Early Detection of Outgoing Spammers in Large-Scale Service Provider Networks

Yehonatan Cohen, Daniel Gordon, and Danny Hendler

Ben Gurion University of the Negev, Be'er Sheva, Israel
{yehonatc,gordonda,hendlerd}@cs.bgu.ac.il

Abstract. We present *ErDOS*, an Early Detection scheme for Outgoing Spam. The detection approach implemented by ErDOS combines content-based detection and features based on inter-account communication patterns. We define new account features, based on the ratio between the numbers of sent and received emails and on the distribution of emails received from different accounts.
Our empirical evaluation of ErDOS is based on a real-life data-set collected by an email service provider, much larger than data-sets previously used for outgoing-spam detection research. It establishes that ErDOS is able to provide early detection for a significant fraction of the spammers population, that is, it identifies these accounts as spammers before they are detected as such by a content-based detector. Moreover, ErDOS only requires a single day of training data for providing a high-quality list of suspect accounts.

Keywords: spam, classification, early detection, email service provider (ESP)

1 Introduction

Email is an important and widespread means of communication used by over 1.8 billion people, often on a daily basis [1]. Due to its widespread use, email has become a fertile ground for cyber-attacks such as phishing, spreading of viruses and the distribution of spam mail, consisting of unsolicited messages mostly of advertisement contents. According to recent statistics, approximately 95.3 trillion spam emails were sent during 2010. This is estimated to be almost 90% of all email traffic [2].

As far as users are concerned, spam [3] is mainly a nuisance, wasting much of their time while having to skim through vast amounts of junk email in search of emails of importance. In addition, spam may contain abusive or dangerous content [4].

Email Service Providers (ESPs) also suffer from spam emails and must therefore combat it. First, vast amounts of email being sent from ESP domains or being sent to these domains overload ESP servers and communication infrastructure [5]. In addition, ESPs from which large numbers of spam messages are sent are likely to become blacklisted, thereby preventing the legitimate users of these

ESPs from exchanging email and disconnecting them from external domains. Indeed, ESPs that fail to deploy effective spam filtering mechanisms provide poor user experience and thus hurt their popularity and reputation [6].

A number of techniques for detecting spamming accounts and mitigating outgoing spam have been developed and are already in use. Content-based filters, i.e. filters that learn and identify textual patterns of spam messages, are used by most ESPs [7, 8]. Unfortunately, spammers have developed their own techniques to circumvent content based filters, such as image-spam mails [9] and hash-busters [10].

Whereas content-based filters consider the properties of individual messages, a different approach examines the social interactions of email accounts, reflected by inter-account communication patterns, since, in many cases, the social interactions of spammers and non-spammer users are significantly different [11–16].

Social interactions can be modeled using a communication graph in which a vertex is generated for each email account appearing in the data set and an edge connecting two nodes appears if and only if there was email exchange between the two accounts represented by these nodes. Communication graphs can be either directed or undirected. Edges may be weighted, e.g. by using a weight function that assigns each edge the number of emails communicated between the two accounts represented by its end-points [13].

After modeling social interactions by communication graphs, network-level features that distinguish between legitimate (non-spamming) accounts and spamming accounts can be extracted. A few methods are based on the assumption that spammers are less likely to receive messages and, in particular, are less likely than legitimate users to receive messages from the accounts to which they send messages [13, 14]. Another approach is based on the assumption that spammers send emails to accounts which do not communicate between themselves. One such feature, Clustering Coefficient (CC) [17], measures the probability that two recipients are "familiar" with each other. Based on such features, a machine learning model is trained in order to detect spamming accounts in future logs.

Several studies synthetically generated outgoing spam traffic, based on the assumption that none of the accounts in their data sets were spammers [11, 14]. Previous studies of outgoing spam also made use of email traffic originating from academic institutes [13, 11, 16] and of log files collected by non-ESP organizations [14].

Lam and Yeung [14] present a machine-learning based outgoing spammer detector that uses inter-account communication pattern features. Our approach is also machine-learning based and uses features based on communication patterns, but there are several significant differences between our work and theirs. First, whereas they train their detector using synthetically-generated spammer accounts, we use real spamming accounts for training, identified as such by a content-based spam filter. Second, Lam and Yeung use a data set taken from a non-ESP organization (Enron). Finally, unlike Lam and Yeung, our goal is to achieve *early detection*, that is, to detect spammers before they are detected by a content-based filter and possibly even if they are not detected by the filter at

all. As established by our empirical analysis, the features we use and our training approach yield significantly better results on our data set than previously published algorithms in terms of both accuracy and early detection.

1.1 Our Contributions

This study is based on a large real-life data set, consisting of both outgoing and incoming mail logs involving tens of millions of email accounts hosted by a large, well-known, ESP. It was made available to us after having undergone privacy-preserving anonymization pre-processing. This data set is much larger than data sets used by previous outgoing-spam detection research.

Using this data set, we evaluated previously published outgoing-spam detection algorithms. Our experimental evaluation finds a large drop in their accuracy on this data set as compared with the results on the data sets used in their evaluation, indicating that algorithms optimized for small and/or synthetic datasets are not necessarily suitable for real-life mail traffic originating from large ESPs. New approaches are therefore needed in order to efficiently detect outgoing spam in large ESP environments.

Our emphasis in this work is on *early detection of spamming accounts hosted by ESPs*. We present *ErDOS*, an Early Detection scheme for Outgoing Spam. The detection approach implemented by ErDOS combines content-based detection and features based on inter-account communication patterns. ErDOS uses novel email-account features that are based on the ratio between the numbers of sent and received emails and on the distribution of emails received from different accounts. By using the output of a content-based spam detector as a means for obtaining initial labeling of email accounts, we manage to avoid the use of synthetically-generated spam accounts as done by some prior work.

