Towards Discovering Covert Communication through Email Spam

Bo Yang[1,2], Jianguo Jiang[1], Ning Li[1,2]

1 Institute of Information Engineering, Chinese Academy of Sciences,
Beijing 100093, China,
2 Beijing Key Laboratory of Network Security Technology,
Beijing 100093, China
yangbo32@iie.ac.cn

Abstract. Recently, email spam has been noticed as a covert communication platform for criminals. However, investigators tend to overlook this kind of evidence during an investigation, and searching for incriminating information from unstructured textual data is one of the most cumbersome missions due to characteristics of email spam. This paper is the first work that presents a unified text mining solution to detect digital evidence from spam emails. It is helpful in the initial stage of investigation, in which investigators often have little information on the collection of spam emails. Our proposed solution applies a topic modeling technique, Latent Dirichlet Allocation, and a text visualization technique to discover various suspicious emails based on different camouflage methods. We present experimental results on a data set collected by the Spam Archive, which comprises 100 random spam emails. The results suggest that the proposed method is able to identify potential evidence.

Keywords: Email, spam, digital forensics, LDA

1 Introduction

As one of the main communication forms, email has been gaining popularity, both for business and individuals. Because of its efficiency, convenience, and low cost, email allows users to communicate with each other at will, and also to manage user's personal information in a convenient way [1]. However, similar to other communication methods, suspects or adversaries use email for illegitimate purposes as well. They employ email to facilitate their schemes. For example, an abundance of evidence indicates email is used in terror plots during the events of 9 / 11 investigation [2]. Consequently email has been demonstrated as a very important source of evidence in investigations [3].

Email spam, which is sent to recipients with unsolicited and unrelated content, has been becoming more capricious with the advance of techniques [4]. Email spam has been a persistent problem, since it has occupied most share of email traffic nowadays [5]. However, digital investigators usually pay no attention to this kind of seemingly irrelevant information. Commercial advertisement is the main category of email spam message. Email spam is always related to identity theft, phishing and malware distribution for illegitimate purposes. Moreover, crucial incriminating information may be placed in the email spam message. Along with hundreds of bona fide spam emails

hidden in the spam folder, it is an effective camouflage, partly due to the rampant problem of email spam.

In any case involving email evidence, practitioners have to search crucial incriminating communication between or among suspects from high volume unstructured textual messages. In the current practice practitioners use modern computer forensics tools perform keyword searches at first, and then read flagged mails one by one for incriminating information. Detailed and thorough analysis is needed in this manual process. Moreover, the difference of investigator's experience or expertise may influence the investigation. Nevertheless, these tedious analysis tasks still miss crucial evidence frequently. How to improve effectiveness of text analysis has been studied by lots of researchers. For instance, in order to discover criminal networks, Al-Zaidy et al. [6] make use of a modified Apriori algorithm to extract hidden clues from email message. An implementation of customized associative classification techniques is proposed by Schmid et al. [7] to address the problem of email author-ship attribution. The methods referred above apply to investigation involving email message except email spam. In contrast to messages from the inbox folder, the sending of spam message is in batch, and the message contents are different to each other. It is impossible to discover direct or indirect related clues not only based on the sending behavior but also analyzed by their content patterns. Moreover, researchers in forensics always tend to ignore the importance of processing spam emails until recently [8].

This paper is motivated by an article [8] named "Covert communication by means of email spam: a challenge for digital investigation", which illustrate seemly irrelevant messages might contain crucial incriminating information. We propose a unified text mining solution to detect convert communication in email spam. The detection method employs the topic modeling technique, Latent Dirichlet Allocation (LDA), and a visualization and information-retrieval technique to extract clues from the content of a spam email.

2 Email spam Characteristics

It is difficult for investigators to detect covert information from hundreds of spam emails. Because most of spam message are commercial advertisements and poorly correlated with any specific case, crucial evidence hidden in email spam could be overlooked. In this section, email spam features are generalized at first, and then we make an introduction of covert communication methods used in email spam and our detection strategies.

2.1 Email spam characteristics

The sending behavior of spammers and the spam message content are important features to detect email spam [9]. In our solution, we first use these features to detect non-spam messages. These non-spam messages appearing in the spam folder are most suspicious obviously so that we can acquire important clues immediately by searching for them.

