Shades of Gray: A Closer Look at Emails in the Gray Area

Jelena Isacenkova
Eurecom
Sophia Antipolis
06410 France
jelena.isacenkova@gmail.com

Davide Balzarotti
Eurecom
Sophia Antipolis
06410 France
balzarotti@eurecom.fr

ABSTRACT

Every day, millions of users spend a considerable amount of time browsing through the messages in their spam folders. With newsletters and automated notifications responsible for 42% of the messages in the user's inboxes, inevitably some important emails get misclassified as spam. Unfortunately, users are often unable to take security related decisions, and tools provide no assistance to easily distinguish harmless commercial messages from the ones that are most certainly malevolent.

Most of the previous studies focused on the detection of spam. Instead, in this paper we look into the often overlooked area of gray emails, i.e., those messages that cannot be clearly categorized one way or the other by automated spam filters. In particular, we analyze real-world emails by grouping them into clusters of bulk email campaigns. Our approach is able to automatically classify and reduce by half the gray emails area with only 0.2% false positives.

Moreover, we identify a number of campaign features that can be used to predict the campaign category and we discuss their effectiveness and their limitations. Our experiments show that a large fraction of emails in the gray area are composed of legitimate bulk emails: newsletters, notifications, and marketing offers. The latter appears to be a large e-marketing business industry that has grown into a complex infrastructure for sending legitimate bulk emails. To the best of our knowledge, this is the first real-world empirical study of such emails.

Keywords

Gray emails, classification, botnet generated campaigns, commercial campaigns, Nigerian scam, newsletters, phishing, challenge response system

1. INTRODUCTION

Nowadays, many antispam filters provide a good level of protection against large-scale unsolicited email campaigns. However, as spammers have improved their techniques to increase the chances of reaching their targets, also antispam solutions have become more aggressive in flagging suspicious emails.

On one side, this arms race has lead to a steady increase in the detection rate. On the other, it also contributed to the increase of the false positives, with serious consequences for the users whenever an important message is erroneously flagged as spam. Moreover, at the border between legitimate user emails and spam lies a gray area of messages that are hard to automatically classify. This area often contains newsletters and commercial offers which were originally solicited, but that are not anymore interesting for the users [9]. More in general, it includes messages that are not flagged by traditional antispam filters, but that are not necessarily wanted by the users. In 2012, Hotmail estimated that gray emails were the source of 75% of all spam complaints. Another Email Intelligence Report published by ReturnPath [28] pointed out that 16% of emails containing advertisements or marketing information are normally flagged as spam and, therefore, never reach user mailboxes. At first glance, many people would consider this "side-effect" as an advantage. However, it has been estimated that only one-third of users consider such messages as spam, while two-thirds prefer to receive unsolicited commercial emails from already known senders [7]. A more recent report shows that despite the mailboxes being overloaded, consumers still read 18% of subscribed marketing emails, and continue to sign up for email offers and mailing lists [28], with the result that newsletters and automated notifications sum up up to 42% of inbox messages. For these reasons, it is a well known fact that most of the users regularly check their spam folder to verify that no important messages have been misclassified by the antispam filter.

Unfortunately, this process is very time-consuming. Antispam solutions are not very helpful in this direction, and do not usually provide any additional information to help users in quickly identifying marketing emails, newsletters, or "borderline" cases that may be interesting for the users.

Even worse, when users skim through their spam messages looking for something that looks legitimate, they need to take a decision on which email can be trusted, which one is just annoying, and which can pose a real security threat. Unfortunately, several studies showed that most users are very bad in taking these kind of security-related decisions [19], and this is one of the reasons why we need automated spam filters in the first place. For example, a recent survey conducted in 2010 [17] by the Messaging Anti-Abuse Working Group reported that 57% of the people who have accessed spam messages admitted to have done so intentionally, be-

ASIA CCS'14, June 4–6, 2014, Kyoto, Japan.

http://dx.doi.org/10.1145/2590296.2590344 .

cause they were unsure whether the suspicious message was spam or not.

While most of the existing research deals with the problem of efficiently and accurately distinguishing spam from ham, in this paper we focus on the thin line that separates the two categories. In particular, we limit our study to the often overlooked area of gray emails [31], i.e., those ambiguous messages that cannot be clearly categorized one way or the other by automated spam filters. We start from the assumption that spam filters are good in detecting most of the spam, and if the filter has "good reasons" to believe that a message is unsolicited or that it contains malicious content (e.g., by employing an antivirus, a black list, or by matching a signature of a known scam message), there would be no reason for most users to double-check that decision.

We start our study by analyzing a real deployment of a challenge-response antispam solution to measure the extent of this gray area. We use system's quarantined emails that already exclude the majority of the ham and spam messages as an approximation of the gray emails category. According to our data, after the obvious spam and ham emails have been eliminated, users still manually check on average five to six messages per day. On average, 1.5% of these messages have an attachment with 9% of them being malicious. However, some of these messages also contain interesting content, as proved by the fact that users read and whitelist an average of 1.5 messages per day. We also confirm the belief that ordinary users are not very good in telling spam and ham apart.

Under these premises, we analyze the messages in the gray area in order to improve our understanding about them and the reasons that make them difficult to categorize. In particular, we adopt a three-phase approach based on message clustering, classification, and graph-based refinement. Extracted email features are applied in a context of email campaigns instead of individual emails. Our technique is able to automatically classify half of the gray emails, corresponding to 15% of all email traffic, with only 0.2% of false positives. Moreover, our results show a number of interesting characteristics of commercial marketing campaigns, which constitute to a large fraction of the gray area. To the best of our knowledge, this is the first empirical study of legitimate bulk emails.

2. BACKGROUND

Email-based marketing is a common practice for advertisement and sales, and it is used both to maintain communication with current customers, as well as to acquire new ones. Unfortunately, when users' inboxes started to get overloaded with various types of bulk messages, mailbox maintenance became highly time-consuming. As a result, email filters were introduced to protect users from unsolicited messages, starting a multi-million anti-spam protection industry and a battle that is far from being over. In fact, even though direct mail marketing has a higher response rate (3.4%) than email marketing (0.12%) [8], the low cost of emails still makes electronic messages a very attractive solution.

Marketers often use professional marketing tools in order to maximize their campaign delivery rates. These tools help to clean non-existing emails from customer lists, to deal with recipient complaints, to avoid hitting spam traps, and even provide detailed campaign delivery statistics. Today running an email marketing campaign is a complex operation, and more and more solicited bulk emails fall into the recipient spam folders. In fact, these folders often contain messages that cannot be clearly categorized by automated spam filters. This gray area is responsible for 75% of the spam complaints [9] and it contains both legitimate and harmless bulk emails, and malicious messages that can result in a computer infection or in stolen personal data. However, users also appear to be ineffective in distinguishing one class from the other [17, 19] and often (70%) take their decision based only on the sender field and subject line.

For clarity, in this study we separate bulk emails into two categories: legitimate and spam. The first includes solicited (subscribed newsletters and notifications) or *potentially* solicited (advertisements) messages sent according to legal regulations (e.g. the CAN-SPAM Act [1] and the E-Privacy Directive [3]). The second category contains instead unsolicited, malicious, or illegal promotion emails.

