Clustering Spam Campaigns with Fuzzy Hashing

Jianxing Chen[1], Romain Fontugne[2,3], Akira Kato[4], Kensuke Fukuda[2]
[1]Saarland University [2]National Institute of Informatics
[3]Japanese-French Laboratory for Informatics [4]Keio University
amikpie@gmail.com, romain@nii.ac.jp, kato@wide.ad.jp, kensuke@nii.ac.jp

ABSTRACT

Identifying spamming botnets is essential to defeat spammers and reduce the harm caused by spam emails. The first step to uncover these botnets is the identification of spam campaigns. Simple methods looking for common identifiers in emails, such as URL or email addresses, are inefficient due to the emergence of obfuscation techniques like URL shortening. In this paper we propose a new method based on fuzzy hashing to cluster spam with common goals into the same spam campaign. Fuzzy hashing allows us to identify emails with similar contents even though usual identifiers are obfuscated. Using the proposed method we process a three year long dataset that consists of 540 thousand spam emails. The efficiency of the proposed method is assessed by inspecting the characteristics of the top 100 campaigns found. Finally, we present typical behaviors of the uncovered spam campaigns and the corresponding botnets.

Categories and Subject Descriptors

C.2.0 [**Computer-Communication Networks**]: General—*Security and protection (e.g., firewall)*

General Terms

Measurement, Security

Keywords

botnet, spam, clustering

1. INTRODUCTION

Several studies estimate the amount of abusive emails, or spam emails, for more than 85% of the daily emails [1, 6]. Because of their abusive use of network resources, the cost of these unsolicited emails is significant for Internet service providers. Moreover, spam emails are both irritating and harmful to Internet end-users as the purposes of these unsolicited emails are essentially malicious and illegal.

Spam is a key medium for scams, phishing, illegal advertising, and malware spreading. To send numerous emails in a stealthy manner, spammers take advantages of large sets of compromised hosts also called botnets. A botnet is controlled from a single entity, the command and control server, that allows a spammer to distribute spam sending tasks across numerous hosts. As each infected host sends a small number of spam, detecting all members of a botnet is particularly hard.

An effective approach to infer spamming botnets is to identify all spam emails from a same campaign. Thereby, the challenge is shifted to the identification of spams from the same campaign [5]. This task is easier because all spams in a campaign share the same goal, hence, they have common features that permit to distinguish distinct spam campaigns.

In this paper, we propose a new methodology based on fuzzy hashing to identify spam campaigns. Fuzzy hashing is an effective technique to measure the similarity of two sequences of characters. We implement fuzzy hashing to compare the content of spams and compute a similarity measure. Spam emails from one campaign have a high similarity score among each other and a low score with other emails. Within a campaign spam emails also share other features such as URL or email address. By combining fuzzy hash results and theses common features, we accurately cluster spam emails into campaigns.

An important contribution of the paper is a detailed analysis of long-duration botnet spam campaign identified in our dataset. The analyzed dataset is collected by a few mail accounts over three years. The major findings of our study include:

- Many spam campaigns are constantly appearing, and last for months.
- Most spam campaigns can be classified into two types; some campaigns are easily detectable as the

AINTEC'14, November 26–28, 2014, Chiang Mai, Thailand

http://dx.doi.org/10.1145/2684793.2684803

corresponding spam emails contain common identifiers (e.g. URL), while more sophisticated campaigns generate different contents to avoid simple spam detectors.

- Different spam campaigns may be initiated from the same botnet.

The rest of this paper is organized as follows. Section 2 discusses related work. In Section 3, we present our approach to extract features and cluster spam emails. Then we describe the results of our analysis in Section 4. Section 5 is the discussion on our study. Finally, we conclude in Section 6.

2. RELATED WORK

Techniques to cluster botnets can be divided into two categories [7]. The first category is to cluster botnets based on the inspection of captured malware [2]. The second category of techniques cluster botnets according to external traces such as flow data collected from a large network, DNS traces [9, 3], or traces of spam email [5]. In this paper, we fall into the second category, using spam email as the external data source to identify botnets[7]. One advantage of this approach is that spam emails are relatively easy to collect and inherently comprehensive. In comparison, DNS probing [9] requires extra queries to DNS servers which can limit the tracking capability.

