex and sex-related drugs were commonly the main theme in advertising. Scam was mostly seen in English spam, but not uncommon in Chinese spam. In English spam, scam schemes were more various, including lottery claims, business proposals, advance fee scam, charity, financial offers, and phishing. In Chinese spam, business proposals and advance fee scam were rarely seen. Japanese spam accounted for 5.72% of the spam emails in this paper, and almost all of it was aimed to promoting pornography or sex-related products. In terms of format, Asian spam did not seem to prefer incorporating image spam, while images were more widely used in English spam. When HTML was used, Asian spam appeared to be less colorful and fancy (i.e., simple text plus hyperlinks), whereas more artistic designs (e.g. color text and background colors) could be found in English spam. Regardless of the language, however, all spam emails were similar in violation of the CAN-SPAM Act.

The analysis performed in this study sheds some light on why spam becomes spam. Only 5 emails qualified for legitimate commercial emails according to the CAN-SPAM Act. The major finding was that it does not seem like spammers really try to legitimize their messages, which corresponds to the literature (Kigerl, 2009). They utilized some techniques in an attempt to bypass spam filters for practical reasons, regardless of the legal requirements. For example, most spam emails do not use deceptive subjects, because they are not afraid to let you know this is a commercial message, not because the law says so. Hence, the assertion that the CAN-SPAM Act would actually assist spammers in evading legal responsibilities (Ford, 2005; Yeargain et al., 2004; Weinstein, 2003) is unfounded; because most spammers do not really care to follow those legal guidelines to make their spam emails look more legitimate.

Conclusion

The findings are consistent with the literature as mentioned in the literature review and discussion section. Most spam was aimed for advertising and sex-related products are the dominant subject. Other popular items included drugs, educational programs, computer software, household items, electronics, and jewelry. Scam and fraud emails were also common. They usually involved lottery claims, business proposals, advance fee scam, bogus charity, financial offers (e.g. loans), and phishing. Some techniques were adopted in an effort to bypass spam filter, such as sending emails from a foreign or third party server, inserting random text to fake uniqueness, using images to avoid keyword search, and intentional misspelling of sensitive words. It is noteworthy that most spam emails did not use any of these techniques and even if they did they were not successful in passing spam filters. In this study, most spam was in English and about 30% of it was in Chinese. One third of the emails originated from USA, followed by Taiwan, China, the UK, and Japan. Most spam emails were written in HTML and images were found in 43% of them. Hyperlinks were almost always embedded in the images or highlighted text. All but five emails failed to meet the legal requirements one way or another. Most salient violations included the lack of a physical address, the absence of an opt-out choice, and dishonest header information.

On the whole, there is no evidence suggesting the CAN-SPAM Act has guided spammers to make spam more legitimate. The disregard for legal requirements may be attributable to the difficulty in law enforcement, which can in turn be attributable to jurisdictional issues and insufficiency of police resources. If we only see spam as sending

annoying emails it may not call for much attention. However, the real cost of spam could stem from the recipient's response to spam. For instance, how many people's health was jeopardized due to buying drugs online? How many people suffer financial losses as a result of online scam? Future research should look into the real life impact of spam more than just how much time it takes to delete spam emails.

Limitations and Implications

Admittedly, the analysis discussed above by no means are representative of any global spam trends, as the spam emails collected in this paper came from only five email accounts. As mentioned, our individual online activity and habits would influence what kind of spam we receive, especially in terms of content and language. Conceivably there are more spam with different content and format left out in the present study. Hence, the findings presented in this article may not apply to other people's experiences of spam. Nonetheless, the findings are largely consistent with the literature. Another major limitation is the inability to confirm the legitimacy of the servers utilized by spammers. To the extent possible, the present study could only detect the use of a third-party server, but it is not clear whether the server owner was aware that it has been used to send spam. Without sufficient information about the servers, especially the foreign ones, it is not possible to determine effective methods to trace spammers. If spammers remain untraceable for the most part, no legislation would ever render any observable impact on spam.

