Behavior analysis of spam botnets

Conference Paper · February 2008

DOI: 10.1109/COMSWA.2008.4554418 · Source: IEEE Xplore

4 authors, including:

Huriin Husna
Gadjah Mada University
3 PUBLICATIONS **33** CITATIONS

SEE PROFILE

Santi Phithakkitnukoon
Chiang Mai University
67 PUBLICATIONS **459** CITATIONS

SEE PROFILE

Ram Dantu
University of North Texas
108 PUBLICATIONS **963** CITATIONS

SEE PROFILE

Behavior Analysis of Spam Botnets

Husain Husna, Santi Phithakkitnukoon, Srikanth Palla, and Ram Dantu
Network Security Laboratory
Department of Computer Science and Engineering
University of North Texas
Denton, TX 76203 USA.
{*hjh0036, santi, svp0009, rdantu*}*@unt.edu*

***Abstract*—Compromised computers, known as bots, are the major source of spamming and their detection helps greatly improve control of unwanted traffic. In this work we investigate the behavior patterns of spammers based on their underlying similarities in spamming. To our knowledge, no work has been reported on identifying spam botnets based on spammers' temporal characteristics. Our study shows that the relationship among spammers demonstrates highly clustering structures based on features such as Content length, Time of arrival, Frequency of email, Active Time, Inter-arrival Time, and Content Type. Although the dimensions of the collected feature set is low, we perform Principal Component Analysis (PCA) on feature set to identify the features which account for the maximum variance in the spamming patterns. Further, we calculate the proximity between different spammers and classify them into various groups. Each group represents similar proximity. Spammers in the same group inherit similar patterns of spamming a domain. For classification into Botnet groups, we use clustering algorithms such as Hierarchical and K-means. We identify Botnet spammers into a particular group with a precision of 90%.**

I. Introduction

In recent years, a wide variety of spam filtering techniques have gained popularity. These techniques, though widely used, cannot be applied to discover the common trends of variations shared by the majority of spammers spamming a given domain. Spam filtering techniques such as statistical analysis, email authentication standards, trust, and reputation techniques will not identify common association patterns among the spammers. In this paper, we identify the underlying spamming patterns by applying Principal Component Analysis (PCA) [5] and Clustering Techniques. PCA allows the identification of association patterns of groups of spammers and the individual behavior of a spammer in a given domain. With the aid of PCA it is possible to find association patterns such as,

1) Spammers spamming at the same time of the day
2) Spammers sending the same content over and over
3) Spammers changing their email id's and spamming the same recepient
4) Spammers sharing contact lists

From this analysis we found that, for multiple recipients' in a domain, the kinds of spam received is diverse enough to render the common association patterns among spammers in such environment insignificant. However, when recipients are analyzed individually, common association patterns become stronger.

We also analyzed the individual spammer association patterns periodically and found a consistent trend. Because most spammers show common association patterns, or a common behavior over the time, we identify such patterns as the eigen-behaviors of spammers. Eigen-behaviors can be used to characterize [1] individual spammers and describe their association behaviors. We applied PCA, clustering techniques [2], [3] (such as, hierarchical and K-means) on three email corpuses to validate our findings with the recipients' preferences.

The remainder of this paper is organized as follow. A description of the problem is given in Sec. II. A brief description of background work is the subject of Sec. III, followed by a detailed discussion on the eigen-behavior of spammers in Sec. IV. In Sec. V we discuss hierarchical and K-means clustering techniques used in identifying and grouping the botnets. Section VI details the hand labeling of the corpus and calculation of precision in clustering botnets. We conclude this paper with a brief discussion of related work and offer our observations in Sec. IX

II. Problem Definition

With the amount of email spam received these days, it is hard to imagine that spammers act individually. Most spams nowadays have been sent from a collection of compromised machines controlled by some spammers. These compromised computers are often called bots. By using them, spammers can send a massive volume of spams within a short time. According to a recent survey [4], spammers sent an estimated 80% of email spam by using zombie PCs. About 30,000 new machines are compromised daily and become bots. One of the most common usages of botnets is to launch massive spams. Spam remains an annoying problem because a majority of spam filtering techniques focus on the content of an email, which is in complete control of the spammers. Most spammers and phishers obfuscate their email content to circumvent spam filters. So, such techniques are not of use as their classification strategies depend upon the message's meaning. Our approach avoids this limitation as we base classification on the individual user's behavior.

The motivation of this work is to understand the behavior of spammers through a large collection of spam mails. To gain this understanding, we analyzed a data set collected over 2.5 years (corpus-I) and developed an algorithm which gives

us the Botnet Features and, then, classifies them into distinct groups. We use principal component analysis (PCA) to analyze the association patterns of groups of spammers and to analyze the behavior of individual spammers within a given domain. These analyses are based on features which capture maximum variance of the information we have clustered.