ErDOS uses the account labels induced by the output of the content-based detector for supervised learning of a detection model based on the features we defined. Empirical evaluation of ErDOS on our data set shows that it provides higher accuracy as compared with previous outgoing-spam detectors. Moreover, by using only a single day of training data, ErDOS is able to provide early detection for a significant fraction of the spammers population.

The rest of this paper is organized as follows. In Section 2, we describe the data set used for this research and the manner in which features are extracted. We describe the features used by ErDOS in Section 3. In Section 4, we describe the structure of ErDOS and the process of model generation. We report on our experimental evaluation in Section 5. The paper concludes with a summary of our results and future work in Section 6.

2 Data Set and Feature Extraction

Figure 1 depicts the manner in which the data set we received is generated by the ESP's mail servers. The data set is composed of two parts - incoming logs and outgoing logs, which store log records of incoming and outgoing email messages,

respectively. The left-hand side of Fig. 1 provides a schematic illustration of how incoming emails are processed.

First, incoming messages originating from blacklisted IPs are filtered out. Incoming messages that are not filtered out are processed by a content-based spam detector that tags incoming messages as either spam or ham (non-spam). Whether or not incoming messages tagged as spam are filtered out depends on the receiver's identity: the ESP has different contracts with different customers, which in some cases mandate that spam messages should still be relayed to the customer, with a "Spam" tag appended to the message subject; in other cases, spam messages are simply discarded. Incoming emails that are relayed to internal customers generate log lines that are written to the incoming mail logs, including an indication of whether the respective message is spam or ham.

The right-hand side of Fig. 1 illustrates how outgoing messages are processed. Unlike incoming messages, outgoing messages are not subject to IP blacklist filtering, but they do undergo content-based spam detection. Outgoing messages are relayed to their destinations and generate log lines that are written to the outgoing mail logs. Messages whose destinations are internal accounts (i.e. accounts hosted by the ESP) are relayed by an outgoing mail forwarded server, whereas messages sent to external accounts are relayed by an outgoing mail server.

Our data set consists of log records collected over a time period of 26 days. During the first 4 days, both incoming and outgoing log records were provided. During the rest of the period (additional 22 days), only outgoing log records were provided. Table 1 compares our data set with the largest data sets used in previous studies of outgoing spam detection.[1]

[1] The details of additional, smaller, such data sets are presented in [15, Table 1].

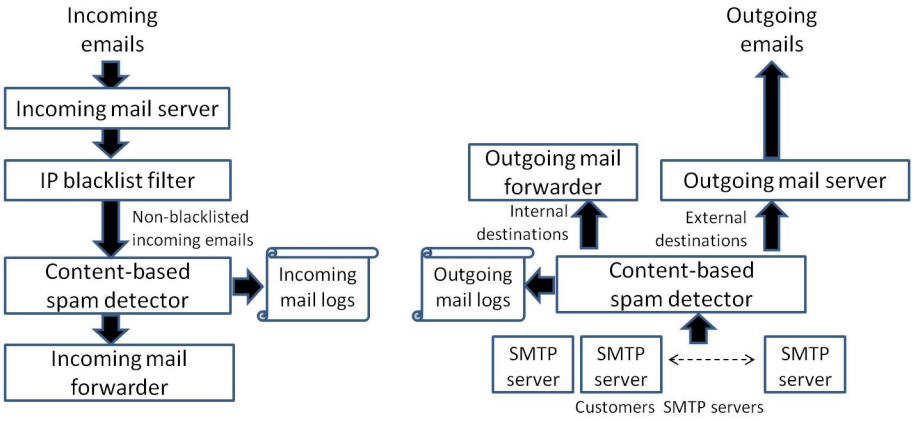


Fig. 1. Data-set collection process.

Table 1. Our data set vs. data sets used by previous studies.

	Our data set		SUNET[15]	NTU[16]	Kossinet et al.[19]	Enron[14, 20]
#mails	9.86E7	2.13E8	2.40E7	2.86E6	1.46E7	5.17E5
#edges	7.40E7	7.40E7	2.16E7	-	-	3.68E5
#accounts	5.63E7	5.81E7	1.05E7	6.37E5	4.35E4	3.67E4
time period	4 days	26 days	14 days	10 days	355 days	3.5 years
contents	spam & ham		spam & ham	spam & ham	-	ham

It can be seen that the rate of email traffic (number of emails sent or received per day) logged in our data set is between 1-3 orders of magnitude higher than that of the other data sets. The reason for this huge difference is that, whereas previous studies of outgoing spam mostly use small sized datasets taken from academic institutes [16, 11, 13, 18] or from Enron's public data set [14], our study uses a data set collected by the mail servers of a large ESP.

Table 2 provides additional information regarding our data set. An account is designated as a *spamming account*, if it sent at least one message tagged as spam by the content-based spam detector. The log files in our data set contain meta-data regarding exchanged emails, such as the date and time of delivery, the IPs they were sent from[2] (for emails originating from external ESPs), etc., but do not contain any information regarding the contents of the email, except for a tag, assigned by the content-based spam detector, indicating whether or not the email is considered by it as spam. Each internal account has a unique identifier associated with it, which appears in every log record that corresponds to a message sent by it, regardless of the address used by the account in the "From" field when sending the email. Table 2 presents information on the four days of outgoing logs for which there are also incoming logs in addition to information on all 26 days of outgoing email logs.

The features used by the ErDOS detector require information regarding the numbers of sent and received emails by the ESP's accounts. Information regarding received emails was extracted from incoming log files, which contain information about every email which was sent from external ESPs to local accounts. Emails which were filtered out by IP blacklists do not appear in these logs. In-

[2] The source IP field is anonymized

Table 2. Dimensions of our data set.

	incoming logs 4 days	outgoing logs 4 days	outgoing logs 26 days
#emails	8.27E7	1.59E7	1.14E8
#spam	2.72E6	3.2e5	2.42E6
#ham	8.00E7	1.56E7	1.12E8
#accounts	5.40E7	2.31E6	4.12E6
#spamming accounts	6.01E5	3,099	1.22E4
size	59.8GB	11.6GB	61.3GB

formation on emails sent by internal accounts was extracted from outgoing logs. Features are extracted by offline processing of the data set's log files.