Stealthy spammers tend to use low-volume spamming hosts (instead of high-volume hosts) hence the detection of spamming sources is made particularly difficult [5]. Previous works have proposed to learn and identify bots' behavior, for example, using spam campaign to identify a spamming botnet. The common approach is to determine identifiers characterizing the spam campaign, e.g., keywords, URLs, or particular contact information. For example. Pathak et al. [5] proposed to use URLs in spam email as the main characteristic to manually identify spam campaign and considered different URLs in the same campaign.

Another approach to characterize spam campaigns is to identify emails with similar content. Zhuang et al.[11] applied a shingling algorithm based on bi-gram to entire mail content to distinguish different spam campaigns in spam traces collected from Hotmail over a period of days to weeks. In this paper, we propose a similar approach by using fuzzy hash in order to compare the similarity between different spam belong to a spam campaign. The proposed method allows us to analyze three years of spam emails.

Previous studies also uncovered characteristics of spam campaigns and behavior of spammers. For example, Ramachandran et al. [8] studied relationship between spam emails and botnets. However, they do not infer botnets from spam data; Their work more focuses on the analysis of spammers at the network-level.

3. METHODOLOGY

The methodology proposed in this paper enables us to analyze a large dataset of raw spam emails collected over three years. The results are a list of clustered spam campaigns and their characteristics, such as size of content, distribution of campaign and spam sources.

The proposed method consists of three main steps:

1. **Cluster spam emails into campaigns.** To avoid spam detectors, spammers slightly alter the content of their spam emails over time. We refer to these groups of similar spam emails from the same spammer as spam campaigns. Figure 1 is the overview of the proposed procedure to cluster spam emails into campaigns. The first step in this procedure is to parse and extract features from spam emails. Next, we calculate the fuzzy hash of each spam content. The basic idea of fuzzy hash is to cut content into many slices and calculate hash value for each slice. Then merging all hashed slices into a single fuzzy hash value. Finally, comparing fuzzy hash values provides a similarity score of spam emails and permits to efficiently cluster spams into different campaigns.

2. **Analyze characteristics of spam campaign.** In a spam campaign, spammers deliberately send spam with varied content. The dynamics of the different spams is of prime importance to characterize and understand the different behaviors of spammers. Analyzing different characteristics of spam campaign, such as content sizes, content encoding types, activity time period, and SMTP path, helps us to identify the behavior of spammers. Also using these characteristics allows one to evaluate the probability of a spam message with new content to belong to previously detected campaigns.

3. **Infer botnets from sender IP address.** Spam campaigns are initiated by one or several botnets. Each spam message contains the IP address and timestamp of each SMTP server relaying the message. Using the complete route path of each spam email, we retrieve the source IPs and activity time of bots from large spamming botnets. Thereby, we can estimate the size of botnets corresponding to the spam campaigns identified in the first two steps. The characterization of botnets also permit to report botnets repeatedly used to send spam emails.

The reminder of this section provides details on the analyzed dataset and further describes the three steps of the proposed methodology.

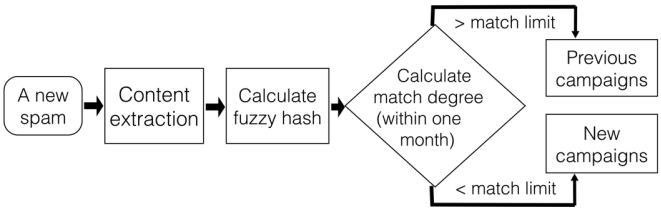

Figure 1: Overview of the identification procedure of spam campaigns

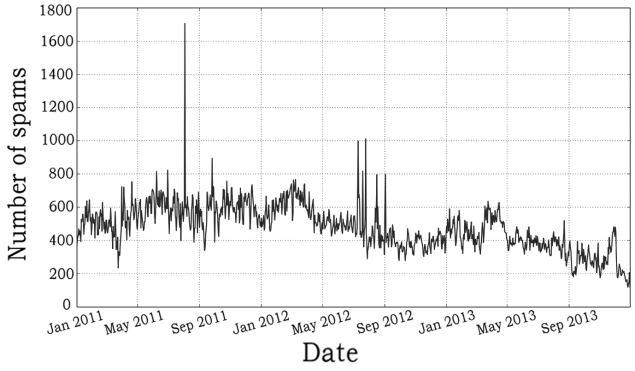

Figure 2: Number of spams from 2011 to 2013

3.1 Dataset

A spam email is an unsolicited email that is sent to many people for different purposes (e.g. phishing, advertisement, malware). In this paper, a spam email dataset is collected from a few email accounts in Japan. Emails are classified as spam manually by the user or automatically by dedicated software. The dataset consists of 540 thousand spam messages collected over three years from Jan. 2011 to Dec. 2013. Figure 2 depicts the number of spams in the analyzed dataset across time. The number of spam emails is rather stable over the years with 180K spam in 2011, 220K in 2012 and 150K in 2013. Overall the daily amount of spam is also stable with about 500 spams. This longitudinal spam dataset is appropriate to reveal the complete lifetime of spam campaigns and permits to analyze the trend of spam campaigns over three years.

