Data Mining Challenges for Electronic Safety: The Case of Fraudulent Intent Detection in E-Mails

Edoardo Airoldi Bradley Malin
Data Privacy Laboratory, School of Computer Science
Carnegie Mellon University, Pittsburgh, PA USA 15213-3890
{eairoldi,malin}@cs.cmu.edu

Abstract

Online criminals have adapted traditional snail mail and door-to-door fraudulent schemes into electronic form. Increasingly, such schemes target an individual's personal e-mail, where they mingle among, and are masked by, honest communications. The targeting and conniving nature of these schemes are an infringement upon an individual's personal privacy, as well as a threat to personal safety. In this paper, we introduce an array of challenges which are ripe for the attention of the data mining research community and are vastly different from those of combating the general problem of spam. We illustrate how state-of-the-art spam filtering systems fail to capture fraudulent intent hidden in the text of e-mails, but demonstrate how more robust systems can be engineered using existing data mining tools. We conclude by examining a specific scheme, the Nigerian 4-1-9 advance fee fraud scam, for which we design a learning system capable of accurately identifying the fraudulent indent within an e-mail. Our system is applicable to fraud detection and can serve as a guide for law enforcement agencies in cyber-investigations.

1. Introduction

Unsolicited communications currently account for over sixty percent of all e-mail sent over the Internet, and experts predict this number will reach the mid-eighties. [1] While much spam is innocuous, a portion is engineered by criminals to prey upon, or scam, unsuspecting people. The senders of scam spam attempt to mask their messages as non-spam and con through a range of tactics, including pyramid schemes, securities fraud, and identity theft via phisher mechanisms (e.g. redirection to faux PayPal or AOL websites).

During 2003, the United States Federal Trade Commission (FTC) received more than a half-million consumer complaints, an increase of 25% on the previous year. [2] Of these complaints, approximately 60% were concerned with various types of fraud. The Consumer Sentinel database, maintained by the FTC, now houses over 1.5 million complaints; one million of which correspond to consumer fraud. The total monetary loss for all fraud victims is in excess of $437 billion, with a median loss of $228. In 58% of the complaints, consumers report being contacted through the medium of the Internet.

A major challenge of Internet-fraud problems is the difficulty in discerning scam from spam and regular e-mail. In fact, scam messages differ from other types of spam for several reasons. First and foremost, the scam's major trait is its hidden criminal intent. In order to lure unsuspecting individuals, the text is engineered to read like regular e-mail, and thus pass successfully through spam filters. Second, messages from the same individual are not necessarily equivalent in text and story. Third, scam messages can be sent out over a longer time period than traditional bulk spam messages. Fourth, scam messages are not necessarily sent via the same physical routes as spam or via the same techniques, such as the commandeering of an open relay.

In this research we study in detail the advance fee fraud, the most infamous of which is the "Nigerian", or 4-1-9, scam. Over the past several years, the number and type of messages imploring readers for monetary assistance today with the promise of future riches, has increased without signs of abating. Several different groups have compiled corpora and made them available online. We take advantage of existing text mining tool and online repositories to build more accurate filters, able to:

- filter scam spam from e-mail with error rates comparable to state-of-the-art spam filters, and
- identify the criminal intent hidden in e-mails.

The remainder of this paper is organized as follows. In the following section we discuss background issues with respect to internet fraud and specific aspects of the Nigerian scam. Additionally, we introduce the new challenges for electronic safety based on concerns expressed by the FTC. In Section 3, we present the technical details of our ScamSlam system. In Section 4, we employ a real world dataset of over 500 Nigerian scam messages to study the filtering and relationship learning capabilities of ScamSlam. Finally, in Section 5 we discuss the limitations of the system, as well as how the ScamSlam system can be validated and applied to a law enforcement setting.

2. Challenges for Electronic Safety

2.1. Spam, Fraud, and E-mail

The concept of spam is not a novelty limited to the electronic world of the Internet. For years, any individual or household with a mailbox in the physical world received their fair share of unsolicited "junk" mail. However, the quantity of junk snail mail sent to individuals is limited by the fact that marginal cost scales linearly with the amount of mail sent. In cyberspace, on the other hand, the current status quo of communication is such that marginal cost is negligible as the quantity of e-mail sent increases. [3] In combination with other factors, including the increased implementation of e-mail as a direct marketing tool, the amount of spam sent over the Internet is continually growing. Statistics compiled by Brightmail Inc., a well-respected antispam company, indicate that as of February 2003 approximately 42% of all messages sent over the Internet was spam. By April 2004 this number had increased to almost 65%, which corresponded to over 96 billion messages filtered during a single month. [4]

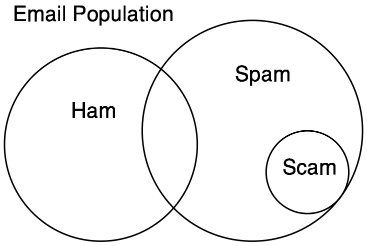

Figure 1. E-mail types and their relationships. Ham corresponds to legitimite e-mail, while spam means non-legitimite. Scam messages are considered a subpopulation of spam.

For this research, we consider e-mail messages to be of three types: ham, spam, and scam. In Figure 1 we depict the relationships between e-mail types. As stated above, spam messages are unsolicited pieces of e-mail. The scam messages are a subset of spam messages which are intelligent in design, such that they attempt to coax the individual to perform some action of illegal purpose beyond a simple "click me". In contrast, "Ham", refers to legitimate e-mail messages. Note, there exist certain messages which are viewed as spam by some individuals and ham by others (e.g. legitimate, but unsolicited advertisments); depicted in the intersection of figure 1.