In this paper, the techniques used by spammers to bypass spam filters largely were not successful but spammers never stopped trying. This ever-lasting endeavor suggests that the law entails little deterrence, while technology is currently the more effective prevention. Thus, perhaps in addition to specifying what makes commercial emails legal, the law needs to also punish the attempt to defeat spam filters, which means spammers, would be required to make sure their messages will end up in the spam folder instead of the inbox. For example, first, a commercial email should not be allowed to be sent from a third party server, even with permission. Second, if a commercial email contains random text or intentional misspelling, it should also incur penalty. Third, if using images, the image should be in certain formats to be recognizable by Optical Character Recognition (OCR). The rules apply as long as the email is unsolicited and is commercial in nature. Therefore, spam when violating these rules can be at least punished for technicality regardless of the content or motive. This way, spammers will have little incentive to even try to violate these rules since most spam emails are already caught by spam filters. It is not very rational for them to adopt techniques that are largely ineffective while risking increased penalty.

Spam is not existent only through email. As the utility of cell phone is rapidly expanding, spam has found a new breeding ground (Fleizach et al., 2007). Future research on spam should devote attention to spam on new media as well.

References

Allenberg, J., Weinberger, K., Dasgupta, A., Langford, J., & Zinkevich, M. (2009). Proceedings from Sixth Conference on Email and Anti-Spam 2009: Collaborative Email-Spam Filtering with Consistently Bad Labels Using Feature Hashing. Mountain View, California USA.

Aradhye, H. B., Myers, G. K., & Herson, J. A. (2005). Proceedings from Eighth International Conference on Document Analysis and Recognition: Image Analysis for Efficient Categorization of Image-Based Spam Email. Menlo Park, CA, USA

Blanzieri, E. & Bryl, A. (2009). A survey of learning-based techniques of email spam filtering. *Artificial Intelligence Review, 29*(1), 63-92.

Chang, J.S. (2008). An analysis of advance fee fraud on the Internet. *Journal of Financial Crime, 15*(1), 71-81.

Delany, S. J., Cunningham, P., Tsymbal, A., & Coyle, L. (2004). A case-based technique for tracking concept drift in spam filtering. *Knowledge-Based Systems, 18*, 187-195.

Dhamija, R., Tygar, J. D., & Hearst, M. (2006). Proceedings of the SIGCHI Conference on Human Factors in Computing Systems: Why Phishing works. New York, NY, USA.

Dredze, M., Gevaryahu, R., & Elias-Bachrach, A. (2007). Learning fast classifiers for image spam. Retrieved Oct 18, 2010 from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.102.8417&rep=rep1&type=pdf

Duan, Z., Dong, Y., & Gopalan, K. (2007). DMTP: Controlling spam through message delivery differentiation. *Computer Networks, 51*(10), 2616-1630.

EC-Council. (2010). Computer Forensics: Investigating Data and Image Files. Cengage Learning: Clifton Park, NY.

Edelson, E. (2003). The 419 scam: Information warfare on the spam front and a proposal for local filtering. *Computer & Security, 22*(5), 392-401.

Federal Trade Commission. (2009). The Can-SPAM Act: A Compliance Guide for Business. Retrieved September 17, 2010 from http://www.ftc.gov/bcp/edu/pubs/business/ecommerce/bus61.shtm

Fleizach, C., Liljenstam, M., Johansson, P., Voelker, G. M., & Mehes, A. (2007). Proceedings of the 2007 ACM Workshop on Recurring Malcode: Can you infect me now?: Malware Propagation in Mobile Phone Networks. New York, NY, USA.

Fogel, J. & Shlviko, S. (2009). Consumers with sexual performance problems and spam e-mail for sexual performance products. *Journal of Internet Banking and Commerce*, *14*(1). Retrieved Oct 18, 2010 from http://www.arraydev.com/commerce/jibc/2009-04/Fogel.doc.pdf

Fogel, J. & Shlivko, S. (2010). Consumers with sexual performance problems and spam email for pornography. *Journal of Internet Banking and Commerce, 15(1).* Retrieved Oct 18, 2010 from http://www.arraydev.com/commerce/jibc/201004/Fogel%20pornography%20sexual%20health.pdf

Ford, R. A. (2005). Preemption of state spam laws by the federal CAN-SPAM Act. *The University of Chicago Review, 72*(1), 355-384.

Gmail. (2010). Gmail uses Google's innovative technology to keep spam out of your inbox. Retrieved September 22, 2010 from http://www.google.com/mail/help/fightspam/spamexplained.html

Goodman, J., Cormack, G. V., & Heckerman, D. (2007). Spam and the ongoing battle for the inbox. *Communications of the ACM, 50*(2), 25-33.