Figure 3 shows the Cumulative Distribution Function (CDF) of the content size for the 540 thousand spam emails. As a pre-processing, encoded contents (e.g., base64) are decoded. We observe that only 10% of the emails contain less than 100 characters, nearly 75% of spam emails have content smaller than 2000 characters and 90% smaller than 4500 characters. From the data, we conclude that most spams are characterized by a small content size whereas some have a large size of content, mainly collapsed GB2312 encoded spams.

3.2 Extracting features

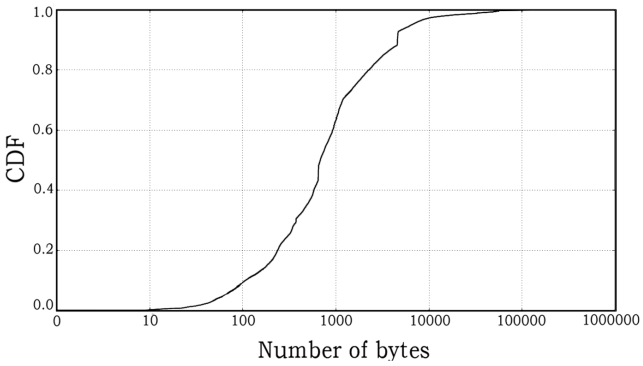

Figure 3: CDF of spam content size

To cluster raw emails in the dataset into different campaigns, we intend to extract features from the raw emails. Each raw-format spam message can be divided into three parts: header information, SMTP servers path, and content. Header information is used to identify the basic property of an email, including sender and receiver's email ID, sending date, title, and MUA (i.e., mail software). The sending date shows the detailed time when the email was sent though it is sometimes forged. To ease the comparison of different spams with distinct timezones, we convert all dates into Unix epoch time. Titles are important for spam, because they are the first piece of information seen by receiver, and might lure them into opening the spam. Spammers need to make their spam titles captivating, but also vary them enough to avoid easy detection. The SMTP servers path includes IP address and receiving time of each hop, hence it uncovers how a spam passes through the network from the sender to the receiver.

The content is the main part of an email, because it is the information exposed to the end-user. The content can consist of several objects. Each object includes content type, encoding type and content body. Content type and encoding type indicate the format of the content body. Indeed, the content body can be formatted in various way, for example plain text, HTML format or binary. As explained before, we extracted base64 encoded contents. We remove possible error and additional messages added by SMTP and ML servers and keep only the content sent by the spammers. This is the key data we inspect to decide if two spam emails are similar and from the same campaign.

In summary, the feature extraction step extracts main characteristics of each raw-format spam. This paper focuses on sending date, router path, spam title and content body, thus the feature extraction module produces a simplified dataset of spam email with only unix timestamp, router path, title, and content MD5.

3.3 Identifying Spam Campaigns

Table 1: Characteristics of common identifiers

Common feature	% of spams
Hyperlinks	72.3%
Email IDs	23.8%
Skype IDs	0.1%

Broadly speaking, we define a spam campaign to be a set of spam emails that are meant to achieve the same spamming purpose [5], for example, to advertise the same product. In this paper, we assume a spam campaign is generated from a single spammer. This assumption is motivated by the fact that previous work showed that each spammer sets up email templates when sending spam email [10]. To evade the increasingly sophisticated content-based spam filters, spammers typically obfuscate the content with more and more sophisticated manners. However, no matter how much the content is obfuscated, the message in a spam campaign has to share a set of "common identifiers" to achieve its initial purpose. It comes in various forms such as similar content [11]. Hyperlink is also widely used; it can be direct or redirect link to a target URL that contains detailed information for completing the spamming purpose. More recently, email IDs and Skype IDs also became main components for spammers [5]. Table 1 lists the characteristics of common identifiers in our dataset. 72% of spams contains at least one URL, thus the rest of them have no key for clustering. Also, email and Skype IDs are likely effective, but they are not always appeared in spams.