2.2. A Collection of New Data Mining Problems

Akin to the spam problem, the phenomenon of fraud is neither new nor trivial. For example, in 2003, the FTC reported the American public lost over $400 million to fraudulent activities. [2] Scams communicated via e-mail and the Internet are on the rise as well. Brightmail reports that over three billion phishing scam e-mails are now sent monthly over the Internet, noting a 50% increase from January to April 2004 alone. [5] In March 2004, Zachary Hill was arrested by the FTC and the Department of Justice for identity theft and illegally attracting people via e-mail to fake websites masquerading as AOL and PayPal. During the tenure of his scam, Hill obtained at least 471 credit card numbers, 152 bank account and routing numbers, and 541 user names and passwords. [6]

To characterize the problems of fraud more specifically, according to the FTC the top ten types of frauds are as follow (percent of total money lost to fraudulent activities): Internet auctions (15%); shop at home, catalog sales (9%); Internet services and computer products (6%); prizes/sweepstakes and lotteries (5%); foreign money offers (4%); advance-fee loans and credit protection (4%); telephone services (3%); business opportunities and work-at-home plans (2%); magazine and buyers clubs (1%); and office supplies and services (1%). [2]

A data mining approach is ideally suited for modeling and solve these problems. Relevant patterns are buried amongst a massive amount of data characterized by a large noise to signal ratio. Several online repositories for various types of spam and scam messages have recently appeared online and as this paper demonstrates, it is possible to adapt data mining methods for fraud analysis. In general, we suspect new filters to solve these novel pattern recognition problems can be designed by adapting or developing supervised learning methods. [7, 8]

Furthermore, the ability to cope with, and eventually limit and prosecute these frauds, is at the heart of privacy and security concerns expressed by government agencies in both the United States and Europe. Prior research in data mining approaches to fraud detection were focused on offline issues, specifically, money laundering and corporate concerns. [9] With respect to Internet fraud, the opportunity for building better detection and investigative tools has again attracted the interest of the corporate world. However, whereas many solutions have been proposed for the spam problem none seems to address Internet fraud.

3. Detecting Fraudulent Intent

The problem we tackle can be stated as one of binary classification. Basically, design a message filter that discriminates between messages which contain patterns of fraudulent intent for the type of scam and other e-mail. The filter is trained to make a Boolean decision on a labeled dataset, where the labels are “scam” and “not scam”. After the filter has been trained, it can be applied to messages incoming to a mail server in real time. This is a new problem, but follows a recent research trend in text mining termed *semantic learning*. Related problems include sentiment identification [10], affect sensing [11], opinion extraction [12], and speech act classification [13].

3.1. The Nigerian 4-1-9 Scam

The dataset studied in this research pertains to one specific type, namely the advance fee fraud (AFF). The AFF is a scheme in which a stranger with an unfortunate story requests an individual for a certain amount of money, usually not a very large sum, to assist in the transfer of a large monetary sum. The hook is that once the requester’s money has been safely transferred, the investor will be paid a percentage of the sum for their assistance, which translates into a much larger amount than initially invested. However, this message being a ruse to bilk the investor out of their money, the return on investment is never realized, much to the investor’s chagrin and frustration. The most well known version of this fraud is the “Nigerian”, or 4-1-9, scam, named after the section of the Nigerian criminal code that explicitly prohibits such actions. The scam has been conducted since at least 1989 in the form of physical mail, fax, and most recently through e-mail. While the fraud is commonly referred to as “Nigerian”, this is partially derivative of the common use of this country in much of the earlier versions of such communicated messages. In actuality, it is quite common for the stranger to claim residence in any number of countries both within and outside the continent of Africa. The scam itself has proven to be quite lucrative, especially over the Internet. In 2003, MessageLabs Inc. reported that the Nigerian scam grossed an estimated $2 billion dollars, ranking it one of the top grossing industries in Nigeria. [14]

3.2. Supervised Poisson Filtering

We begin our model with a short description of the filtration process. Briefly, a filter is a function that takes as input the word counts observed in a message and some parameters (to be defined below) and returns a decision about whether or not the message is scam. Specifically, our Poisson filter labels a message as scam if the probability of the message being scam given the counts of the words it contains is greater than the probability of the message not being scam given the counts.

More formally, we start with a corpus of p messages, $M = \{m_1, m_2, \ldots, m_p\}$, which are labeled as belonging to one of two categories, $C = \{Scam, Not\text{-}Scam\}$, so that $M = \cup_{c \in C} M_c$ is the union of disjoint sets of messages (M_c) in different categories. From M we extract a vocabulary of x unigrams, $V = \{v_1, v_2, \ldots, v_x\}$, defined as contiguous strings of letters. Let X_{mv} be a random variable denoting the counts for unigram v in message m. We assume that the counts for X_{mv} occur according to a Poisson distribution as in [15]:

$$p(x_{mv}|\omega_m, \mu_{vc}) = \frac{e^{-\omega_m \mu_{vc}}\ (\omega_m \mu_{vc})^{x_{vm}}}{x_{mv}!} \quad \text{s.t.} \quad \omega_m > 0,\ \mu_{vc} > 0,\ x_{mv} \geq 0 \tag{1}$$

where ω_m is the length of message m in thousands of words, and μ_{vc} is the Poisson rate for unigram v in category c. The Poisson rate is the number of unigrams we expect to see in an arbitrary block of a thousand consecutive words of text from a messages of category c. During training, we assign a value to the parameter μ_{vc} of the Poisson model for both categories of messages by computing maximum likelihood estimates according to the following formula:

$$\hat{\mu}_{vc} = \frac{\sum_{m \in M_c} x_{mv}}{\sum_{m \in M_c} \omega_m}, \qquad \text{for each } c \in C. \tag{2}$$