We classified each spammer's behavior based on features of its Header Contents. Spammers obfuscate their spam emails' content; however, because our analysis does not focus on an email's content, such content is irrelevant to our results.We categorized each email spammer based on features like IP address, Content Length, time of arrival, frequency of spamming and content type, e.g., 'MIME-Version' - (used for encoding binary content as attachments.) Because we are analyzing spammers' behaviors, other parameters such as reciprocity, read emails, and storage time do not apply, as we assume that users do not read telemarketing emails.

First, we developed a feature set for each Spammer, as a data set matrix. We then applied PCA to the feature set to identify features which captures most of the variance in the data set. Further, we clustered these spammers into groups based on their behavior patterns. We considered the possibility that a spammer might spam multiple receivers within the same domain. To examine this we identified common behaviors for a group of spammers by using the proximity between the senders (using a distance metric) and by applying clustering algorithms.

To verify our approach's accuracy, we hand-labeled the data and compared those results to the automatic botnet identification of each corpus. Accuracies around 90% have been achieved. Thus, using the proposed technique, we may effectively block those spammers as groups instead of blocking them individually.

III. Proposed Approach

Identifying the behavior patterns of spam botnets is based on three phases which are described next.

Phase I: Feature selection using Principal Component Analysis

The goal of Phase I is to reduce the data set and extract only relevant data through the traditional use of eigen analysis. After performing Eigen analysis, we select the feature set which can be used for grouping and which captures maximum variance in the data. The matrix input is NxM; where N is the number of spammers and M is the number of features in the set.

Phase II: Proximity between Senders

This phase is used to find the association pattern between spammers. The association pattern is evaluated using the Euclidean Distance defined between pairs of senders. Input given to this stage is the NxL matrix from Phase I. The output is a NxN Proximity Matrix (a dissimilarity matrix); it gives us the proximity relation between each pair of senders. The lesser the value the more closely associated the particular pair of spammers are.

Phase III: Grouping Methods

In Phase III, we want to cluster spammers who exhibit similar patterns of spamming behavior. Now, using the proximity matrix generated in Phase II, we group spammers with similar proximity values into one cluster, using algorithms such as Hierarchical and K-means clustering. The output at this stage gives us clusters of spammers. Thus, based on the closeness of each sender, we have groupings of compromised computers, i.e., bots for each spammer.

IV. Background

The approach proposed here relies on two existing methods. The first, Principal Component Analysis, is used to extract the components that have a higher impact for a given data set. The second method, clustering, enables classifying individuals into groups evincing similar patterns. A brief description of both techniques is given next.

A. Principal Component Analysis

Typically, [5] PCA is used to reduce the dimensionality of a data set consisting of a large number of interrelated variables while retaining as much as possible of the variation present in the data set. This is achieved by transforming to a new set of variables, the principal components (PCs), which are uncorrelated, and ordered so that the first few retain most of the variation present in the original variables.

Suppose that X is a vector of p random variables, and we want to infer about the variances of the p random variables and the structure of the covariances or correlations between the p variables. One observes the p variances and $\left\{\frac{1}{2} \times p(p-1)\right\}$ correlations or covariances, which increase in complexity with the increase in size of X. An alternative is to observe for a few derived variables $(<< p)$ [5] that preserve most of the information given by these variances and correlations or covariances.

1) Choosing a Subset of Principal Components: In this study we use the scree test, developed by Cattell [6], to decide how many PCs should be retained to account for most of the variation in X.

Principal components are successively chosen to have the largest possible variance [5]. Suppose the variance of the k_{th} PC is l_k, scree test involves looking at a plot of l_k against k and deciding at which value of k the slopes of lines joining the plotted points are 'steep' to the left of k, and 'not steep' to the right. This value of k, (defining an 'elbow' in the graph), is taken to be the number of components m to be retained.

B. Hierarchical Clustering

Hierarchical algorithms [7] can be agglomerative ("bottom-up") or divisive ("top-down"). Agglomerative algorithms begin with each element as a separate cluster and merge them into successively larger clusters. Divisive algorithms begin with a whole set and divide the set into successively smaller clusters.

We have used an agglomerative algorithm for clustering. This method builds the hierarchy from the individual elements by progressively merging clusters. In this approach, clustering begins with each element as a separate cluster. We continue

grouping based on correlation until we have formed one cluster.

C. K-Means Clustering

K-Means clustering provides a simple procedure to classify a data set through a certain number of fixed clusters(assume k clusters). The idea is to define k centroids, one for each cluster. These centroids must be placed carefully because location will affect the results. A better choice is to place the centroids as far as possible from each other. Then, we can take each point belonging to a given data set and associate it with its nearest centroid.