3 Features Used by the ErDOS Detector

We extracted and evaluated multiple features whose goal is to differentiate between spamming and legitimate (non-spamming) email accounts. We have found that a combination of multiple features, encompassing various aspects of an account's behavior, yields significantly better results than those obtained by any single feature by itself. In this section, we provide a detailed description of the features that are used by our detector.

3.1 Ratio of Numbers of Sent and Received Emails

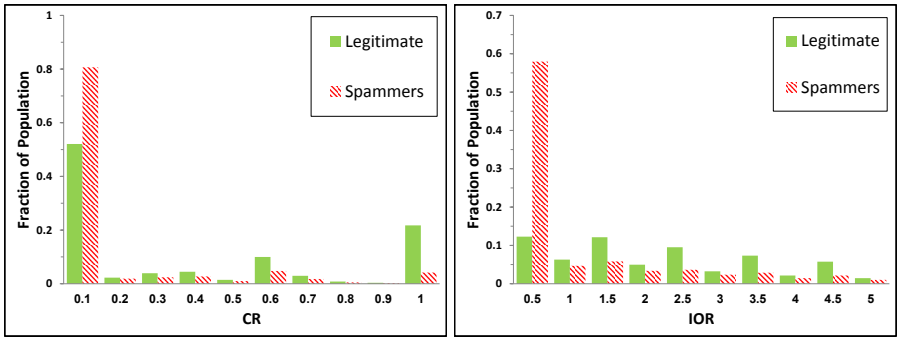
Let a be an account. We denote by $\mathcal{I}(a)$ and $\mathcal{O}(a)$ the number of incoming messages received by a and the number of outgoing messages sent by a, respectively. A feature that we found to be indispensable for early identification of spamming accounts is the *Incoming Outgoing Ratio* (IOR) defined in Equation 1.[3]

$$IOR(a) = \frac{\mathcal{I}(a)}{\mathcal{O}(a)}. \tag{1}$$

Empirical analysis of our data set establishes that spamming accounts have a significantly lower IOR than legitimate accounts. The average IOR of spamming accounts is 1.02, whereas for legitimate accounts it is 8.63. The key reasons for the difference between the IOR values of legitimate and spamming accounts are the following. First, legitimate users often belong to mailing lists and receive messages sent to these lists, whereas accounts dedicated to spamming typically do not. Second, and as observed also by prior works (e.g., [14]), legitimate users are typically involved in social interactions and hence many of the messages they send are responded to. An outgoing spam message, on the other hand, seldom leads to the receipt of an incoming message: even if the spam message is not filtered and arrives at its destination, users rarely respond to such messages; and even if they try to respond, they often can't, since the sender's email address is, in most cases, spoofed.

An outgoing spam detection feature similar in spirit to the IOR feature defined above is *Communication Reciprocity* (CR), presented by Gomes et al.[13]. The CR of an account a quantifies the fraction of accounts with which a had bi-directional communication out of the accounts to which a sent emails. More formally, let $\mathcal{RA}(a)$ denote the set of a's *recipient accounts*, that is the set of

[3] Only accounts which have sent at least a single message are considered when the model is built and when detection is performed. Consequently, the denominator of the IOR feature, as well as of the other features described in this section, is always positive.



(a) (b)

Fig. 2. CR and IOR features' value distributions.

accounts to which a sent emails, and let $\mathcal{SA}(a)$ denote the set of a's *sender accounts*, that is the set of accounts from which a received emails. The CR feature is defined in (2).

$$CR(a) = \frac{|\mathcal{RA}(a) \cap \mathcal{SA}(a)|}{|\mathcal{RA}(a)|}. \tag{2}$$

Figure 2 shows the value distributions of the IOR and CR features across our data set's accounts. It can be seen that low IOR values separate spamming and legitimate accounts better than low CR values. Specifically, only 2.7% of legitimate accounts have an IOR of 0 as compared with 35% of the spamming accounts, whereas almost 30% of legitimate accounts have a CR of 0 as compared with 54.3% of the spamming accounts. Our detector does not use the CR feature.

3.2 Internal/External Behavior Consistency

Although a large fraction of spamming accounts are characterized by very low IOR values, a non-negligible fraction of these accounts have relatively high IOR values which makes it difficult to tell them apart from legitimate accounts based on the IOR feature alone. We next define the *Internal/External Behavior Consistency* (IEBC) feature, which allows us to identify some of these latter accounts.

The rationale behind the IEBC feature is the following. Accounts may communicate with accounts inside the ESP's domain (internal accounts) or with accounts outside of it (external accounts). For legitimate users, the values of IOR for communication with internal and external domains are expected to be similar, since both reflect the characteristics of an account's social interactions. For spamming accounts, however, for the reasons described in Sect. 3.1, incoming and outgoing communications are mostly uncorrelated. Consequently, the internal and external IOR ratios for spamming accounts are expected to vary significantly more than those of legitimate accounts.

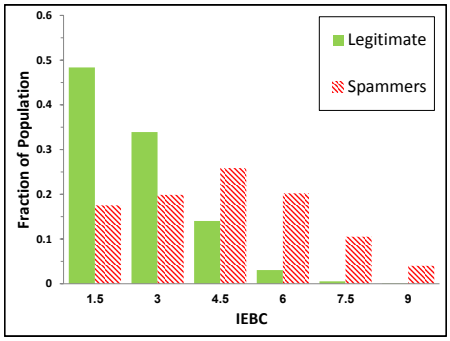


Fig. 3. IEBC values distribution

We let $\mathcal{I}_I(a)$ and $\mathcal{I}_E(a)$ denote the numbers of a's incoming messages from internal and external domains, respectively. Similarly, we let $\mathcal{O}_I(a)$ and $\mathcal{O}_E(a)$ denote the numbers of a's outgoing messages to internal and external domains, respectively. The IEBC feature is defined formally in Equation 3.

$$IEBC(a) = \left| \log_2 \frac{\big(1 + \mathcal{I}_E(a)/(1 + \mathcal{O}_E(a))\big)}{\big(1 + \mathcal{I}_I(a)/(1 + \mathcal{O}_I(a))\big)} \right|. \tag{3}$$

The nominator and denominator express the IOR ratios with external and internal domains, respectively, where one unit is added to each factor for avoiding division by 0. A log of the ratio is taken and then the absolute value is computed, in order to map large discrepancies between the external and internal IOR values to large IEBC values, regardless of whether the external IOR is significantly larger than the internal IOR or vice versa.