Our filter is based on several simplifying independence assumptions. First, the random variables that represent unigram counts in a message, X_{vm}, are independent from one another. Second, the position of the random variables are independent within the text of the message. In our framework, we use the following ratio r_m to determine if it is probabilistically more likely that a message $m \in M$ is $Scam$ or not:

$$r_m = \frac{\prod_{v \in V} p(X_{mv} \mid \hat{\mu}_{v\,Spam})}{\prod_{v \in V} p(X_{mv} \mid \hat{\mu}_{v\,No\text{-}Spam})} \tag{3}$$

When r_m is greater than 1, we classify a message as $Scam$, otherwise it is classified as $Not\text{-}Scam$.

3.3. A Law Enforcement Perspective

In this section, we introduce a system which not only allows detection of fraudulent intent in e-mails, but can be used by law enforcement officials to provide guidance in cyber-investigations. Before delving into the technical details, we provide a brief sketch of the ScamSlam system. The ScamSlam system consists of three main components, as depicted in Figure 2: 1) a trained scam filter, 2) a message normalizer via a vector space projection method, and 3) an intelligent clustering engine.

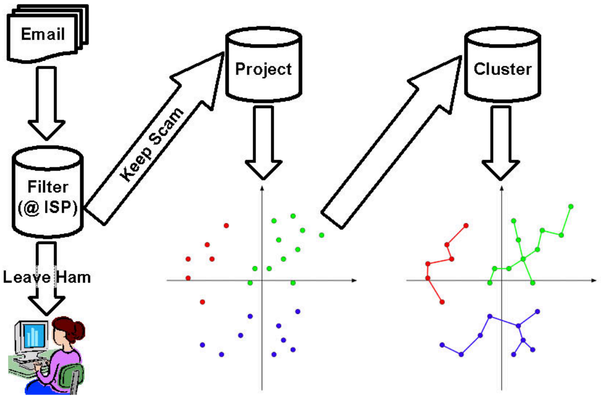

Figure 2. General overview of the ScamSlam system. *Step 1)* **Incoming messages fare filtered for scams.** *Step 2)* **Scam messages projected into Euclidean space for vector representation.** *Step 3)* **Messages clustered based on similarity.**

The first part of the system would be the filter we discussed above, which is trained to make a Boolean decision on a labeled dataset, where the labels are "scam" and "not scam". Next, the scam messages are projected into a common space of representation. More specifically, the ScamSlam system converts a scam message into a normalized vector of words. For each message, each word is assigned a weight that captures information about the frequency with which the word occurs in the message and in the set of scam messages under scrutiny. Once the documents have been normalized by the re-weighting and representation process, the documents are clustered based on similarity using a hierarchical clustering technique, specifically single linkage, which partitions the vector space into clusters of similar messages. The clustering method proceeds in a stepwise manner and terminates when no linkages can be constructed at a minimal level of message similarity. The minimal level, or threshold, is derived using a novel heuristic based on empirical observations of the studied scam messages.

In the following subsections, last two components is described in further detail.

3.3.1 Message Representation

After filtering the scam spam messages, we project them into a normalized multi-dimensional space, the details of which are as follow. Recall that we represent the corpus of messages as a set $M = \{m_1, m_2, \ldots, m_p\}$, from which we extract the vocabulary $V = \{v_1, v_2, \ldots, v_x\}$, which is the set of distinct unigrams, or strings of contiguous letters, found in the messages. Each message $m_i \in M$ is converted into a vector model, such that each message is represented as a n-size vector, $\vec{m} = [x_{m1}, x_{m2}, \ldots, x_{m|V|}]$, where each value x_{mv} corresponds to the observed number of times that term v appears in message m. [16]

Each vector is then re-weighted, or normalized, to account for the relative frequencies of terms in the set of messages M. The weights, components of a normalized vector, represent the term frequency - inverse document frequency scores. With respect to message m, term frequency (tf) corresponds to the number of times a term v is observed in a message, normalized by the maximum frequency term in m, such that term frequency for term t in message m is $tf_{mv} = \frac{x_{mv}}{max_t x_{mt}}$. While the term frequency weight accounts for the relative frequency of a term within a message, the inverse document frequency (idf) accounts for the relative frequency of a term among messages. Specifically, let obs_v represent the number of messages that term v is observed in, the inverse document frequency score idf_v equals $log(\frac{|M|}{obs_i})$. Combining term frequency and inverse document frequency, we re-weighted messages are represented as the $\vec{m'} = [w_{m1}, w_{m2}, \ldots, w_{m|V|}]$, where $w_{mv} = tf_{mv} \times idf_v$.

We measure the similarity between a pair of messages $\vec{m_i}$, $\vec{m_j}$ using the cosine of the angle between the two vectors as explained in the following section.

