Uncovering Social Spammers: Social Honeypots + Machine Learning

Kyumin Lee
Department of Computer Science and Engineering
Texas A&M University
College Station, TX, USA
kyumin@cse.tamu.edu

James Caverlee
Department of Computer Science and Engineering
Texas A&M University
College Station, TX, USA
caverlee@cse.tamu.edu

Steve Webb
College of Computing
Georgia Institute of Technology
Atlanta, GA, USA
steve.webb@gmail.com

ABSTRACT

Web-based social systems enable new community-based opportunities for participants to engage, share, and interact. This community value and related services like search and advertising are threatened by spammers, content polluters, and malware disseminators. In an effort to preserve community value and ensure long-term success, we propose and evaluate a honeypot-based approach for uncovering social spammers in online social systems. Two of the key components of the proposed approach are: (1) The deployment of social honeypots for harvesting deceptive spam profiles from social networking communities; and (2) Statistical analysis of the properties of these spam profiles for creating spam classifiers to actively filter out existing and new spammers. We describe the conceptual framework and design considerations of the proposed approach, and we present concrete observations from the deployment of social honeypots in MySpace and Twitter. We find that the deployed social honeypots identify social spammers with low false positive rates and that the harvested spam data contains signals that are strongly correlated with observable profile features (e.g., content, friend information, posting patterns, etc.). Based on these profile features, we develop machine learning based classifiers for identifying previously unknown spammers with high precision and a low rate of false positives.

Categories and Subject Descriptors: H.3.5 [Online Information Services]: Web-based services; J.4 [Computer Applications]: Social and behavioral sciences

General Terms: Design, Experimentation, Security

Keywords: social media, social honeypots, spam

1. INTRODUCTION

The past few years have seen the rapid rise of Web-based systems incorporating social features – from Web-based social networks (e.g., Facebook, MySpace) to online social media sites (e.g., YouTube, Flickr) to large-scale information sharing communities (e.g., Digg, Yahoo! Answers). These social systems have attracted a tremendous amount of media and research interest [1, 10, 18]. One of the key features of these systems is their reliance on users as primary contributors of content and as annotators and raters of other's content. This reliance on users can lead to many positive effects, including large-scale growth in the size and content in the community, bottom-up discovery of "citizen-experts", serendipitous discovery of new resources beyond the scope of the system designers, and new social-based information search and retrieval algorithms. Unfortunately, the relative openness and reliance on users coupled with the widespread interest and growth of these social systems has also made them prime targets of *social spammers*.

In particular, social spammers are increasingly targeting these systems as part of phishing attacks [14], to disseminate malware [5] and commercial spam messages [7, 26], and to promote affiliate websites [17]. In the past year alone, more than 80% of social networking users have "received unwanted friend requests, messages, or postings on their social or professional network account" (Source: Harris Interactive, June 2008). Unlike traditional email-based spam, *social spam* often contains contextual information that can increase the impact of the spam (e.g., by eliciting a user to click on a phishing link sent from a "friend") [7, 12, 14].

Successfully defending against these social spammers is important to improve the quality of experience for community members, to lessen the system load of dealing with unwanted and sometimes dangerous content, and to positively impact the overall value of the social system going forward. However, little is known about these social spammers, their level of sophistication, or their strategies and tactics. Filling this need is challenging, especially in social networks consisting of 100s of millions of user profiles (like Facebook, MySpace, Twitter, YouTube, etc.). Traditional techniques for discovering evidence of spam users often rely on costly human-in-the-loop inspection of training data for building spam classifiers; since spammers constantly adapt their strategies and tactics, the learned spam signatures can go stale quickly. An alternative spam discovery technique relies on community-contributed spam referrals (e.g., Users A, B, and C report that User X is a spam user); of course, these kinds of referral systems can be manipulated themselves to yield spam labels on legitimate users, thereby obscuring the labeling effectiveness. And neither spam discovery approach can effectively handle *zero-day* social spam attacks for which there is no existing signature or wide evidence.