V. Eigen-behavior of Spammers

Typically, incoming emails consist of emails from numerous senders. For instance, emails may originate from telemarketers, fraudsters, family, friends and opt-in senders. As the percentage of these unsolicited emails increases in the incoming email traffic, annoyance or nuisance increases, resulting in a loss of productivity. To verify this, we categorized the incoming email traffic collected at an enterprise's mailserver by asking the recipients to hand-label their emails. We have examined a corpus of 1496 emails, all from spammers. Based on the spammers' locations, we categorized the traffic profile of the botnet groups. Identifying the spammers' physical locations cannot be achieved using the originating location of the spam as spammers use compromised machines (Bots). Often the spammer is physically located elsewhere. Here we define spammers-feature matrix and group-feature matrix (mix of spam & legitimate emails), over which we performed PCA to identify the association patterns among spammers. Spammers-feature matrix and group-feature matrix are m by n matrices, where m is the total number of senders and n is the number of considered features.

$$Spammers - feature = \begin{pmatrix} e^1_{11} & e^1_{1,2} & \cdots & e^1_{1n} \\ e^1_{21} & e^1_{2,2} & \cdots & e^1_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ e^1_{m1} & e^1_{m2} & \cdots & e^1_{mn} \end{pmatrix}$$

$$group - feature = \begin{pmatrix} e^2_{11} & e^2_{1,2} & \cdots & e^2_{1n} \\ e^2_{21} & e^2_{2,2} & \cdots & e^2_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ e^2_{m1} & e^2_{m2} & \cdots & e^2_{mn} \end{pmatrix}$$

In spammers-feature matrix m, corresponds to the total number of spammers whereas, in group-feature matrix m is a mix of spammers and legitimate senders. The values e^1_{ij} and e^2_{ij} for each entry in the column vector are measurements corresponding to parameters, listed below, extracted from the header of the e-mails.

1) Time of arrival
2) Inter-arrival Time
3) Active Time
4) Content length
5) Frequency

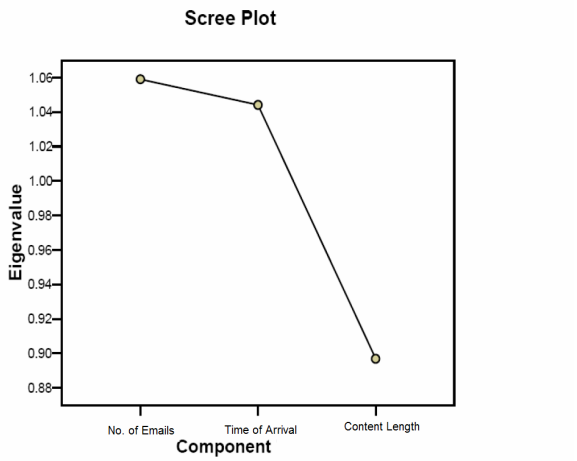

Fig. 1. Scree plot for corpus-I.

6) Content type

The above features have been selected based on PCA. We can see at this point that we do not need to use all the eigenvectors. We represent the data in terms of vectors where the eigen-values are higher than a threshold (pre-specified at the time of analysis). In the following sections we will discuss the strategies used in choosing a subset of principal components and get a consistent representation of the underlying patterns.

A. Components Selection

As a first step in the analysis, we decide how many principal components to retain. This helps identify the predominant features common between the spammers spamming a specific recipient. To achieve this, we retain only components with eigen-values above 1.0 [1], [5]. That is, we drop any component that accounts for less variance than does a single variable. We used scree test [6] to determine the most significant eigenvectors, those account for most of variation.

Figure 1 is a scree plot obtained by performing PCA on corpus-I. From Fig. 1 one can observe that four principal components (Active time, Time of Arrival, Frequency, and Content length) account for most of the email corpus's variation. Therefore, we retained four components for further analysis. The plot also provides a visual aid for deciding at what point including the additional features no longer increases the amount of variance accounted for by a non-trivial amount.

For the corpus-I feature set, the first four components have eigen values greater than 1.0. Component 1:Active Time, component 2:Content Length, component 3:Frequency cumulatively account for 65% of the variance whereas the inclusion of Time of Arrival increases the cumulative variance to 81%. The features Active Time, Content Length, Frequency and Time of Arrival account for the maximum variation.

B. Component Plot in Rotated Space

Another matrix of interest is the component matrix, also known as the feature pattern matrix. Loadings, the entries in this component matrix, are correlations between the components and the variables (in our case the features) between various parameters.

TABLE I
COMPONENT MATRIX.

Component	eigen Value	% of Variance	cumulative %
1	1.798	29.971	29.971
2	1.094	18.231	48.202
3	1.014	16.906	65.108
4	0.994	16.572	81.634
5	0.897	14.954	96.634
6	0.202	3.366	100.00

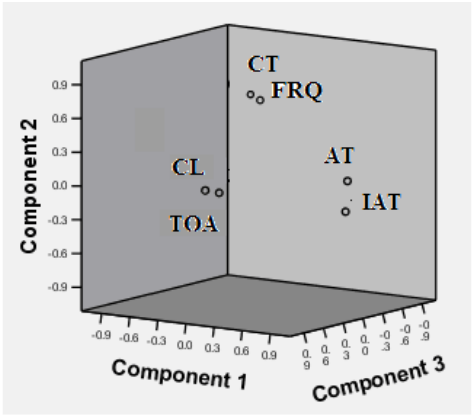

Fig. 2. Component Plot in Rotated Space

Each principal component represents an orthogonal dimension. We retained three dimensions, so that we can plot them on a 3-D plane. But, later in our study we also retain more than three components so that we can examine at several pairwise plots.