3.3.2 Scam Clustering

ScamSlam clusters messages using single linkage over the corresponding weighted vector representations. Single linkage is a hierarchical clustering technique that targets messages which display high similarity between pairs. [17] As clustering proceeds, each message belongs to one and only one cluster at any particular time during the clustering process. The way clustering proceeds is as follows. Let $thresh$ be a threshold of similarity which defines the boundary at which two messages can be considered to belong to the same cluster or not. Initially, each message is a singleton cluster consisting of only itself, so there exist $|M|$ clusters. As clustering proceeds, two arbitrary clusters l_i and l_j are merged into a single cluster if there exists one message m_a in l_i and one message m_b in l_j such that the distance between them does not exceed $thresh$. ScamSlam uses a distance measure, $dist(\vec{m_i}, \vec{m_j})$, induced by the cosine similarity:

$$dist(\vec{m_i}, \vec{m_j}) = 1 - \frac{\sum_{k=1}^{n} w_{ik} \times w_{jk}}{\sqrt{\sum_{k=1}^{n} w_{ik}^2} \times \sqrt{\sum_{k=1}^{n} w_{jk}^2}}. \quad (4)$$

The choice of single linkage addresses one of the observed means by which scam spam authors operate. Specifically, a very useful component of single linkage clustering is its ability to permit messages within a cluster to be very different from each another. Over time, the writers of scam spam can change any number of features, such as the motive for money transfer of the name and title subject of who is in need of help. Moreover, sections of the story or plead may change as well, such as when a paragraph of the message is removed or added. It is not uncommon to find that over time, there is a continual tweaking of the scam, where a part of the scam is changed while keeping most parts in common.

3.3.3 An Exploratory Tool for Cyber-Investigations

A method of scoring and clustering provides law enforcement officials with the capabilities to pursue two strategies for searching and persecuting criminals. More formally, we use distance as a threshold parameter for our model and term it the maximum distance of membership D_*. Consider then two cases; in the presence of evidence from a criminal group, progressive clustering via an increasing value for D_* provides an ordered list of suspects by ranking the messages closest to the cloud of messages that constitute the evidence. In the absence of evidence, law enforcement officials can increase the minimum distance D_* and grow clusters, each of which can be regarded as a possible pocket of criminal activities worthy investigating further, again ranked by similarity. An aspect of interest is a good heuristic to decide whether there is enough evidence in the data to justify the fusion of small pockets of illegal activity. In order to answer this question we use the following metric F_D:

$$F_D = \frac{\sum_{i=1}^{|M|}\sum_{j=i+1}^{|M|} \phi(dist(m_i, m_j))}{\frac{|M|(|M|-1)}{2}}, \qquad (5)$$

$$\text{where } \phi(x) = \begin{cases} 1, & \text{if } x \leq D \\ 0, & \text{otherwise} \end{cases}$$

which measures the fraction of all message pair distances within threshold D. This measure leverages the geometry of the vector space of messages. More specifically, F_D measures how clusters grow, and we set D_* at the point where the growth rate is slow or stagnant for a period of time. The intuition behind this heuristic is that if there are defined clusters, we will discover them when D_* equal to approximately the radius of the majority of the clusters, but less than the distance needed for these well defined clusters to merge. Thus, even if after the period of stagnancy there is an increase in the rate of growth, we suspect that this growth is due to the merging of clusters which should remain independent will begin merging. The lack of growth in cluster sizes is found by minimizing a smoothed version of the first derivative of the F_D.

4. Experiments

For our experiments, we used five different datasets, one for the scam messages, and two for each of the remaining types of messages, spam and ham. The scam corpus consists of 534 messages posted to the Nigerian Fraud E-mail Gallery.[1] [18] Each message was previously been classified as the Nigerian 4-1-9 scam by the proprietor of the website. The messages dates span the time period from April 2000 to April 2004 and are distinct, such that no two messages are duplicates. The spam-A and ham-A corpora were collected and supplied by a graduate student at Carnegie Mellon University, who collected the messages over a four month period. There are 2944 spam and 7651 ham messages. The spam-B corpus was collected by Dr. Latanya Sweeney (Carnegie Mellon University); it contains 2532 spam messages. Finally, we assembled the ham-B corpus by selecting 75 posts from each of seven newsgroups, for a total of 525 ham messages. There are approximately 200,000 distinct unigrams in the combined spam-B and ham-B corpus.

To further validate our findings in a more controlled environment, the we evaluate our methods on the SpamAssassin public mail corpus. This corpus is a selection of mail messages, suitable for use in testing spam filtering. The corpus contains 6047 messages, with about a 31% spam ratio.[2]

4.1. Poisson Filtering vs. Spam-Assassin

Before studying the relationships within a set of scam spam messages, we must address how one goes about filtering scam messages from the deluge of messages flowing through the Internet. We performed a preliminary study to assess how well widely used spam filters would be at recognizing scam messages as spam. To do so, we subjected the combined scam, spam-A, ham-A corpus to analysis and classification by SpamAssassin$^{\text{TM}}$, the popular open source spam filter. [19] SpamAssassin uses a set of rules and a Bayesian classifier to determine if a message is spam or not. It ultimately assigns a message with a total score which denotes the degree to which SpamAssassin considers a message as spam. The more negative a SpamAssassin score is, the lower the probability that the message is spam.

The messages were scored using SpamAssassin. While users of SpamAssassin are afforded with the ability to set their threshold for spam classification, the default value for SpamAssassin is 5.0. Thus, if the score for a spam or scam message was less than 5.0 we consider the message to be misclassified. Similarly, for ham messages that score greater than or equal to 5.0. Side-by-side histograms of the

[1] The corpus is publicly available and can be found at http://potifos.com/fraud/

[2] The SpamAssassin public mail corpus is available at http://spamassassin.org/publiccorpus/.

