Uncovering Social Spammers: Social Honeypots + Machine Learning(Kyumin Lee; James Caverlee ;Steve Webb)

Yi, Muxia (muxiayi2)

April 30, 2018

Abstract

Spammers, content polluters and malware disseminators threatened a new community value and related services such as search and advertising. This paper come up with a honeypot-based approach for uncovering social spammers and evaluated it so as to preserve community value and ensure it success. The proposed approach contained the deployment of social honeypots and statistical analysis of the properties. This paper discovered that the deployed social honeypots had low rate of false positives when recognizing spammers while signals were correlated with observable prole features which enable us to design machine learning based classiers. These classigiers recognized unknown spammers with high accuracy and low false positive rates.

1 Introduction

Relying on users as main contributors, annotators and raters of content is a key feature of social system that can lead to positive effects as well as be targets of social spammers. So successfully preventing social spammers is important, however we know little about spammer such as their level of sophistication, their strategies and tactics. Depending on human-in the-loop inspection of training data for building spam classiers and community-contributed spam referrals are two major traditional techniques to find spam users. Unfortunately since zero-day social spam attacks with no existing signature or wide evidence, they cannot be dealt by these traditional spam discovery approach.

To address these problems, this paper provided an approach which aims to automatically collect spam proles, develop strong statistical user models and positive lter out unknown spammers such as zero-day. Getting some creative idea from preview researches of honeypots to observe and analyze malicious activity [4], generating intrusion detection signatures [2], and observing email address harvesters [3]), the paper deleloped social honeypots for trapping evidence of spam profile behavior. Test showed that the approach proposed in this paper is able to offer generalized and efficient social spam detection after evaluating two different communities (MySpace and Twitter).

2 Methodology

The social spam detection problem is defined as follow:

A classifier $c : u_i - [spammer, legitimateuser]$ approximates whether $u_i$is a spammer. Since social spam detection is more disruptive to spammers, in this paper, social honeypots which defined as information system resources was suggested to monitor activity and behaviors of spammer, to log information, and to specically target community-based online activities.

The overall social honeypot architecture in this paper was showed as figure 1. Social honeypot consisting of a legitimate prole and an associated bot to detect social spam behavior would collect evidence of the spam candidate when it detects suspicious user activity. Based on new observations, it would optimize and update for the particular community. The researchers extracted remarkable features from the collected candidate spam proles since social honeypots collectted spam evidence. The initial training set of a spam classier contains spam and legitimate proles which were more populous and easy to extract from social networking communities. Iteration refines the selected features and used classier to optimize the spam classier over the known spam and legitimate proles. The next step was to explore wider space of unknown proles. My Space had 100s of millions of proles including spam. Since harvested social honeypot data had develop some calssifiers, so detect these proles to explore new identified spammers in social honeypot. In addition, human inspectors were included in-the-loop in order to validate the quality of extracted spam candidates while inspectors only need to validate a few spam candidates recommended by the learned classiers and proveide feedback. Then update the spam classiers by the new evidence and continue the process.

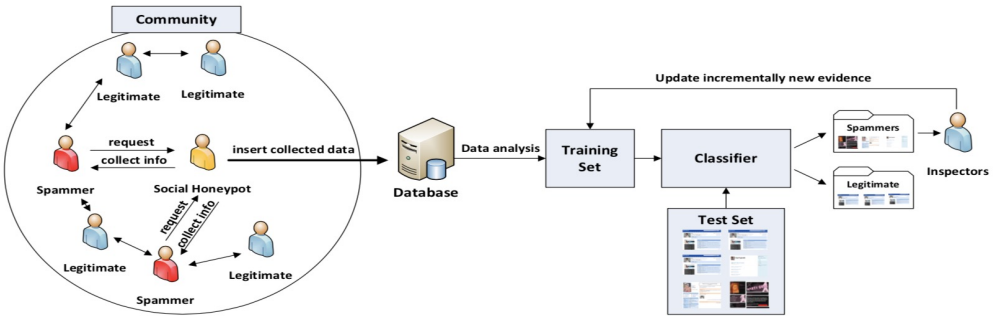

Figure 1: Overall Framework of Social Honeypot-based Approach

Figure 1:

The classication implemented by using 10-fold cross validation while the original sample was randomly divided into 10 equally-sized subsamples, because when a dataset was not large, it was common to use 10-fold cross-validation to achieve statistically precise results. And standard metrics like precision, recall, accuracy, the F1 measure, false positive and true positive were used to measure effectiveness of classiers. All these were defined as : the precision (P) of the spammer class is a/(a + c). Recall (R) is of the spammer class which was calculated as a/(a + b). Accuracy was the proportion of the total number of predictions that were correct. The accuracy was determined as (a+d)/(a+b+c+d). F1 measure of the spammer class was calculated as 2P R/(P + R).