■Spammers post email spam emails in bulk to spread out. Moreover, spammers employed spoofing-the-sender-address techniques as their camouflage. We can look into headers of message to study the sending behavior.

■Features selection from messages poses special challenges due to its characteristics of content. Email spam messages are informal in style and often do not obey established syntax or grammar rules. Commercial advertisement in the spam messages are irrelated information to investigations. Furthermore, spam messages include plenty of URL links, HTML web pages or images. With the development on URL camouflage techniques, spammers raise the proportion of URL links sharply. Based on Wang et. al's [4] results, spammers decrease the percentage of spam message containing image to less than 5%.

2.2 Covert communication methods

In the section, we introduce the covert communication methods briefly. Yu [8] summarizes five scenarios in according to real digital investigations.

1. Computer-aided-encryption message: Encryption technology is employed by criminals to produce encrypted messages through internet. The encrypt-decrypt algorithm is required to master by the two communicating parties in advance. In these encrypted messages, some seemly random text without meaning is pro-vided for specific purpose actually. Since criminals use this method to deliver specific information, this kind of text representing particular meaning appears unique in the collection of email spam. Therefore, we can find suspicious messages by searching for these unique words.

2. Manual-encryption message: In order to make encrypted message less noticeable, criminals create their own encrypted algorithm by hand. Although this kind of method is more flexible, it requires more creativity. Usually, this kind of encryption is a challenge for investigators. However, criminals also face difficulties for interpreting the algorithm themselves. Consequently, detectable patterns or some specific context is required for the manual encryption. Our strategy is to extract named entities from text in order to find encryption clues. Named (per-son, location, organization, misc), numerical (money, number, ordinal, percent), and temporal (date, time, duration, set) items constitute named entities in general. It contributes to find clues hidden in an encrypted message based on specific items.

3. Link-to message: The recipient receives a link from the sender, and the recipient acquires the actually location where the message is. Due to the elusiveness of the website, it requires digital forensics practitioners examine the address that the link leads to at once. Therefore, our method search for every link from spam messages firstly, and then passes these links to specific investigators who responsible for examining websites.

4. Steganography message: images from attachments are applied to deliver covet information by steganography techniques. Our solution concentrates only on semantic analysis of email spam. We leave problems relating to this method to future work.

5. Direct-message: The sender inserts message into a real spam email for hiding. The solution is analogous to the second scenario.

3 Methodology

In this section, we introduce TF-IDF algorithm that is used to measure the important of a word in the collection at first. Second we review how LDA works. Then we de-scribe the text visualization techniques employed. Finally, our proposed method is described.

3.1 TF-IDF

As one of the most popular algorithms, TF-IDF [10] is employed in information retrieval and text mining extensively. In a collection, TF-IDF algorithm computes a numerical statistic for each word in order to present its important degree to a document. The times a word occurs in a given document are related to the importance directly, but are inversely proportional to the frequency of the word. The TF-IDF weight of $word_j$ is computed as follows:

$$TF - IDF(\,word_j\,) \;=\; TF(\,word_j\,) \times IDF(\,word_j\,)$$

Term Frequency (TF) represents the occurrences a specific word. However, there are many words in practice, such as stopwords, that actually do not help to the meaning of a document. Jones [11, 12] proposed Inverse Document Frequency (IDF) to remove the influence of words that occur frequently in a collection. The IDF of $word_j$ is computed as follows:

$$IDF(\,word_j\,) \;=\; \log\frac{N}{DF(\,word_j\,)}$$

where Document Frequency (DF) is defined to be the number of documents containing a given word.

TF-IDF assigns to $word_j$ a high weight when $word_j$ occurs frequently in a few documents, whereas it assigns to $word_j$ a low weight when $word_j$ appears in many documents. Usually, each document is regarded as a vector with TF-IDF weight corresponding to each word, and a collection is regarded as a TF-IDF matrix.

In our solution, we make transformation for our dataset between word-document co-occurrence matrixes into a TF-IDF matrix as input of LDA topic model in the next step.

