Spam Campaign Cluster Detection Using Redirected URLs and Randomized Sub-Domains

Abu Awal Md Shoeb, Dibya Mukhopadhyay, Shahid Al Noor, Alan Sprague, Gary Warner
Department of Computer and Information Sciences
University of Alabama at Birmingham, AL, USA
{shoeb, dibya, shaahid, sprague, gar}@uab.edu

Abstract

A substantial majority of the email sent everyday is spam. Spam emails cause many problems if someone acts or clicks on the link provided in the email body. The problems may include infecting users personal machine with malware, stealing personal information, capturing credit card information, etc. Since spam emails are generated as a part of a very limited numbers of spam campaigns, it is useful to cluster spam messages into campaigns, so as to identify which campaigns are the largest. This enables investigation to focus this attention on the largest as the most significant clusters. In this paper, we present a method to cluster spam emails into spam campaigns. In our approach, the redirected URL has been chosen as the primary field for cluster formation. Our study shows that, a huge number of URLs arriving in spam email eventually points to a much smaller set of redirected URLs. Our multilevel clustering method grouped 90% of our half million spam emails into 4 spam campaigns. In addition to redirected URLs, we also use randomized sub domains, which come as a given URL in email body, for campaign identification. We believe that our model can be applied in real time to quickly detect major campaign.

1 Introduction

Spam email identification, also called filtering is an essential concern of numerous internet security companies. According to the Kaspersky Security Bulletin in 2013, about 70% of all emails sent today is spam [16]. A spam message is originated from a spam campaign. A campaign is a collection of messages that are generated from a single message template. The two primary method of filtering spam emails are content-based and blacklist-based [2]. Content based approach considers several factors such as, number of words in page tittle and body along with their average length, fraction of visible content and globally popular words, compressibility, n-gram likelihoods etc. during spam detection [3]. On the other hand, in blacklist-based approach, the well-known spamming hosts are detected, blacklisted, and blocked [5].

Several researchers considered IP address of the botnets to detect spam emails [4, 6, 7]. Even though detecting blacklisted IP is a naive approach, which can be done using limited computing resources, compiling and maintaining in such approach is challenging. Often the attackers change the host IP address or patch an already compromised host [8]. In contrary, some researchers proposed white list IP approach which maintains a list of trusted IP while anything other than those IP will be considered spam [9]. However, detecting white listed IP is not easy and often a legitimate email is categorized into a blacklisted email list due to the presence of a large numbers of emails [10]. Challenge response technique is another popular method for detecting spam where sender has to prove its authenticity by replying on the challenge sent by the recipient [11]. However, when both party implement this approach, the system can be in deadlock [18].

For many reasons, spam campaign identification is a challenging problem. First of all, spam has to be identified on the fly as the attributes of a spam campaign namely, email subject, URL, IP addresses, message body, and etc. change very frequently. Secondly, it is hard to catch the pattern or template of the spam without having a large collection of spam data set. Finally, analyzing the different parameters of a spam campaign requires lots of analysis that is time consuming and requires high computing resources.

The University of Alabama at Birmingham maintains the UAB Spam Data Mine. This data mine receives one million spam emails daily, and contains nearly one billion spam emails in total[20].

The process of mining spam data is time consuming. It requires going through every email to mine and cluster them based on their multiple attributes. A spam email has many attributes in common as a regular email. However, it contains some additional properties that help to identify them properly. In order to detect spam mail, we consider the given URLs, URLs embedded within the spam emails, along with its subject. Proper selection of an attribute is a key part of real time campaign identification. We process more than a half million of spam emails based on their URLs and subject. We find all redirected URLs by processing their given URLs. We also utilize the sub-domain of the given URLs by splitting them into several parts. The subject of the email has been used as a secondary attribute. As a result, our approach has come to minimize the clustering time. Since we have chosen redirected URL as our primary attribute of clustering, a major campaign containing URL can be identified within a very short period of time.

Since a large campaign can affect a larger mass of users, a large campaign can be way more harmful compared to

the smaller campaigns if it is not detected in time. So in this paper, we devise an algorithm that has two primary goals. The first and most important goal is to identify large campaigns with highest possible accuracy. To achieve this goal, redirected URLs, URL with randomized sub-domains, and exact subject match have been used. The second goal is to identify medium and smaller campaigns with subject matching.

Contributions: The contributions of this paper are as follows:

- We propose a multi-level clustering model to cluster spam emails in real time to detect large scale spam campaigns.
- Our approach efficiently identifies approximately 4 major campaigns out of half a million spam emails.
- To effectively find out the large scale spam campaigns, our model uses two most important parameters namely: Redirected URLs and Randomized Sub-Domain based URLs.
- Fetching the redirected URLs from the spam URLs; identification and utilization of a special class of URLs having randomized host-names or sub-domains with a fixed domain, to cluster spam email successfully with possibly good accuracy.
- To maximize the use of our model in gaining information about spam campaigns, we have attempted to merge the small spam clusters to find out if they form significant spam campaign(s).

The rest of the paper is organized as follows. In Section 2, we provide some background information and describe the challenges of detecting spam emails. In Section 3, we outline about our data structure and cleansing mechanism before applying our algorithm. In Section 4, we provide our multi-level clustering algorithm and implementation procedures in detail. Results are demonstrated in section 5. The final outcome of our implementation is also discussed in section 5. Section 6 describes the related work, and finally, we conclude our work in Section 7.

2 Background

In this section, we discuss the challenges and motivation behind our work.

2.1 Challenges

There are several problems that make clustering of related campaigns difficult in case of spams.