We rotate these axes so that the three dimensions passed more nearly through the major clusters. By rotating them, (preserving their perpendicularity), one axis passes through or near the one cluster, the other through or near the other cluster. Table I is the loading matrix after rotation and Fig. 2 gives the component plot in rotated space.

One can see from Figure 2 that different variables load well on different components, Active time shows its highest positive loadings towards the first component and so does Inter-arrival time. Thus component 1 has a strong affinity towards these two parameters. Similarly the next two parameters, Content Length and Time are positively loaded on component 3. Using this qualitative approach, we provide quantitative based result, where based on two parameters grouping was obtained.

C. Eigen-Clustering of Spammers

Once we obtained the principal components, we identified clusters of spammers sharing similar spamming patterns. This is done by rotating the components and plotting the points in a 3-D plane. Figure 3 displays the corpus-I subjects in a 3-D plane formed by rotating the first three eigen-vectors.

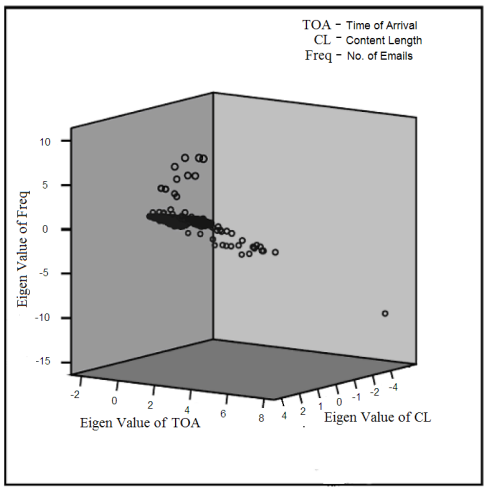

Fig. 3. Cluster Of Spammers sharing Similar Patterns. The Fig. gives a 3-D view of an eigen cluster for corpus-I by rotating the first 3 Eigen Vectors. The cluster of Spammers are well separated from the clusters of legitimate senders.

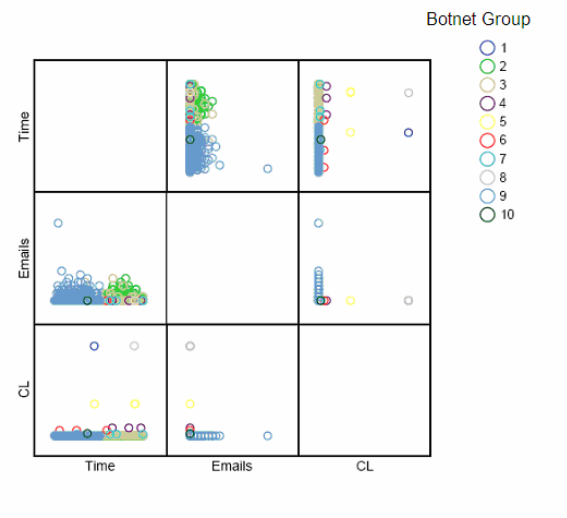

Fig. 4. Cluster of spammers in corpus-I based on their association patterns. Each Color Ring Represents a Group and Each Quadrant describes the similarity of features for a botnet

In the Fig. 3 one can notice the cluster of spammers well separated from the clusters of legitimate senders. We further studied the cluster of spammers and classified these spammers based on the similarities in their association patterns. Figure 4 displays the various sub-clusters of spammers having similar association patterns. Overall results obtained from this analytical study were satisfying. We were able to cluster the subjects in corpus-I with a precision of 91.86% with few false positives and false negatives.

D. Clustering

We performed K-means [8], [9] clustering to corpus 1. As a result, we clustered the spammers into three prominent clusters, (Table II.) We also found 18 missing cases while performing K-means. Missing data values [10] can occur for

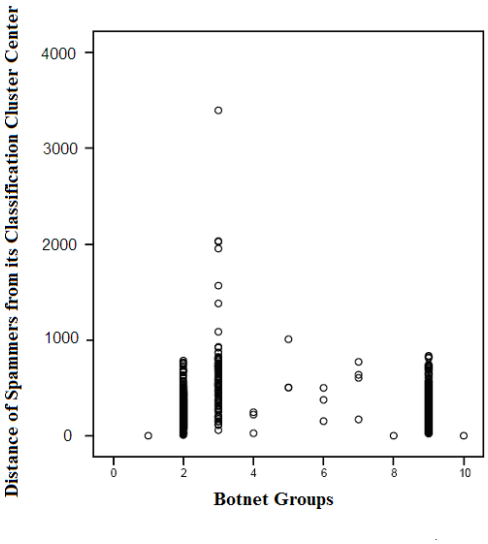

Fig. 5. Results of the application of the k-means clustering algorithm for k=10. The manual analysis of the clusters confirmed that densely populated clusters (Clusters 2, 3, and 9) were indeed botnets.