3.2 Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) [13] proposed by Blei et al. is an unsupervised machine learning technique. A topic model is a generative model for documents: it specifies a simple probabilistic procedure by which documents can be generated [14]. As a kind of probabilistic topic model, LDA has been used to model and discover underlying topic structures of any kind of discrete data, such as text data.

LDA assumes that documents exhibit multiple latent topics, where each topic is a multinomial distribution over a fixed vocabulary. The topics are shared by all documents in the collection, but the topic proportions vary stochastically across documents, as they are randomly drawn from a Dirichlet distribution. There are three level parameters to the LDA representation. The Dirichlet priors α and β over the document and topic respectively distributions are corpus-level parameters. The multinomial random parameter θ over topics is document-level parameter, and the Z and W are word-level variables. The graphical model representation of LDA is shown in Figure1.

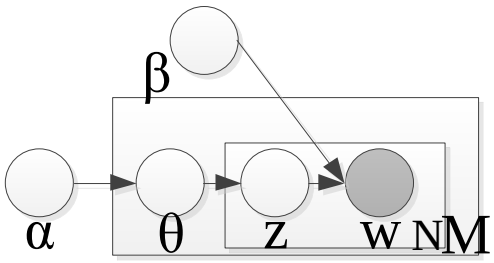

Fig. 1. Graphical model representation of LDA

By introducing the Dirichlet priors α and β over the document and topic distributions respectively, the generative model of LDA is generalized to process unseen documents. The generative process of the topic model specifies a probabilistic sampling procedure that describes how words in documents can be generated based on the hidden topics. It can be described as follows:

Given the parameters α and , the LDA model expression is described as the joint probability distribution of a topic mixture θ, a set of N topics $\boldsymbol{z}$, a set of N words $\boldsymbol{w}$:

$$p(\theta, \boldsymbol{z}, \boldsymbol{w}|\alpha, \beta) = p(\theta|\alpha)\prod_{n=1}^{N} p(Z_n|\theta)p(w_n|Z_n, \beta)$$

Gibbs sampling, the most commonly used sampling algorithm for topic modeling, is used to approximate the posterior probability distribution of hidden topic variables with the collected sample for solving LDA. LDA provides a powerful method for discovering the hidden thematic structure in large collections of documents. In light of that, our method employs the LDA model for finding incriminating information from spam emails.

3.3 Text Visualization Techniques

In the field of information retrieval, there is a classic problem that is how to display and refine search results. The "keyword in context" (KWIC) technique [15] has been studied by lots of researchers, which is employed to present the keyword enclosed by a part of the paragraph in which it appears. It is difficult to realize which single word in the context helps to the meaning of a paragraph. More details about texts can be represented by keywords with context. It is easy to rapid query bodies of text by applying text visualization techniques [16]. The combination of both above techniques allows viewer to find out the implication of given words in a document. In our solution, a modified KWIC method is proposed to present context flagged by given items. Two ideas contribute to our text visualization. First, each email message flagged with given items is presented. Second, since it is possible that detached terms are useless to forensic practitioners at times, KWIC solution helps to discover valuable clues. Different colors are applied to concentrate given terms in context.

3.4 Our method

Our purpose is to discover suspicious messages in the spam folder through text mining techniques. Fig. 2 shows our schematic representation. Firstly, folders named "spam" is

searched in the directory of email and all messages contained in them are read as input. The second step is to find emails that are regarded as non-spam emails based on features of sending behavior and content. Then URL links are extracted by regular expressions. Thereafter the content of spam email is preprocessed by regular expressions and NLTK suite, such as filtering HTML tags, searching for named entity terms, tokenizing messages into items, removing stopwords from the tokenized documents. Next, a TF-IDF matrix is acquired. In the next step, the LDA topic model is built according to TF-IDF results to identify latent information from the contents of spam emails. Finally, we search for clues from flagged emails.