Figure 3 shows the distribution of IEBC values for spamming and legitimate accounts. The IEBC values of legitimate accounts are significantly smaller, indicating, as suspected, that their internal and external IOR values are more correlated than those of spamming accounts.

3.3 Characteristics of Sender Accounts

Boykin and Roychowdhury [12] comment that "spammers don't spam each other". In our dataset, however, spammers do spam each other. Moreover, *spamming accounts are much more likely than legitimate accounts to receive a large portion of their messages from (other) spamming accounts.* More specifically, approximately 32% of the emails received by spamming accounts originate from spamming accounts, as compared with only about 0.3% of emails received by legitimate accounts! We hypothesize that the reason is that legitimate accounts seldom send emails to spamming accounts, whereas techniques such as dictionary attacks, used by spammers to harvest email addresses, cause spammers to spam each other.

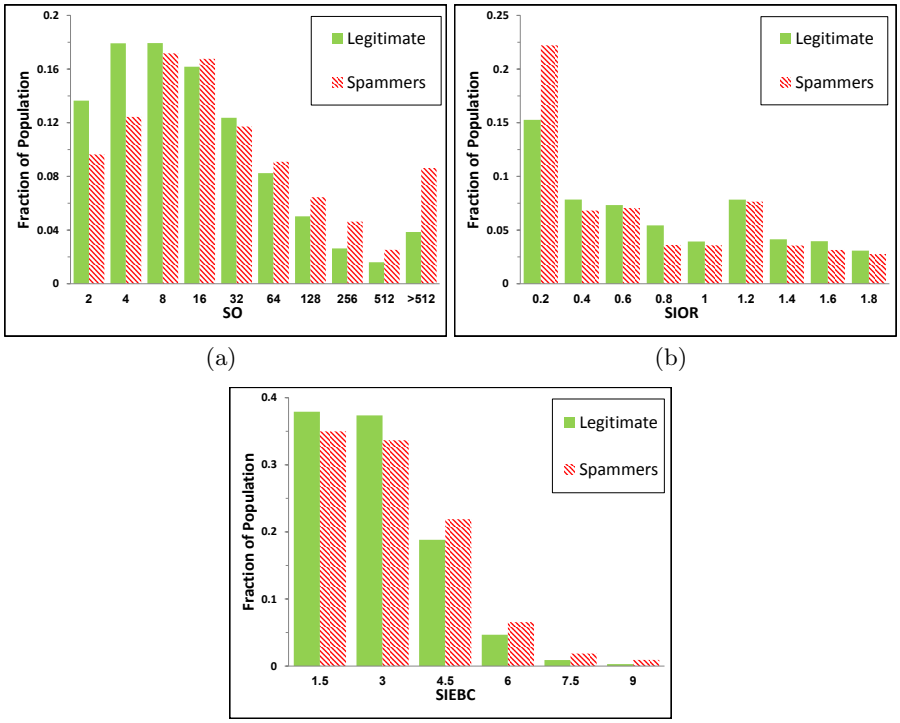


Fig. 4. Sender account features value distributions.

To model the above observation, we introduced per-account features that characterize its sender accounts (we remind the reader that these are the accounts from which an account receives emails). For each account a, we compute its sender accounts' weighted average number of outgoing emails (SO), IOR value (SIOR) and IEBC value (SIEBC). Formally, let a be an account and let $\mathcal{S}(a)$ denote the set of its sender accounts. Also, for accounts a and s, let $\mathcal{I}(a,s)$ denote the number of a's incoming messages that were sent by s. The SO, SIOR and SIEBC features are defined in Equation 4.

$$SO(a) = \sum_{s \in \mathcal{S}(a)} \frac{\mathcal{I}(a,s)}{\mathcal{I}(a)} \cdot \mathcal{O}(s), SIOR(a) = \sum_{s \in \mathcal{S}(a)} \frac{\mathcal{I}(a,s)}{\mathcal{I}(a)} \cdot IOR(s)$$
$$SIEBC(a) = \sum_{s \in \mathcal{S}(a)} \frac{\mathcal{I}(a,s)}{\mathcal{I}(a)} \cdot IEBC(s) \quad (4)$$

Figure 4 shows the value distributions of the SO, SIOR and SIEBC features. As expected, spamming accounts tend to have bigger SO and SIEBC values and smaller SIOR values, as compared with other accounts.

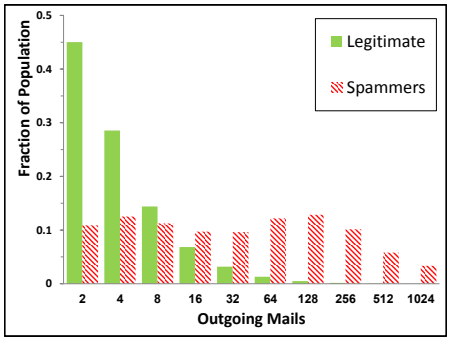


Fig. 5. Distribution of number of outgoing mails

The number of outgoing messages sent by an account is used to compute some of the features we defined above. We use it also as an independent feature as was done in some prior works [13, 14, 16]. If a message was addressed to multiple recipients, we refer to each recipient as an additional outgoing email, e.g. if an email was sent to 10 recipients we consider it as 10 outgoing emails. As can be seen in Fig. 5, spamming accounts send significantly more emails than legitimate accounts. However, a significant fraction of the spamming accounts cannot be distinguished from legitimate accounts solely based on this feature, which is the reason why more elaborate features, such as those we defined above, are required.