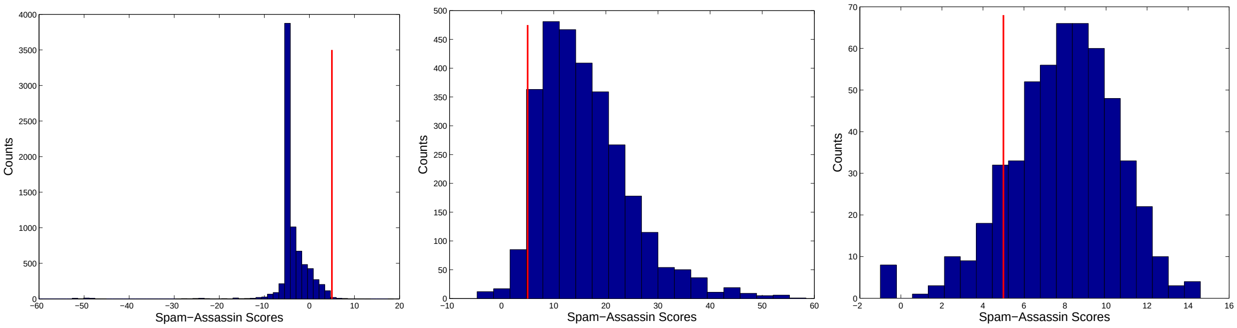

Figure 3. Distribution of SpamAssassin scores for test corpora. Scores for *left*) **ham,** *center*) **spam, and** *right*) **Nigerian scam corpus. The thin vertical line at $x = 5$ represents the default threshold value for which messages are considered spam, (*i.e.* a message with a score greater than 5 is considered spam). We notice an increase in the "falsely classified as ham" rate from $\approx 4\%$ for spam to $\approx 12\%$ for scam.**

resulting scores are depicted in Figure 3 with the threshold score depicted by a thin vertical line. The classification and misclassification rates are provided in Table 1. Based on the observed scores, SpamAssassin does very well at classifying ham as ham, However it has a more difficult time classifying the other message types and disproportionately so for spam versus scam. As seen in Figure 3, SpamAssassin misclassifies about $\approx 4.1\%$ and $\approx 11.8\%$ of the spam and scam messages as ham, respectively. Similar results were observed for the other corpora.

		SpamAssassin Prediction	
		Ham	**Spam**
Reality	**Ham**	99.65%	0.35%
	Spam	4.14%	96.86%
	Scam	11.8%	88.2%

Table 1. Average confusion matrix for SpamAssassin on corpus A (*7651 ham, 2944 spam, and 534 scam messages***).**

It appears that whereas SpamAssassin performs extremely well on the task it was engineered for, separating spam from ham, it is not able to accurately distinguish scam messages from ham and spam.

We then explored the problem of scam classification. In order to do so, we trained and tested a Poisson classifier [15] using a balanced 5-fold cross-validation scheme[3], and performed an additional set of experiments.

In the first Poisson classification test, the ham-B corpus was considered as one class and we combined both the scam and spam-B corpora for the second class. With classes defined as such, this classification experiment is equivalent to the traditional spam filter (or spam classification problem). In the second Poisson experiment, we consider the problem of directly filtering scam from the general population of e-mail messages. Therefore, the first class consists of both ham-B and spam-B messages, while the second class consists solely of scam messages. Using 5000 unigrams, we observe the results as shown in tables 2 and 3.

		Poisson Prediction	
		Ham	**Spam+Scam**
Reality	**Ham**	98.29%	1.71%
	Spam+Scam	2.41%	97.59%

Table 2. Average confusion matrix of Poisson classifier obtained via 5 fold cross validation, on corpus B (*525 ham, 534 scam, and 2532 spam***).**

We chose to use 5000 unigrams in both our experiments, since this number minimizes the cross-validated misclassification error (ham erroneously tagged as spam or scam) as shown in the right panel of Figure 4. It is worth noting that a decision about the number of words, or equivalently about the threshold for SpamAssassin, is essentially a policy decision about which type of mistake is more important. The cross-validated misclassification error plots in figure 4 decreases sharply as more strongly discriminating words

[3] This means that all messages are split into two classes A and B, each of which is partitioned into 5 equal-sized exclusive sets of messages (i.e. $A_1, A_2, \ldots, A_5$, such that $A_1 \cup A_2 \cup \ldots \cup A_5 = \emptyset$). The classifier was trained on eight of the partitions, four from each class, and we tested the trained classifier on the remaining two classes. This scheme was used to test the classifier in five separate runs, such that each of the partitions for each class is tested one time.

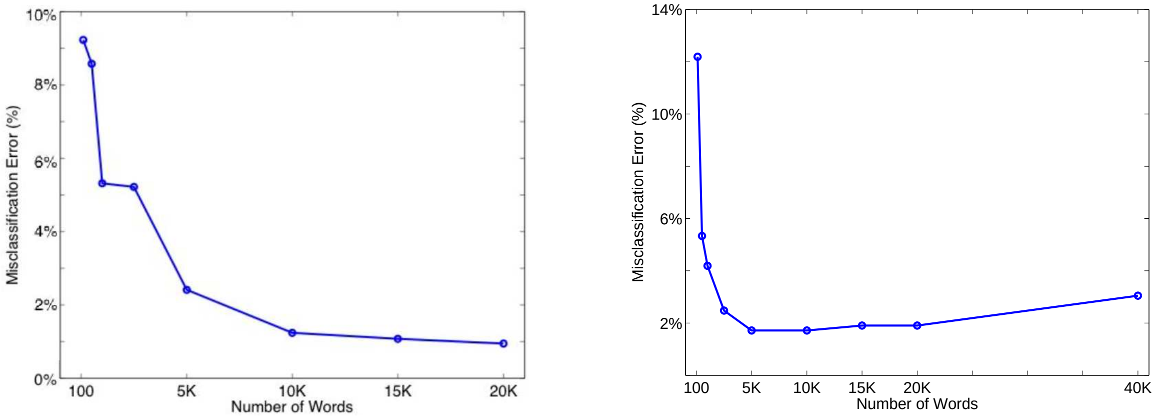

Figure 4. Poisson misclassification ham as one class and spam and scam (spam+scam) combined as a second class. *left*) **Spam+Scam misclassified as ham.** *right*) **Ham misclassified as spam+scam.**