- Emerging Threats - Spam can deliver a previously unknown malware, or use a previously unknown infrastructure to deliver malware.
- Botnets - In order to hide their spamming infrastructure, criminals infect home user computers and use them to send their spam.
- Unique Subjects - A wide variety of subjects are used to send URLs for a single campaign so it is hard to prevent spam from blocking email on exact subject matches.
- Unique URLs - Spammers use a wide variety of URLs. They use hostname wildcards or customized paths for the same website. In that case, web filters for blocking URL destinations do not work well.

Another challenge of spam campaign detection is immediate extraction of URLs. All URLs that are included with the spam emails need to be extracted as soon as possible. Otherwise the obtained information might not be helpful in detecting spam campaigns. There are plenty of reasons to extract the given URL immediately.

- Firstly, spammers choose to shut down or change a given URL after a specific period of time. So it is obvious that one valid URL, which is currently accessible, might not be accessible after a certain period of time. Spammers change their host URLs very frequently so that they can hide their malicious activities behind these randomly chosen URLs and continue their business by keeping their identities obfuscated.
- Secondly, spammers use other popular domains for redirecting to their spam sites. For example, spammers may get access to the Nike site and can place a URL (www.nike.com/men/shoes/new/shaAHidAktaFaul) under the Nike subdirectory. Instead of changing the main directory, they change a subdirectory cleverly to avoid Nike's attention to the hacked directory. The main intention of spammers is to deceive users into believing that their fake and malicious URL is a valid one. But, higher the popularity of the host website, the higher is the probability of the hacked page to get caught within a very short span of time. As a result, the advertised URLs in the spam emails might not work if it is visited later.
- Lastly, a campaign should be identified as soon as possible before it reaches all target clients and become successful by reaping their profits by affecting the users.

2.2 Motivation

For several commercial web sites, the rise in search engine referral implies a rise in sales and revenue. The amount of US e-Commerce sales in 2004 was 69.2 billion, which is 1.9% of the total US sales meanwhile web-based e-Commerce increases steadily at 7.8% per year [12].

Forrester Research estimated that online US business-to-consumer acquisition of goods, which compromises travel and auctions, will rapidly rise. For commercial web sites to suitably benefit from this rapidly increasing market, they need to multiply their traffic that will make them up in the first couple of search engine results [13]. As a consequence, the number of spam email is also growing abruptly. Recently, it is observed that around 90% to 97% of emails are some kinds of spams [15].

In case of advertised products, users are asked to enter their personal and billing information to purchase designated products. Often the malicious sites record the information provided into the website and initiate some unwanted activities that might harm users. Therefore, the need for detecting spam emails before they reach the users is indispensable. The naive approaches either consider IP, domain, subject or content for detecting spam. However, blocking an IP or the domain based on recent history might not be an effective solution for preventing spam. An attacker often attacks some victim IP or domain and spread their campaign. Therefore, blocking based on IP or host will prevent a valid source to communicate in future when they will get rid of the attacker. On the other hand, checking redirected URLs before blocking will never prevent any legitimate source to send email. Some of the clustering algorithm just considers subject matching. But those clustering algorithms might generate false negative results. Sometimes two different companies can use same type of subjects for their product campaign. However, considering redirected URL in first level will ensure more accuracy during clustering as each campaign is associated with only a specific redirected URL.

3 Our Implementation

Real-time spam data collected from various sources is not always ready for use; the data often requires some preprocessing before clustering. In our experiments, at first, we perform some basic cleansing operations describes in 3.2. Secondly, we find the redirected URLs based on the given URLs in the email body. The redirected URL is a derived attribute that was not available directly in the email body. Lastly, we apply our multi-level clustering to identify a spam campaign.

3.1 Dataset

We utilized the UAB Spam Data Mine for the purpose of our research. The UAB Spam Data Mine is a research project under The Center for Information Assurance and Joint Forensics Research (CIS-JFR)[21]. The Center is responsible for gathering information about currently ongoing campaigns by spammers. It archives spam emails received from a wide variety of sources and honey-pots, and collects approximately 1 million spam emails each day. The archived spam emails are collected batch-wise at every 15 minute time intervals during the day. All general users on the internet, mark spam email and forward it to the honey-pot email address to archive. Moreover, other different honey-pots are placed at different points in the network, which dedicatedly archive spam emails upon reception. The archive extracts several attributes from each spam. From the current database design, we pick following attributes for each spam email:

- Subject - The subject of the spam email message.
- Sender IP - IP address of the machine that sent the spam email to our system
- Machine - Web address of the machine that is used in the URL embedded in the spam email message. Most of these machine addresses are genuine and they are exploited to fool the users and make them click on URLs containing these innocuous-looking websites. For example: mindshare-media.com, alpeturizm.com, etc.
- Path - A particular page in the given machine site (explained in the point above) that has been added to the machine site to add malicious content in it. For Example: /jUijHkHpwEd1P/, /etTgz6tFNhh2q/, etc.
- URL - The concatenation of the machine and the path forms the URL. For example: http://mindshare-media.com/jUijHkHpwEd1P/, http://alpeturizm.com/etTgz6tFNhh2q/ , etc.

The UAB Data Mine collects a million spam emails per day from various feeds. Among those data, we are given a 6 hours worth of data for April 15, and a whole day's worth of data for August 20, and 21, 2014. This data contains emails from both single or multiple feeds. Some emails might conatin multiple URLs in email body. However, each spam email we are provided contains only a single URL.

