Technologies to Defeat Fraudulent Schemes Related to Email Requests

Edoardo Airoldi, Bradley Malin, and Latanya Sweeney

Data Privacy Laboratory, School of Computer Science
Carnegie Mellon University, Pittsburgh, Pennsylvania 15213-3890
{airoldi, malin, latanya}@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. We argue that 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 starting from existing AI tools. We illustrate how to design a learning system capable of accurately identifying the fraudulent indent within an e-mail in order to tackle, for example, the advance fee fraud scam. Further, we propose data structures, as well as statistical tests for them, which capture evolutionary patterns within e-mails that are not likely to be due to chance. Last, our system can serve as a guide for law enforcement agencies in cyber-investigations.

Email Populations and Spam Filters

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. In combination with other factors, including the increased usage 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 anti-spam company, indicate that as of February 2003 approximately 42% of all messages sent over the Internet were spam. By April 2004 this number had increased to almost 65% of over 96 billion messages, filtered during a single month.

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 and 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 that are viewed as spam by some individuals and ham by others (e.g. legitimate, but unsolicited advertisements); depicted in the intersection of figure 1.

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 a combined scam, easy/hard spam, ham corpus to analysis and classification by SpamAssassin, the popular open source spam filter. 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 that 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.
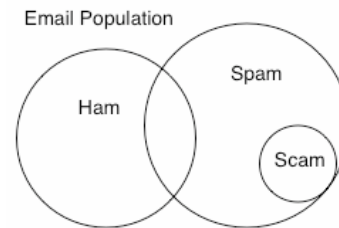

Fig 1. Email populations.

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 resulting scores are depicted in figure 2 with the threshold score depicted by a thin vertical line. Based on the observed scores, SpamAssassin does very well at classifying ham as ham. However SpamAssasin has a more difficult
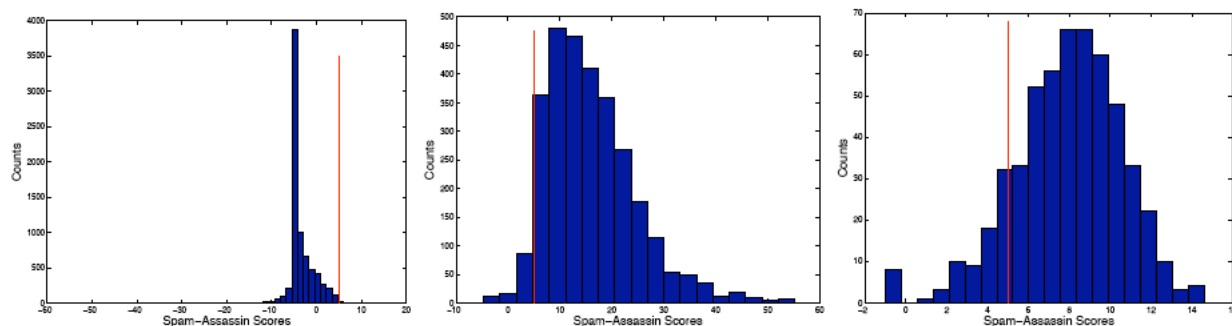
Fig 2. SpamAssassin scores on Ham, Spam and Scam emails.

time classifying the other message types and disproportionately so for spam versus scam. As seen in figure 2 SpamAssassin misclassifies about approximately 4.1% of the spam and approximately 11.8% of the scam messages as ham, respectively. Similar results were observed on other corpora. 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.

Overview of the X-RAY System

Our system is called *X-RAY*, for eXtraction of fRaudulent Activities against You, and consists of methods for identifying fraudulent intent by analyzing the semantics of the content of electronic messages. In particular, *X-RAY* has two components:

a) X-TEXT: In prior research, we developed methods to recognize advance fee fraud intent (Airoldi and Malin 2004a). Yet, different types of story-based fraud exist, such as pyramid schemes or phisher frauds. We propose to extend and apply the methodology underlying *X-TEXT*, which learns to recognize fraudulent intent within texts by measuring similarity to known fraudulent messages. We will test *X-TEXT* with our own (Sweeney 2003), and corporate affiliates' (ChoicePoint), repositories of emails.

b) X-ACT: Fraudulent emails entice users to perform certain actions, such as respond, visit a trap website, click a hyperlink, or supply personal information. We propose to design and evaluate *X-ACT*, which will identify action requests within emails, akin to part of speech detection.

Email analysis is an emerging application domain of machine learning and intelligent systems, however, there has been minimal investigation into preemptive efforts to combat identity theft. Current techniques combat spam, viruses, and malware, but do not address the problem of detecting *intent* or *action-requests*. In this respect, we present the only approaches to prevent identity theft before it begins by detecting fraudulent intent prior to the user taking action. In our research we verified that a direct approach is more quantitatively accurate than other

attempts, such as Bayesian spam filtering (Airoldi and Malin 2004).

Core Technology

At the core of *X-TEXT* existing technology for fraudulent intent detection is the exploitation of novel statistical models (Airoldi and Cohen 2004, Airoldi, Cohen, and Fienberg 2004, Bai, Padman, and Airoldi 2004, 2005). These models take advantage of labeled samples of fraudulent email messages and malicious executables. Their strength is rooted in causal reasoning and adaptive learning theory. Intuitively:

- **Representative Sample.** We train our system using labeled fraud examples. By separating the training phase from the testing phase, we avoid over-fitting in the estimation of the parameters underlying our models. We are able to produce reliable decisions regarding the intent of new, unseen messages.

- **Statistical Models.** We model words that occur at different frequencies with different statistical models. Our research shows that mimicking the distributions according to which terms are generated consistently improves decision accuracy (Airoldi, Cohen, and Fienberg 2004, 2005). Further, the models we developed for medium-to-high frequency terms are appropriate for applications to fraudulent emails, as they tend to be lengthy (Airoldi 2003).