Spammers usually obfuscate the spam email content such that each spam content has slightly different text from the others. One common technique is altering frequent filtered words or inserting obfuscated words to evade detection. For HTML-based spam, spammers can insert legitimate invisible code or special characters. However, no matter how the content is altered, spam messages still share common characteristics. Consequently, we design a method to cluster similar spam emails into the same spam campaign.

The proposed algorithm to cluster similar spam emails must be robust to the different obfuscation methods of different spammers. As we discussed above, in a spam campaign, all spam content is similar because those spam emails share the same spamming purpose. Thus, we use context triggered piecewise hashes (CTPH) [4], also called fuzzy hashing, to cluster similar spam messages from the same campaign. Fuzzy hashing relies on two hash functions, a traditional hash function (e.g. MD5 or Fowler/Noll/Vo (FNV) hash) and a rolling hash function. The rolling hash slices the input into arbitrary size pieces that are hashed with the traditional hash function. The concatenation of the hashes from all pieces constitutes the fuzzy hash value of the given input (i.e. the signature of spam content in our case). Two distinct fuzzy hash values are compared using a weighted edit distance, the resulting score indicates the similarity of the two corresponding inputs. In our experiments, we employ the same fuzzy hash implementation as the one presented in [4]. The traditional hash function used is the Fowler/Noll/Vo (FNV) hash, the rolling hash is based on the Alder32 checksum and the similarity scores range between 0 and 100; 0 means that two spams are totally different and 100 means that they are identical. To compare spams with multiple contents (i.e. MIME multiparts), we define the similarity score of two multipart spams as the maximum similarity score between one part of each spam.

Consequently, we compare spams using computed fuzzy hash value. We consider each spam email as a node; for each new spam, we calculate the similarity by comparing the new spam fuzzy hash value with those for other nodes if their difference of received time is smaller than a threshold. We empirically set this to one month. It is a reasonable time period, as Pathak et al. [5] showed that spam campaigns are not bursty in nature and they continue on for months. The advantage of choosing a time window is that we can cluster spam into campaign more accurately and it improves the algorithm computational time. An arbitrary similarity threshold discriminates if two spams belong to the same spam campaign. To classify a new spam, we choose spams that have a similarity score higher than the similarity threshold value (called match degree) with this new spam. Then, from those spams we choose the spam with most common features (hyperlinks, email IDs and Skype IDs) with this new spam and cluster new spam into the same campaign the chosen spam belongs to.

Compared to previous work, the proposed approach can effectively and automatically clusters spam campaign. For example, with the URL-based approach proposed in [5] to cluster spam, the shortcoming is that the sent URL can be easily changed by spammers through redirect URL (e.g. URL shortening). Our approach is reasonable and accurate because of two facts. First, spam emails messages in a same campaign have similar content. Second, common features (hyperlinks, email IDs and Skype IDs) are appeared in nearly all spams.

3.4 Identifying campaign botnets

A botnet is a set of compromised hosts (called bots) that are administrated by a botmaster to involved one or more spam campaigns. In this work we infer botnets from the list of IP addresses accountable for a same spam campaign.

The extraction step of the previous section lists all SMTP server related information for each spam email message including the source server IP address, the destination server IP address, and the received date and

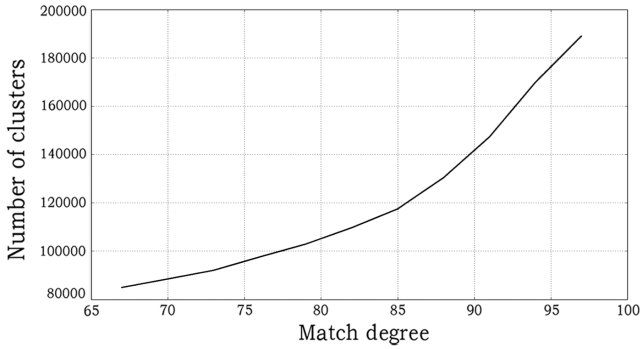

Figure 4: Dependency of match degree

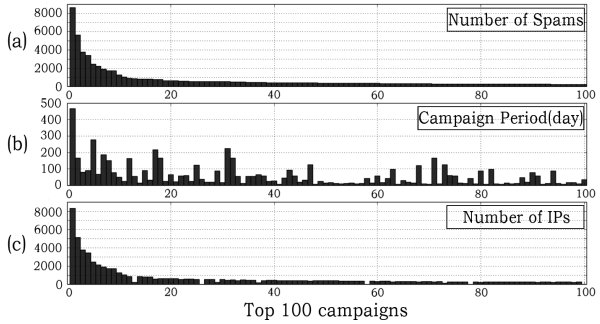

Figure 5: Number of spams, time period and number of IPs for top 100 campaigns

time. The path followed by the email is reconstructed by chaining the information reported by the different SMTP servers.