3.2 Data Cleansing and Pre-Processing

First, we exclude some special characters from subject for the purpose of making our operation faster. These special characters include symbols, local languages in different encoding systems. However, in the future, we will treat them as a special attribute to form clusters. Second, we find redirected URLs for the given URLs of each spam email. In order to find a redirected URL, we combine the machine and the path together to form a complete URL.

3.3 Finding Redirected URLs

Spammers change given URLs very frequently to deceive the target user. Sometimes they use different path with same domain name and sometimes they use popular domain names that are already hacked with a modified path. Moreover, they use popular domain names as URLs that actually link to a completely different URL. However, in most of the cases, all URLs redirect to very few unique URLs. Since redirected URL is one of the primary attributes of our clustering, we had to find all the redirected URLs that are fetched by the given URLs, available in email body. The redirected URL can remain same and active for longer period whereas the given URL changes after a very short period of time. We use Mechanize, a very useful and popular library for emulating browser in Python, to find a redirected URL for a given URL. We set 10 seconds as the timeout time to find a redirected URL. If an URL does not reach its redirected URL within this time period, we move to next entry. The timeout value could be less than a second but we set 10 seconds on a trial and error basis as it works well in our case. We consider the network condition, Internet speed, etc. to find a reasonable timeout value. This process time to find a redirected URL has been reduced by means of using multi-thread programming to process all the entries and generate the redirected URLs from them.

3.4 Identifying Randomized Sub-Domain based URLs

Spammers nowadays are using a new kind of URLs to propagate their spams. The URLs have a randomized subdomain (or hostname) but a specific domain with no path, ending in a ".in" or ".ru". The hostnames here exhibit the behavior of being a "wildcard" hostname. In other words, any random word or set of characters could be prepended to the domain name and the same webpage would be the result. For example, consider some domain names like, bestdealpills.ru, bhoomirealty.in, bulkmailbox.co, burra.ru, calag.ru, cialt.ru, datae.ru, entam.ru. Any character or group of characters can be placed before these domain names list and the result will be a Canadian Health & Care Mall.

3.5 Spam Cluster Formulation

The primary objective of our algorithm is to cluster a huge collection of spam emails collected over a span of some days, for identifying the major campaigns involved. In order to perform this, we have defined a cluster with certain attributes. The attributes are as follows:

- **Key -** The key of a cluster is an attribute that identifies that cluster. That is, a cluster A can be identified by its Key Ka. Depending on the attribute or variable for clustering at each level, as new clusters are formed and/or old clusters merge the keys of the clusters keep changing. A cluster's key can be a URL or a Subject String depending on the attribute that has been taken into account during the creation of that particular cluster.
- **Set of given URLs -** For each cluster, we save a set of given URLs for that cluster. This set contains the set of all the given machine sites of the spam entries that went into that particular cluster. This set is named as the *given_url* set.
- **Set of Subjects -** For each cluster that is formed, we save a set of *Subject Strings*. This set contains the subjects each saved as a separate String for all the spam emails that form that particular cluster.This set is named as the *Subject* set.
- **Bag of words -** For each cluster, we save a set of all the words found in the subjects of the spam entries that formed the cluster. This set is named as the *subject_words* set.
- **Set of IPs -** For each cluster, we save a set of IPs for that cluster. This set contains a set of the sender IP addresses for the spam entries that went into that particular cluster. This set is named as the IP set.
- **Level Flag -** This flag for a cluster is a numerical value that represents a level of clustering. This flag saves the information when a new cluster is formed or an existing cluster is modified. At each level of clustering, as new clusters are formed and/or older clusters disappear by merging with existing clusters, the Level flag of the clusters, which are modified or created, are updated to represent the value of the current level of clustering. For example, if cluster A is created at Level 1 and is updated at Level 2, then the Level flag of cluster A at level 2 will be set to 2 and this value will get changed as and when cluster A gets updated.

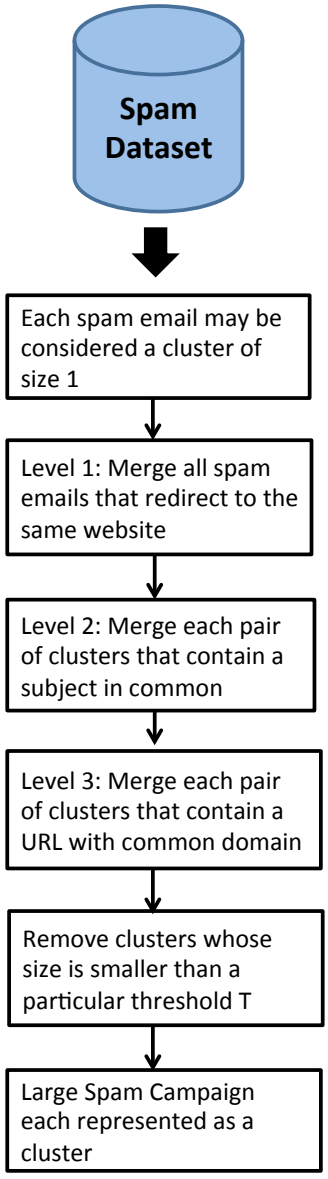

Figure 1: Block Diagram of the Multi-Level Clustering

4 Algorithm

Figure 1 describes the algorithm of the model we have proposed for our multi-level clustering. We have skipped the pre-processing step in our algorithm since we are assuming that the input to our algorithm would be a pre-processed and cleansed data file. The details of each step of the algorithms will be presented now.