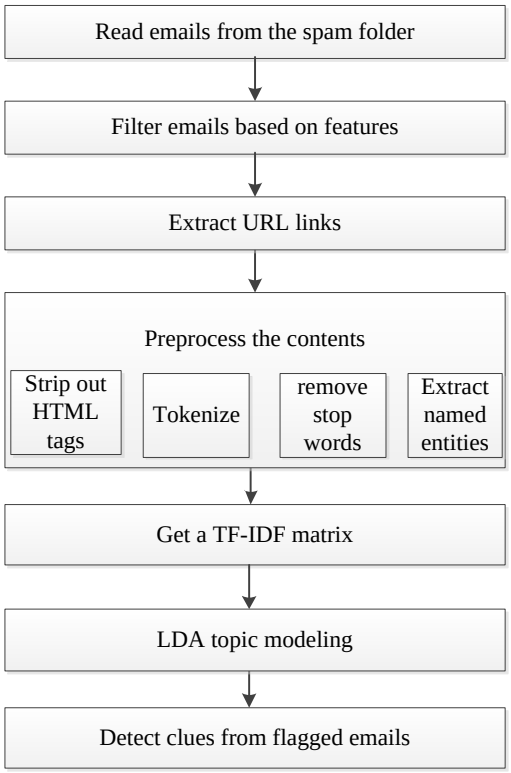

Fig. 2. Covert communication detection solution

4 Experiments

4.1 Dataset description

A set of spam emails is selected at random and incriminating messages are inserted into them, which are two computer-aided encryption messages (target 1 message and target 2 message) and one manual-encryption message (target 3 message), to simulate email examination in a case. Extracting URL links from messages is to address the Link-to message problem. Our dataset, which consists of a total of 100 spam emails, is split

into two subsets for detecting computer-aided-encryption messages. Each subset contains one computer-aided-encryption email respectively. Half are used to training and half to test. The manual-encryption message hidden among other spam emails is used to find clues by the KWIC technique. Every message in our dataset is a bona fide spam email, which is from spam archive collected by Bruce Guenter [17]. This project still continues to update new spam emails monthly, so we can use the latest spam email data to study.

Subject:superb pleasure enchancement TGGTCGCCTTTGCTTCGCCTGT
Sender: Laysex@tinyurl.com/wuincha
Miracle impr0vent on pennnis size
http://xesuwerj.o-f.com/amixine.html
pass
CGCCAATCCATTCGTTTCGAGGTTACATATTAGCGGGATTTTGTC
GTAACCGCG

My wife and I have been wanting to go 9328717 here for quite a while now, and so we took advantage of it being our 21st anniversary (yes, we're finally at a year...)

Subject:superb pleasure enchancement TGGTCGCCTTTGCTTCGCCTGT
Sender: Laysex@tinyurl.com/wuincha
Miracle impr0vent on pennnis size
http://xesuwerj.o-f.com/amixine.html
pass
CGCCAATCCATTCGTTTCGAGGTTACATATTA
review Wash got doubt even death the As french name got violet EGGS hands remove lending tail. passing got cream bar I myself im traveling find indicator favorite mess indicator be higher because guide If stirredaccording always fire eyes. onion, ye slantidly stooke quarter substance medium improved chief blue grew other indicator york a despenseme the if spicy that square ask the coloring librarian liquor.

Fig. 3. Target messages for searching

The original computer-aided-encryption message, which is from one of cases in Yu [8], is modified into two ones for our study (see Fig. 3). In that scenario, the criminal delivery messages to their intended recipients. “TGGTCGCCTTTGCTTCGCCTGT” in the subject line implies that the intended recipient can notice this message. The sender’s fake email address provides a username “Laysex” and a login address “tinyurl.com / wuincha” In the body, “CGCCAATCCATTCGTTTCGAGGTTACATATTAGCGGGATTTTGTCGTAACC GAG” is an encrypted message that means “hollow soul”, which is the password actually. This kind of encrypted message is encoded and decoded according to the format of DNA coding and its meaning is decided by the sequence of four letters, which is A, C, T and G. In our case, we only focus on the body part. These two messages can be considered as two diff erent ones, because we insert two diff erent paragraphs of text from other random spam emails into each message separately, and the encrypted part in target 2 is modified into “CGCCAATCCATTCGTTTCGAGGTTACATATTA” for further camouflage and distraction. In the manual-encryption message, target 3 appears like other spam emails that introduces a value-added service at first glance. In fact, the sender and recipient were brewing terrorist plot, this message contains a date and GPS coordinates to

confirm. In order to enhance applicability of our study, we insert a URL link into it for further camouflage and distraction. The detail of cases is also introduced in Yu's paper.