A Receiver Operating Characteristics (ROC) curve was used to measure the discrimination power as well. In the research, the range of ideal ROC curve was from 0 to 1. When handle text data, the performance measured by each word basing on$tf - idfweighting : tf - idf_{t,d} = log(1 + tf_{t,d}) * log(N/df_t)$, where $tf_{t,d}$ means term frequency of term t in a prole, N is the number of proles, and df_t is the number of proles which includes term t. The length in bytes of the content was also measured. on Twitterthe content similarity was calculated as $similarity(a, b) = \frac{V(a) \cdot V(b)}{||V(a)||||V(b)||}$, so the total average content similarity was: $\sum \frac{similarity(a,b)}{|setofpairsintweets|}$. When we were able to calculated Spam precision as: $SpamPrecision = \frac{truepositives}{truepositives+falsepositives}$

3 Experiment/Results

In Myspace experiments, data sets were randomly collected 388 legitimate proles from MySpace and 627 deceptive spam proles. Because the data set was collected from public proles duplicated proles was removed, the MySpace data was either spammer or legitimate. In addition, nal experiments was done by a large database: a data set including about 1.5 million of public proles collected in 2006 and 2007[1].Since both data set of experiments were random and one of them was large enough, they were representative. The major metrics to measure the performance were several representative user features: number of friends, age, marital status, gender, and some text-based features, and their meaning had been explained at section Methodology. The best classier was Decorate because its accuracy is 99.21 percent, F1 measure is 0.992, and false positive rate is only 0.7 percent. Different training mixtures of spam and legitimate training data showed that the metrics are strong across these changes in training data and all of the classiers were successful.

In Twitter experiments, 104 legitimate users were randomly collected as data sets from a Twitter data set of 210,000 users with 61 spammers and 107 promoters were selected as two classes of spam users. Additional, nal experiment was done with a large database collected from Twitter including 215,345 user proles, 4,040,415 tweets. So the data set were very presentive. To model users features, metrics were: the longevity of account on Twitter, the average tweets per day, the ratio of the number of following and the number of followers, the percentage of bidirectional friends, some features of the tweets sent as well as the average content similarity. To model the content in tweets, some text-based features were metrics. The meaning of metrics was the same as in section Methodology. The meta classiers had better performance and the best classier was Decorate. Considering different training mixtures of spam and legitimate training data the classication, metrics are strong and healthy.

These experiments showed that social honeypots can achieve effective social spam detection, though they did not show the degree of traditional email and web spam approaches. Additional, initial results provided positive evidence for the robustness of the method designed by this paper due to the constant adaptation of spammers, so social honeypots can address social spam attacks.

4 Conclusions

In this paper, a new social honeypot-based approach was designed to address social spam detection problem. So coupled with investigating techniques and developping effective tools for automatically detecting and filtering spammers who target social systems, this new approach was propased after examming the traditional techniques. Its efficiency was assessed by MySpace and twitter. The experiments clearly provided positive evidence for the robustness of this new method, which could be used to solve uncovering social spammers problems. However, the paper did not answer that what degree traditional email and web spam can approach. Additional, how and why social honeypot-based approach can handle social spam detection was not answered as well. So for a more comprehensive study, that how and why social honeypot-based approach can handle social spam detection would be an interesting topic .

References

[1] J. Caverlee and S. Webb. A large-scale study of myspace: Observations and implications for online social networks, 2008.

[2] C.KreibichandJ.Crowcroft. Honeycomb:creatingintrusiondetection signatures using honeypots, 2004.

[3] L. A. a. M.B.Prince, B.M.Dahl. Understanding how spammers steal your e-mail address: An analysis of the first six months of data from project honey pot, 2005.

[4] L. Spitzner. The honeynet project: Trapping the hackers, 2003.