		Poisson Prediction	
		Ham+Spam	**Scam**
Reality	**Ham+Spam**	99.57%	0.43%
	Scam	0%	100%

Table 3. Average confusion matrix of Poisson classifier obtained via 5 fold cross validation, on corpus B (*525 ham, 534 scam, and 2532 spam***).**

are used, and eventually starts increasing after too many weakly discriminating words are used. The Poisson classifier makes a decision by weighting and composing into a linear combination the probabilities of each message being of one category rather then the other; ideally we would want few strongly discriminating words pushing the sum in one direction or another, whereas too many small terms introduce confusion and, in the end, misclassification errors. In our experiments we assessed how good of a discriminator each unigram was on the training set, for each fold, according to their information gain, and that is the ordering that we used for the X axes in figure 4.

We repeated the same experiments on the SpamAssassin corpus, augmented with the 534 scam messages, and obtained similar results. Specifically, Spam-Assassin has an error rate bigger than 10% in identifying scam, whereas the Poisson filter keeps the error rate below 1%.

4.2. Clustering Analyses

For the following unsupervised clustering experiments, we continue with the Nigerian scam corpus. All header information was removed so that clustering was performed with only the text of the messages. One of the assumptions that we incorporate into this analysis is that messages which form clusters are scattered at nonuniform levels of density in the vector space of tf-idf weights. Since, the measure F_D captures the density of message clustering in the vector space, we empirically tune D_* according to the observed growth rate. We observe in Figure 5 the growth rate of F_D is minimized at a distance of 0.6. Though the global minimum is realized at the boundary point, this is an artifact of the fact that all messages are clustered at distance equal to 0.9. While the growth rate in messages clustered continues to grow beyond 0.6, this is mainly due to the uniform distribution of single message clusters. At this point we begin to observe that large clusters which are well defined at a relatively low threshold (below 0.6) begin merging.

At D_* equal to 0.6, we uncover approximately 20 clusters of size 5 or larger, where the largest cluster consisted of 40 messages. These clusters account for approximately half of the total corpus.

4.2.1 Temporal-Trees: A Compelling Hypothesis

While the scam dataset is devoid of the reality regarding relationships, the temporal aspect of our hypothesis permits validation via an alternative route. If scam messages are both reused and changing over time, then it is possible that scam clusters can be modeled as an evolutionary process. That is, the spam message within a cluster can be partially ordered on the dates messages were sent. We introduce a data structure, termed a *temporal-tree*, for studying the temporal ordering of a cluster of nodes. An example of a temporal-tree is shown in figure 6.

Definition (Temporal-Tree) *A temporal-tree is a tree data structure. Nodes correspond to independent observations. Edges correspond to linkages of observations given by a*

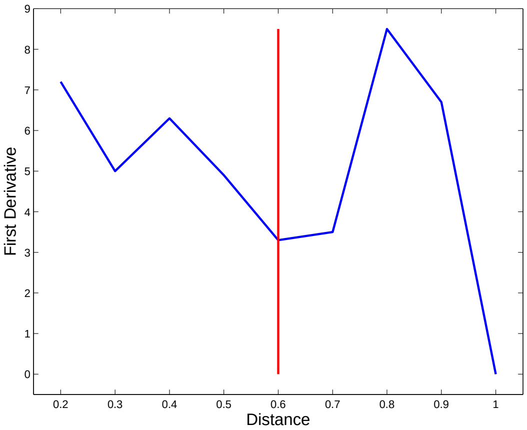

Figure 5. The first derivative of F_D versus D in the Nigerian dataset implies $D_* = 0.6$.

single linkage criteria. The root node corresponds to the observation with the earliest date.

If the cluster is indeed an evolutionary process, then we expect several features will be observable. First, linkages within clusters will adhere to a partial ordering on the dates the scam messages were sent. The temporal ordering is the result of a continual changing of messages, such that each scam message is augmented to yield a child in the temporal-tree. Second, as in many evolutionary processes there exist bifurcations in the family tree of scam. Such bifurcations will manifest when a single scam message is used as the basis for two or more lines of message augmentation. Subsequently, each of the children can sustain an independent lineage of evolution. It is interesting to note that the single linkage criteria provides an ideal setting for analyzing such patterns since the returned clusters represent spanning trees over a set of messages.

4.2.2 Statistical Validation of Temporal-Trees

The hypothesis we test is whether, and to what degree, the temporal-trees of e-mail messages are arranged as evolutionary processes. In order to do so, we perform a sign test, which measures how much a certain configuration of pluses and minuses differs from random. [20] To obtain the signs, first note that a temporal-tree naturally entails the notions of parent and children, as well as ancestor and descendant of a node, along paths from the root to the leaves. We assign a plus sign to nodes that have a time stamp later than that of their nearest ancestor, otherwise we assign a minus sign. Thus, the root node is always assigned a plus sign in this labeling scheme. We correct the total number of plus signs by decrementing the number of plus signs for each temporal-tree by 1. Once labeling and correcting the number of pluses we are ready to perform a sign test.

