Spam Miner: A Platform for Detecting and Characterizing Spam Campaigns*

Pedro H. Calais Guerra
Univ. Federal de Minas Gerais
Belo Horizonte, Brazil
pcalais@dcc.ufmg.br

Douglas E. V. Pires
Univ. Federal de Minas Gerais
Belo Horizonte, Brazil
dpires@dcc.ufmg.br

Marco Túlio C. Ribeiro
Univ. Federal de Minas Gerais
Belo Horizonte, Brazil
marcotcr@dcc.ufmg.br

Dorgival Guedes
Univ. Federal de Minas Gerais
Belo Horizonte, Brazil
dorgival@dcc.ufmg.br

Wagner Meira Jr.
Univ. Federal de Minas Gerais
Belo Horizonte, Brazil
meira@dcc.ufmg.br

Cristine Hoepers
Network Information Center
NIC.br, São Paulo, Brazil
cristine@cert.br

Marcelo H. P. C. Chaves
Network Information Center
NIC.br, São Paulo, Brazil
mhp@cert.br

Klaus Steding-Jessen
Network Information Center
NIC.br, São Paulo, Brazil
jessen@cert.br

ABSTRACT

This demo presents Spam Miner, an online system designed for real-time monitoring and characterization of spam traffic over the Internet. Our system is based on high-level abstractions such as spam message attributes, spam campaigns and spamming strategies. A campaign is a cluster of messages that are generated from a single message template; campaign identification is a challenging problem because it has to handle spammer evolution, while seeking for a spam similarity function that combines different message characteristics and for strategies that efficiently process large volumes of spams. Moreover, spam campaigns need to be identified on-the-fly, to allow incident response teams and security specialists to react to the threat adequately. Spam Miner addresses campaign identification as a data clustering problem and campaigns are identified dynamically using a novel incremental approach based on the concept of Frequent Pattern Tree. Spam Miner is being used by NIC.br (Brazilian Network Information Center) and mined more than 350 million spam messages, detecting meaningful clusters and patterns, and helping the organization to better understand the spam problem in Brazil and how the Brazilian Internet infrastructure is being abused by spammers.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous; H.2.8 [**Database Applications**]: Data mining

*This work was partially supported by UOL, NIC.br, CNPq, CAPES, FAPEMIG and FINEP.

KDD'09 June 29-July 1, 2009, Paris, France.

General Terms

spam, clustering, data mining

1. INTRODUCTION

Together with the development and popularization of the Internet, spam has become one of the biggest sources of unwanted traffic [4]. Some ISPs report that between 40% and 80% of the messages received by their mail servers are unwanted, and other studies estimate in billions of dollars the losses of American companies due to spam [5]. Motivated by its highly negative impact, researchers from different areas have been working towards detecting and mitigating spam and unwanted traffic in general. Data mining techniques such as Bayesian Filters, SVM Classifiers and Decision Trees have been applied to detect and filter spams with relative success. However, a notable characteristic of the spam problem is the *spam arms race* phenomenon, in which both spammers and anti-spammers evolve, trying to beat each other efforts [3]. As a consequence, continuous monitoring and measurement of spam traffic characteristics is necessary. Indeed, this scenario poses an interesting and challenging data mining problem, as highlighted in [2]: how to mine a continuous flow of spams, detecting and anticipating new threats and patterns in an evolving and changing environment.

Spam Miner is a platform based on data mining algorithms to monitor and characterize spam traffic. To address the huge amount of spam messages to be analyzed, as those collected by honeypots [6], we employ the concept of **spam campaign**, which can be seen as a cluster of spam messages that are originated from a single campaign template. To disseminate a campaign, spammers obfuscate and insert randomness on each message's header and body to evade detection. Spam Miner undoes this process and turns the fragmented spammer behavior spread on spam messages into meaningful, representative behavior carried on spam campaigns, identifying the different strategies used by the spammer to obfuscate content and distribute them through the network. This approach provides higher level abstractions we can work with, like spam campaigns and campaign

strategies, instead of forcing us to consider each message in isolation. To the best of our knowledge, we are the first to build an online data mining system for spam characterization purposes.

Spam campaign identification is a difficult problem and has been addressed in the literature as a near-duplicate text problem, and by image similarity [8] and URL similarity [9, 10] algorithms. Each technique considers different spam features, and we propose a more generic framework that combines those heterogeneous features into a single clustering approach.

The difficulty on determining spam campaigns efficiently and online comes from a set of reasons. The majority of the techniques previously mentioned are designed for post-mortem analysis: all spams are collected and then grouped. Our approach is to detect campaigns in real-time, which brings the challenge that we are not able to observe a full pattern and later decide whether it corresponds to a new spam campaign or not; detecting a campaign *while* it is being disseminated is a harder problem.