4 The ErDOS Detector

Here we present ErDOS - an Early Detection scheme for Outgoing Spam, which is based on the features described in Sect. 3 and on the rotation forest classification algorithm [21] which generates an ensemble of classification trees. The inputs to ErDOS are one day of incoming and outgoing email logs, based on which ErDOS returns a list of accounts most suspicious of sending spam, by evaluating the behavioral patterns of each account. A flow diagram of the ErDOS detector is presented in Fig. 6.

The first three stages described in Fig. 6 are part of the preprocessing phase. First, using a single day of incoming/outgoing email logs, features are extracted for all email accounts which sent at least one email during that day. Next, all accounts which sent at least one email recognized as spam by the content-based detector are assigned the spam tag, while the rest of the accounts are tagged as legitimate. The last preprocessing step extracts a training set to be used for learning a model, by selecting all spamming accounts from the data set and undersampling the legitimate accounts such that the number of spam and legitimate accounts in the training set is equal. Undersampling is required as our data set exhibits great imbalance with an average of 1750 spamming accounts each day as compared with 1.14E6 legitimate accounts, which causes bias towards the

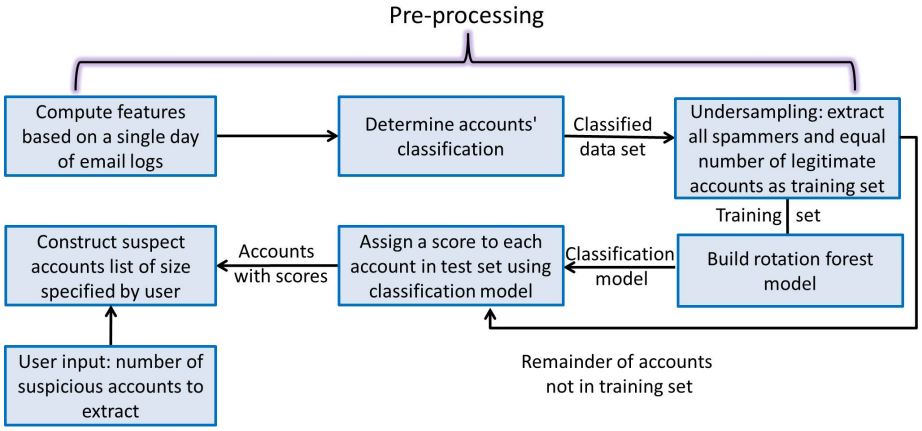


Fig. 6. Flow diagram of ErDOS

larger class (legitimate accounts), while our focus is set on detecting spammers. Therefore we undersample the class of legitimate accounts by randomly selecting accounts, to ensure both classes are of equal size. The remainder of the accounts not used for the training set will be evaluated by the model to identify spamming accounts which have evaded the content-based detector.

After the data has been prepared, the training set is used for training a rotation forest classification model. We used the implementation of rotation forest packaged in WEKA [22]. We used the default configuration, which splits the feature space into 3 random partitions while building the rotated feature space and builds 10 classification trees (C45). Next, all accounts not used during the training phase are examined and assigned a score by the obtained model. The score assigned indicates how suspicious the account is. The score is an average of the scores assigned by each of the trees in the ensemble where the score assigned by each classification tree is a function of the ratio of spamming accounts out of all accounts which reached the leaf during the training phase [23]. Last, the most suspicious accounts, i.e. those with the highest scores, are returned as potential spammers which should be further investigated. The exact number of suspicious accounts to return is a user defined parameter.

A question which may arise is, how can we train a model to distinguish between spamming and legitimate accounts based on our data, if we hypothesize that some of the legitimate accounts are actually sophisticated spammers that manage to evade the content-based detector? Our answer is that we assume that the ratio of spamming accounts out of the entire legitimate population is very small. Therefore, the ratio of spamming accounts which managed to evade the content-based detector out of the legitimate accounts selected for the training phase will also be very small, and will not significantly influence the learning process.

5 Experimental Evaluation

In this section, we describe the experiments we conducted to validate the effectiveness of the ErDOS detector and their results. In our first experiment, we checked whether a single day of incoming/outgoing email logs is sufficient for building an accurate classification model using the features presented in section 3. The purpose of this experiment is to assess the ability of ErDOS to identify spamming accounts after a very short learning period. In our second experiment, we evaluate the early detection capabilities of ErDOS and compare it with two of the previously published outgoing spam detection algorithms.

5.1 Single-Day Training

In this test we show that, using the features defined in Sect. 3, it is possible to build an accurate classification model using only a single day of incoming/outgoing logs. This is important for early detection, as it allows the detection of new spammers who would otherwise not be identified until enough data has been collected. In addition, having to process multiple days of email logs introduces computational challenges as the total size of the data may be tremendous (the average size of the logs per day from our ESP is 18GB).

Our data set contains four days for which we have both incoming and outgoing email logs. We conducted our evaluation for each day separately, using 10-fold cross validation [24][4]. Due to the huge imbalance between the numbers of spam and legitimate accounts, we undersample the data set, as described in Sect. 4. We combine the two preprocessing steps (under sampling and splitting into 10 folds) by first creating the 10 fold data sets and then undersampling each data set separately.

In the single-day training experiment, we do not use the scores assigned by the classification model nor do we extract a suspect accounts list. Rather, we use *binary classification* and calculate accuracy measures using all accounts in the test set.

First, we evaluated the accuracy of ErDOS with different machine learning algorithms, to find the approach best fitted for our data. We evaluated the C4.5 classification tree, SVM and rotation forest algorithms and found that the rotation forest performed slightly better than both SVM and C4.5. The descriptions that follow refer to the evaluation of ErDOS using the rotation forest algorithm.