- **Cause/Effect Patterns.** We learn statistically significant dependencies among terms relevant to intent detection. We enrich the dependency structure by adding cause and effect patterns. This two-stage process allows us to obtain a parsimonious vocabulary and a complex, but interpretable, structure of patterns among terms (Bai, Padman, and Airoldi 2004, 2005).

- **Adaptive Learning.** The user's final action (e.g. click / no click) is compared to prediction of the automated decision process; the model is updated accordingly. The history of decisions, along with other characteristics such relationships between messages, provide training data for our online learning algorithms. As a result, the
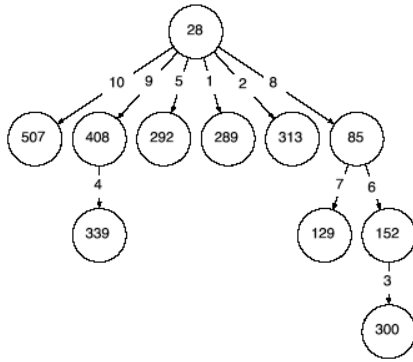
Fig 3. Time tree with p-value ≈ 0.0215.

Fig 4. Model for cyber-investigations.

protection model evolves over time and automatically adapts to account for new perpetrations.

We have performed an in-depth analysis of the advance fee fraud, and were able to accurately detect the fraudulent intent. The parameters underlying our models can be learned in linear time.

For *X-ACT* we will use our information extraction language (Cohen 2004) to identify messages that require the recipient to act upon them by sending some sort of information to the sender and to extract what are the semantics of the information requested. Depending on the sensitivity of the information requested, and on the typical/ historical behavior of the user, messages will be assigned, an assessment about the potential economic damage.

Performance Assessment. We are in possession of a large representative corpus of real-world fraud examples and non-malicious email messages and executables that will allow for an unbiased assessment of the performance of our methodology. We use two metrics to evaluate the accuracy of our methodology. The cross-validated accuracy, which quantifies the expected accuracy for a fixed sensitivity threshold of the decision process, relies on a large number of experiments where testing is performed on out-of-training-sample examples. The area under the ROC curve, which provides an assessment of the overall quality of the classification methods, informs us about the cross-validated accuracies at different sensitivity thresholds.

Concluding Remarks

We approach the problem of detecting the fraudulent intent from a text classification perspective 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, our document forensic architecture, *X-RAY*, can guide cyber-investigations (figure 4) 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
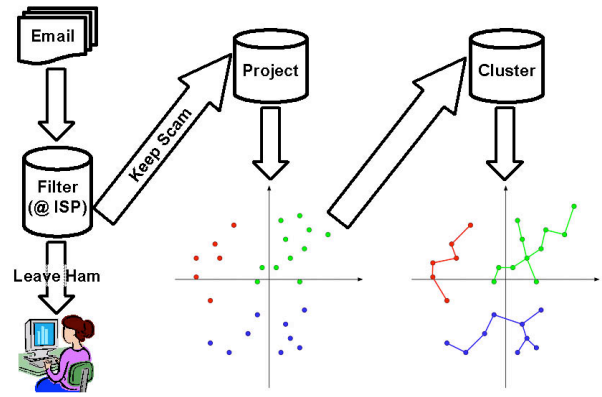
streams of scam messages as evolutionary processes. Such a model is validated using tree-based data structures and a statistical test to determine how well the learned relations correlate with an evolutionary process (figure 3) of scams sent over the Internet. *X-RAY* provides the basis for a forensic tool to assist law enforcement agencies track criminals about whom some evidence has been gathered in the form of electronic content.

References

Airoldi, E., and Malin, B. 2004a. Scam-Slam: An Architecture for Learning the Criminal Relations Behind Scam Spam, Tech. Report, CMU-ISRI-04-121, School of Computer Science, Carnegie Mellon Univ.

Airoldi, E., and Malin, B. 2004b. Data Mining Challenges for Electronic Safety: the Case of Fraudulent Intent Detection in E-mails. In *Proceedings of the Workshop on Privacy and Security Aspects of Data Mining, IEEE International Conference on Data Mining*. Brighton, England.

Sweeney, L. 1996. Replacing Personally-Identifying Information in Medical Records, the Scrub System. In *Proceedings of the 1996 American Medical Informatics Association Annual Symposium*, 333-337. Washington, DC. Hanley & Belfus, Inc.

Airoldi, E., Cohen, W., and Fienberg S.E. 2004. Statistical Models for Frequent Terms in Text, Tech. Report CMU-CALD-04-106, School of Computer Science, Carnegie Mellon Univ.

Airoldi, E., Cohen, W., and Fienberg, S.E. 2005. Bayesian Models for Frequent Terms in Text. *Manuscript*.

Bai, X., Padman, R., and Airoldi, E. 2004. Sentiment Extraction From Unstructured Text Using Tabu Search-Enhanced Markov Blanket. In *Proceedings of the Workshop on Mining the Semantic Web, ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Bai, X., Padman, R., and Airoldi, E. 2005. On Learning Parsimonious Models for Extracting Consumer Opinions. In *Proceedings of the IEEE Hawaii International Conference on System Sciences*, 75b.

Sweeney, L. 2003. Email Archives for Combating Spam. http://privacy.cs.cmu.edu/courses/pad1/assign/lab10/archives.html.

Airoldi, E. 2003. Who Wrote Ronald Reagan's Radio Addresses?, Technical Report CMU-STAT-03-789, Statistics Department, Carnegie Mellon University.

Cohen, W. 2004. Minorthird.: methods for identifying names and ontological relations in text using heuristics for inducing regularities from data. http://minorthird.sourceforge.net.