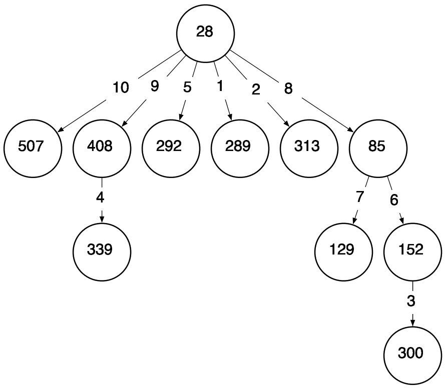

Figure 6. Temporal tree for an observed cluster. Each node number refers to the ordinal time (based on the date) of a message. Numbers on edges indicate the order messages are connected via single linkage.

Zooming in at cluster level, the sign tests are not reliable for smaller clusters due to the fact that we are dealing with small statistics. However, the sign tests support our hypothesis for the few bigger clusters. The corrected sign test (124 pluses and 77 minuses) leads to a p-value of 0.00112 which strongly supports our overall hypothesis. For example in figure 6, we plot a cluster of 11 messages, that entails 10 pluses and a minus. The corresponding corrected sign test with 9 pluses and 1 minus yield a p-value of 0.0215.

5. Discussion

Though our methods achieve a certain level of discriminative and learning capabilities, there exist alternative techniques used in similar problems which one might expect to be equally feasible. However, before we begin critiquing our own methodology, we briefly report on some notable approaches that do not seem to be as useful as expected. In order to classify scam e-mails we first analyzed the usage of alpha-numeric characters and the writing skill level (e.g. the Flesh-Kincade scoring system currently implemented in Microsoft Word's grade writing level evaluator). [21] These methods approximate the complexity of words and sentences, but the only discovery made is that older scam messages use capitalization, and that writing skills are not generally informative. One of the main reasons for this failing is that the story elements of the scams may change completely, and thus a new e-mail may bear little specific resemblance to it's predecessor. Next, we attempted a more general technique and applied principal components analysis to search for a representation of the scam messages in terms of a few interpretable *dimensions*. However, this too yielded little success. Last, we extracted semantic features from the messages using DocuScope [22] to discover that such features could be used to describe the variability of the messages, but they were not useful for discriminating the intents.

What is common to the human eye though is the fraudulent intent hidden between the lines, which can be recovered using a few high and medium frequency words, both non-contextual, as well as contextual but *orthogonal*, to the fraudulent intent in some sense. This intuition led us to the choice of the Poisson filter. [23]

5.1. Fraudulent Intent and Spam Filters

It is important to assess the degree to which current spam filters can cope with fraudulent intent. In this paper, we evaluated SpamAssassin, a popular spam filter which recently received attention for being very accurate, as a benchmark in our experiments. However, as our results demonstrate, SpamAssassin has difficulty in filtering scam from ham as opposed to spam from ham. This difference is significant, given that we observe a threefold difference in SpamAssassin's false classification of such messages. Based on these findings it is clear that a different type of system is necessary for filtering scam messages from the general population of e-mail. This is not overly surprising since one would expect the typical scam message used in our studies to be much more similar to the average ham than spam message. Moreover, the overall goal of SpamAssassin is the classification of spam in general, of which scam is only a fraction. This is supported by the disproportional misclassification rate and by the distribution of scores observed in the SpamAssassin filtering experiments as depicted in Figure 3.

5.2. Single Linkage and Temporal Trees

In general, the sole criteria single linkage clustering requires is there must exist a logical path of data points between any two data points in the cluster. As a result, clusters learned via single linkage tend to have a bias to be more elongated in the vector space than clusters learned through other clustering criteria. In certain settings this is considered a limitation, however, this method is a preferred representation for a hypothesis regarding how scam messages are used by groups of authors. Recall that our hypothesis of scam authorship is that scam messages are reused, such that each time the message is recycled a certain component of the message is changed, but not the whole of the message. With each change, the new scam message deviates a little further from the previous version of the initial scam message.

In addition, the temporal aspect of e-mail may assist in the design of useful heuristics for clustering. For example, one simple heuristic based on time is to incorporate the message date as a feature for measuring the distance between messages. Caution and intuition must be used with such a heuristic since it may predispose messages to cluster in a manner such that authorship relations are eroded. This would more likely be the case if date was considered as part of the cosine measure of distance. Used in this way, clusters would bias toward messages of similar time points, which may not necessarily help to discern between criminal groups perpetrating during the same time period. Rather, it seems more feasible that such a heuristic would be more useful to guide the addition of messages already assigned to a particular cluster, possibly as a tie-breaking criteria. For instance, if a message is equidistant from two or more other messages in the same cluster, then it appears more intuitive to link the documents closer in time.

5.3. Open Data Mining Problems

The results reported in this research are based on a particular scam, the advance fee fraud. Scam messages of this type are susceptible to analysis by text mining partially as a result of being several paragraphs in length and somewhat verbose. The combination of these characteristics permits the use of a significant number of discriminative features for learning the hidden fraudulent intent. Based on our finding, we expect similar results with other types of e-mail scams, such as securities and bank fraud. An extension to our analyses is to determine the usefulness of the ScamSlam system with types of e-mail fraud that communicate much less information in the message body. Given the rise of phisher fraud e-mail over time, the AOL, PayPal, and Ebay scams are of particular interest. However, even though phisher frauds may communicate less information in an e-mail, the websites which they redirect individuals to are amenable to study via text mining as well. This is because ScamSlam, at its core, is basically a text analysis tool, which permits analysis on e-mail messages, webpages, or any other type of information communicated via text.