With these challenges in mind, we propose and evaluate a novel honeypot-based approach for uncovering social spammers in online social systems. Concretely, the proposed approach is designed to (i) automatically harvest spam profiles from social networking communities, avoiding the drawbacks of burdensome human inspection; (ii) develop robust statistical user models for distinguishing between social spammers and legitimate users; and (iii) ac-

SIGIR'10, July 19–23, 2010, Geneva, Switzerland.

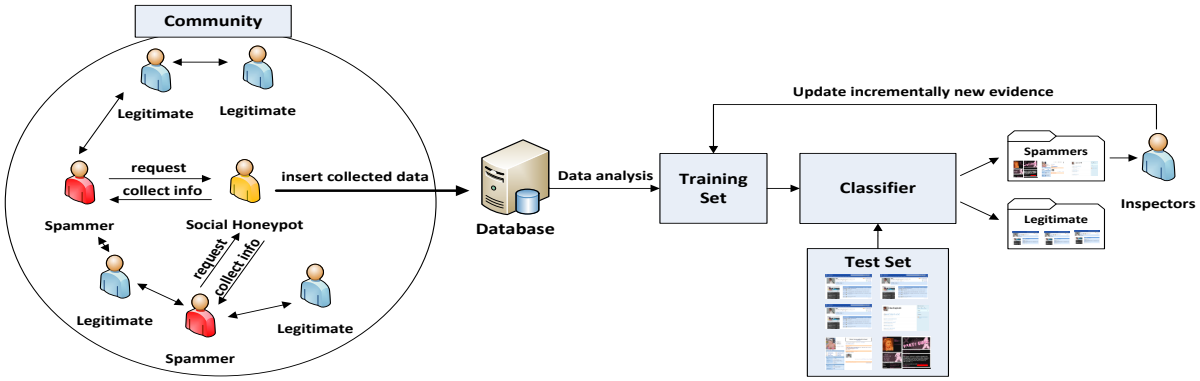


Figure 1: Overall Framework of Social Honeypot-based Approach

tively filter out unknown (including zero-day) spammers based on these user models. Drawing inspiration from security researchers who have used honeypots to observe and analyze malicious activity (e.g., for characterizing malicious hacker activity [22], generating intrusion detection signatures [16], and observing email address harvesters [20]), we deploy and maintain *social honeypots* for trapping evidence of spam profile behavior, so that users who are detected by the honeypot have a high likelihood of being a spammer (i.e., low false positive rate). Over two distinct communities (MySpace and Twitter), we find that the proposed approach provides generalizable and effective social spam detection.

2. OVERALL FRAMEWORK

In this section, we present the conceptual framework of the proposed honeypot-based approach and outline the research questions motivating our examination of this framework.

2.1 Problem Statement

In social networking communities like MySpace and Facebook, there are a set of k users $U = \{u_1, u_2, \ldots, u_k\}$. Each user u_i has a profile p_i. Profiles are self-describing pages that are created and controlled by a given user. For example, users typically include information such as their name, gender, age, and so on in their profiles. Each community has its own profile format, but most fields in the formats are the same.

The **social spam detection problem** is to predict whether u_i is a spammer through a classifier c when p_i is given. A classifier

$$c : u_i \rightarrow \{spammer, legitimate\ user\}$$

approximates whether u_i is a spammer. To build c, we need to extract a set of m features $F = \{f_1, f_2, \ldots, f_m\}$ from U. For example, we can extract F_{u_i} from p_i of u_i.

Whereas traditional email spam detection has focused on identifying *spam messages* which are of relatively low individual value to the spammer (and whose identification typically doesn't threaten the ongoing ability of a spammer to send new messages), social spam detection is focused on identifying and eliminating *spam accounts* themselves. This detection is potentially more disruptive to spammers, since these accounts typically represent a more expensive investment by the spammer (through email and social media account registrations).

2.2 Solution Approach

We propose to monitor spammer activity through the creation of social honeypots. We define social honeypots as information system resources that monitor spammers' behaviors and log their information (e.g., their profiles and other content created by them in social networking communities). Social honeypots and traditional honeypots (e.g., in domains such as network systems and emails [16, 20, 22]) share a similar purpose in that they both monitor and log the behaviors of spammers or attackers. However, traditional honeypots typically target network or systems-level behavior, whereas social honeypots specifically target community-based online activities.