One technology for spammers to evade IP-based detection is sending spam through virtual network with legitimate IP which can be generated by software. Considering this, we select the IP sending the spam to the first mail server but not listed in a white list and use it to identify bot. This approach is robust for most obfuscated path. When identifying a spam campaign in the previous step, we have clustered all spam emails into different spam campaigns. By retrieving source IPs of all spams from the same campaign, we can cluster campaign botnets.

Multiple spam campaigns can be triggered by one botnet. Therefore, we merge individual spam campaign botnet into a set of botnets. If a large number of bots participate in both spam campaigns, we cluster such two spam campaign botnets into a large botnet.

4. RESULTS

Here, we present results on identification of spam campaign and analysis of detected spam campaigns.

4.1 Identification of spam campaigns

Before we cluster the spam campaigns, we first intend to investigate a dependency of the threshold parameter called match degree in fuzzy hashing on the distribution of size of spam campaign. Figure 4 shows the change of the number of clusters for different match degrees. X-axis is a threshold value of match degree that represents the similarity of two fuzzy hash, and y-axis is the number of obtained clusters. Only two spams with a similarity score more than match degree can be considered into the same campaign cluster. We observe that the number of clusters increases for a large match degree, as expected. It means that spam campaigns are divided into small pieces of clusters for a large value whereas they contains mixture of spam campaigns for a small value. However, we confirm a rapid change of the amount of clusters around 85 of match degree. Thus, we empirically choose this value as the threshold through this paper since such change is a sign of a change of structure of clusters. After applying the clustering technique to the spam emails, we obtained 118760 spam campaigns in 550K spam data. Actually, it is a huge number of campaign clusters. A reason of this is that our spam dataset is obtained from a few mail accounts; the collection is comprehensive and random.

To characterize spam campaigns we analyze top 100 campaigns in terms of the number of spams in a campaign. Those campaigns contain enough spam messages to extract relevant behavior of spam campaigns. Figure 5 demonstrates three main characteristics of top 100 campaigns: the number of spams, campaign periods, and the number of IP addresses.

In figure 5(a), x-axis indicates top 100 campaigns and y-axis shows the number of spams in a campaign. We can confirm two big spam campaigns with more than 5000 spams in our data, though the distribution is stretched. For the rest of campaigns, the number of spam is about 500 spams. Because those spams are collected by a few email accounts, it can be regarded as a biased and sampled data with all network spams. As a sampled data, however, we still obtained hundreds of spams in a campaign, meaning that they are still useful in quantifying spam campaigns.

4.2 Campaign lifetime

Investigating temporal behavior of spam campaigns, we plot active periods of top 35 campaigns in Figure 6; x-axis is the date when spam messages are sent, and y-axis represents the top 35 campaigns. Thus, a horizontal line indicates the lifetime of one spam campaign. From the distribution of the top 35 campaigns, we find that each campaign lasts for a few months, consistent with the past literature. Some separated dots are likely due to lack of the complete spam dataset. One notable point is that we confirm some synchronized campaigns such as pair of campaigns 19 and 26, or campaigns 22

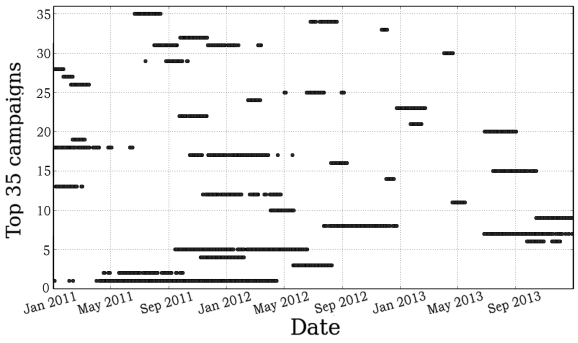

Figure 6: Time evolution of top 35 campaigns

and 32. They are different campaigns in terms of contents, but they can be temporally correlated, suggesting an existence of some intentions by spam originators or botnet owners. We will again discuss this in section 4.4 where we investigate the source IP addresses.