We compare the ErDOS detector with two previously published outgoing spam detection schemes that are based on accounts' behavioral patterns. Lam and Yeung [14] presented a method using a number of features quantifying social behavior such as CR and CC and utilizing the k-nearest-neighbors algorithm for classification. We will refer to this method as *LY-knn*. Tseng and Chen [16]

[4] We emphasize that 10-fold cross validation is not an integral part of the ErDOS detector; rather, it is only done for assessing the accuracy of the models generated by ErDOS.

Table 3. Detection results per day

day	ErDOS		LY-knn		MailNET	
	TP (%)	Suspect accounts (%)	TP (%)	Suspect accounts (%)	TP (%)	Suspect accounts (%)
1	72.4	7.1	78.2	48.0	20.5	35.9
2	68.5	9.7	76.2	50.8	21.4	48.7
3	74.3	5.9	77.9	36.5	22.0	39.7
4	68.9	12.7	73.1	56.0	26.5	52.6
average	71.0	8.9	76.3	47.8	22.6	44.2

proposed a method using similar features but learning only from *pure accounts*, i.e. accounts which sent only spam or only ham emails. The algorithm they used for classification is SVM [25]. The name they assigned to their approach is MailNET.

Table 3 presents accuracy measures of the models built on each of the four days of data. Two performance measures are used. The first is true positive (TP), which measures the number of accounts correctly classified as spam accounts out of the total number of spam accounts.

The second measure calculates the percentage of accounts that are classified by the model as spammers, even though they were not identified as such by the content-based classifier. We emphasize that accounts that are not identified as spammers by the content-based detector are not necessarily legitimate, and therefore should not be considered as false positives of the detection scheme, as content-based classifiers can be circumvented. Therefore, accounts that are characterized by suspicious behavioral patterns but which did not send any emails detected as spam by the content-based detector cannot be ignored off-hand as false positives. Indeed, as established by the results we present in Sect. 5.2, a substantial number of these are spamming accounts which were not detected by the content-based detector during a specific day but were detected on later days.

The results shown in table 3 clearly indicate that, when using only a single day of data for training a model, the ErDOS detector outperforms both LY-knn and MailNET. Although LY-knn has TP values that are higher by between 3.6%-7.7%, this comes at a high price, as its percentage of suspect accounts is extremely high. The single-day models generated by the MailNET detector indicate that it is inappropriate for our data set, since they result in very low TP values and a very large percentage of suspect accounts.

Analysis of the models built by ErDOS during the training phase reveals that the most dominant features are the number of outgoing emails, IEBC and IOR (see Section 3), confirming the importance of these features. Moreover, the impact of the least dominant feature (SIOR) used by ErDOS on the generated models was more than half the impact of the most dominant factor (number of outgoing emails), implying that all the new features defined in Section 3 have significant impact.

The classification accuracy obtained by the LY-knn and MailNET algorithms on our data set is significantly lower than the results reported by [14, 16]. The large discrepancy can most probably be attributed to two factors: the number of days of email transactions used for training, and the different characteristics of the data sets.

In our experimental set-up, only a single day of email logs is used during the training phase. In contrast, MailNET was evaluated using two or more training days and LY-knn used the Enron data set, which contains 3.5 years of email transactions. At least some of the features used by the LY-knn and MailNET algorithms require a relatively long training period.

One example of such a feature is communication reciprocity (CR), used by both LY-knn and MailNET to distinguish between spammers and legitimate accounts. As described in Section 3, the CR value of an account a equals the number of accounts which reply to messages sent by a. Whereas a large fraction of a legitimate account's recipients typically reply to its messages over a long period of time, it is not necessarily the case that they reply within a time-window of a single day, as demonstrated by Fig. 2.

Another example of such a feature is *Clustering Coefficient* (CC), which is also used by both LY-knn and MailNET. The CC feature measures the friends-of-friends relationship between accounts [14]. Similarly to CR, whereas a significant fraction of a legitimate account's "friends" typically communicate with one another over a long period of time, it is less probable that they do so within a time-window of a single day

The second factor to which the discrepancy in the results of the LY-knn and MailNET algorithms may be attributed is the difference in data set characteristics. Our data set was collected from a large ESP which provides email services to a heterogeneous population of users, including a large number of home users as well as numerous companies of various sizes. In contrast, the data sets used by LY-knn and Mailnet are of a homogeneous population (a single company and a single university).

5.2 Early Detection of Spammers

In this section, we evaluate the practical usefulness of ErDOS as an outgoing-spammer detector scheme that is complementary to content-based spam detection. We show that ErDOS can be used to detect accounts, exhibiting suspicious behavior, which manage to evade the content-based detector.

Outgoing spam detectors must provide very low false positive rates: generating suspect account lists that are too long would make it impractical to further investigate these accounts (which is often a manual process) and runs the risk of having the results of the detector ignored altogether.

Although the percentage of accounts suspected by ErDOS is small relative to the *LY-knn* and MailNET detectors, it still results in huge lists of suspect accounts, as even 5.9% of the accounts population (the size of ErDOS suspected accounts list on the third day) amounts to approximately 67,000 accounts. To alleviate this problem we use the scores assigned by the classification model for

generating shorter lists, containing only those accounts whose behavior is the most suspicious. The size of the output suspect accounts list is a user defined parameter.

For the early detection test, we used both the four days of data for which we have outgoing/incoming logs and the additional 22 days of data for which we have only outgoing logs. We applied the ErDOS detector to the logs of each of the four days separately. We extracted a short list of 100 accounts that were assigned the highest scores by ErDOS. We now define the quality criteria that we use in the early detection test.

Quality criteria : Let a be an account in a suspect accounts list produced by a detector on day d. We say that a is an *early-detected account*, if no messages sent by a before or during day d are tagged by the content-based detector as spam, but at least one message sent by a at a later day is tagged as spam.

The following quality criteria are used in our evaluation.