While social honeypots alone are a potentially valuable tool for gathering evidence of social spam attacks and supporting a greater understanding of spam strategies, it is the goal of this research project to support ongoing and active *automatic* detection of new and emerging spammers (See Figure 1). In practice, we deploy a social honeypot consisting of a legitimate profile and an associated bot to detect social spam behavior. If the social honeypot detects suspicious user activity (e.g., the honeypot's profile receiving an unsolicited friend request) then the social honeypot's bot collects evidence of the spam candidate (e.g., by crawling the profile of the user sending the unsolicited friend request plus hyperlinks from the profile). What entails *suspicious user behavior* can be optimized for the particular community and updated based on new observations of spammer activity.

As the social honeypots collect spam evidence, we extract observable features from the collected candidate spam profiles (e.g., number of friends, text on the profile, age, etc.). Coupled with a set of known legitimate (non-spam) profiles which are more populous and easy to extract from social networking communities, these spam and legitimate profiles become part of the initial training set of a spam classifier. Through iterative refinement of the features selected and the classifier used (e.g., SVM), the spam classifier can be optimized over the known spam and legitimate profiles.

Based on these developed classifiers, we can then explore the wider space of unknown profiles. On MySpace alone there are 100s of millions of profiles, of which some unknown fraction are spam. Using the classifiers based on the harvested social honeypot data, we iteratively explore these profiles "in-the-wild" to detect new spammers that have yet to be identified by a social honeypot directly. In our design of the overall architecture, we include human inspectors in-the-loop for validating the quality of these extracted spam candidates. Instead of inspecting the entirety of all profiles, these inspectors are guided to validate just the few spam candidates recommended by the learned classifiers. Based on their feedback, the spam classifiers are updated with the new evidence and the process continues. Given the overall architecture, we address three important research challenges in turn in the rest of the paper:

- *Research Challenge #1 [RC1]:* Do social honeypots collect evidence of spam with low false positives? In other words, do honeypots really work in practice? A poorly performing social honeypot will negatively impact the spam classification approach, leading to poor performance.
- *Research Challenge #2 [RC2]:* Can we build classifiers from the harvested social honeypot profiles and known legitimate profiles to effectively identify spam profiles? Since social honeypots are triggered by suspicious user *activity*, we must explore if the harvested spam data contains signals that are strongly correlated with observable profile features (e.g., content, friend information, posting patterns, etc.). It is our hypothesis that spammers engage in behavior that is correlated with observable features that distinguish them from legitimate users.
- *Research Challenge #3 [RC3]:* Finally, can the developed classifiers be effectively deployed over large collections of unknown profiles (for which we have no assurances of the degree of spam or legitimate users)? This last question is important for understanding the promise of social honeypots in defending against new and emerging spam as it arises "in-the-wild."

3. RC1: STUDY OF HARVESTED SPAM USERS

Based on the overall social honeypot framework, we selected two social networking communities – Myspace and Twitter – to evaluate the effectiveness of the proposed spam defense mechanism. Both MySpace and Twitter are large and growing communities and both also support public access to their profiles, so all data collection can rely on purely public data capture.

MySpace Social Honeypot Deployment: In previous research [23], we created 51 generic honeypot profiles within the MySpace community for attracting spammer activity so that we can identify and analyze the characteristics of social spam profiles. To observe any geographic artifacts of spamming behavior, each profile was assigned a specific geographic location (i.e., one honeypot was assigned to each of the U.S. states and Washington, D.C.). Each honeypot profile tracks all unsolicited friend requests. Upon receiving a friend request, we store a local copy of the profile issuing the friend request, extract all hyperlinks in the "About Me" sections (we find that these typically point to an affiliate spam website), and crawl the pages pointed to by these hyperlinks. Based on a four month evaluation period (October 2007 to January 2008), we collected 1,570 profiles whose users sent unsolicited friend requests to these social honeypots.

Twitter Social Honeypot Deployment: Similarly, we created and deployed a mix of honeypots within the Twitter community – some of them had personal information such as biography, location and so on, while others did not have this personal information. Some social honeypots posted tweets, while others did not post them. We omit some of the concrete details of the Twitter honeypot deployment due to the space constraint. From August 2009 to September 2009, these social honeypots collected 500 users' data.