Our proposed model takes as input a spam file and clusters the spam file using attributes like given URL, redirected URL (extracted by using given URLs as mentioned in section 3.5) and spam email subject. A detailed explanation

Algorithm 1 Level-1 Clustering

```
set ClusterSet C = All spams from the dataset
For each spam S in C
if S has a redirected URL then
    Fetch a redirected URL R
    if R is not in C then
        form a cluster C1 considering redirected URL as cluster key
        merge S to C1 and delete S from C
        C1.LevelFLAG=1
        C=C1 ∪ C
    else
        C=merge(C,S)
    end if
end if
```

of each step of our algorithm is as follows:

Level 1: Clustering based on Redirected URL : In this level of clustering, we start with a spam file having a list of parsed spam email records each denoting a cluster. For all the spam records from the input file, the redirected URLs if exists are extracted. For a fetched redirected URL,the clusters that having that particular redirected URL are merged together to form a new cluster. The key of that cluster becomes that redirected URL and the Level flag of the cluster is set to 1 (the structure of a cluster is described in section3.5).

At the end of level one clustering, we are left with a set of clusters formed from the redirected URLs, each consisting of the spam emails that had that particular redirected URL as its attribute. The rest of the spam emails that did not have any redirected URLs are left as it is and are considered a single cluster points that would be merged in the subsequent levels.

Redirected URLs are nothing but the addresses of the actual websites that spammers are using to run their business. The products hosted in these sites are the products that the spammers want the users to buy. Since, these websites are primarily focused on selling and marketing one kind of products, so we can conclude with a high degree of confidence that spams containing a common redirected URL would essentially belong to the same spam campaign. This is our assumption behind Level 1 clustering.

Level 2: Subject-Based Clustering and Subsequent Merging: In this level, the existing clusters are further are merged together based on the subjects of the spam emails. Our objective in this level is to merge clusters with matching subjects. Each pair of clusters formed in the previous levels are checked for matching subjects and clusters having at least one common subject are merged together and the key of the cluster is set as that common subject string. When one of the matching clusters is a redirected URL based cluster though, the components clusters are merged but the key of the resulting cluster remains as the redirected URLs. The level flag of the clusters that merged in this level, are set as 2.

At the end of the second level, we have two kind of clusters. First kind are the ones whose key are redirected URLs and were merged in Level 1 and and some of them might have been modified in Level 2. And the second kind are the ones whose keys are subjects and were merged in Level 2.

Algorithm 2 Level-2 Clustering

```
ClusterSet C=set of all the existing clusters
For each cluster E from C
if E.Subject is not in C then
    form a cluster C1 considering E.Subject as cluster key
    merge E into C1 and delete E from C
    C1.LevelFLAG=2
    C=C ∪ C1
else
    C=merge(C1,E)
    delete E from C
end if
For each cluster M in ClusterSet C where M.LevelFLAG=1
For each cluster N in N.LevelFLAG=2
if M.Subject_set ∩ N.Subject_set=∅ then
    continue for next value N
else
    merge(N, M)
    delete M from C
    set M.LevelFLAG=2
end if
```

Algorithm 3 Level-3 Clustering

```
ClusterSet C=set of all the existing clusters
For each cluster E from C
if E.Doman is not in C then
    form a cluster C1 considering E.Domain as cluster key
    merge E into C1 and delete E from C
    set C1.LevelFLAG=3
    C=C ∪ C1
else
    C=merge(C,E)
    delete E from C
end if
```

The assumption behind this level of clustering is, spams having exact same subjects must fall into the same spam campaign. It is highly unlikely that spam emails belonging to different spam campaigns will have the same subject and even though it happens, then such spams will rather be considered as noise.

Level 3: Randomized Subdomain URL based Clustering: In this level, our motive is to use a portion of the given URLs to further merge some more clusters that were created in Level 2. For this purpose, we extract the domain information of the URLs from the given_URL set of the clusters formed in Level 2. By comparing each pair of domains we merge the two clusters that have at least one domain in common. The cluster keys of the merged clusters still remains the subject as before ; their level flags are set as 4.

At the end of this step we have a set of spam clusters identified by a key which can be either subject or redirected URL and each of these clusters represent a spam campaign.

In practicality, spam emails having a particular domain with randomized sub-domains will be part of a single spam campaign. This is because, these sites hosted by the spammers are all possibly residing at the same locations as their domain names are the same. And, contents that are hosted in a common physical machine are most likely to belong to a single spam campaign.

Table 1: Number of Clusters (clst) in Different Levels

Clustering Level	Case 1	Case 2	Case 3	Case 4
# of spam	60995	249389	247922	497311
# of clst: L1	17253	156077	166645	322722
# of clst: L2.1	1198	5171	7965	11910
# of clst: L2.2	1086	5145	7938	11878
# of clst: L3	289	1044	1037	1670

Table 2: Number of Clusters (clst) for Different Threshold

	Case 1	Case 2	Case 3	Case 4
# of clusters after L3	289	1044	1037	1670
Threshold 10	64	314	330	462
Threshold 20	49	211	202	324
Threshold 50	24	117	112	180
Threshold 100	15	70	69	117
Threshold 200	9	44	35	76
Threshold 500	4	18	15	26

Thresholding: After the final clustering is done, many small and medium sized spam campaigns are formed that do not have a significant number of spam emails inside them. These spam campaigns are usually not very prominent or malicious and can be treated as noise. And in some cases, these might be consist of some valid emails that got wrongly classified as spams. So, we apply a thresholding function to the spam campaign list such that when a campaign's size is below a certain threshold, that spam campaign is deleted and hence not taken into consideration.