For other types of scam which make use of images rather than text, different data mining approaches are needed. Nonetheless the availability of free online repositories regarding such scams is a starting point for thinking about how to adapt existing tools or for developing new tools that can be embedded into a filtering system.

6. Conclusions

This paper introduced several challenges to electronic safety recently raised by the Federal Trade Commission and discusses why such problems are open to the data mining community. Specifically, in this paper we concentrated on the case of advance fee frauds. The problem was approached from a text classification perspective; the identification of fraudulent intent in e-mails. Our experiments demonstrate that current filters tailored to spam are not well suited to identify targeted scams. In comparison, we were able to implement a system capable of filtering scam spam from e-mail with error rates comparable to state-of-the-art spam filters.

Furthermore, we oriented the problem from a law enforcement perspective and introduced a forensic architecture, ScamSlam, that can guide cyber-investigations through intuitive distance measures between scam e-mails. With respect to criminal relations behind scam e-mail, we proposed a generative model for scam messages, which models streams of scam messages as evolutionary processes. Such a model is validated using tree-based data structures and a statistical test to determine how well learned relations their correlate with an evolutionary process (i.e. temporal ordering of related e-mails) of scams sent over the Internet. Our findings suggest ScamSlam provides the basis for a forensic tool to assist law enforcement agencies track criminals for which some evidence has been gathered in the form of electronic content.

Finally, for additional types of scam which make use of images rather than text, it is clear that other data mining approaches are needed. Yet, the availability of publicly available online repositories regarding such scams provides an opportunity for studying and adapt existing tools from a data mining perspective for the protection of the general Internet user from being preyed upon by online criminals.

Acknowledgements

The authors wish thank the members of the Data Privacy Laboratory, especially to Dr. Latanya Sweeney, for their insightful suggestions, discussion, and support. For a portion of our experiments we used Minorthird [24], courtesy of Dr. William Cohen. The authors also wish to thank the anonymous reviewers for their comments. This research was supported by the Data Privacy Laboratory in the School of Computer Science at Carnegie Mellon University.

References

[1] MessageLabs Inc. How effective is current legislation? *April Monthly Report*, April 2004.

[2] Federal Trade Commission. National and state trends in fraud and identity theft: January - December 2003. *Federal Trade Commission*, January 22, 2004.

[3] D.R. Shiman. When e-mail becomes junk mail: the welfare implications of the advancement of communications theory. *Review of Industrial Organizations*, 11(1), 35-48, 1996.

[4] Brightmail, Inc. Spam statistics. Available online at http://www.brightmail.com/spamstats.html

[5] B. Warner. Billions of “phishing” scam e-mails sent monthly. *Reuters News Service*, May 6, 2004.

[6] FTC vs. Zachary Keith Hill. *United States District Southern Court of Texas*. File No. 032-3102, December 8, 2003.

[7] P. Turney and M. Littman. Unsupervised learning of semantic orientation from a hundred-billion-word corpus. *Technical Report EGB-1094*, National Research Council, Canada, 2002.

[8] E.A. Erosheva and S.E. Fienberg. Bayesian mixed-membership models for soft clustering and classification. Manuscript, 2004.

[9] D. Jensen. Prospective assessment of AI technologies for fraud detection: a case study. *In Workshop on Artificial Intelligence Approaches to Fraud Detection and Risk Management*, July 1997.

[10] X. Bai, R. Padman, and E.M. Airoldi. Extracting consumer sentiments from unstructured text using tabu search-enhanced Markov blanket. To appear in *Proc. Hawaii International Conference on System Sciences*, Hawaii, 2005.

[11] H. Liu, H. Lieberman, and T. Selker. A model of textual affect sensing using real-world knowledge. *In Proc. International Conference on Intelligent User Interfaces*, 2003.

[12] K. Dave, S. Lawrence, and D.M. Pennock. Mining the peanut gallery: opinion extraction and semantic classification of product reviews. *In Proc. International Conference on World Wide Web*, 2003.

[13] W.W. Cohen, V.R. Carvalho, and T. Mitchell. Learning to classify e-mail into “Speech Acts”. *In Proc. Conference on Empirical Methods in Natural Language Processing*, 2004.

[14] D. Leonard. E-mail threats increase sharply. *IDG News Service*, December 12, 2002.

[15] E.M. Airoldi and W.W. Cohen. Bayesian models for frequent terms in text. *Technical Report No. CMU-CALD-04-106*, Carnegie Mellon University, July 2004.

[16] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison Wesley. New York. 1999.

[17] R. Duda, P. Hart, and D. Stork. *Pattern Classification*, second edition. John Wiley & Sons, New York, 2001.

[18] B. Sullivan. Nigerian scam continues to thrive. *MSNBC News*, March 5 2003.

[19] SpamAssassin$^{\text{TM}}$. Available online at: http://www.spamassassin.org/

[20] J.D. Gibbons. *Nonparametric Statistical Inference*, second edition. Dekker, 1985.

[21] T. Qin, J. Burgoon, and J.F. Nunamaker Jr. An exploratory study on promising cues in deception detection and application of decision tree. *In Proc. Hawaii International Conference on System Sciences*, Hawaii, 2004.

[22] J. Collins and D. Kaufer. Docu-Scope: a Java application for statistical literary style modeling. *Technical Report*, Carnegie Mellon University, 2001.

[23] E.M. Airoldi, W.W. Cohen, and S.E. Fienberg Statistical models for frequent terms in text. *Manuscript*, 2004.

[24] W.W. Cohen. Minorthird. Available online at http://minorthird.sourceforge.net, 2004.