Another major challenge concerning the detection of spam campaigns resides in the fact that spammers are always evolving and creating new tricks, strategies and forms of abuse. Thus, any technique that looks for pre-defined campaign generation patterns will become less effective over time. Previous campaign identification patterns rely on fixed obfuscation patterns: messages will be grouped into a campaign if they share several terms, their URLs share the same domain, and differ just on CGI parameters, for example. Our contribution is to see campaign identification problem as a genuine data mining problem and thus the obfuscation patterns are mined and not fixed.

A third challenge arises from the fact that it is very difficult to define a distance metric – required by most clustering algorithms – that combines heterogeneous spam message attributes such as language, URL tokens and text tokens. Which set of spams is more similar, the one which exhibits similar URLs or the one that shares a similar subject? Our technique eliminates the need for defining a distance measure.

Finally, spam monitoring results in a huge amount of messages and thus an efficient strategy for grouping messages into campaigns is required.

Spam Miner implements a novel data clustering algorithm that addresses these problems. Our technique employs a **Frequent Pattern Tree** and is **incremental**, being able to process a continuous flow of spam traffic, **efficient**, as it does not compare messages pairwise, and **generic**, as it is not tied to a given content obfuscation technique and is able to deal with the constant spammer's evolution.

In the next sections, we present the architecture and the main features of the system, which is currently a prototype being used by the Brazilian Network Information Center (NIC.br) to unveil information from more than 350 million spam messages during 15 months of operation.

2. THE SPAM MINER SYSTEM

Our system architecture is shown in Figure 1. The **Data Collection module** collects spam traffic from a set of low-interaction honeypots [6], which are systems designed to collect, detect, deflect, or in some manner counteract attempts at unauthorized use of information systems. Our honeypots emulate open proxies and open mail relays, which are machines over the Internet traditionally abused by spammers [6]. Honeypots have been suggested as tools that can play a significant role in providing early warning in the case of network attacks and so they are perfect for our purpose of building a system that monitors spamming activity.

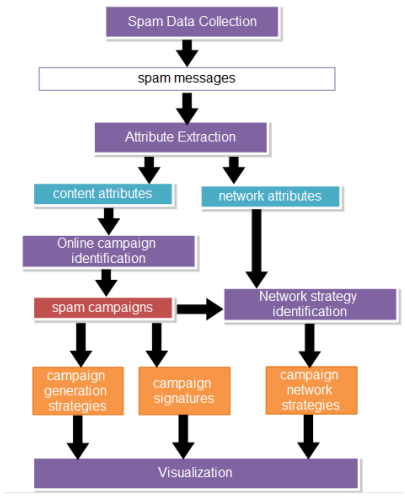

Figure 1: Spam Miner System Architecture

All strategy identification activity in Spam Miner is focused on campaign attributes, gathered by the **Attribute Extraction module**. We chose message and network attributes that have been identified as key to the analysis of spam campaigns, and that are related to characteristics often obfuscated by spammers.

In terms of message attributes, currently we consider the layout, subject, language, and encoding, as well as the URLs found in the text. These features contain useful information to discriminate and describe spam campaigns. For example, message layout (a codification of message formatting properties to a sequence of characters) has been an important identifying feature, since it usually is not obfuscated by template-based bulk mailers. These data are fed into the **Online Campaign Identification module**, which determines spam campaigns, as detailed in Section 2.1.

Network attributes collected include the traditional connection identifiers: timestamp, source and destination IP addresses, and TCP ports of the connections that reached the honeypots to deliver the messages. From IP addresses we derive and store the associated country codes. The **Network Strategy Identification module** uses that information, along with network attributes, to output the network strategies employed by spammers using data mining algorithms. Once campaigns and strategies are identified, they are stored and fed to the **Visualization module**, responsible for presenting the information to users.

Our system is extensible in the sense that we handle messages' attributes generically. As soon as new attributes, which may improve campaign identification or make the characterization more complete, are identified, they can be readily incorporated into our framework. In the next sections we describe how campaigns are identified and how it is performed incrementally, as well as their visualization.

2.1 Online Campaign Detection using FP-Tree

Spam campaign identification can be seen as a data mining clustering problem — the process of partitioning a set of messages into meaningful groups. The messages in each cluster (campaign) should be similar to one another and thus they can be treated collectively as a group.