5 Results and Discussion

In this section, we discuss the results that we obtained by clustering using our proposed model. Here, Case 1 refers to the spam dataset from April 15, 2014, Case 2 refers to the spam dataset from August 20, 2014, Case 3 refers to the spam dataset from August 21, 2014 and Case 4 refers to the spam dataset combined from August 20 and August 21, 2014.

5.1 Result of the Multi-level Clustering

Table 1 shows the result of our multi-level clustering model in terms of number of cluster obtained in different levels. In first level clustering, we obtain the number of cluster based on redirected URLs. In all cases, we found 3 clusters that have three specific redirected URLs. In general, almost 50% of the emails become members of these clusters. The rest of the emails are treated as a singular cluster point and yet to be merged in the next subsequent levels. For example, if we look at the result of 21 Aug 2014, Level 1 (L1) shrinks 247922 emails down to 166645 emails and 3 clusters based on redirected URLs.

In Level 2(L2.1), remaining emails have been clustered based on their subjects. Spam emails that have exact same subject are placed into same cluster. As a result, for the data 21 Aug 2014, 166645 emails formed 7965 clusters that also include 3 clusters formed in Level 1. We obtain a small amount of improvement in Level 2.2 (L2.2). Here we merge certain Level 2.1 clusters that have at least one subject in common with 3 redirected URLs' clusters. By looking into the results of Case 3, we observe that only 27 clusters have been reduced from Level 2.1 to Level 2.2. Total 7965 clusters are reduced to 7938 which is a 0.33% improvement.

Our last level of clustering, based on randomized subdomains, shows a significant improvement for all data sets. In this level, we match domain of the given URLs of the clusters formed in Level 2.1 and 2.2 amongst each other. Spammers create a bunch of different URLs, where prefix (i.e., sub-domain) of the URL is randomized and suffix (i.e., domain) of the URL is kept same if they are originated from the same spam campaign. In this phase, we truncate a portion of the prefix and matched the domain with each other to merge the clusters formed in Level 2.1 and 2.2. As a result, we acquire almost 70% improvement from previous level to this level. The results of Case 3 improve by nearly 86% in Level 3 (L3); 7938 clusters have been merged and to form 1037 clusters.

Finally, after performing Level 3 algorithm, we observe that there a quite many number of clusters that have very negligible number of members. A statistics about the cluster size is given in the Table 2. By applying the threshold of 100, in Case 3, we find 69 clusters which is a commendable gain in last levels. The number of the cluster in after thresholding is the number of large spam campaigns indeed.

5.2 Prominent Large Spam Campaigns

In Table 2, the number of campaigns left after applying different thresholds is shown. From the numbers in the table, it is evident that as we increase the threshold even by a small number, the total number of campaigns decrease by a way larger number. For example, for Case 2, after applying a threshold of 10, the number of campaigns (or clusters) remaining at the end of the multi-level clustering, reduces by 70%(approximately) and this percentage goes upto 97%(approximately) as the threshold value goes upto 200. Hence, there is a need to focus on the large and most prominent spam campaigns and ignore the less important ones that are removed after thresholding. As mentioned at the beginning of this section and also in Section 3.1, we are

using spam data from 3 different days namly : 15th April 2014, 20th August 2014 and 21st Aug 2014. The large spam campaigns found in these three days are as follows:

5.2.1 Campaigns on 15th April:

In Figure 2, the size distribution of the largest six spam campaigns is plotted. By doing a careful inspection of the graph, it is clear that the largest four campaigns constitute approximately 93% of the total number of spam emails received during a period of that day. A small description of each of these four campaigns are as follows:

- The first and the largest campaign detected during this day that constitutes upto 36% of the total spams is related to anti-aging pills. All the emails in this campaign got redirected to fjxnewsdaily.com. This advertises and sells anti-aging pills. So this spam campaign can be pointed out as a **"Anti-Aging Pills"** campaign.
- The second largest campaign detected in this day comprising of about 32% of the total spams is related to weight-loss pills. The emails in this campaign got redirected to wghtnews.com . This website promotes and sells a brand of weight loss pills. Hence, this campaign can be concluded as a **"Weight-Loss Pills"** campaign.
- The third largest campaign of this day, which constitutes about 18% of the total spams. These spams are related to weight loss as well. The emails in this campaign are redirected to mscdailynews.com. This website advertises about certain weight-loss products and solutions. Though this campaign is somewhat similar to the previous campaign, since both of them talk about weight loss, they are advertised by different sets of spammers. So, the products advertised in this website is different from that of the previous site. We can identify this campaign as **"Weight-Loss Solution"** campaign.
- The fourth largest campaign that forms around 5% of the total spams is related to viagra. These emails have a common domain in their URLs namely, oizcuscp.in. All the URLs embedded in these emails go to a particular website that sells and advertises Viagra products. So, this campaign can be pointed out as **"Viagra"** campaign.

5.2.2 Campaigns on 20th August:

In Figure 3, the size distribution of the largest seven spam campaigns has been shown. From this figure, it is clear that the largest four campaigns constitute approximately 85% of the total number of spam emails received that day. A brief description of each of these four campaigns are as follows:

- The largest campaign forms 36% of the total number of spams. This campaign is formed of spams that got redirected to dietscoop24.com. This website talks about and advertises weightloss pills. This campaign can be identified as a **"Weight-Loss Pills"** campaign for this day.