1. **Early true positive (e-TP)**: This is the fraction of the accounts in the detector's daily suspect accounts list that are early-detected accounts.
2. **Enrichment factor (EF)**: Compares the e-TP of a list of suspicious accounts returned by a detector with that of a randomly generated list. More formally: EF for day d compares the e-TP of the list produced by a detector for day d with that of a list of the same length whose accounts are randomly selected from the entire population of email accounts that have sent no messages tagged as spam up to (and including) day d.

$$EF = \frac{\text{e-TP}(\textit{detector list})}{\text{e-TP}(\textit{random list})} \tag{5}$$

   The higher the score, the stronger is the indication of a large proportion of early-detected accounts in the detector's list in comparison with a random list.
3. **Contribution of complementary method (CCM)**: This is a daily measure of how beneficial a detector is when used along side the content-based detector. CCM is calculated by dividing the number of early-detected accounts in the detector's daily suspect accounts list by the number of new detections made by the content-based detector that day (that is, the number of accounts a message of which is tagged as spam for the first time during that day). The higher the score. the stronger is the indication that the detector's contribution to the content-based detector is substantial.

The left-hand part of Table 4 shows the daily and average e-TP and enrichment factors obtained by ErDOS. On average, 9% of the accounts in the suspects list are early-detected accounts. We note that this is in fact a lower bound on the actual detection rate, since it is plausible that additional listed accounts either send spam that is detected by the content-based detector only at a later period for which we have no data or manage to entirely evade it.

Table 4. e-TP and enrichment factors for 4 different days.

day	List of suspect accounts			Entire legitimate population			Enrichment factor
	accounts	early detections	e-TP (%)	accounts	non-detected spammers	e-TP (%)	
1	100	11	11.0	1,155,236	6964	0.60	18.2
2	100	14	14.0	1,128,121	6355	0.56	24.8
3	100	2	2.0	1,130,701	6026	0.53	3.7
4	100	9	9.0	1,085,796	4894	0.45	20.0
average	100	9.0	9.0	1,124,963	6060	0.53	16.9

We tested whether the average of daily e-TP values obtained by ErDOS is statistically significant with respect to that of a randomly selected suspect list of the same size, using a one-sample test of proportions [26]. The resulting p-value was smaller than 0.001, establishing high statistical significance in the average proportions of early-detected accounts between lists produced by ErDOS and randomly selected lists.

The right-hand side of Table 4 shows the total number of legitimate accounts each day (accounts that were not detected as spammers up to and including that day), and the total number and proportion of these accounts that do send spam on later days. On average, 0.53% of legitimate accounts turn out to be spammers in later days. Enrichment factors are shown in the rightmost column of Table 4. Based on the average e-TPs of the suspect lists produced by ErDOS and of the entire legitimate accounts population, the average enrichment factor is 16.9.

We conducted a comparison of the early detection quality measures of the ErDOS detector with those of the LY-knn and MailNET algorithms with a list of suspects of length 100. The e-TP and enrichment factor values are shown in Table 5. These results show that, on average, ErDOS provides e-TP and EF values that exceed those of MailNET by a factor of 4 and those of LY-knn by a factor of 7. In addition, we compared the e-TP, enrichment factors and CCM values of all three methods on varying sizes of suspect lists. Graphs of the average results of all four days are presented by Fig. 7, showing the advantage of ErDOS over LY-knn and MailNET for all suspect list lengths. These results

Table 5. Comparison of detectors using early detection criteria

day	ErDOS		LY-knn		MailNET	
	e-TP (%)	EF	e-TP(%)	EF	e-TP(%)	EF
1	11.0	18.2	2.0	3.3	3.0	5.0
2	14.0	24.8	1.0	1.8	1.0	1.8
3	2.0	3.7	1.0	1.9	3.0	5.6
4	9.0	20.0	4.0	2.2	3.0	6.6
average	9.0	16.9	1.2	2.3	2.5	4.7

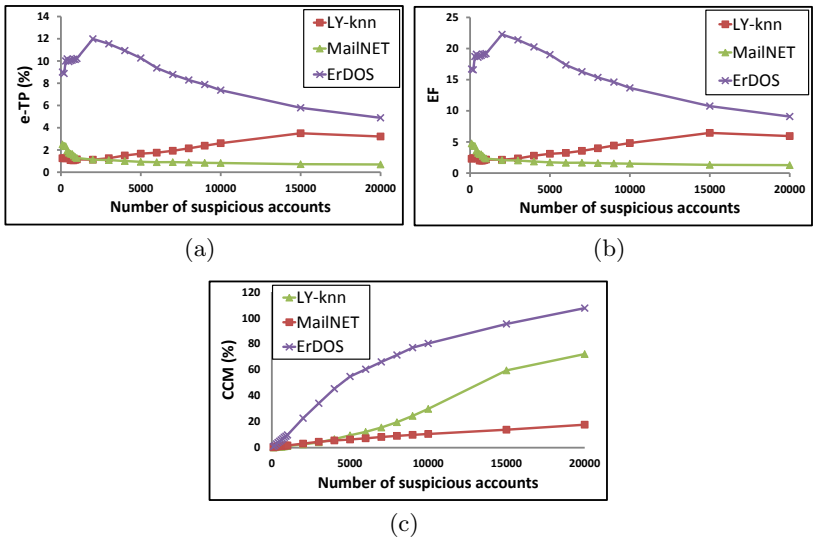


Fig. 7. Contribution of complementary methods.

establish that the early detection capability of the new detector on our data set is significantly superior to that of the LY-knn and MailNET algorithms.

6 Conclusions

In this work, we presented *ErDOS*, an Early Detection scheme for Outgoing Spam. The detection approach implemented by ErDOS combines content-based detection and features based on inter-account communication patterns. ErDOS uses novel email-account features that are based on the ratio between the numbers of sent and received emails and on the distribution of emails received from different accounts. By using the output of a content-based spam detector as a means for obtaining initial labeling of email accounts, we manage to avoid the use of synthetically-generated spam accounts as done by some prior work.