Goodman, J., Heckerman, D., & Rounthwaite, R. (2005). What can be done to stanch the flood of junk email messages? Retrieved Oct 18, 2010 from http://davidjf.free.fr/new/Xscientific%20american_%20stopping%20spam.pdf

Hayahi, P. & Potdar, V. (2008). Proceedings from the 10th International Conference on Information Integration and Web-based Applications & Services. Evaluation of spam

detection and prevention frameworks for email and image spam: A state of art. Linz, Austria: ACM.

Holt, T. J. & Graves, D. C. (2007). A qualitative analysis of advance fee fraud e-mail schemes. *International Journal of Cyber Criminology, 1*(1), 137-154.

Kigerl, A.C. (2009). CAN SPAM Act: An empirical analysis. *International Journal of Cyber Criminology, 3*(2), 566-589.

Kraut, R. E., Sunder, S., Telang, R., & Morris, J. (2005). Pricing electronic mail to solve the problem of spam. *Human-Computer Interaction, 20*, 195-223.

Kumaraguru, P., Rhee, Y., Acquisti, A., Cranor, L.F., Hong, J., & Nunge, E. (2007). Proceedings of the SIGCHI Conference on Human Factors in Computing Systems: Protecting people from phishing: The design and evaluation of an embedded training email system. New York, NY, USA.

Lai, C. C. (2007). An empirical study of three machine learning methods for spam filtering. *Knowledge-Based Systems, 20*(3), 249-254.

Lee, Y. (2005). The CAN-SPAM Act: A silver bullet solution? *Communications of the ACM, 48*(6), 131-132.

Lueg, C. (2003). Spam and anti-spam measures: A look at potential impacts. Retrieved Oct 18, 2010 from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.4.6385&rep=rep1&type=pdf

Maggs, P. B. (2006). Abusive advertising on the Internet (SPAM) under United States law. *The American Journal of Comparative Law, 54*, 385-394.

McAfee, (2009). The carbon footprint of email spam report. Retrieved Oct 08, 2010 from http://resources.mcafee.com/content/NACarbonFootprintSpam

Nhan, J., Kinkade, P., & Burns, R. (2009). Finding a pot of gold at the end of an Internet rainbow: Further examination of fraudulent email solicitation. *International Journal of Cyber Criminology, 3*(1), 452-475.

Rabinovitch, E. (2007). Staying protected from "social engineering". *Communications Magazine, IEEE, 45(9)*, 20-21.

Rogers, K. M. (2006). Viagra, viruses and virgins: A pan-Atlantic comparative analysis on the vanquishing of spam. *Computer Law & Security Report, 22*, 228-240.

Spammer X. (2004). *Inside the SPAM Cartel: Trade Secrets from the Dark Side*. Syngress Publishing Inc.: Rockland, MA

Stolfo, S. J., Hershkop, S., Wang, K., Nimeskern, O., & Hu, C. W. (2010). Behavior profiling of email. *Intelligence and Security Informatics, 2665*, 74-90.

Symantac. (2010). Global spam categories. State of Spam & Phishing: A Monthly Report. Retrieved on Oct 6, 2010 from http://www.symantec.com/content/en/us/enterprise/other_resources/b-state_of_spam_and_phishing_report_09-2010.en-us.pdf

Taylor, B. (2006) Sender reputation in a large webmail service. Retrieved Oct 18, 2010 from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.122.3561&rep=rep1&type=pdf

Ting, A. S. H. (2004). Penis enlargement, cheap Viagra, and noni juice: Evaluating and winning the war against health product spam. Retrieved Oct 18, 2010 from http://leda.law.harvard.edu/leda/data/665/Ting.html

Wall, D. S. (2004). Digital realism and the governance of spam at cybercrime. *European Journal on Criminal Policy and Research, 10*, 309-335.
Weinstein, L. (2003). Spam wars. *Communications of the ACM, 46*(8), 136.
Yeargain, J. W., Settoon, R. P., & McKay, S. E. (2004). Can-Spam Act of 2003: How to spam legally. *Journal of Strategic E-Commerce, 2*(1), 15-30.