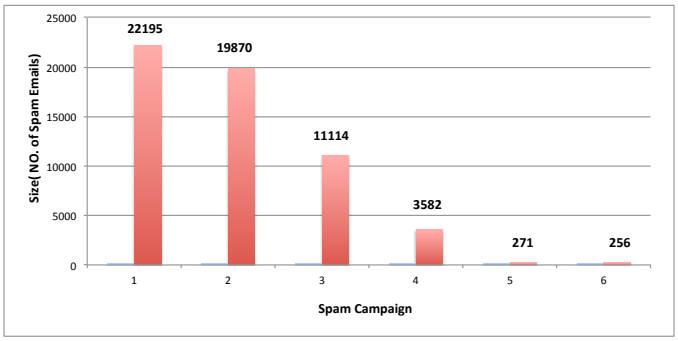

Figure 2: Spam Campaign Size Distribution on April 15, 2014

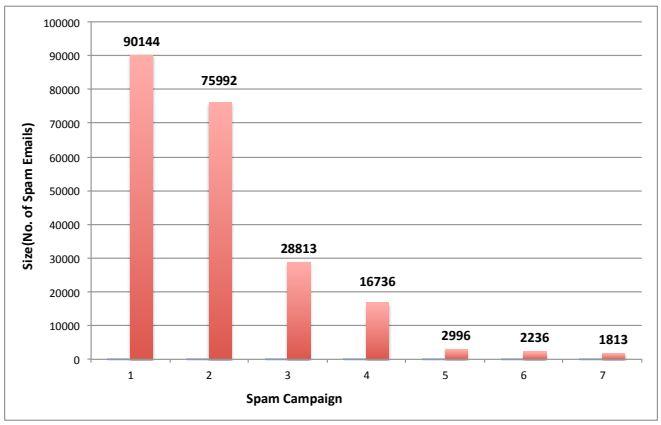

Figure 3: Spam Campaign Size Distribution on August 20, 2014

- The second largest campaign forms 30% of the whole spam set for the day. All the emails in this campaign are redirected to skinnewsdaily7.com. This website also talks about weight loss products and sells such products. But this might be a different set of spammers advertising for a different brand of products. Hence, this campaign can be identified as **"Weight-Loss_2"** campaign for that day.
- The third largest campaign forms 11% of all the spams. All the emails in this campaign have a common domain in their URLs namely, smartdrugservices.be and they go to a single website. This website is related to viagra. So, this spam campaign can be named as a **"Viagra"** campaign.
- The fourth largest campaign of this day forming around 6% of all the emails talks about weight loss as well. But all these emails have a common subject, "At Least 15lbs OFF Within 3-4 Weeks!". This spam campaign can be noted as a **"Weight-Loss_3** campaign for the day. Since these weight loss products are of different brands and most probably are hosted by different set of spammers, hence we will consider this as a separate campaign.

5.2.3 Campaigns on 21st August:

In Figure 4, the size distribution of the largest seven spam campaigns has been shown. From this figure, it is clear that the largest four campaigns constitute approximately 87% of

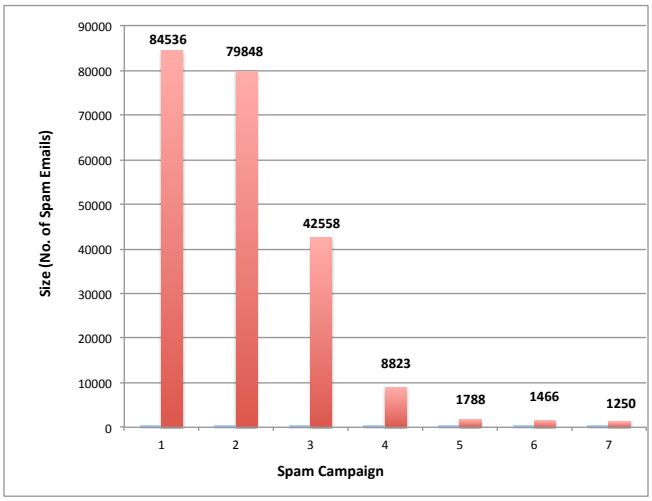

Figure 4: Spam Campaign Size Distribution on August 21, 2014

the total number of spam emails received that day. A short description of each of these four campaigns are as follows:

- The largest campaign forms 34% of total number of spams. All the emails from this campaign are redirected to skinnewsdaily7.com. Hence this campaign is same as the **"Weight-Loss_2"** of 20th August.
- The second largest campaign forms 32% of the whole spam set for the day. All the emails from this campaign are redirected to dietscoop24.com. Hence this campaign is same as the **"Weight-Loss Pills"** of 20th August.
- The third largest campaign forms 17% of all the spams. All the emails in this campaign have a common domain in their URLs namely, rhin.ru and all of them go to the same website. This website is related to viagra. So, this spam campaign can be named as a **"Viagra"** campaign.
- The fourth largest campaign of this day forms around 3% talks about weight loss pills. All the emails in this campaign have a common domain in their URLs namely, mediczier.ru. Any randomized subdomain prepended before this domain name will go to a particular website that sells those products. Hence, this can be identified as **"Weight-Loss_3** for this day.