This study was done using a very large data set, collected by a large ESP, that contains no information on the contents of email messages except for a tag assigned by a content-based detector. A key challenge we faced was to extract meaningful and succinct information from a data set which, on the one hand, is very large, but on the other hand spans a relatively short period of time.

Our goals were to design and implement a detector that is able to detect spamming accounts that evade a content-based detector based on their communication patterns. To this end, we defined a set of new account features that are able to characterize the behavior of spamming email accounts based on a single day of incoming/outgoing data.

Our empirical evaluation of ErDOS establishes that it provides higher accuracy as compared with previous outgoing-spam detectors when applied to our data set. Moreover, by using only a single day of training data, ErDOS is able to provide early detection for a significant fraction of the spammers population, significantly better than the algorithms with which we compared it.

In the future, we plan to further improve the accuracy and early-detection capabilities of ErDOS by evaluating new features for characterizing additional aspects of an account's behavior. We also plan to check what is the minimum amount of data required for training an accurate model for ErDOS, as it is possible that effective models can be generated based on even less than one day of data.

7 Acknowledgments

We would like to thank Lior Rokach and Eitan Menachem for their useful insights and for helpful discussions on machine learning best practices.

References

1. Radicati, S.: Email statistics report. Technical report, The Radicati Group, Inc. (2010)
2. Pingdom: Internet 2010 in numbers. `http://royal.pingdom.com/2011/01/12/internet-2010-in-numbers/`
3. Fallows, D.: Spam: How it is hurting email and degrading life on the internet. Pew Internet and American Life Project (2003) 1–43
4. Clayton, R.: Stopping spam by extrusion detection. In: First Conference on Email and Anti-Spam (CEAS 2004), Mountain View CA, USA. (2004) 30–31
5. Venkataraman, S., Sen, S., Spatscheck, O., Haffner, P., Song, D.: Exploiting network structure for proactive spam mitigation. In: Proceedings of 16th USENIX Security Symposium on USENIX Security Symposium, USENIX Association (2007) 11
6. Taylor, B.: Sender reputation in a large webmail service. In: Proceedings of the Third Conference on Email and Anti-Spam (CEAS). Volume 27. (2006) 19
7. John, J., Moshchuk, A., Gribble, S., Krishnamurthy, A.: Studying spamming botnets using botlab. In: Proceedings of the 6th USENIX symposium on Networked systems design and implementation, USENIX Association (2009) 291–306
8. Sahami, M., Dumais, S., Heckerman, D., Horvitz, E.: A bayesian approach to filtering junk e-mail. In: Learning for Text Categorization: Papers from the 1998 workshop. Volume 62., Madison, Wisconsin: AAAI Technical Report WS-98-05 (1998) 98–105
9. Aradhye, H., Myers, G., Herson, J.: Image analysis for efficient categorization of image-based spam e-mail. In: Document Analysis and Recognition, 2005. Proceedings. Eighth International Conference on, IEEE (2005) 914–918
10. Krawetz, N.: Anti-honeypot technology. Security & Privacy, IEEE **2**(1) (2004) 76–79
11. Bouguessa, M.: An unsupervised approach for identifying spammers in social networks. In: Tools with Artificial Intelligence (ICTAI), 2011 23rd IEEE International Conference on, IEEE (2011) 832–840

12. Boykin, P., Roychowdhury, V.: Leveraging social networks to fight spam. Computer **38**(4) (2005) 61–68
13. Gomes, L., Almeida, R., Bettencourt, L., Almeida, V., Almeida, J.: Comparative graph theoretical characterization of networks of spam and legitimate email. Arxiv preprint physics/0504025 (2005)
14. Lam, H., Yeung, D.: A learning approach to spam detection based on social networks. In: Proceedings of the Fourth Conference on Email and Anti-Spam (CEAS), 2007
15. Moradi, F., Olovsson, T., Tsigas, P.: Towards modeling legitimate and unsolicited email traffic using social network properties. In: Proceedings of the Fifth Workshop on Social Network Systems, ACM (2012) 9
16. Tseng, C., Chen, M.: Incremental SVM model for spam detection on dynamic email social networks. In: Computational Science and Engineering, 2009. CSE'09. International Conference on. Volume 4., Ieee (2009) 128–135
17. Watts, D., Strogatz, S.: Collective dynamics of 'small-world' networks. nature **393**(6684) (1998) 440–442
18. Gomes, L., Cazita, C., Almeida, J., Almeida, V., Meira, W.: Workload models of spam and legitimate e-mails. Performance Evaluation **64**(7) (2007) 690–714
19. Kossinets, G., Watts, D.J.: Empirical analysis of an evolving social network. Science **311**(5757) (January 2006) 88–90
20. Shetty, J., Adibi, J.: The Enron email dataset database schema and brief statistical report. Information Sciences Institute Technical Report, University of Southern California **4** (2004)
21. Rodriguez, J.J., Kuncheva, L.I., Alonso, C.J.: Rotation forest: A new classifier ensemble method. Pattern Analysis and Machine Intelligence, IEEE Transactions on **28**(10) (2006) 1619–1630
22. of Waikato, U.: Weka 3: Data mining software in Java. `http://www.cs.waikato.ac.nz/ml/weka/`
23. Rokach, L., Maimon, O.: Data mining with decision trees: theroy and applications. Volume 69. World Scientific Publishing Company Incorporated (2008)
24. Kohavi, R., et al.: A study of cross-validation and bootstrap for accuracy estimation and model selection. In: International joint Conference on artificial intelligence. Volume 14., Lawrence Erlbaum Associates Ltd (1995) 1137–1145
25. Cortes, C., Vapnik, V.: Support-vector networks. Machine Learning **20**(3) (1995) 273–297
26. Kirk, R.: Statistics: an introduction. Wadsworth Publishing Company (2007)