3.1 MySpace Observations

After analyzing the harvested spam profiles from MySpace, we find some interesting observations (more fully detailed in [23]): (1) The spamming behaviors of spam profiles follow distinct temporal patterns. (2) The most popular spamming targets are Midwestern states, and the most popular location for spam profiles is California. (3) 57.2% of the spam profiles copy their "About Me" content from another profile. (4) Many of the spam profiles exhibit distinct demographic characteristics (e.g., age, relationship status, etc.). (5) Spam profiles use thousands of URLs and various redirection techniques to funnel users to a handful of destination Web pages. Through manual inspection, we grouped the harvested spam profiles into five categories:

- *Click Traps:* Each profile contains a background image that is also a link to another Web page. If users click anywhere on the profile, they are directed to the link's corresponding Web site.
- *Friend Infiltrators:* These nominally legitimate profiles befriend as many users as possible so that they can infiltrate the users' circles of friends and bypass any communication restrictions imposed on non-friends. Once a user accepts a friend request from one of these profiles, the profile begins spamming the user through existing communication systems (e.g., message spam, comment spam, etc.).
- *Pornographic Storytellers:* Each of these profiles has an "About Me" section that consists of randomized pornographic stories, which are book-ended by links that lead to pornographic Web pages. The anchor text used in these profiles is extremely similar, even though the rest of the "About Me" text is almost completely randomized.
- *Japanese Pill Pushers:* These profiles contain a sales pitch for male enhancement pills in their "About Me" sections. According to the pitch, the attractive woman pictured in the profile has a boyfriend who purchased these pills at an incredible discount.
- *Winnies:* All of these profiles have the same headline: "Hey its winnie." However, despite this headline, none of the profiles are actually named "Winnie." In addition to a shared headline, each of the profiles also includes a link to a Web page where users can see the pictured female's pornographic pictures.

3.2 Twitter Observations

Similarly, we discovered various types of spam users in the harvested data from Twitter. In many cases, spammers inserted malicious or spam links into their tweets. Since most Twitter links use a form of URL-shortening, users clicking on these links have no assurances of the actual destination.

- *Duplicate Spammers:* These users post a series of nearly identical tweets. In many cases, the only different content from tweet to tweet is the inclusion of different @usernames (or @replies). The insertion of these @usernames essentially delivers the tweet to the *username*'s account even if the spammer has no relationship with the intended target.
- *Pornographic Spammers:* Their data such as user image, profile URL, and text and links in tweets contains adult content.
- *Promoters:* These users post tweets about several things such as online business, marketing and so on. Their posting approach is more sophisticated than duplicate spammers since spam tweets are randomly interspersed with seemingly innocuous legitimate tweets.
- *Phishers:* Similar to promoters, these users use a mix of strategies to deliver phishing URLs to targets on Twitter.
- *Friend Infiltrators:* Much like their counterparts on MySpace, these users have profiles and tweets that are seemingly legitimate. They follow many people and intend to accumulate many followers; then they begin engaging in spam activities like posting tweets containing pornographic or commercial content.

These observations indicate that social honeypots can successfully attract spammers across fundamentally different communities (MySpace vs. Twitter), and the results suggest that building automatic classifiers may be useful for identifying social spam.

4. RC2: EMPIRICAL EVALUATION OF SOCIAL SPAM SIGNALS

We next explore whether there are discernible spam signals in the harvested spam profiles that can be used to automatically distinguish spam profiles from legitimate profiles. Since social honeypots are triggered by spam *behaviors* only, it is unclear if the corresponding profiles engaging in the spam behavior also exhibit clearly observable spam signals. If there are clear patterns (as our observations in the previous section would seem to indicate), then by training a classifier on the observable signals, we may be able to predict new spam even in the absence of triggering spam behaviors.