In order to quantify lifetime of spam campaign, we define the duration of spam campaign as the time between the first spam email message and the last one appeared in a campaign. Figure 5(b) represents the distribution of lifetime for top 100 spam campaigns. It is clear that lifetime for different campaign is quite different. A long lifetime campaign can last around a year such as campaigns 1 and 5. However, short lifetime campaigns are more usual and last less than three months such as campaigns 3 and 4. Also, we observe a long lifetime in small campaigns like campaigns 50-100, suggesting that a campaign with a larger number of spams does not always mean a longer lasting campaign, i.e., lifetime and density of spams are orthogonal.

4.3 Campaign features

In order to quantify identified spam campaigns, we focus on three spam features: MD5, title, and URL.

Figure 7 (a), (b) and (c) shows the distribution of the number of unique MD5s, the number of unique titles, and the number of unique URLs, respectively. The distribution of MD5s highlights a strong contrast between two different distributions: One is characterized by thousands of MD5s. The other one only has a few MD5s. The first category corresponds to a campaign where the contents of all spams are slightly different from each other. Note that the fuzzy hash would classify two contents that are largely different into two different campaigns. Comparing to the results in Figure 5(a), we can find that the number of MD5s is almost double to the number of spams for certain campaigns. For example, campaign 6 only consists of 2222 spams, however it has unique 4444 MD5s. Those spams have multipart contents (e.g. text, html and image) and a MD5 is computed for each content. For the second category, the number of MD5s is quite small. Spam in such campaign usually shares the same content body.

Next, as shown in Figure 7(b), most of campaigns are characterized by a wide variety of titles to alter since title of spam is easy to be detected. However, we still can find that the number of titles is much smaller than the number of spams in campaign. This is a natural consequence that spammer has an intention to send spams; a randomly generated title is not appealing to end-users.

Now, we focus on the distribution of URLs in top 100 campaigns in Figure 7(c). URLs are the most essential and common feature to characterize campaigns because they point contents that spammers intend to appeal. It is clear that many campaigns share only a few URLs; changing websites or images frequently is a laborious task for spammers. Even though most URLs are redirection links, spammers can not alter URLs in a large scale dataset. However, we also observe some campaigns with a large number of URLs. In this case, one URL is shared by a few number of spams. For those spam campaigns, URLs do not point to the same website. Those URLs have different spamming purposes. Thus, botnets need to work frequently to generate new content message.

For further investigation of big spam campaigns, we list the details of top 15 campaigns in Table 2. As previously explained, in campaign 6, the number of MD5s is twice as many as the number of spams. In contrast, campaigns 3, 4 and 5 only have respectively 6, 8 and 6 different MD5s, meaning that most of spams share the same message body. Results are similar with the number of URLs; A few number of spams share the same URL in the former campaigns, while hundreds of spams have the same URL in the latter. In campaign 7 each spam has a different MD5 (thus a different content) but they all share 3 unique URLs.

By manually inspecting spam content bodies in detail, we found two campaign types. First type consists of spam formatted with certain template (e.g. HTML template). Spammers do not send specific URLs, on the contrary, they send a large number of different URLs, and the purpose of these spam campaigns is obfuscated such as in campaigns 1 and 2. Other feature of this campaign type is the content body that usually includes multiple parts, so these campaigns contain much more spamming information. Accordingly to the change of URLs, their titles are also frequently altered. In this way, by changing URLs and titles constantly, campaigns can effectively evade spam detection. Such campaigns are usually lasting longer than other types. Because the template usually conveys no information about the spammers purposes, one can strip spams of the template, then reiterate the proposed clustering algorithm to get more insights for this type of campaign. This kind of hierarchical clustering is but left for future works.

The second type of campaign is based on certain iden-

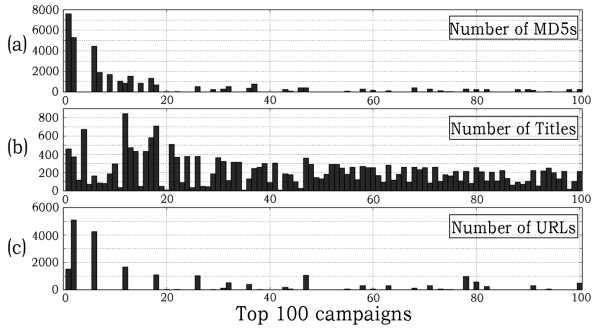

Figure 7: Characteristics of top 100 campaigns

tifiers (e.g. URL or email) and is the most common campaign type. Such campaign has a specified purpose, for example to show an advertisement. For example, in campaigns 3, 4, and 5 there are many spams sharing the same MD5s and same URLs. Campaigns 7, 9 and 11 spams have different MD5 but spams in these campaigns share the same URLs, meaning that within this campaigns the same URL is advertised for a long time, but contents or titles are altered to evade the detection. Campaign 6 and 13 are other examples of this type of campaign but where the common identifiers are email addresses. Another characteristic of this campaign type is that the average message size is much smaller than the campaign using templates. Spams in this campaign type usually highlight specific identifiers in a concise message. Overall the lifetime of these campaigns is usually smaller since the spamming purpose has a short effectiveness.