The distribution of legitimate marketing campaigns became a large business where specialized companies provide as a service professional email marketing tools, and also sell categorized email lists for marketers looking for new clients. The collection of such lists is legal: when users subscribe to some services and fill out a form, they might – by choice or by default – agree to share their information with the third-parties. Hence, at some point users do agree to receive the advertisements.

Related work

Many filtering solutions, often used in combination with each other, exist to detect and mitigate spam.

In our approach we focus on the analysis of the senders behavior. Pathak et al. [20] suggested to analyze the sending behavior of spammers, while Ramachandran et al. [27, 26] used behavioral blacklists to classify sender IP addresses based on their behavior. Ramachandran observed that spammers exhibit recognizable sending patterns, based on which behavior fingerprints can be build. Qian et al. [25] proposed to rely on the reputation of IP clusters, e.g. BGP clusters, and combine it with DNS information, improving the precision of public IP-based blacklists by 50%. Hao et al. [10] built an automated reputation engine, called SNARE, aiming at distinguishing legitimate senders from spammers based on a number of non-content email features. West et al. [30] built a reputation model to predict the behavior of spammers. The model relies on spatial and temporal features that can be especially useful in partial-knowledge situations. The model managed to classify up to 50% of the spam emails that were not identified by the blacklists. The advantage of such network-level detection techniques is that they tend to react faster to spam campaigns than typical blacklisting services.

The study closest to our research was performed by Qian et al. [24], where the authors proposed a content-based unsupervised email campaign clustering algorithm, and also recognized the problem of classifying legitimate campaigns in a real-world dataset. In particular, they tried to filter out legitimate bulk emails by using certain keywords and a threshold of IPs per campaign. While the latter can be efficient when dealing with campaigns sent by botnets, it is ineffective against other malevolent campaigns sent from webmail accounts. We demonstrate in our study that such campaigns tend to mimic the traits of legitimate campaigns and are difficult to identify based only on sender characteristics.

To our knowledge, there are no known studies performed on legitimate bulk emails, and only few have studied the gray emails phenomenon [32, 31, 6]. Yih et al. [31] argued that filtering gray emails with even an optimal spam filter is a very difficult task. Therefore, the authors proposed to treat gray emails separately and rely on user feedback to label messages. Their experiments, performed on a dataset that is similar to the one we used in our study, showed that classifying emails on per-campaign basis yielded a higher precision and data coverage compared to a per-email treatment. However, the email or campaign class may depend on the user [6, 7]. Therefore, Chang et al. [6] studied how to combine user feedback with user preferences to improve the classification results. Youn et al. [32] proposed an ontology-based technique to provide personalized gray email filtering based on user behavior. Although we agree that the personalization of gray email is crucial, our results also suggest that user feedback might be unreliable for class prediction.

As spam is primarily sent in bulk emails, many studies try to identify it through the analysis of bulk email campaigns ([16, 21, 29, 13]). Kanich et al. [13] studied spam campaigns by infiltrating a botnet and evaluated their conversion rate from a marketing perspective. Clustering spam emails by URLs and their redirections was first proposed by Li et al. [16]. Pathak et al. [21] tried to cluster spam campaigns using URLs, but it proved to be a challenging task due to URL obfuscation. Thomas et al. [29] confirmed the problem and proposed a new technique to filter URLs in real-time. Finally, Qian et al. [24] identified email campaigns based on their content similarity and Pitsillidis et al. [23] proposed to automatically extract spam campaign templates from regular expressions extracted from the messages.

As we mentioned at the beginning, most of the previous work on the field is focused on identifying spam and its campaigns. In this paper, we exclude most of the spam and legitimate messages and focus instead on the borderline area between them.

3. METHODOLOGY

This section presents the dataset we used in our experiments and the techniques we adopted to process and analyze the email messages. Since it would be impossible to classify each email in isolation, we adopted a multi-layered approach to group them into similar campaigns (a solution proven to be effective by several previous studies [31, 24, 21]). In particular, we start by clustering them based on the email headers. We then extract a set of features based on a number of campaign attributes and we use them to train a classifier in order to predict the campaign class. Finally, we employ a graph-based refinement technique to further increase the coverage and precision of our classification.

3.1 Data Collection

The amount and diversity of the available data is crucial in order to successfully identify email campaigns. Messages should be collected from multiple feeds, cover numerous recipients, several organizations, and for a long period of time [21, 22]. Our email dataset fulfills these requirements as it was collected from a commercial Challenge-Response (CR) spam system deployed in tens of different organizations. A CR filter is a software that automatically replies with a challenge (in our case a CAPTCHA) to any previously-unknown sender of incoming emails. If the sender solves the challenge, the message is delivered to the recipient and the sender is added to a whitelist; if not, it remains in a quarantined folder, where its recipient can manually view and whitelist/blacklist it. Since in our study we want to focus on the borderline area that contains the emails that cannot be easily classified as legitimate or spam, we installed a sensor in the CR system to intercept any quarantined message. These emails have successfully passed through a number of traditional antispam filters including virus scanners, reverse DNS, and DNS blacklisting verification. Moreover, users never had any previous conversation with the sender. Therefore, we can consider this dataset as pre-filtered from obvious legitimate and spam emails.

Sometimes this set is referred to as a gray zone [6] that stores emails of uncertain class. Email categories often found in this group include traditional spam and scam messages, automated notifications, newsletters, and commercial offers. Due this variety, users need to manually check these messages from time to time looking for any interesting or missing email.

We also instrumented the CR-system to collect additional information (see Table 1): opened emails by the users, and whitelisted messages (thus showing that the user manually classified them as legitimate). This provides insights on the users ability to distinguish harmless from harmful messages. Finally, our sensor collected the delivery status information, e.g. sent, bounced, and delivered, for each challenge email sent back by the CR system.

In our experiments we relied on statistical email data that we collected from companies of different sizes. The monitoring period covered 6 months, from August 2011 to January 2012. During this period around 11 million messages were delivered to the monitored mail servers (Table 1). 29.4% of them belonged to the class of gray messages. To protect the privacy of both the users and the companies involved in the study, the data we used in our experiments did not include the email bodies, and the headers were sanitized and analyzed in an aggregated form.

3.2 Email Clustering

The task of grouping emails into campaigns has already been covered by several previous studies ([15, 23, 16, 24, 21]). Previous results were very successful in identifying email campaigns, but, unfortunately, often relied on the content of the email body. Our dataset is limited to the email headers, thus forcing us to use a different approach based only on the email subjects. The main limitation of this technique is that the email subjects have to be long enough to minimize the chances of matching different messages by coincidence.

The obvious solution for grouping similar subjects would be to apply some text mining algorithm, but our input text is short and it is important to preserve the word order. Hence, we decided to use a simple approach based on "almost exact" text matching, extended to include subjects with a variable part. The latter could be a varying phrase in the subject, including random words, identifiers, or user names. We use word n-grams of a decreasing length (between 70 and 8), with a sliding window that permits to skip over varying parts of the subjects. Our implementation is based on an existing n-grams extraction library (Ngram Statistics Package [4]), a standard list of stop-words, and a number of custom scripts to match the extracted n-grams and assign them to clusters.