4.1 Classification Approach and Metrics

As part of this empirical evaluation of social spam signals, we consider four broad classes of user attributes that are typically observable (unlike, say, private messaging between two users) in the social network: (i) user demographics: including age, gender, location, and other descriptive information about the user; (ii) user-contributed content: including "About Me" text, blog posts, comments posted on other user's profiles, tweets, etc.; (iii) user activity features: including posting rate, tweet frequency; (iv) user connections: including number of friends in the social network, followers, following. For MySpace and Twitter we select a subset of these features to train the classifier.

The classification experiments were performed using 10-fold cross-validation to improve the reliability of classifier evaluations. When a dataset is not large, it is common to use 10-fold cross-validation to achieve statistically precise results. In 10-fold cross-validation, the original sample is randomly divided into 10 equally-sized sub-samples. 9 sub-samples are used as a training set and the remaining one is used as a testing set; the classifier is evaluated, then the process is repeated for a total of 10 times. Each sub-sample is used as a testing set once in each evaluation. The final evaluation result is generated by averaging the results of the 10 evaluations. In practice, we evaluated over 60 different classifiers in the Weka [24] machine learning toolkit using 10-fold cross-validation with default values for all parameters. Classification results are presented in the form of a confusion matrix as in Table 1.

Table 1: Confusion matrix example.

		Predicted	
		Spammer	Legitimate
Actual	Spammer	a	b
	Legitimate	c	d

To measure the effectiveness of classifiers based on our proposed features, we used the standard metrics such as precision, recall, accuracy, the F_1 measure, false positive and true positive. Precision is the ratio of correctly predicted users as a class to the total predicted users as the class. For example, the precision (P) of the spammer class in Table 1 is $a/(a+c)$. Recall (R) is the ratio of correctly predicted users as a class to the actual users in the class. The recall of the spammer class in the table is $a/(a+b)$. Accuracy is the proportion of the total number of predictions that were correct. The accuracy in the table is $(a+d)/(a+b+c+d)$. F_1 is a measure that trades off precision versus recall. F_1 measure of the spammer class is $2PR/(P+R)$. A false positive is when the actual Y class users are incorrectly predicted as X class users. The false positive of the spammer class is c. A true positive is when actual X class users are correctly predicted as X class users. The true positive of the spammer class is a.

To measure the discrimination power between spammers and legitimate users of each of the proposed features, we generate a Receiver Operating Characteristics (ROC) curve. ROC curves plot false positive rate on the X axis and true positive rate on the Y axis. The closer the ROC curve is to the upper left corner, the higher the overall accuracy is. The ideal ROC curve includes the coordinate (0, 1), indicating no false positives and a 100% true positive rate.

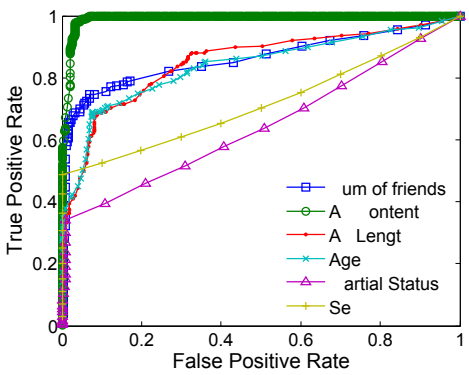


Figure 2: MySpace – Feature Comparison

4.2 MySpace Spam Classification

We randomly sampled 388 legitimate profiles from MySpace (which were labeled by us) and 627 deceptive spam profiles from the 1,570 deceptive spam profiles collected by our social honeypots. When we sampled the profiles, we considered several conditions. Profiles have to be public, and marital status, gender, age, and "About Me" content in the profiles have to be valid (i.e., a non-empty value). In addition, we removed duplicated profiles among the 1,570 deceptive spam profiles in the case that a spammer sent a friend request to several social honeypots. The goal of spam classification over the MySpace data is to predict whether a profile is either spammer or legitimate.

We considered several representative user features: number of friends, age, marital status, gender, as well as some text-based features modeling user-contributed content in the "About Me" section. Specifically, we consider a bag-of-words model in which we remove punctuation, make all letters lowercase, tokenize each word, remove stopwords, and do stemming for each word using the Porter stemmer [19]. We assigned weights to each word based on tf-idf weighting: $\text{tf-idf}_{t,d} = \log(1+tf_{t,d}) \times \log(\frac{N}{df_t})$, where $\text{tf}_{t,d}$ means term frequency of term t in a profile's "About Me", N is the number of profiles, and df_t is the number of profiles which includes term t. We also measured the length in bytes of the "About Me" content.