2 reasons: either a measurement is made and then lost or a measurement cannot be made at all. We signify missing values by using the symbol N in the data matrix. In our analyses, missing data is just 1.7% of the total matrix values, so it does not have impact on the clustering method. When we standardize the matrix, we act as if the Ns are not present. That is, we skip over them.

TABLE II
NUMBER OF CASES IN THE CLUSTERS.

Cluster	Number of Spammers
1	1
2	412
3	96
4	3
5	3
6	3
7	4
8	1
9	498
10	1
Valid	1022
Missing	18

Figure 5 shows the number of clusters on the X axis. The Y axis shows the distance of spammers from their classification cluster's center. Spammers having their centroids close to a classification cluster are classified as a group (represented by a vertical line of spammers who are closely associated in Fig. 5). Thus we can say that the spammers having similar features will have their centroids exactly or close to the classification cluster.

We also have few outliers present in K-means clustering; These senders (outliers) have a pattern of spamming which doesn't fall into any of the clusters. Hence, they comprise a separate cluster, since it's an iterative process any new spammers will be associated with their corresponding clusters.

After performing K-means clustering we further analyzed the results and categorize individuals in each cluster based on similar time of arrival, similar Content Length and similar Frequency. In addition, these individuals also shared other similarities such as possessing the same network ID and having a close (or the same) geographic location. Thus, we identified Bots which were present in each cluster. We discuss the accuracy of these identifications in Sec. IX.

We have represented the association patterns based on the feature set, consider Fig. 4 showing categories of spammers having similar: time of arrival & content length, time of arrival & number of emails, content length & number of emails. since we have set k=10 we will have 10 different cases which belong to one of the block having atleast 2 similar feature in the feature set.

1) Hierarchical Clustering: We also used a hierarchical clustering algorithm which produced similar results. The initial 5 hierarchical levels are depicted in Fig. 7. There are known advantages and disadvantages of using different clustering approaches.

Hierarchical Clustering plays an important role in our observations, as it provides a tree-like structure called a Dendogram which presents hierarchical relations between clusters. Using hierarchical clustering, we capture a concentric cluster,which is not the case in K-means clustering.

Hierarchical clustering is less efficient than K-means as one has to compute at least n x n similarity coefficients and,then, update them during the clustering process. If a data set is very large, efficiency is a key issue. Because K-means was conceptually the simpler method, we used it first on the corpus-I as its results were often sufficient for our analysis. An additional problem associated with an hierarchical clustering approach is that it is not easy to define levels for clusters.

As for the dendograms produced, one must prune the tree structure as per the hypothesis; so, deciding the level might be difficult. K-means algorithm has low complexity $O(nkt)$ where t= no. of iterations. One basic disadvantage of the K-means algorithm is that we must specify k number of clusters to be formed initially. This may lead to erroneous results when we specify less than the cluster groups. Hence, clusters are sensitive to initial assignment of centroids.

The relation between objects is shown in proximity matrix (See Figure. 6) in which rows and columns correspond to objects. In our research, using the Euclidean distance measure, we computed the Proximity matrix, which represents how close the senders are from each other. This is a matrix of: No. of senders $\times$ No. of senders. Where each sender shows his proximity with all the other senders in the corpus. It gives a dissimilarity matrix, i.e lower the value, closer the sender. Proximity Values range from (0- 425) for example, the range of proximities for level=1 is from (0-0.818).

In the Fig. 7, height of the vertical lines and the range of the (dis)similarity axis give visual clues about the strength of the clustering. Long vertical lines indicate more distinct separation

Proximity Matrix

Case	IP Adresses	Squared Euclidean Distance				
		1 xxx.xx.xxx.23	2 yyy.yy.yyy.235	3 zzz.zz.zzz.136	4 ppp.pp.ppp.230	5 qqq.qq.qqq.101
1	xxx.xx.xxx.23	.000	2.396	2.265	7.630	2.314
2	yyy.yy.yyy.235	2.396	.000	.003	6.760	.806
3	zzz.zz.zzz.136	2.265	.003	.000	6.719	.785
4	ppp.pp.ppp.230	7.630	6.760	6.719	.000	2.939
5	qqq.qq.qqq.101	2.314	.806	.785	2.939	.000
6	rrr.rr.rrr.20	1.560	.089	.067	6.554	.724
7	sss.ss.sss.229	2.569	.004	.010	6.817	.842
8	ttt.tt.ttt.210	3.271	.726	.734	3.189	.112
9	uuu.uu.uuu.229	5.040	11.968	11.712	7.070	8.149
10	vvv.vv.vvv.67	2.448	.001	.004	6.776	.817

Fig. 6. Table describes a sample of each IP address (Sender) and its proximity with all the other senders in the corpus. The value indicates the dissimilarity; this means the lesser the value, the more closeness between the two senders.