Appendix 1: Spam Example 1

Hello,

I've added you to my qojmj group at Yahoo! Groups, a free, easy-to-use service. Yahoo! Groups makes it easy to send and receive group messages, coordinate events, share photos and files, and more. Description of the group:

---

i02ww0ad50rvi
Complete your Yahoo! Groups account:

Your email address has been added to the email list of a Yahoo! Group. To gain access to all of your group's web features (previous messages, photos, files, calendar, etc.) and easier control of your message delivery options, we highly recommend that you complete your account by connecting your email address to a Yahoo account. It is easy and free.
Please visit:http://groups.yahoo.com/convacct?email=funnyass%40gmail.com&list=qojmj
Important information about the qojmj group

---

* To send a message to the members of this group, send an email to:
qojmj@yahoogroups.com
* To leave the group, you can unsubscribe by replying to this message, or by sending an email to: qojmj-unsubscribe@yahoogroups.com
Regards,
Moderator, qojmj
Report abuse:

Because Yahoo! Groups values your privacy, it is a violation of our service rules for moderators to add subscribers to a group against their wishes. If you feel this has happened, please notify us:
http://help.yahoo.com/l/us/yahoo/groups/original/members/forms/abuse.html
You may also change your email preferences to prevent group owners from adding you to their groups. To do so, please go here:
http://groups.yahoo.com/s?tag=_x1x1428ewguczhZqyeLd1e9s07FmH-VJTW9tAlfgB4nbgi9icEom0SOar45GOWoOEAVlydV0ug
Your use of Yahoo! Groups is subject to:
http://docs.yahoo.com/info/terms/

Appendix 2: Spam Example 2

Message 1:

Subject: website"WorldPharmacy"
Sender: Nola@dfgbifedji.abvgdejka.com
http://ivaxitumen.mindnmagick.com/hilymig.html
Hate not at the first harm.
==========

Message 2:
Subject: RE:SiteWorldPharmacy
Sender: Eloy@cbhaahggbj.pilluliradosti.com
website, which you asked:http://xesuweryj.o-f.com/amixine.html
A clean hand wants no washing.

Appendix 3: Spam Example 3

Subject : You Have Everything To Gain! 3-4 inche more to your Penis, 100% MoneyBack Guaranteeed lxcy

You Have Everything To Gain!

Incredib1e gains in length of 3-4 inches to yourPenis, PERMANANTLY

Amazing increase in thickness of your Penis, up to 30%
BetterEjacu1ation control
Experience Rock-Hard Erecetions
Explosive, intense Orgasns
Increase volume of Ejacu1ate
Doctor designed and endorsed
100% herbal, 100% Natural, 100% Safe
The proven Natural Penis Enhancement that works!
100% MoneyBack Guaranteeed

Appendix 4: Spam Example 4

Appendix 5: Example of Spam 5

Subject: Greeting in the name of God
Greeting in the name of God
My name is Mrs. Emily Adams. I am a sick woman who has decided to donate what I have to charity through you. You may be wondering why I chose you. But someone has to be chosen. I am 59 years old and was diagnosed for cancer about 2 years ago, immediately after the death of my husband who had left me everything he worked for when he was working in our Embassy for nine years before ended up poisoning him to death my relatives on the 19th November 2006. I am suffering from a long time cancer, which also affected my heart.

My husband and I are Christians, but quite unfortunately, he died by poisoning. l will be going in for an operation, and I pray that I survive the operation. I have decided to WILL / Donate the sum of $4.5 Million United State Dollars to charity through you for the good work of the lord, and to help the motherless, less privileged and also for the assistance of the widows. At the moment I cannot take any telephone calls, due to the fact that my relatives are around me and I have been restricted by my doctor from taking telephone calls because I deserve all the rest I can get.
I inherited a total sum of $4.5 million dollars from my late husband, this money which is concealed in a metallic trunk box is deposited with a security and finance company in Cote D' Ivoire Due .to This deposit was coded under a secret arrangement as a family treasure. This means that the security company does not know the content of this trunk box.

Presently, I have informed the security company about my decision in WILLING this trunk box to you. I wish you all the best and may the good Lord bless you abundantly, and please use the funds well and always extend the good work to others. If you are interested in carrying out this task, I will kindly want you to write me back for further direction and on how to carry out this task.

I know i have never met you but my mind tells me to do this with me and I hope you act
sincerely.

Your beloved sister in Christ,
Mrs. Emily Adams