4.4 Botnet inference

In a spam campaign, spams share similar contents. Considering the fact that the number of distinct subnets (/24) is almost equal to the number of IPs in a campaign as shown in Table 2, it is natural to consider that campaigns are mainly invoked by botnet rather than by spammer with own pooled IP addresses. However, it is possible that one spammer can use multiple botnets, or multiple spammers use the same botnet. In order to quantify these botnet behaviors, we intend to compare an overlap of sender's IP addresses between two campaigns. Figure 5(c) demonstrates the distribution of suspicious unique IP addresses belonging to top 100 campaigns. It is clear that a large number of IPs appear in spam campaigns, as expected. Compared to Figure 5(a), the number of IPs is positively correlated to the number of spams. A large spam campaign requires an ability to send many spams, however, it is better that the workload of senders should be low to evade campaign detection. Thus, it is plausible that the number of spams per IP is relatively small.

Next, to check the overlap of IP addresses among

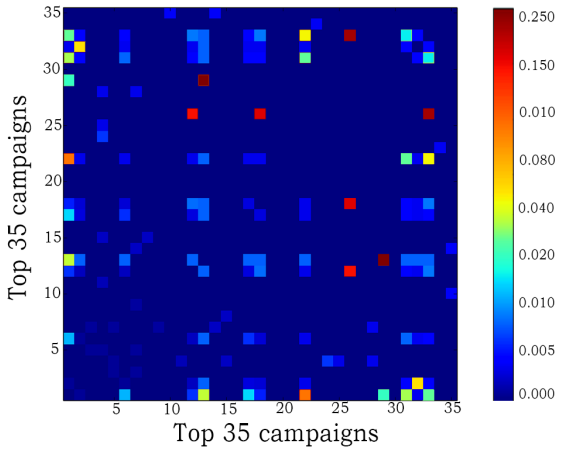

Figure 8: Proportion of shared bots between top 35 campaigns

campaigns, we calculate the proportion of shared IPs between two given sets of IP addresses. For campaign A and campaign B, we define the proportion of shared IPs as $\frac{\#IPs(A \cap B)}{\#minIPs(A,B)}$. Figure 8 illustrates a heatmap of the proportion of shared IPs for top 35 campaigns. Red color indicates more overlap though blue color represents less overlap. We confirm that some pairs of campaigns such as campaigns 12 and 26, campaigns 19 and 26, and campaigns 26 and 33 share more than 20% of their IP addresses. Recall also that campaigns 19 and 26 are synchronized in lifetime (see Figure 6), we can conclude that they are different spam campaigns in contents, but they share the same infrastructure to deliver spams. On the contrary, the IP overlap between campaigns 22 and 32 is almost zero though they are temporally correlated. They are different campaigns in contents but with the similar active period, and their infrastructures are also totally different.

5. DISCUSSION

In our present study, fuzzy hashing is an effective method to identify spam content message with a hash value. Comparing with a shingling algorithm with bigram of characters, the computation cost is low because the method relies on a simple hash, although it still requires pair-wise comparison among two hashes to obtain a similarity score.

Spam campaigns usually consist of a large number of spams/hosts in a limited lifetime. According to the distribution of campaigns, we find that spam campaigns are not bursty in volume and usually last for months. However, some special campaigns may work in a short time, just about one week.