Table 1: General statistics

Mail servers	13	White emails	2,806,415	Challenges solved	166,279
Active users	10,025	Black emails	5,066,141	Users whitelisted emails	42,384
Total messages	11,203,905	Gray emails	3,331,349	Users viewed emails	104,273

Table 2: Cluster features

Group A	
Sender IPs	Distribution of network prefixes (/24)
Sender names	Distribution of email sender names
Sender add.domain	Distribution of email domain names
Sender add.prefix	Distribution of email prefixes
Group B	
Rejections	Percentage of rejected emails at MTA
White emails	Percentage of whitelisted emails
Challenges bounced	Percentage of bounced challenges
CAPTCHAs solved	Percentage of solved challenges
Unsubscribe header	Percentage of Unsubscribe headers
Group C	
Number of recipients per email	Normalized number of unique recipients per email
Recipient's header	Location of recipient's email: To/Cc/Bcc/Mixed
Countries	Distribution of countries based on originating IPs

The process starts by searching for the longest n-gram (70) and then decreasing the length until enough similar matches (with a threshold of 30 emails per cluster) are found to create a cluster. This algorithm is efficient on long subjects but problematic on short ones, thus limiting our analysis to subjects containing at least 10 characters and 3 words. In this phase we successfully clustered 50% of all emails in 12,250 clusters. Cluster sizes varied between 30 and 8,468 messages.

3.3 Feature-based Classification

To be able to differentiate and classify the identified clusters, we extract a set of eleven features grouped in three categories (see Table 2).

Group A: Features in this group reflect the similarity of a certain feature inside a cluster. The values are expressed as a range between 0 and 1, where 0 indicates a high distribution (low data similarity) and 1 indicates a low distribution (high data similarity) in the cluster. The feature similarity is defined as:

$$a(C) = 1 - u/t$$

where u is the number of unique or similar feature values, and t is the number of total emails. This group contains four features measuring the similarity of sender IP prefixes and email addresses, and the similarity of the sender names. In particular, we split the email domain address into two parts: the *email prefix* and the *email suffix*. The suffixes are grouped by removing numerical differences (e.g., between `abc10.com` and `abc22.com`). When similar suffixes are found, they are merged until there are no similar values left. *Email prefixes* are instead compared using a variation of the Levenshtein distance algorithm in which a threshold is computed based on the length of the email prefix itself. In this way, the similarity score is normalized to account for the fact that, for example, a two-chars difference for short strings is somehow equivalent to a six-chars difference for longer ones.

Group B: Features of this group reflect the percentage of messages in a cluster that have a certain feature value. There are five features in this group: *CAPTCHA solved*, *rejections*, *white emails*, *challenges bounced*, and *unsubscribe header*. The first measures the percentage of challenges that were solved by the senders. The *challenges bounced* are instead emails not delivered because the recipient did not exist, or did not accept emails from the sender. Whenever an email was sent to multiple recipients, we were also able to compute the percentage of *white emails* (i.e., the percentage of recipients that had already whitelisted the sender) and the percentage of incoming email *rejections* (i.e., the percentage of recipients that were rejected by the Mail Transfer Agent - normally because the corresponding addresses did not exist on the server). Finally, the *unsubscribe header* feature evaluates the percentage of emails that contained the unsubscribe header. The latter is generally used by commercial messages and notifications providing the users an option to unsubscribe from the list.

Group C: Features in this groups are computed in different ways. *Recipients per email* estimates the average number of recipients per email. The *Recipient's header* feature indicates the location of email recipient address in the email headers: *To*, *Cc*, *Bcc*, or *Mixed* when multiple locations are used in the same campaign. Finally, the *countries* feature reflects the number of countries (based on the sender IP geolocation) in the cluster.

Manual Labeling

Before performing our classification, we need to build a training set. Obviously, the result of our manual labeling process depends on the actual definition of spam that we adopt in our experiments. By definition, spam is an unsolicited email that is usually sent in bulk. However, there is no reliable way to verify if a certain email is solicited, i.e., if the recipient has subscribed to it or not. Moreover, the notion of spam is somehow subjective and it may not be the same for all the users. Some commercial campaigns are probably unsolicited, and therefore *could be* considered as spam. However, when such emails are sent by professional marketing companies according to the country regulations, it becomes unclear how they should be treated by antispam filters. This is also the main reason why they are considered as gray emails in the first place.

In this paper we take a conservative approach, and flag as spam only campaigns with potentially illegal content that may involve malicious, fraudulent or illegal online activities. This includes different "business models": illegal product sellers, malware spreading emails, personal data and credential thieves, or advanced fee fraud specialists. Finally, we consider any email belonging to a commercial marketing campaign as legitimate (in the sense that general antispam

filters should not block them, unless they are specifically instructed to do so by the user).

Although email labeling might be difficult even with the full email content, it can be facilitated by enriching emails with aggregated campaign features. All the campaign features are stored and viewed in an aggregated form, thus never providing access to any distinct email information. A particular case is represented by the *email subject*, a textual information that would be difficult to aggregate without textual data. As we group emails based on subject similarity, we also keep an aggregated copy of the campaign subject.

During the sampling, we relied on the domain knowledge of the analyst and on the additional information (e.g., average number of recipient per email, and number of originating countries), that would not be available to a user reading only one message at a time. Often a subject is enough to make a labeling decision, but in cases when it is not, aggregated header information is used by the analyst. For example, if the message subject resembles a private communication but the email has been sent in 50 identical copies to different recipient, this is more likely to be a scam than a real personal message. In the same way, a message promoting a new product or services online, sent in thousands of copies from over 30 different countries and with multiple recipient per email is probably an illegitimate campaign.

To build the training set, we randomly selected 2,000 campaigns and performed a manual labeling of them. We labeled 1,581 (79%) as legitimate and 419 (21%) as spam campaigns. This preliminary classification confirms that the majority of spam was already filtered out from the gray dataset.

Classification

Using the eleven features presented above, we trained a binary classifier. To select a classifier we referred to the results presented by Kiran et al. [14], in which the authors demonstrated that, on spam datasets, ensemble classifiers perform better than single classifiers. Based on this conclusion, for our classification task we decided to use a supervised Random Forest ensemble classifier.

We first performed a cross validation test in which we randomly split the sampled data into two groups including respectively 70% and 30% of the data. We then trained the Random Forest classifier (configured with 500 trees and three random variables per split) on the first group, and we tested the extracted model on the second one. For each cluster, the algorithm returned a score ranging between -1 (for spam) and 1 (for legitimate). A score close to zero indicates that the classifier was uncertain about the sample.