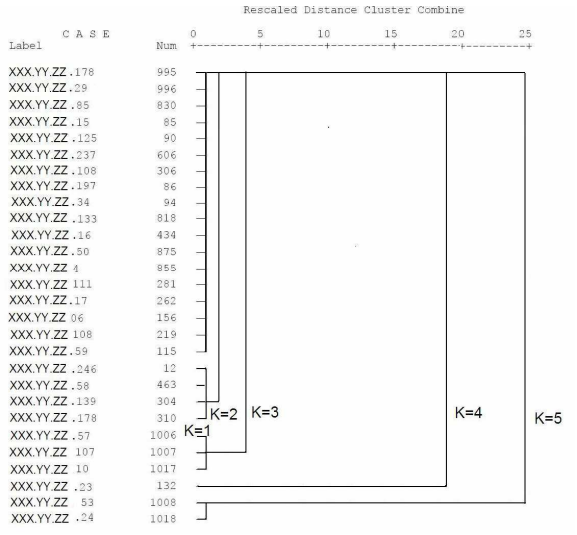

Fig. 7. A sample of Hierachy levels of the algorithm up to level k=5

between the groups. Long vertical lines at the top of the dendogram indicate that the groups represented by those lines are well separated from one another. Shorter lines indicate groups that are not distinct.

Each level of the tree in Fig. 7 represents a partition of the input data into several (nested) clusters or groups. All the IP addresses which are closely related are nested under one level (k) , if the distance of the similarity increases then it jumps altogether to another level (say k=2) indicating less association between them.

VI. Clustering of active spammers and timing analysis

The most threatening spammers are the botnet spammers and within the botnet we want to catch the ones who are most active and are highly serious in spamming. We relate seriousness term to the one who is having high frequency of spamming within a given short period of time. While it is difficult to estimate the total number of systems that participate in botnets at any point in time, but during the process of sub clustering, we could get hold of the number of spammers within the same botnet group who at the same time and in a similar pattern spam a domain. Active spammers are categorized as those senders who send spam email within

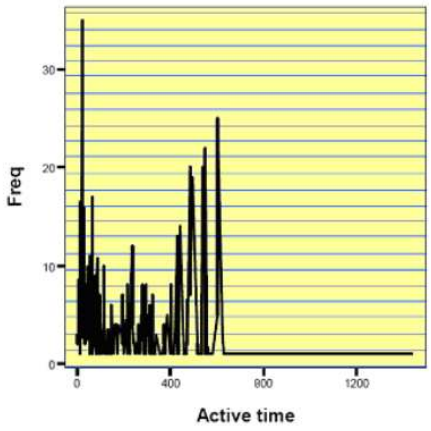

Fig. 8. This figure includes analysis of all the botnet groups in the spamming domain. Graphs describes the first 1200 minutes of active spammers. Dense activity means botnet groups are active for a short period of time. From our analysis we can see that majority of spammers lie within the active time of first 400 minutes.

the same range of time. So, spammers in the same group are active for same amount of time. Furthermore, the magnitude of the botnet threat is now widely recognized to be compounded by the emergence of an active botnet economy that is likely to be funded by an organized crime. We further analyze the structure and behavior of Botnets, (e.g., the network effects and impact of multiple bots communicating and reacting to botmaster commands) will grow. We can analyze further from our results that within a botnet group we have sub-clusters and these sub clusters are having similar trend of spamming in terms of frequency and inter-active time, but their location of spamming is different. In our analysis the bot machines from US are spamming during the daytime whereas from a different sub cluster from the same botnet group is spamming in a similar trend from China/Australia. So one can say that though these guys have separate locations and separate active time they are managed by the same Bot Master which is sitting at a common place. We were also able to track bots located at different places, by capturing their trend of spamming and by studying their timing patterns.

A. Botnet Propagation Based on Time of arrival

The graph represents a particular botnets timing patterns based on their Time of spamming. Each bot is assumed to be a programmed machine or a compromised machine which will spam within a time slot. We see bots spamming in burst and then they are inactive throughout the day. But several bots are mastered together to keep spamming the entire day. Botnet 9 has a trend of sending most of emails within the time range of 10 am to 9 pm. There are few spam emails not falling in that category but they are treated as false positives and negatives in our case. Whereas on the other hand botnet groups 2 and 3 have patterns of spamming during late nights and early mornings so we assume them to be probably set up somewhere outside USA (usually we observe a trend of receiving spam

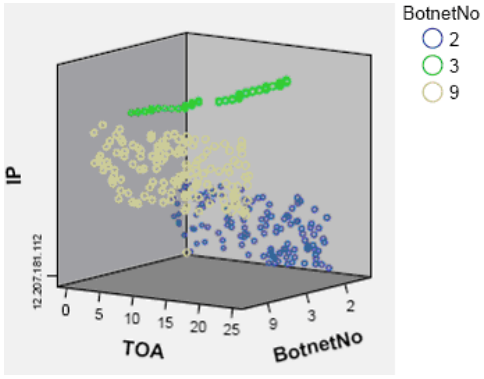

Fig. 9. Looking at all 3 botnet groups at a time: Where on the x-axis we have time of arrival in minutes based on 24 hour scale, y axis all the spammers and z axis has the botnet group they belong to.

mails early morning when we check our email in-box).