As the results, we differentiate two different campaign types: one is characterized by templates to send a large

Table 2: Detailed characteristics of the top 15 campaigns

Camp. No	Nb. Spam	Nb. IPs	Nb. Subnets	Nb. MD5	Period(days)	Nb. Titles	Nb URLs	Nb. Emails
1	8630	8281	7959	7570	467	462	1337	0
2	5640	5107	4683	5304	166	374	5074	0
3	3760	3756	3753	6	80	113	5	0
4	3419	3419	3418	8	91	673	6	0
5	2454	2454	2454	6	278	71	6	0
6	2222	2142	2115	4444	69	169	2261	122
7	1909	1909	1909	1909	186	92	3	0
8	1715	1714	1714	4	152	89	2	0
9	1706	1705	1704	1706	78	191	3	0
10	1276	1276	1276	2	49	298	3	0
11	1031	1030	1030	1031	25	34	5	0
12	917	864	856	829	164	864	1399	10
13	850	225	213	1550	56	483	0	198
14	838	838	838	1	16	432	2	0
15	828	828	827	828	91	37	2	0

number of different spam email messages. The other one contains a simple URL to achieve a spamming purpose whose main message in the content is constant. We also find that different campaigns may use the same infrastructure to deliver spam messages. From the proportion of shared IPs, we can infer if two campaigns work at the same botnet.

As our dataset is collected from a few mail accounts, the results can be biased. However, we emphasize that detected campaigns consecutively send spams to them. Thus, without a large amount of accounts, still it is possible to identify existence of campaign, though correctly estimating an impact of the campaign may be difficult. However, considering the results by our longitudinal spam dataset, we believe that those findings are also appeared without lack of generality.

6. CONCLUSION

In this paper, we presented an analysis of spam campaign that focuses on the main characteristics of campaign behavior. The main idea of clustering relies on a technique called fazzy hashing that enables us to cluster similar spam contents as a group. We applied the method to 550K spams arriving at a few email addresses over three years, and characterized longitudinal spam campaign behavior. As a detailed analysis, we focus on the top 100 campaigns in terms of the number of spams, and quantified the size and period of spam campaigns, and their features. We extracted two types of spam campaigns from our analysis: template based complicated campaigns and URL based simple campaigns. Considering a possibility that more sophisticated spam campaign appears in future, we believe that tracking spam campaign behavior is an important and appropriate approach for spam campaign detection.

As a future work, we plan to extend our method to support online update and detection.

7. ACKNOWLEDGMENTS

The authors thank the NII Internship program. Also, this research has been supported by the Strategic International Collaborative R&D Promotion Program of the Ministry of Internal Affairs and Communication, Japan, and by the European Union Seventh Framework Programme (FP7/2007-2014) under grant agreement No. 608533 (NECOMA).

8. REFERENCES

[1] Messaging Anti-Abuse Working Group (MAAWG), Email metrics program: The network operators' perspective. Report #15 - first, second and third quarter 2011. Nov. 2011.

[2] D. Dagon, C. C. Zou, and W. Lee. Modeling botnet propagation using time zones. In *NDSS 2006*, pages 1–15, San Diego, CA, Feb 2006.

[3] N. Jiang, J. C. Y. Jin, L. Li, and Z.-L. Zhang. Identifying suspicious activities through dns failure graph analysis. In *ICNP 2010*, pages 144–153, Vancouver, Canada, Oct 2010.

[4] J. Kornblum. Identifying almost identical files using context triggered piecewise hashing. *Disital Investigation*, 3:91–97, 2006.

[5] A. Pathak, F. Qian, Y. C. Hu, Z. M. Mao, and S. Ranjan. Botnet spam campaigns can be long lasting: evidence, implications, and analysis. In *SIGMETRICS 2009*, pages 13–24, Seattle, WA, Jun 2009.

[6] C. Porter. *Email Security with Cisco IronPort*. Networking Technology: Security Series. Cisco Press, 2012.

[7] M. A. Rajab, J. Zarfoss, F. Monrose, and A. Terzis. My botnet is bigger than yours (maybe, better than yours). In *HotBots 2007*, pages 1–5, Cambridge, MA, Apr 2007.

[8] A. Ramachandran and N. Feamster. Understanding the network-level behavior of spammers. In *SIGCOMM 2006*, pages 291–302, Pisa, Italy, Aug 2006.

[9] A. Ramachandran, N. Feamster, and D. Dagon. Revealing botnet membership using DNSBL counter-intelligence. In *SRUTI 2006*, pages 1–8, San Jose, CA, Jul 2006.

[10] B. Stone-Gross, T. Holz, G. Stringhini, and G. Vigna. The underground economy of spam: A botmaster's perspective of coordinating large-scale spam campaigns. In *LEET 2011*, pages 1–8, Boston, MA, Mar 2011.

[11] L. Zhuang, J. Dunagan, D. R. Simon, H. J. Wang, I. Osipkov, G. Hulten, and J. D. Tygar. Characterizing botnets from email spam records. In *LEET 2008*, pages 1–9, San Francisco, CA, Apr 2008.