Since our set includes classes of very different sizes, we use the Matthews Correlation Coefficient (MMC) to measure the classification quality. Our model achieved MCC of 0.75, where the value is between [-1,..1], and 1 represents a perfect prediction. The model produced 0.9% false positives (i.e., legitimate campaigns being misclassified as spam) and 10% false negatives (i.e., spam being misclassified as legitimate). These rates suggest that the set of attributes we identified are effective in separating the two types of campaigns. We also noticed that while our classifier identified legitimate campaigns well, it had a higher probability of misclassifying spam campaigns. A further interpretation of this phenomenon is described in section 5.

Finally, we applied the model extracted from our training set to predict the classification of the remaining unlabeled campaigns. Results are presented in Table 3.

Table 3: Campaign classification results

Campaign type	Manual sampling	%	Unlabeled	%
Legitimate	1,581	79%	8,398	81.9%
Spam	419	21%	1,852	18.1%
Total	2,000		10,250	

Table 4: Attribute values per campaign category

Attribute	Legitimate	Spam	Gray
		Min / Mean / Max	
Countries	1 - 1.2 - 6	7 - 29 - 123	1 - 5 - 80
IPs	0.13 - 0.9 - 1	0 - 0.06 - 0.82	0 - 0.7 - 1
Sender email domain	0.2 - 0.98 - 1	0 - 0.3 - 1	0 - 0.85 - 1
Sender email prefix	0.03 - 0.98 - 1	0 - 0.09 - 1	0 - 0.81 - 1
Senders	0 - 0.98 - 1	0 - 0.3 - 1	0 - 0.8 - 1
Unsubscribe header	0 - 0.5 - 1	0 - 0 - 0.3	0 - 0.3 - 1
Bounced	0 - 0 - 1	0 - 0.1 - 1	0 - 0.1 - 0.9
CAPTCHAs	0 - 0 - 1	0 - 0 - 1	0 - 0.1 - 1
White emails	0	0	0.001
Rejections	0 - 0 - 0.4	0 - 0.23 - 1	0 - 0.1 - 0.7
Rec.per email	1 - 1 - 1.1	1 - 3 - 16	1 - 1.1 - 8
Recipient header	*To*, *Bcc*, *Mixed* shares		
	0.76 - 0.04 - 0.2	0.3 - 0.1 - 0.6	0.4 - 0.33 - 0.3

3.4 Graph-based Refinement

Although we achieved a relatively high accuracy using our classifier, we still found that for some campaigns our algorithm gave uncertain results. Luckily, the vast majority of the campaigns are located at the extremes of the classifier scores, either close to 1 (legitimate), or to -1 (spam). Campaigns become much more scarce in the range between [-0.8..0.8]. This gray area inside the gray area represents cases for which our technique was unable to automatically assign a definitive category.

Using these two thresholds, we can refine our classification and split the data into three classes: legitimate (77% of the total campaigns), spam (16%), and gray (6.4%). The minimum, average, and maximum values for each attribute in the three classes are summarized in Table 4. Since most of the false positives and false negatives are located in the gray area, we focused on improving the classification of those messages by using a graph-based technique.

In particular, we built a graph in which nodes represent campaigns and edges model the fact that two campaigns share a combination of sender IP address and email domain name. These links created networks of campaigns sent from the same mailing infrastructure. To avoid false connections that might appear between campaigns when they use webmail providers (spoofed or not), we removed those links from the graph.

The resulting graph contained 9,891 connected campaigns and 608 isolated subgraphs. By visually looking at the subgraphs, we noticed that the majority consisted of a predominant class (either spam or legitimate nodes) sometimes intermixed with gray nodes (see an example in Figure 1). This seems to suggest that gray campaigns also belong to the same class as the other nodes in the same group, since they are sent using the same infrastructure.

Additionally, our graph contains a Giant Component – a graph linking together 52% of all the campaigns – for which it is impossible to decide which class it belongs to. There-

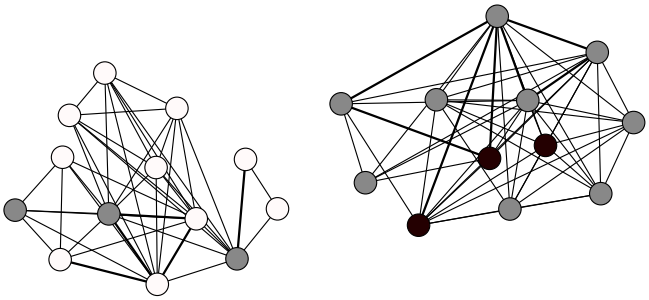


Figure 1: Subgraphs with mixed campaign classes: white for legitimate, gray for gray, black for spam

Table 5: Refining the campaign classification using graph analysis. Classification errors evaluated on 2,000 sampled campaigns

	RandomForest	Graph analysis
False Positives	0.9%	0.2%
False Negatives	8.6%	7.6%
Gray area	6.4%	2.9%

fore, we apply a community finding algorithm [5] that groups all the nodes into interconnected communities, also called groups, decomposing the Giant Component into smaller parts. We end up with 660 groups, for most of which we can accurately associate a single class. When gray campaigns are in the same group with any other class, we assign gray campaigns to the class of its group.

While this technique works well for most of the groups, some noise is still introduced in the results by the presence of loosely connected nodes. These are nodes that get erroneously connected to a group due to emails reusing the subjects of legitimate campaigns. To remove these connections, for each node we compute a graph metric called *clustering coefficient*. The coefficient for loosely connected nodes is equal to 0, whereas it approaches 1 for tightly connected nodes. As a result, we re-classify all the gray nodes with a clustering coefficient greater than zero and that belong to a group of either legitimate or spam campaigns. To decide on the class of the group, we compute the mean of classifier score of all nodes in the group: groups above 0.2 are considered legitimate, and groups below this threshold are considered spam.

Using this approach we were able to re-classify over half of the gray campaigns (427). This reduced the false positives from 0.9% to 0.2% (see Table 5 for more information). The entire dataset is now split into legitimate (80%), spam (17%) and gray (2.9%) messages (an increase of 3% for legitimate campaigns and 1% for spam). Again, our method performs better with legitimate messages. This is due to legitimate campaigns forming stronger networks (reusing the same mailing infrastructure over time) than malicious campaign.

4. ATTRIBUTE ANALYSIS

In this section we analyze the characteristics of spam and legitimate campaigns, and compare our findings to the ones presented in previous spam studies [21, 24].

The Random Forest classifier provides some information about the relevance of each feature. Interestingly, the least important attributes are the ones in Group B, and in particular the percentage of already whitelisted emails in the cluster. The most important ones are the distributions of countries and IP addresses, followed by the average number of recipients, and the sender email address similarity. The latter proved to be useful because spammers often change sender emails, while legitimate campaigns use a single or several repetitive patterns.

In particular, we found the number of originating countries to be the most indicative parameter, whereas previous research often relied on the IP address distribution (e.g. [24]).

4.1 The Role of IPs and Geolocation

IP address variation is often regarded as a strong indicator of botnet activity and often used as a reliable metric to detect spam. However, it is unclear what should be adopted as a threshold for this metric, how many different IPs should alert us of a distributed malicious activity, or how accurately we can classify email campaigns simply by looking at their IP address distribution.