VII. Hand labeling & Precision

If an instance (here an, email) is unwanted (spam or phishing) and classified as unwanted, then, it is counted as true positive, "TP". If an instance is wanted (legitimate) and classified incorrectly as unwanted, it is counted as false positive "FP". Let "P" and "N" be the total number of positive and negative instances in a corpus; we determined the precision, true positive rate "tp_{rate}" and false positive rate "fp_{rate}" of researched classifier as:

$$Precision = \left\{ \frac{TP}{TP + FP} \right\} \quad (1)$$

$$tp_{rate} = \left\{ \frac{TP}{P} \right\} \quad (2)$$

$$fp_{rate} = \left\{ \frac{FP}{N} \right\} \quad (3)$$

To calculate the efficiency of our clustering technique and botnet classification, we required a means of measuring the precision. So, we hand labeled the spam data set and calculated the accuracy of the technique. The most significant feature for classification of Botnet grouping was the Network Id and Location from where the spam email was sent. We can observe from the relevant work [11], [12], [13] that Bots possess a similar network Id and they may have same location of spamming, thus, because of the nature of bots, spamming usually inherits similar behaviors when spamming a domain. Table 3 describes hand labeling of botnet 9 and we can see that the spammers display similar behavior and characteristics. They have their spamming time very close to each other.

During hand labeling, we also correlated characteristics of spam emails, such as, country, city, network ID, Time of Arrival, Active Time, Content Length and Frequency. The following properties are observed during our experiments and hand labeling:

1) Spammers send large numbers of emails in a short period of time.

TABLE III
Precision of analyses performed on corpus-I using K-means. Please note that due to discrepancy in clustering algorithms, the classified number of emails in each technique are slightly different. However, majority of the emails were classified correctly by both the techniques.

Corpus-I	Cluster Analyses	True Positives	False Positives	False Negatives	Precision Hits
Botnet2	417	379	34	4	91.76%
Botnet3	97	81	11	5	88.04%
Botnet9	484	436	41	7	91.40%

TABLE IV
Precision of analyses performed on corpus-I using Hierarchical Clustering.

Corpus-I	Cluster Analyses	True Positives	False Positives	False Negatives	Precision Hits
Botnet2	371	242	60	15	80.10%
Botnet9	257	222	30	5	88.09%
Botnet3	396	298	91	7	77.00%

2) Spammers send the same content repeatedly (with different IDs).
3) Spammers send the same number of emails in a given day.
4) Spam from a botnet arrives at its destination at the same time although spammers are located in geographically distributed locations.
5) Based on our corpus,major source of spam is from USA, followed by Malaysia and China. Moreover, due to the time difference, Asia is actively sending spam while USA is sleeping.
6) A small number of botnets account for most spam.
7) Spammers and legitimate users share SMTP paths and relays [14].
8) In the email corpus of a single user, most spam was generated from a few botnets and the emails were highly correlated.
9) Hand labeling of botnets for a large corpus is humanly difficult and requires behavior-based automation (e.g., the techniques described in this paper).

We used the true positive rate and false positive rate during Receiver Operating Characteristics [15] analysis performed to optimize the performance. Tables IV shows the precision of our filter.

VIII. Related wok

To our knowledge, only one work reports on group-based anti-spam strategies. Li and Hsieh, [16] conducted an empirical study on the clustering behavior of spammers and detected groups of spammers. Majority of the spammers emails content is related to money. Authors found 2% of the spammers accounted to 20% of the spam, whereas 68% of the spammers sent only one spam.

Authors converted ASCII characters in the URL into binary data and calculated complementary cumulative distribution function (CCDF) [16] of spam scores of the spam-groups.

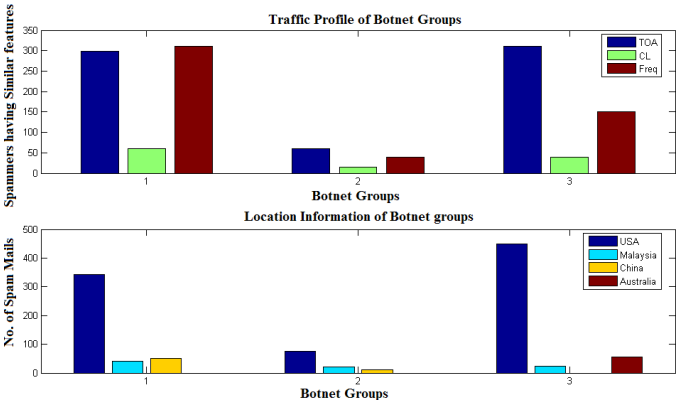

Fig. 10. Email traffic based on the geographic locations of spammers present in the clusters 2,3 & 9. In our experimental corpus-I, United States contribute the highest amount of bot spam followed by Malaysia. Also, the majority of bots share similar time of arrival.