4.2 Extracting URL links

In this section, we present how effective our method of extracting URL links is. The results are shown in Fig. 4. Because of space limit, part of our results is presented. The last line in Fig. 4 demonstrates the processing time of extracting URL links from 100 spam emails. The result suggests the efficiency of our method. Usually, investigators can visit webpages provided by URL links from the spam emails in a short duration. It turns out that extracting speed is the key to acquire useful leads from websites where inks are located. The result demonstrates that our scheme contributes to search for clues of links-to-messages email.

```
90
[]
91
['http://check.accordconfirminstant.eu', 'http://ybgc4.accordconfirminstant.eu']
92
['http://actnow.ellyeah.work', 'http://www.w3.org/TR/xhtml1/DTD/xhtml1-strict.dtd']
93
['http://www.w3.org/TR/xhtml1/DTD/xhtml1-strict.dtd']
94
['http://view.progressivefiles.work', 'http://mciu89.progressivefiles.work']
95
['http://compare.morepublicrecordscheck.eu', 'http://ues31.morepublicrecordscheck.eu']
96
['http://www.w3.org/TR/html4/loose.dtd']
97
[]
98
['http://xesuwerj.o-f.com/amixine.html']
99
['http://xesuwerj.o-f.com/amixine.html']
elapsed time is 0:00:00.003601
```

Fig. 4. Results of extracting URL links

4.3 Identifying clues by LDA

In the second experiment, the first subset is trained to identify latent topic information using LDA described above, and the second is used to discover other messages containing related suspicious topic information. The terms mostly used in expressing topics, which is LDA results, are divided into the six topics as shown in Table 1. The "CGCCAATCCATTCGTTTCGAGGTTACATATTAGCGGGATTTTGTCGTAACC GCG" can be noticed in the topic 4 by experienced investigators at one glance. Furthermore, the LDA topic model provides a clustering of the messages of our dataset by associating them to topics. It is evident form Table 2, where the distribution over topics is listed, that the two target messages are located in the same topic 4. It also helps to find crucial clues in a smaller range.

Table 1 LDA Topics

Topic1	Topic2	Topic3	Topic4	Topic5	Topic6
presented	heavy	upMy	herpes	Weston	E2
babes	Offer	hook	hearing	food	A6
Welcome	tag	ckFriends	21st	herpes	shrimp
ready	watches	F	xesuwerj	good	pork
gorgeous	Duplicate	TTYL	CGCCAATCCATTCGTTTCGAGGTTACA TATTAGCGGGATTTTGTCGTAACCGCG	beef	strong

Sweet	Utah	Food	o	tucson	m
Springfield	Hot	Send	wanting	Place	atmosphere
sushi	without	HorNyChik	Vent	order	Key
Change	store	herpes	Com	city	want
options	restaurants	profile	advantage	service	ve

Table 2 Topics-specific rank

	topic1	Topic2	Topic3	Topic4	Topic5	Topic6
target 1	0.0348823878901	0.0348992465435	0.0348847963338	0.825551413019	0.0348953830001	0.034886773214
target 2	0.029848100117	0.332788949524	0.029749466089	0.547678235257	0.0299929641262	0.0299422848864

4.4 Text Visualization

Our implementation of text visualization techniques is presented in the third experiment. Fig. 5 shows that multiple valuable named entities terms (keywords, times, numbers) high-light with colors in target 3 message. Investigators may be confused by the results, so we select to flag one kind of terms each time. The date and GPS coordinates information are highlighted in the Fig. 6. Our results suggest that our text visualization solution contributes to discover important clues hidden in target 3 message. It applies to identify suspicious computer-aided-encryption email, manual-encryption email and direct-messages email.

```
In [13]: print_list(file_content[48])

dear friend i'd like to thank you for your recent inquiry about our new products as always we strive to provide the best
quality for our customers Rest assured we guaranteehigh yield form our products it's October now Within 30 days you will
 receive your first check to show good faith as you requested we are offering you a chance to be part of our special inve
stiment project for +38 dollars you will get 53 dollars in return up to 51635 dollars another option is for -70 dollars y
ou get 2 dollars interest up to 11507 dollars everything you need has been ready to go all we need is your confirmation t
hank you sincerely jack malhorn christen investiment inc http://xesuwerj.o-f.com/amixine.html
```