In a previous study of spam campaigns, Qian et al. [24] used a threshold of 10 IPs per campaign to separate spam campaigns from legitimate ones. To evaluate this threshold, we applied it on our gray dataset as shown in Figure 2 (a). The graph plots the distribution of unique IP prefixes for both spam and legitimate campaigns. Around 90% of the legitimate campaigns are indeed below the 10 IP threshold, while 90% of the spam is above - resulting in a global error rate of 9.2% (to be precise, our measure is based on /24 subnetworks and not on single IP addresses, and therefore the real error rate is much higher than 9.2%). In comparison, this error is 5 times higher than the one of our classifier.

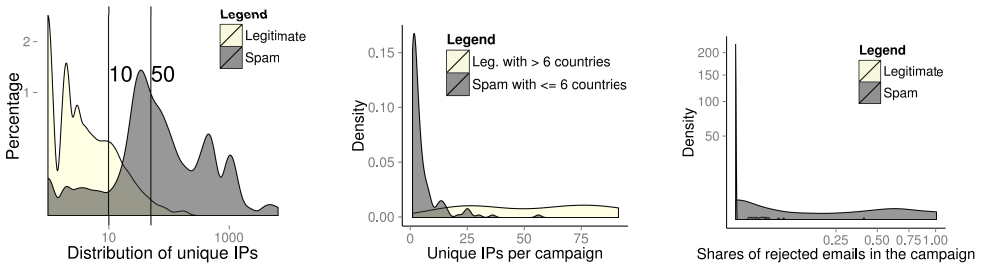
By looking at Figure 2 (a), we notice that above 50 IP prefixes there are few legitimate campaigns left and 99.8% of legitimate campaigns are below this threshold. However, half of the spam campaigns are located above the threshold and another half in between the two thresholds (10-50). This suggests that there is not a single value that separates the two classes with an acceptable error rate.

When we look at IP country distribution, the results improve considerably as some legitimate campaigns have many IP prefixes, but originate from few countries. This could be explained by one commercial campaign being distributed by several marketing companies in different locations. In contrast, the vast majority of spam campaigns originate from multiple IP prefixes *and* multiple countries. In fact, by using a six-countries threshold (the one chosen by our classifier) we misclassify only 0.4% of legitimate and 12% of spam campaigns - resulting in a total error rate of 2.8%. Figure 2 (b) shows the classification error.

Finally, we investigate closer this group of spam campaigns with few origins. Interestingly, the classifier for most of them gave a weak score, between 0 and -0.5. The graph refinement was ineffective for them, because these campaigns did not appear at all in our graph. At a closer manual inspection, these cases mainly corresponded to phishing and Nigerian scams. Several of these campaigns are sent in low volume and for short periods of time using webmail accounts, thus hiding under benign IP addresses.

4.2 Recipient-Oriented Attributes

The email recipient can be specified in three different headers: *To*, *Cc*, and *Bcc*. Interestingly, we found no campaigns using the *Cc* header, and some campaigns that seem to randomly change the location of the recipient over time



(a) Logarithmic plot of unique IP prefix distribution

(b) Campaigns with a threshold of 6 countries

(c) Emails rejections per campaign

Figure 2: Attribute distributions in campaigns

Table 6: *To/Bcc/Mixed* recipient header distribution

	To	*Bcc*	*Mixed*
Legitimate	75%	5%	20%
Spam	30%	12%	58%
Gray	20%	53%	27%

(we categorize them as *Mixed*). We also looked at the number of recipients per incoming email and at the number of non-existing email accounts (rejected at MTA-in because of non-existent user) in multiple recipient emails. We look at these three features together as they are often more informative when combined than when taken individually.

Around 75% of the legitimate campaigns use the *To* header (Table 6), whereas spammers often mix different headers in the same campaign. The *Bcc* header is adopted by both campaign types, although less frequently. However, it is very common among gray campaigns: in fact, half of them use exclusively this header to specify the recipient. Again, this is very common between the previously mentioned scam campaigns.

Since the campaigns located in the gray zone often use the *Bcc* field, they have shorter recipient lists including on average only 1.2 recipients per email. In contrast, 94% of legitimate campaigns have a single recipient, while spammers tend to include an average of at least three recipients per email.

However, these features alone cannot be used to reliably separate spam from legitimate messages. For example, 36% of spam campaigns used only one recipient per email, and in 30% of the cases specified in the *To* header. Interestingly, by combining these two criteria with the fact that these campaigns also have high IP prefix distribution, we can deduct that they originate from infected machines or botnets.

When some of the messages in a campaign are rejected, it is an indicator that the sender's recipient list was unverified or not up-to-date. Although sometimes users make typos while providing their email addresses, a higher rejection ratio, as shown in Figure 2 (c), along with multiple recipients is a good indicator of spammer activity. In fact, only 1% of spam campaigns sent with two recipients per email have a rejection ratio lower than 0.1. Thus, the combination of these two characteristics performs well for campaign classification.

4.3 Newsletter Subscription Header

Table 7: *Unsubscribe* header presence in campaigns

Campaigns	Header present	Missing header
Spam	225 (10%)	2,013 (90%)
Legitimate	5,064 (51%)	4,948 (49%)
Emails		
Spam	2,710 (0.6%)	482,133 (99%)
Legitimate	506,352 (43%)	668,153 (57%)

One of our features counts the presence of the *List-Unsubscribe* header in the emails. This header is intended specifically to indicate bulk email senders in order to treat such emails separately, and normally points to a URL or email address that can be used to unsubscribe from a mailing list[1]. This header is recommended to be used by regular bulk senders. Another recommendation for bulk email is to use the *Precedence: bulk* header. However, since in our dataset this header was used only in a few messages, we focus on the more common *List-Unsubscribe* header.

Figure 3 shows the percentage of each campaign type that uses the unsubscribe header. Only 10% of the spam campaigns adopt the header, counting only for a total of 0.6% of the spam messages. While legitimate campaigns tend to use the header in most of their emails, around half of the campaigns do not use it at all. This is due to several different email marketing companies advertising the same campaign, where some include the header, and some do not. In total, around half of the legitimate campaigns include the header (Table 7), and 27% of all legitimate campaigns have the header present in all messages.

In conclusion, we find it uncommon for spammers to use the *Unsubscribe* header, but at the same time legitimate campaigns use it in only half of their emails. While this attribute seems to be a good feature to identify marketing campaigns, spoofing the *Unsubscribe* header is extremely easy and could be done with minimal additional costs for spammers.

5. EMAIL CAMPAIGNS

In this section we present four categories of email campaigns that we identify in the gray area. We already separated spam from legitimate campaigns. We further divide

[1]In general an unsubscribe option is also included in the body of the message, but we could not check for this case since we had no access to email bodies.