Based on the extracted attributes discussed previously, our campaign identification strategy determines the invariant parts of the spam message attributes and uses a dynamic Frequent-Pattern tree (introduced by the FP-Growth pattern mining algorithm [7]) to organize them hierarchically. In our case, the FP-Tree is a prefix tree where each node after the root represents a feature extracted from the spam messages, which is shared by the sub-trees underneath, and the root itself is empty. Each path in the tree represents a set of features that co-occur in messages, in non-increasing order of frequency of occurrence. Therefore, messages that have frequent features in common (such as language, subject and layout) and differ only w.r.t. infrequent features will share a common path in the tree. We delimit campaigns by analyzing the occurrence of random and infrequent fragments that appear in a path. We identify such fragments by a significant frequency change considering the ancestor or a sudden increase in the number of children, or both. In other words, all messages that are in the sub-trees under a node where a significant increase in the number of children is observed are grouped into the same campaign. In practice, the tree allows the detection of messages that share invariant (frequent) characteristics and differ due to obfuscated (random) features. The power of the technique comes from the fact that we do not pre-define obfuscation patterns to be matched against; such patterns are naturally identified when the tree is constructed.

Figure 2 illustrates a small portion of the tree, with three large campaigns in the center (differentiated by distinct obfuscation patterns, determined by the different sequence of colors among tree levels). Figure 3 shows another portion of the tree representing a single campaign. We can notice a sequence of invariants, which are the features each spammer kept static, followed by a sequence of obfuscated features. The regular aspect of each sub-tree, in fact, determines a spam campaign template, which defines a campaign's signature. The different signatures unveiled represent different spam generation strategies.

From our observations, the most common infrequent feature found in the tree is usually a URL fragment randomly generated, while message layout, type and language emerge as strong content invariants. Notice that these differences were naturally found by our technique. If a spammer starts obfuscating a new attribute, the new pattern will be promptly detected by the FP-Tree. This is how we deal with the evolutionary aspect of the spamming activity.

Our experiments showed that we are able to reduce by thousands of times the amount of elements to consider in our analysis: 350 million messages with 6 million unique URLs were reduced to 60 thousand distinct campaigns. This reduction directly enables the use of other data mining algorithms that can now be applied to the campaigns identified, eliminating the need for processing hundreds of millions of messages. As the spammer behavior is summarized through their spam campaigns, many other complex analysis become possible, like crawling and analyzing the web pages pointed by spam URLs. Determination of spam campaigns also creates new dimensions associated with each campaign's spam traffic, such as volume of messages and duration of abuses, which can then be correlated and analyzed. For network administrators and security professionals, monitoring spam traffic through spam campaigns is also an easier task than examining each spam message individually.

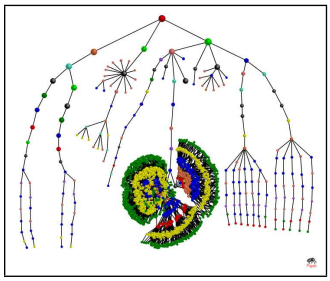

Figure 2: *FP-Tree showing distinct spam campaigns*

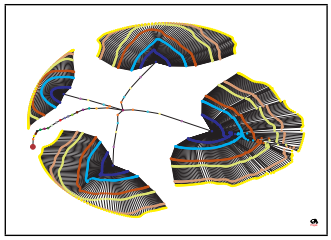

Figure 3: *branches of the FP-Tree showing a single campaign*

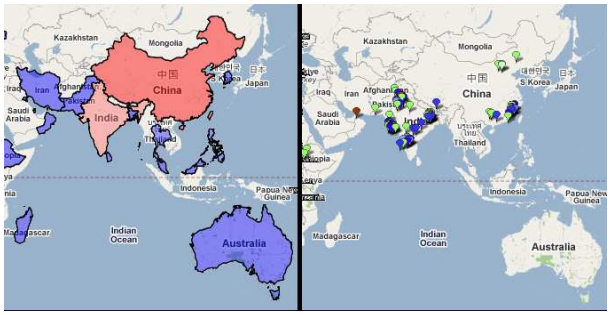

Figure 4: Different views offered by Spam Miner: Heat Map and markers representing campaigns

2.2 Online, incremental FP-Tree

The key point to build the Frequent Pattern Tree online is to note that it is not possible anymore to compute the global frequency of each spam message feature, which will be known only when the dataset is fully processed (which may never happen, if the spam traffic flow is continuous). Our solution to this problem is to build the tree incrementally, i.e., an algorithm that is able to process spam traffic as it is collected and to use the knowledge acquired from the traffic previously observed, avoiding to process again the same transactions and operations every time new data is available.

In the *incremental FP-Tree*, frequencies are computed as spam messages arrive and their characteristics are extracted. They are inserted in the Tree as explained earlier, but now it is not possible to guarantee that the FP-Tree's properties are maintained as new spam data is processed. A child may become more frequent than its parent and, then, the FP-Tree key property (infrequent features are near the leaves) is violated. We implemented a set of operations which permute, join and divide nodes to reconstruct the tree periodically.

The incremental FP-Tree scales well, since processing each message requires only a traversal of the tree, so messages are not compared pairwise, what would lead to a quadratic complexity on the number of messages.