The authors claim to block 70-90% of email based on the URL-based group approach. However, these results are limited to content-based filtering, in particular for URL and money-specified emails. Moreover, they observed money-amount-based clustering structures may not be effective for group-based, anti-spam strategies. We believe that spammers share email lists and use this list for various kinds of advertisements. Therefore, content-based grouping may not work for blocking all types of contents.

In contrast, we grouped spammers based on their behavior and transmission patterns. These patterns show high correlation between group members irrespective of geographic location, network ID, content, and kind of receivers. In addition, we verified our results using hand labeling. To our knowledge, there is no work reported on detecting botnets based on their behavior and in particular transmission patterns.

IX. Conclusion

An in-depth understanding of botnet behavior is a precursor to building effective defenses against this serious and fast growing threat in emails and in future it would be the voice over IP applications. Using our technique we were able to perform a range of experimental study on new methods and tools for characterizing, comparing, identifying, tracking, dismantling, and preventing botnets.

In this work we investigated the clustering structures of spammers based on spam traffic collected over a period of 6 months. Our analysis shows that the relationship among spammers demonstrate highly clustering structures based on features such as Content Length, Time of Arrival and Frequency of an email. We extracted many features like Content Type and Storage Time but did not use them because the eigen values were very low and these features were eliminated during Scree plot.

The inter-arrival time of spam from the same group of spammers exhibits long-range dependence in the sense that the spam from the same group of botnets often arrives in-burst. We also observed that spammers associated with multiple groups tend to send more spams in the near future. We need to emphasize that group-based method may not be highly effective as a stand-alone approach as some groups may have only one member. Some botnet groups had 1-10 spammers and we categorized them as outliers in our analysis.

Using the described clustering techniques, we could accurately identify Botnets as Bots usually inherit similar behavior when spamming a domain. We also could identify by hand labeling that these botnets indeed fall into a particular cluster with a precision closet to 90%. We will continue to explore interesting properties of the clustering structures of telemarketing spammers as our future work and also deploy the above techniques as a complementary tool for existing anti-spam tools.

Acknowledgment

This work is supported by the National Science Foundation under grants CNS- 0627754, CNS-0516807, CNS-0619871, and CNS-0551694. Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

[1] K. Yeung, W. Ruzzo, "An empirical study on Principal Component Analysis for clustering gene expression data," *Bioinformatics*, November, 2000.

[2] K. Xu, Z. Zhang, S. Bhattacharyya, "Profiling and Clustering Internet Hosts," *Sprint ATL Research Report RR05-ATL-020777*, Tech. Rep., February 2005.

[3] J. Erman, M. Arlitt, A. Mahanti, "Traffic Classification Using Clustering Algorithms," *SIGCOMM06 MineNet Workshop*, Pisa, Italy, September 2006.

[4] *Cybercrime Primarily Originates in the U.S., Report Says,Symantec Corporation* http://www.thenewsmarket.com/CustomLink /CustomLinks.aspx?GUID=053d4530-3ffd-4423-9089-ceae4ca598e9, March 19, 2007.

[5] I.T. Jolliffe, *Principal Component Analysis*, 2nd ed. Springer Series in Statistics, 1986, New York, USA.

[6] M. McDermeit, R. Funk, M. Foss, M. Dennis, "Exploratory Factor Analysis with alpha method and varimax rotation," LI Analysis Training Series, 2000.

[7] C. Ding and X. He, "Cluster merging and splitting in hierarchical clustering algorithms," *IEEE International Conference on Data Mining (ICDM'02)*, Japan, 2002.

[8] H. Spth, "Cluster Analysis Algorithms for data reduction and classification of objects," Ellis Horwood, 1980.

[9] R. Dubes, A. Jain, *Algorithms for Clustering Data*, Prentice-Hall Advanced Reference Series, 1988.

[10] H. Charles, *Cluster analysis for Researchers*, Lulu Press, 2004, North Carolina, USA.

[11] E. Cooke, F. Jahanian, D. McPherson, "The Zombie Roundup: Understanding, detecting, and disrupting botnets," *SRUTI Workshop*, July 7, 2005.

[12] D. Dagon, C. Zou, W. Lee, "Modeling Botnet Propagation Using Time Zones," *In Proceedings of the 13th Network and Distributed System Security Symposium NDSS*, February 2006.

[13] A. Ramachandran, N. Feamster, D. Dagon, "Revealing Botnet Membership Using DNSBL Counter-Intelligence," *USENIX, SRUTI*, 2006.

[14] S. Palla, "A Multi-Varaite Analysis of SMTP Paths and Relays to restrict spam and phishing attacks in Emails," Masters Thesis, Masters in Computer Science, *University of North Texas*, December, 2006.

[15] F. Tom, *Notes and practical considerations for researchers*, Kluwer Academic Publishers,The Netherlands, 2004.

[16] L. Fulu, M. Hsieh, "An Empirical Study of Clustering Behavior of Spammers and Group-based Anti-Spam Strategies," *CEAS*, 2006.