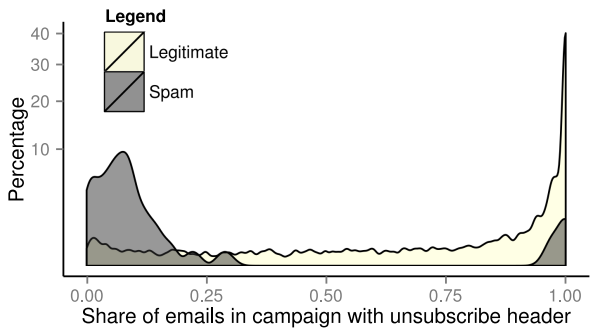


Figure 3: Newsletter subscription header distribution. Only the cases where the header is present are plotted

the spam in two categories: the one generated by distributed and dynamic infrastructures (likely sent by botnet or infected machines) from the smaller campaigns sent by few IPs.

We also split the legitimate campaigns into two groups. The first sent by private marketing companies as a service to distributes legitimate bulk advertisements, i.e., commercial campaigns. The second including newsletters that are sent to the users subscribed to a web services or mailing lists, and the automatically generated notifications (e.g. for online registrations). Again, the first ones are delivered by large infrastructures, while the second ones are often sent from a limited and constant set of IP addresses.

To identify these four categories in our dataset, we adopt a number of simple heuristics. As *commercial campaigns* we mark legitimate campaigns that belong to the biggest interconnected component of the graph described in Section 3.4. These are campaigns that are spread over many different networks and domain names as such campaigns are sometimes sent by several different e-marketing services providers, thus forming a large graph of interconnected campaigns. We consider the remaining scattered legitimate campaigns as *newsletters and notifications* as they rely on more static and isolated email delivery infrastructures. Botnet-generated campaigns are instead approximated by the spam clusters that are sent from more than six different countries and by more than 20 unique /24 IP prefixes. Finally, we manually sample over 350 of the remaining spam campaigns to identify *scam* and *phishing campaigns*.

All the categories are visualized in Figure 4, and the mean values of their features are summarized in Table 8.

5.1 Commercial Campaigns

This is the largest category in our dataset covering 42% of the identified campaigns, with an average of 148 emails each. By looking manually at these clusters, we confirm that these messages are mainly generated by professional email marketers sending. We were able to identify some of the main players (both national and international), and often confirmed that they actually run a legal business. On their websites, they repeatedly underline the fact that "they are not spammers", and that they just provide to other companies a way to send marketing emails within the boundaries of the current legislation. In fact, they also offer an online procedure for users to opt-out and be removed from future communications. These companies also use wide IP address ranges to run the campaigns, probably to avoid being blacklisted. Moreover, we find quite interesting that some of these companies also provide a pre-compiled list of emails (already categorized by user interests) that can be used to acquire new clients.

Therefore, email recipients can be taken both from *cold lists* (i.e., people who are not yet customers), or from current customer lists. As a result, different marketers send many different email campaigns, thus forming a large interconnected network of campaigns, as captured by our graph. As the senders also rely on cold lists, it is crucial to ensure that recipients can unsubscribe from the unsolicited advertisements. Indeed, commercial campaigns have the highest rate of *unsubscribe* headers.

On average, this class of campaigns lasts for 26 days, but some also continue for several months. Different email marketing companies are often involved in sending a single campaign, where each company is only active during a certain time frame. Also, each marketing service provider has its own dedicated range of IP addresses, which explains sometimes high IP address variance and high geographical distribution of campaigns in this group. As a comparison, newsletters (Figure 4, upper-left part) use on average three times less of unique IP addresses than a professional marketer.

To conclude, commercial campaigns can be highly distributed, but, at the same time, they often adopt consistent email patterns with similar sender names and email addresses.

Table 8: Feature mean values per campaign category. **Note**: User actions were evaluated only on campaigns with actions

Attribute	Commercial	Newsletter	Botnet	Scam
Countries	1.4	1.14	28.2	2.74
Recipients per email	1.00	1.00	2.80	1.16
Recipient *To*:	0.75	0.77	0.31	0
header (%) *Bcc*:	0.07	0	0.12	0.83
Mixed:	0.18	0.22	0.57	0.17
Sender email prefix	0.97	0.98	0.12	0.94
Sender email domain	0.96	0.99	0.31	0.97
IP distribution	0.84	0.94	0.08	0.86
Unique IPs	6	2	172	5
Rejections	0	0	0.24	0.02
Senders	0.97	0.98	0.34	0.95
Bounced	0.01	0.02	0.09	0.14
Unsubscribe header	0.59	0.39	0.01	0
CAPTCHAs	0.006	0.007	0	0.007
White emails	0.007	0.004	0.004	0.02
Period (days)	28	19	59	41
Viewed emails	3.6	6	7.3	2.9
Whitelisted emails	2.9	4	1.26	2.25
CAPTCHAs solved	19	26	1.7	7.6
Campaigns	5,113	3,597	2,107	150

5.2 Newsletter Campaigns

The newsletter senders rely mostly on static and small mailing infrastructure. The sender is often the actual company distributing the emails, with typically a small and fixed IP address range. This category contains half of the emails of the previous one (probably because most of the legitimate mailing lists do not get into the quarantined area as they are already whitelisted by their customers) and covers around

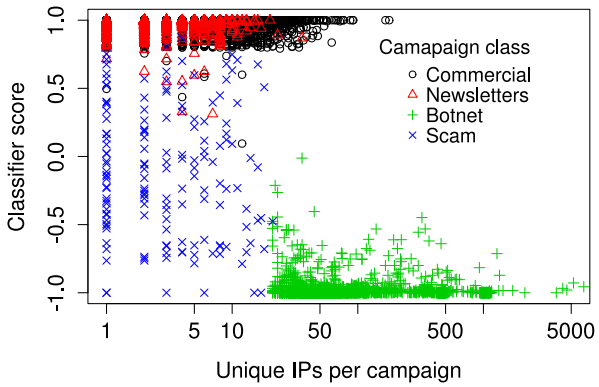


Figure 4: Email campaign classes distribution

30% of the total campaigns with an average size of 90 emails each.

A manual inspection seems to confirm that these campaigns consist mainly of notifications and newsletters sent by online services to which users have subscribed in the past. The senders are geographically localized (we encountered only one exception of a distributed newsletter campaign) and have extremely consistent sending patterns. Since we cluster campaigns based on their subjects, newsletters tend to last for very short periods of time. In addition, they normally use valid email recipient lists, and exhibit the lowest IP address, country, and sender email address variations. Only the use of the *Unsubscribe* header seems inconsistent, as only 39% of emails use it. However, this can be explained by the fact that notification emails normally do not use this header - only newsletters are subject to optional subscription. The consistent patterns in the email headers of this category indicate that the senders are making an effort to build a reputation and successfully deliver their correspondence. Not surprisingly, this is also the category that is whitelisted most often by the users.

5.3 Botnet-Generated Campaigns

Unsurprisingly, Table 8 shows that botnet-generated campaigns have highly dynamic attribute values, making them the easiest category to identify automatically. This category contains clusters that accounts for only 17% of all campaigns (also because most of the spam emails were already excluded from the gray emails by other antispam filters). Botnet campaigns have the highest geographical distribution as they are sent by infected computers from all over the world: 172 unique /24 networks per campaign, spread on average over 28 countries. Another prevalent characteristic is the use of multiple recipient emails sent using unverified email lists. Consequently, this leads to the highest email rejection rates (24%), and highest bounced CAPTCHA requests. The *Unsubscribe* header is rarely used, and sender email addresses have low similarities.