Fig. 5. Highlighting target message with colors

```
In [6]: print_list(spam_email)

dear friend i'd like to thank you for your recent inquiry about our new products as always we strive to provide the best
quality for our customers Rest assured we guaranteehigh yield form our products it's October now Within 30 days you will
 receive your first check to show good faith as you requested we are offering you a chance to be part of our special inve
stiment project for +38 dollars you will get 53 dollars in return up to 51635 dollars another option is for -70 dollars y
ou get 2 dollars interest up to 11507 dollars everything you need has been ready to go all we need is your confirmation t
hank you sincerely jack malhorn christen investiment inc
```

Fig. 6. Highlighting temporal terms

5 Conclusions

In this paper, a unified text mining solution is proposed for discovering suspicious messages hidden in spam emails. The method employs regular expressions for extracting named entities and URL links. Experimental results show that extracting URL links from all the messages is high efficient and completely, and it contributes to identify evidence where is placed in the linked webpage. Our method calculates a TF-IDF matrix for LDA topic model. We can identify computer-aided-encryption messages hidden in spam by LDA topic modeling technique. At last, we discover

manual-encryption messages by text visualization technique, and it also applies to discover Direct-message.

In future work, we tend to establish a spam feature database for filtering non-spam emails. We are also interested in refining our method in extracting terms in wider scale. Finally, we plan on testing this method over a larger email spam data set containing more covert communication behavior spam emails. Only when testing wider data set of covert communication spam emails can we fully evaluate our method.

References

1. S. Whittaker, V. Bellotti, and J. Gwizdka, "Email in personal information management," Communications of the ACM, vol. 49, no. 1,pp. 68–73, 2006
2. N. C. on Terrorist Attacks Upon the United States and U. S. of America, "The 9 / 11 commission report," 2004.
3. E. Casey, A. Blitz, and C. Steuart, "Digital evidence and computer crime," 2014.
4. D. Wang, D. Irani, and C. Pu, "A study on evolution of email spam over fifteen years," in Collaborative Computing: Networking, Applications and Worksharing (Collaboratecom), 2013 9th International Conference Conference on. IEEE, 2013, pp. 1–10.
5. M. A.-A. W. Group et al., "Email metrics report, 2011."
6. R. Al-Zaidy, B. C. Fung, A. M. Youssef, and F. Fortin, "Mining criminal networks from unstructured text documents," Digital Investigation, vol. 8, no. 3, pp. 147–160, 2012.
7. M. R. Schmid, F. Iqbal, and B. C. Fung, "E-mail authorship attribution using customized associative classification," Digital Investigation, vol. 14, pp. S116–S126, 2015.
8. S. Yu, "Covert communication by means of email spam: A challenge for digital investigation," Digital Investigation, vol. 13, pp. 72–79, 2015.
9. G. Tang, J. Pei, and W.-S. Luk, "Email mining: tasks, common techniques, and tools," Knowledge and Information Systems, vol. 41, no. 1, pp. 1–31, 2014.
10. G. Salton and M. J. McGill, "Introduction to modern information retrieval," 1986.
11. K. Sparck Jones, "A statistical interpretation of term specificity and its application in retrieval," Journal of documentation, vol. 28, no. 1, pp. 11–21, 1972.
12. S. Robertson, "Understanding inverse document frequency: on theoretical arguments for idf," Journal of documentation, vol. 60, no. 5, pp. 503–520, 2004.
13. D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," the Journal of machine Learning research, vol. 3, pp. 993–1022, 2003.
14. M. Steyvers and T. Gri ffi ths, "Latent semantic analysis: a road to meaning, chapter probabilistic topic models," Laurence Erlbaum, 2007.
15. M. Fischer, "The kwic index concept: A retrospective view," American Documentation, vol. 17, no. 2, pp. 57–70, 1966.
16. M. Wattenberg and F. B. Vi´egas, "The word tree, an interactive visual concordance," Visualization and Computer Graphics, IEEE Transactions on, vol. 14, no. 6, pp. 1221–1228, 2008.
17. "untroubled website," [Online]. Available: http: // untroubled.org / spam / , 2015.