2.3 Visualization

Our system offers a web-based visual interface, whose main screen presents a world map (using Google Maps), as shown in Figure 4. The interface shows a heat map (left side), which displays measures such the as number of spam campaigns disseminated by each country. On the right side of the same figure, another view provided by our system can be seen: colored markers pop on the map when spam campaigns are detected, and each marker represents an IP address that abused one of our honeypots, mapped to its location through a geolocation service. Each color identifies a different campaign; that makes it possible to assess the geographical dispersion of spamming abuses. If the user selects a campaign marker, a detailed pop-up frame is shown with specific information for that campaign, such as the campaign obfuscation strategy, as recorded by the FP-Tree.

3. EXPERIENCE WITH SPAM MINER

We have developed Spam Miner for NIC.br (Brazilian Network Information Center), which aims to better understand the spam problem in Brazil and how the Brazilian Internet infrastructure is being abused by spammers. The system has been key to enable the analysis of a huge volume of data and to better understand the characteristics of the spam traffic in Brazilian networks.

Together with the campaign identification technique, Spam Miner applies an association rule mining algorithm on campaigns' attributes. It identified relationships between the language and the source of spam, as well as relationships between the origin of the abuses and the intended recipients of spam messages. Based on those observations, an interesting finding was that most of the observed traffic was originated from abroad and was aimed at recipients in other countries, using the Brazilian networks simply as a stepping stone, a way to hide their real origin.

The campaign identification process have identified campaigns in which spammers are able to obfuscate the web site domain of the URLs embedded in message bodies; this pattern would not be identified by URL similarity techniques such as [9] and [10], which look for some pre-defined URL obfuscation patterns (such as obfuscation of URL CGI parameters).

Additional results of applying the campaign characterization approach to characterize spam traffic were presented in a previous work [1].

4. CONCLUSIONS AND OUTLOOK

Understanding spam campaigns and being able to identify them as they evolve is an important activity. In this work, we presented Spam Miner, an online spam campaign monitoring and characterization system that processes spam traffic and applies data mining techniques to unveil meaningful abstractions such as spam campaigns, strategies and spam signatures. Due to its online and real-time nature, our system helps network administrators to monitor how their networks are abused. A prototype of the system is being used by NIC.br (Brazilian Network Information Center), and a demo is available at `http://spammining.speed.dcc.ufmg.br/`.

Our future plans include extending the system to consider other data mining algorithms for online spam trend detection, differentiating typical campaigns from campaigns which exhibit evolving or brand new patterns when compared to previously known spamming strategies. We are also working on new visualization schemes, such as displaying an online animation of the Frequent Pattern Tree being constructed dynamically. We are also considering other application domains that may be suitable for applying our incremental Frequent Pattern Tree clustering approach.

5. REFERENCES

[1] P. H. Calais, D. Pires, D. Guedes, J. Wagner Meira, C. Hoepers, and K. Steding-Jessen. A campaign-based characterization of spamming strategies. In *Proceedings of the 5th Conference on e-mail and anti-spam (CEAS)*, Mountain View, CA, 2008.

[2] T. Fawcett. "in vivo" spam filtering: a challenge problem for kdd. *SIGKDD Explor. Newsl.*, 5(2):140–148, 2003.

[3] J. Goodman, G. V. Cormack, and D. Heckerman. Spam and the ongoing battle for the inbox. *Comm. ACM*, 50(2):24–33, 2007.

[4] B. Hayes. Spam, spam, spam, lovely spam. *American Scientist*, 91(3):200–204, May–June 2003.

[5] J. C. Sipior, B. T. Ward, and P. G. Bonner. Should spam be on the menu? *Commun. ACM*, 47(6):59–63, 2004.

[6] K. Steding-Jessen, N. L. Vijaykumar, and A. Montes. Using low-interaction honeypots to study the abuse of open proxies to send spam. *INFOCOMP Journal of Computer Science*, 2008.

[7] P. Tan, Steinbach, M., and V. Kumar. *Introduction to Data Mining, (First Edition)*. Addison-Wesley Longman Publishing Co., 2005.

[8] Z. Wang, W. Josephson, Q. Lv, M. Charikar, and K. Li. Filtering image spam with near-duplicate detection. In *Proc. of the Fourth Conference on Email and Anti-Spam (CEAS). Mountain View, CA.*, 2007.

[9] Y. Xie, F. Yu, K. Achan, R. Panigrahy, G. Hulten, and I. Osipkov. Spamming botnets: signatures and characteristics. *SIGCOMM Comput. Commun. Rev.*, 38(4):171–182, 2008.

[10] C.-C. Yeh and C.-H. Lin. Near-duplicate mail detection based on url information for spam filtering. In *Information Networking. Advances in Data Communications and Wireless Networks*, pages 842–851. Springer Berlin / Heidelberg, November 2006.