On average, botnet campaigns are the ones lasting the longest, with one drug-related campaign sent slowly over the entire six-months period of our experiments. Pathak et al. [21] also studied the length of *spam* campaigns, reporting a maximum length of 99 days over a dataset spanning 150 days. Our campaigns are substantially longer than that, maybe due to different datasets (we collected directly from user mail servers, not from open-relays), different email grouping methods (similar subject vs. URLs), or to changes in the behavior of spammers over time.

Despite the easily recognizable characteristics of these campaigns, users show a surprisingly high interest in these emails. This category has the highest number of email views per campaign, suggesting that users are often curious about products promoted and sold on the black market [18].

5.4 Scam and Phishing Campaigns

These campaigns contain phishing and Nigerian scam emails. Fraudsters trick their victims using threatening messages or by trying to seduce them with huge monetary gains. The characteristics of this category largely resemble those of commercial campaigns, thus making it difficult to automatically separate these campaigns without analyzing the email body. In fact, most of these campaigns belong to the gray area of our classifier. This is the reason why we needed to verify this set manually. These kind of threats are more likely to be identified by content-based detection techniques, e.g., by looking at email addresses and phone numbers [12], or URL [21, 29] included in the body.

We found only 12,601 of such emails, with an average campaign size of 84 emails. Phishing campaigns often spoofed the email addresses using well known company names (e.g. banks, eBay, Paypal), whereas Nigerian scammers relied mostly on webmail accounts [12]. In this case, many senders solved the CAPTCHA challenge – confirming that there is usually a real person behind these kinds of scams. The IP addresses from where the CAPTCHAs were solved are mostly located in West-African countries, like Nigeria or Ivory Coast. None of the messages in this category include an *Unsubscribe* header.

Unfortunately, users seemed to often fall victims to this type of attack, as they opened and even whitelisted messages in these campaigns.

6. USER BEHAVIOR

Our dataset also provides information about which actions were performed by the users on the quarantined emails. In particular, we collected information regarding the messages that were read, added to a user whitelist or blacklist, and the CAPTCHA that was later solved by the sender. These data can give us some useful insights on the ability of average users to identify suspicious emails.

Table 9 presents three user action statistics. As expected, user activity involves mainly legitimate and gray campaigns. In fact, the main reason for users to go through the emails in this folder is to spot missed notifications or undelivered benign messages. However, a large fraction of users also opened spam messages, maybe attracted by some deceiving subjects. As shown in Table 8 and Figure 5 (a), the highest campaign viewing rates are produced by botnet-generated campaigns, overpassing even newsletters. Over 3,888 spam emails were viewed by users during our six-month experiments, resulting in the fact that *one out of five users* has viewed at least one spam message, and, on average, *opened 5 of them*[2].

After a manual inspection of botnet-generated campaigns where the emails were read and whitelisted, we confirmed that those campaigns were promoting illegal products, e.g.

[2]Unfortunately, from our dataset we are unable to tell how many users downloaded attachments or followed links included in the message body.

Table 9: User actions performed on campaigns

	Viewed	Whitelisted	CAPTCHA solved
Legitimate	42%	12%	3.5%
Spam	25%	6%	0.2%
Gray	40%	17%	10%

drugs and pirated software. This may suggest two things: either users have problems in distinguishing legitimate emails from harmful, or some users are genuinely interested in the products promoted by spammers. It is difficult to draw conclusions as both hypotheses might be true for different users, but, clearly, most of them are unaware of the security threats involved in opening malicious emails.

Meanwhile, we should compare the reported statistics of viewed emails with the number of emails that actually got whitelisted – an action that could be interpreted as the equivalent of clicking the "Not Spam" button provided by several webmail services. The number of whitelisted emails per botnet-generated campaign (1.26 emails, Table 8) is the lowest among all the categories, suggesting that most users successfully differentiate them. However, we notice that scam/phishing campaigns have almost the same number of emails being whitelisted per campaign as commercial campaigns (2.25 vs 2.9). This suggests that users might have difficulties in differentiating these categories. It is important to remember that this category was manually sampled by domain experts, which is not the case for the typical users as most of them are untrained and are more likely to fall for these kind of fraud.

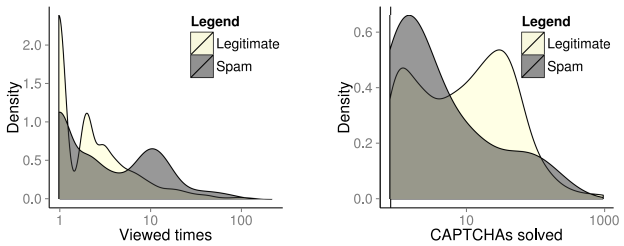
To further measure how significant this phenomenon is, we compute that there is a 0.36% probability that a certain user whitelists a legitimate email and 0.0005% that she whitelists a spam message. These numbers may seem low, but they rapidly increase when multiplied by the number of users and the number of messages received. In total, an average of 3.9 emails get whitelisted per legitimate campaign compared to 1.1 emails per spam campaign.

The last question we want to answer is whether the fact that the senders solve some CAPTCHAs in a campaign could be a good indicator of its legitimacy. Unfortunately, it is not and for two reasons. First, most of the legitimate campaign senders are automated tools, since a large potion of gray emails consists of newsletters, online notifications, and marketing emails. Secondly, although the general tendency is that users solve more CAPTCHAs within legitimate classes (Figure 5 (b)), as shown in Table 8, also scam and phishing campaigns have high CAPTCHA solved rates (7.6) compare with other categories. Finally, the rare cases of botnet-generated campaigns solving few CAPTCHAs correspond to challenges delivered to spoofed addresses by spammers as previously described by Isacenkova et al. [11].

To conclude, user-generated actions on gray emails are erroneous and thus are inaccurate to use for prediction. They often open even potentially dangerous emails, ignoring security risks. These results are in line to what has been tested in a user study conducted by Onarlioglu et al. [19].

7. UNCLUSTERED EMAILS

Our campaign classification covers half of the emails in the quarantined area, with 0.2% false positive rate. One may wonder what is inside the remaining 50% that is left outside our clustering approach. Qian et al. [24] concluded that the majority of legitimate emails should not be classifiable into clusters because of the content uniqueness generated by humans. Additionally, since most of the spam and legitimate emails were already filtered out from our dataset, the exact proportion may be different.



(a) Viewed emails (b) CAPTCHAs solved

Figure 5: Number of user actions taken per campaign

We could try to approximate the content of the unclustered part by assuming that the legitimate campaign senders always rely on a stable hosting infrastructure (as described in Section 3.4). In this case, for every legitimate campaign we can try to find messages sent by the same subnetwork and domain name in the unclassified email set. Using this technique, we found that 26% of the emails were sent from senders that were also responsible for legitimate campaigns. Almost 40% were sent from webmail providers. The spam set had a very low number of matches in the unclustered set, which is expected since most of these emails are sent from compromised machines that change over time.