5.3 Prominent Redirected URLs

A large portion of about 70% to 80% of the spams in our spam datset corresponds to redirected URLs. So, there should be a mention of the mot prominent redirected URLs found in the dataset over 3 days. They are as follows:

- http://wghtnews.com - This page advertises about weight loss pills. It has referred to an issue of renowned magazine "Womenś Health". The cover story of that issue being the secret of famous celebrity Oprah to lose weight quickly. This site also refers to a well-known show called "The Dr. Oz show" and how a popular host of this show advises everyone to take the "miracle pills" to "beat the bulge". To make this article more convincing, they have added a video clip demonstrating the actions of the pill and this clip claims claimed to have been telecasted in a program called TV Doctor. Needless to say, all these claims are in fact false. The Womenś Health magazine has even published disclaimers of never publishing anything about weight-loss pills and advises its women readers to be careful of such fraudulent articles and advises them against taking any such dietary supplements. Some of the subjects of the spam emails that lead to website are: Win the battle of the bulge, Lose 20 pounds with just 2 pills, Indulge and still lose your bulge,etc.
- http://mscdailynews.com - This page is currently blocked by all the leading browsers because it might have got black-listed. This site advertised about weight loss solutions as well. Some of the subjects of the spam emails that lead to website are: Weight loss has never been so easy, Get ready for a new you, Weight loss without the workout,etc.
- http://fjxnewsdaily.com - This page advertises about an anti-aging skin-care product that made Martha Stewart look 20-year younger almost overnight. This site also exploited the names of well known magazines like "Cover Girl", "New You", etc. and of well known shows like "The Dr. Oz show" to sell this product to users. And as with the previous spam sites, this site was also fake with fake "Before-After" pictures, fake claims and fake video clips. Some of the subjects of the spam emails that lead to website are: Looking young has never been so cheap, Refresh, rejuvenate, return to youth, Your skin has never looked so vibrant, etc.
- http://dietscoop24.com - This page advertises about weight-loss pills. In this website they claim to show a clip from the "Dr. Oz. Show". It is needles to say that all these are fake. The spammers have put up this content to advertize and sells their brand of weight-loss pills.
- http://skinnewsdaily7.com - The page has been black-listed and it does not open on browsers anymore. But the emails in these campaign have subjects like: Achieve your goal weight this month, Dramatic results in under a month, etc. From these subjects, one can conclude that these spammers are selling weight loss products that they claim can help people lose weight in a month.

The first three URLs in this list appeared in April 15th's spam feed while the bottom two URLs appeared on 20th August and 21st August spam feeds. The second URL though in small number, has appeared in 20th and 21st August feeds also. Speedy detection of these redirected URLs can help the security enforcing agencies to close down spam campaigns very efficiently.

5.4 Prominent Randomized Sub-Domain URLs

The spam campaigns formed by matching domain of URLs form a considerably large portion of the entire spam

feed. About 18% - 20% of the entire spam emails, got clustered using randomized sub-domain and fixed domain URLs. Some of the notable randomized subdomain URLs on 20th Aug and 21st Aug are: rhin.ru, mediczier.ru, lgmmi.mediczier.ru, mediczier.ru, cfwmedic.ru, etc. While some of the important randomized subdomain URLs on 15th April are oizcuscp.in, enhairations.de, etc. Spammers use a random sub-domain URL to obfuscate their identity. The identification of the pattern of the randomized subdomain URLs and their subsequent detection can help to curb the activities of a large group of spammers.

A study of the results show that redirected URLs and randomized sub-domain based URLs can be highly effective to cluster current spam feeds. Spammers keep changing the redirected URLs names frequently and also use several different subjects to hide their identity. But, spam clustering based on the our proposed model can detect spams very effectively despite these challenges.

6 Related Work

To date, several teams of researchers have proposed several methods to detect spam. Carlos et. al. proposed a method for detecting spam on website by combining both the link and content information [26]. They introduced a topology of the web graph by exploiting the dependencies among the various web pages. Pedro et. al. design a real time spam detection system by analyzing the several attributes of online traffic [1]. They consider an incremental approach where the order of attribute selection is changed over time. A very similar kinds of method is proposed by Rasib et. al. However, their proposed method reduces the time of detecting a spam campaign significantly by identifying the hot zone using data sampling [19]. However, all the above method overlooked one important attribute, which is the redirected web site. We know that the objective of spam is to campaign about a product. Hence, we consider the redirected web information as our primary source of clustering followed by given URL and subject information.

Ramachandran and Feamster [24], in an influential early paper, investigated the distribution of senders of spam in IP-space, and suggested the important role of botnets in distributing spam. Wei et al. clustered spam by spam campaign, and noted that hosting IPs were much longer-lived than domains, and wildcard DSN entries were prevalent [25]. It was also noted that when the volume of spam of one campaign decreased, the volume of some other campaign increased, suggesting that perhaps some botnet was alternating between sending spam from one campaign and from a different campaign.

Wei et al. [25] performed clustering on sending domains, based on both the set of IPs and the set of subjects associated with each domain. These authors observed that the largest 5 clusters contained 80% of the domains. In [23] Calais-Guerra et al. organize spam emails in a FP-tree, and let the FP-tree direct the formation of clusters. This allows the clusters to be defined in a data-dependent manner, based on the attributes that will show the most distinct clusters. In a later work [22], they study the use of open proxies and open relays by spammers; such a study aids in the detection of compromised end-user machines.

7 Conclusion

Our implementation showed a significant progress in clustering based on only email URLs and subjects. The key attribute, redirected URLs, plays a magical role to put all spam emails in to a very small number of clusters. The technique of sub-domains in last level clustering has also improved the performance of campaign identification. Our model demonstrates impressive results against all spam emails that are already collected from various sources. We believe that our model will show the same result, better result in some cases, if it is applied to spam emails as soon as they are received.

8 Acknowledgement

The research conducted in this paper initially began to fulfill the requirements for the course work of CS663 (Knowledge Discovery and Data Mining) at UAB. Later, it expanded and furnished by changing clustering methods to obtain more accurate and improved result. We are grateful to our instructor, Dr. Alan Sprague, for his guidance throughout the research. We are also grateful to Gary Warner for providing his valuable comments and real time spam data for our research.