Even though this heuristic can only provide a rough approximation of what is inside the remaining 50% of the messages, it can still be used (as part of a more complex system) to automatically separate marketing campaigns from more dangerous forms of spam.

8. DISCUSSION AND CONCLUSIONS

In this paper we presented a system to identify and classify campaigns of gray emails. As an approximation of this set, we chose to use the quarantined folder of a challenge-response antispam filter, since it is already clean from obvious spam and ham messages.

Our analysis unveiled the most and the least predictive email campaign class attributes. We also demonstrated that previous techniques used for email campaign classification [24] did not provide acceptable results in our settings, confirming that the gray area contains the hardest messages to classify. Additionally, we confirmed and extended some of the findings of previous studies regarding botnet campaigns [21].

Our system could be used in different ways. First of all, it can help understanding how large commercial campaigns work, how they originate, and how they differ from other unsolicited emails. It could also serve as an input to automatically place marketing campaigns and newsletters in a separate folder, so that users can clearly differentiate these messages from other forms of spam. In fact, the users in our study often opened botnet-generated emails and were especially prone to errors when dealing with scam and phishing messages; we believe that a separate folder dedicated to legitimate bulk emails would create an extra layer between the users and the malicious messages, thus allowing users to focus on the bulk folder when looking for missing and misclassified emails. Interestingly, after we completed our

study, a similar solution was deployed by GMail [2], to place user newsletters, notifications, and other commercial email into distinctive categories.

We also found out that our classification method based on sender behavior works well for any campaign except scam. We believe that the latter would benefit largely from content-based email analysis, e.g. URL or emails/phones clustering. Finally, we demonstrated that by using a graph-based refinement method, legitimate email campaigns can often be identified based only on sender information, and can be categorized as newsletters or commercial advertisement. This is a particularly promising result in the direction of empirical study of legitimate bulk emails.

9. ACKNOWLEDGEMENTS

The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement nr. 257007. We also thank MailInBlack for providing the data that was used in our study.

10. REFERENCES

[1] CAN-SPAM Act: Controlling the Assault of Non-Solicited Pornography And Marketing Act of 2003.

[2] Inbox tabs and category labels. `http://gmailblog.blogspot.fr/2013/05/a-new-inbox-that-puts-you-back-in.html`.

[3] Directive 2002/58 on Privacy and Electronic Communications, concerning the processing of personal data and the protection of privacy in the electronic communications sector (Directive on privacy and electronic communications). , 2002.

[4] S. Banerjee and T. Pedersen. The design, implementation and use of the ngram statistics package. *ITPCL*, 2003.

[5] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, 2008.

[6] M.-W. Chang, W.-T. Yih, and R. Mccann. Personalized spam filtering for gray mail. *CEAS*, 2008.

[7] D. Fallows. Spam: How it is hurting email and degrading life on the internet, 2003.

[8] Direct Marketing Association. Response Rate Report, 2012.

[9] D. M. A. DMA. Email deliverability review whitepaper, 2012.

[10] S. Hao, N. Syed, N. Feamster, A. Gray, and S. Krasser. Detecting spammers with SNARE: Spatio-temporal network-level automatic reputation engine. In *Proc. of the 18th conference on USENIX security symposium*, pages 101–118. USENIX Association, 2009.

[11] J. Isacenkova and D. Balzarotti. Measurement and evaluation of a real world deployment of a challenge-response spam filter. *ACM SIGCOMM, IMC*, 2011.

[12] J. Isacenkova, O. Thonnard, A. Costin, D. Balzarotti, and A. Francillion. Inside the scam jungle: A closer look at 419 scam email operations. *IWCC*, 2013.

[13] C. Kanich, C. Kreibich, K. Levchenko, B. Enright, G. M. Voelker, V. Paxson, and S. Savage. Spamalytics: An empirical analysis of spam marketing conversion. *CCS*, 2008.

[14] P. Kiran and I. Atmosukarto. Spam or not spam–that is the question. *Tech. rep., University of Washington*, 2009.

[15] A. Kolcz and A. Chowdhury. Hardening fingerprinting by context. *CEAS*, 2007.

[16] F. Li and M. han Hsieh. An empirical study of clustering behavior of spammers and groupbased anti-spam strategies. *CEAS*, 2006.

[17] MAAWG. Email Security Awareness and Usage Report, 2012.

[18] D. McCoy, A. Pitsillidis, G. Jordan, N. Weaver, C. Kreibich, B. Krebs, G. M. Voelker, S. Savage, and K. Levchenko. Pharmaleaks: Understanding the business of online pharmaceutical affiliate programs. *USENIX*, 2012.

[19] K. Onarlioglu, U. O. Yilmaz, D. Balzarotti, and E. Kirda. Insights into user behavior in dealing with internet attacks. *NDSS*, 2012.

[20] A. Pathak, Y. Hu, and Z. Mao. Peeking into spammer behavior from a unique vantage point. In *Proc. of the 1st Usenix Workshop on Large-Scale Exploits and Emergent Threats*, pages 1–9. USENIX Association, 2008.

[21] A. Pathak, F. Qian, Y. C. Hu, Z. M. Mao, and S. Ranjan. Botnet spam campaigns can be long lasting: Evidence, implications, and analysis. *SIGMETRICS*, 2009.

[22] A. Pitsillidis, C. Kanich, G. M. Voelker, K. Levchenko, and S. Savage. Taster's choice: a comparative analysis of spam feeds. *IMC*, 2012.

[23] A. Pitsillidis, K. Levchenko, C. Kreibich, C. Kanich, G. M. Voelker, V. Paxson, N. Weaver, and S. Savage. Botnet judo: Fighting spam with itself. *NDSS*, 2010.

[24] F. Qian, A. Pathak, Y. C. Hu, Z. M. Mao, and Y. Xie. A case for unsupervised-learning-based spam filtering. *SIGMETRICS*, 2010.

[25] Z. Qian, Z. M. Mao, Y. Xie, and F. Yu. On network-level clusters for spam detection. In *NDSS*, 2010.

[26] A. Ramachandran and N. Feamster. Understanding the network-level behavior of spammers. In *ACM SIGCOMM Computer Communication Review*, volume 36, pages 291–302. ACM, 2006.

[27] A. Ramachandran, N. Feamster, and S. Vempala. Filtering spam with behavioral blacklisting. In *Proc. of the 14th ACM conference on computer and communications security*, pages 342–351. ACM, 2007.

[28] Return Path. Email Intelligence Report, Q3 2012.

[29] K. Thomas, C. Grier, J. Ma, V. Paxson, and D. Song. Design and evaluation of a real-time url spam filtering service. *IEEE Symposium on Security and Privacy*, 2011.

[30] A. G. West, A. J. Aviv, J. Chang, and I. Lee. Mitigating spam using spatio-temporal reputation. Technical report, DTIC Document, 2010.

[31] W.-t. Yih, R. McCann, and A. Kolcz. Improving spam filtering by detecting gray mail. *CEAS*, 2007.

[32] S. Youn and D. McLeod. Spam decisions on gray e-mail using personalized ontologies. *SAC*, 2009.