References

[1] Pedro Calais, Douglas E. V. Pires, Marco Ribeiro, Dorgival Guedes, Wagner Meira Jr., Cristine Hoepers, Marcelo Chaves, and Klaus Steding-Jessen. Spam miner: A platform for detecting and characterizing spam campaigns. In KDD, 2009.

[2] Zhiyun Qian, Z. Morley Mao1, Marco Ribeiro, Yinglian Xie2, On Network-level Clusters for Spam Detection, In Proceedings of the Network and Distributed System Security Symposium (NDSS), 2010.

[3] Alexandros Ntoulas , Marc Najork , Mark Manasse , Dennis Fetterly, Detecting spam web pages through content analysis, Proceedings of the 15th international conference on World Wide Web, May 23-26, 2006, Edinburgh, Scotland

[4] Kelly Jackson Higgins, Dark Reading, Botnets Battle Over Turf. `http://www.darkreading.com/document.asp?docid=122116`, Apr. 2007.

[5] Wesley Pronk, Real-time Blacklisting of Bots based on Spam Analyis , 15th Twente Student Conference on IT June 20th, 2011, Enschede, The Netherlands.

[6] A. Ramachandran and N. Feamster. Understanding the Network-Level Behavior of Spammers. In Proc. ACM SIGCOMM, Pisa, Italy, Aug. 2006.

[7] Cook D., Hartnett J. et al., "Catching spam before it arrives: domain specific dynamic blacklists", Proceedings of the 2006 Australasian workshops on Grid computing and e-research - Volume 54, Hobart, Tasmania, Australia, pp. 193-202, 2006.

[8] S. Sinha, M. Bailey, and F. Jahanian. Shades of Grey: On the Effectiveness of Reputation-based Blacklists, In Malware 2008.

[9] Pfleeger S. L. and Bloom G., "Canning Spam: Proposed Solutions to Unwanted Email", IEEE Security and Privacy, vol. 3, no. 2, pp. 40-47, 2005.

[10] Golbeck J. and Hendler J., "Reputation Network Analysis for Email Filtering", CEAS, pp. 1-8, 2004.

[11] O'Brien C. and Vogel C., "Spam filters: bayes vs. chi-squared; letters vs. words", Proceedings of the 1st international symposium on Information and communication technologies, Dublin, Ireland, pp. 291-296, 2003.

[12] U.S. Census Bureau. Quarterly Retail E-Commerce Sales 4th Quarter 2004. http://www.census.gov/mrts/www/data/html/04Q4.html (dated Feb. 2005, visited Sept. 2005)

[13] C. Johnson. US eCommerce: 2005 To 2010. http://www.forrester.com/Research/Document/Excerpt/0,7211,37626,00.html (dated Sept. 2005, visited Sept. 2005)

[14] LEVENSHTEIN, V. I. 1966. Binary codes capable of correcting deletions, insertions and reversals. Soviet Physics Doklady. 10, 8 (Feb), 707710.

[15] Jacqui Cheng, Microsoft: 3% of e-mail is stuff we want; the rest is spam, http://arstechnica.com/information-technology/2009/04/microsoft-97-percent-of-all-e-mail-is-spam/

[16] D Gudkova, http://securelist.com/analysis/kaspersky-security-bulletin/58274/kaspersky-security-bulletin-spam-evolution-2013/

[17] JACCARD, P. 1901. Distribution de la flore alpine dans le bassin des Dranses et dans quelques regions voisines. Bulletin de la Socit Vaudoise des Sciences Naturelles 37, 241272.

[18] Alkahtani, H. S., Gardner-Stephen, P. et al., "A taxonomy of email SPAM filters", The 12th International Arab Conference on Information Technology (ACIT), Riyadh, Saudi Arabia, p. 351-356, 201.

[19] Rasib Khan, Mainul Mizan, Ragib Hasan, and Alan Sprague, Hot Zone Identification: Analyzing Effects of Data Sampling on Spam Clustering, ADFSL Conference on Digital Forensics, Security and Law Richmond, VA, USA, May 2014

[20] The University of Alabama at Birmingham, UAB phishing data mine, https://cis.uab.edu/uab-spam-data-mine/, [Last Accessed Oct 14th, 2014].

[21] UAB Center for Information Assurance and Joint Forensics Research, The University of Alabama at Birmingham, http://thecenter.uab.edu/.

[22] P. Calais-Guerra, D. Guedes Neto, W. Meira Jr., C. Hoepers, M. Chaves, and K. Steding-Jessen. Spamming chains: A new way of understanding spammer behavior. In: 6th Conf. on e-Mail and Anti-Spam (CEAS), 2009.

[23] P. Calais-Guerra, D. Pires, D. Guedes Neto, W. Meira Jr., C. Hoepers, and K. Steding-Jessen. A campaign-based characterization of spamming strategies. In: Fifth Conf. on Email and Anti-Spam (CEAS), 2008.

[24] A. Ramachandran and N. Feamster. Understanding the network-level behavior of spammers. Proc. ACM SIGCOMM, 2006.

[25] C. Wei, A. Sprague, G. Warner, and A. Skjellum. Clustering spam domains and targeting spam origin for forensic analysis. J. Digital Forensics, Security, and Law, 5: 2010.

[26] Jacob Abernethy and Olivier Chapelle and Carlos Castillo, Webspam Identification Through Content and Hyperlinks. Proc. Adversarial Information Retrieval on Web, 2008.