Before evaluating the effectiveness of our classifiers, we investigated the discrimination power of our individual classification features. Recognizing that social spam classification is an example of an adversarial classification problem [11], we wanted to evaluate the robustness of our features against an adversarial attack. To show the discrimination power and robustness of each feature, we generated ROC curves for each feature using an AdaboostM1 classifier. The results are shown in Figure 2. Marital status and sex are the least discriminative features, which is unsurprising because they are represented by a small set of predefined values (e.g., "Married", "Single", "Male", etc.) that will inevitably appear in both legitimate and spam profiles. On the other hand, the bag of words features extracted from "About Me" content (AMContent) are the most discriminative. This is a very encouraging result because it means our classifier was able to distinguish between legitimate and

Table 2: Performance results of top 10 classifiers

Weka Classifier	Accuracy	F_1	FP
Decorate	**99.21%**	**0.992**	**0.7%**
SimpleLogistic	99.01%	0.99	0.9%
FT	99.01%	0.99	0.9%
LogitBoost	99.01%	0.99	1.3%
RandomSubSpace	98.72%	0.987	1.1%
Bagging	98.62%	0.986	1.2%
J48	98.42%	0.984	1.5%
OrdinalClassClassifier	98.42%	0.984	1.5%
ClassBalancedND	98.42%	0.984	1.5%
DataNearBalancedND	98.42%	0.984	1.5%

spam "About Me" content with a high degree of accuracy. Therefore, if spammers begin varying the other features of their profiles (to appear more legitimate), our classifiers will still be effective. Additionally, the "About Me" content is the most difficult feature for a spammer to vary because it contains the actual sales pitch or deceptive content that is meant to target legitimate users.

In Table 2, the performance results for the top 10 classifiers are shown. The table clearly shows that all of the classifiers were successful. Each classifier generated an accuracy greater than 98.4%, an F_1 measure over 0.98, and a false positive rate below 1.6%. Overall, meta-classifiers (Decorate, LogitBoost, etc.) performed better than tree classifiers (FT and J48) and a function-based classifier (SimpleLogistic). The best classifier is Decorate, which is a meta-learner for building diverse ensembles of classifiers. It obtained an accuracy of 99.21%, an F_1 measure of 0.992, and a 0.7% false positive rate. We additionally considered different training mixtures of spam and legitimate training data (from 10% spam / 90% legitimate to 90% spam / 10% legitimate); we find that the metrics are robust across these changes in training data.

4.3 Twitter Spam Classification

To evaluate the quality of spam classification over Twitter, we randomly selected 104 legitimate users (labeled by us) from a previously collected Twitter dataset of 210,000 users. We additionally considered two classes of spam users: the 61 spammers and the 107 promoters sampled from 500 users' data collected by the social honeypots. For each user, we collected the user profile, tweets (status update messages), following (friend) information and followers' information. The goal of spam classification over the Twitter data is to predict whether a profile is either spammer, a promoter, or legitimate. When we sampled users' data, we considered two conditions: the profiles did not have a *verified account badge* and the number of tweets had to be over 0. The *verified account badge* is one way Twitter ensures that profiles belong to known people (e.g, Shaquille O'Neal and not an impersonator).

Unlike MySpace profiles which emphasize on longer-form personal information sharing (e.g., "About Me" text) and usually have self-reported user demographics (e.g., age, gender), Twitter accounts are noted for their short posts, activity-related features, and limited self-reported user demographics. For user features, we consider the longevity of the account on Twitter, the average tweets per day, the ratio of the number of following and number of followers, the percentage of bidirectional friends ($\frac{|following \cap followers|}{|following|}$), as well as some features of the tweets sent, including:

- The ratio of the number of URLs in the 20 most recently posted tweets to the number of tweets ($|URLs|/|tweets|$).
- The ratio of the number of *unique* URLs in the 20 most recently posted tweets to the number of tweets ($|unique\ URLs|/|tweets|$).

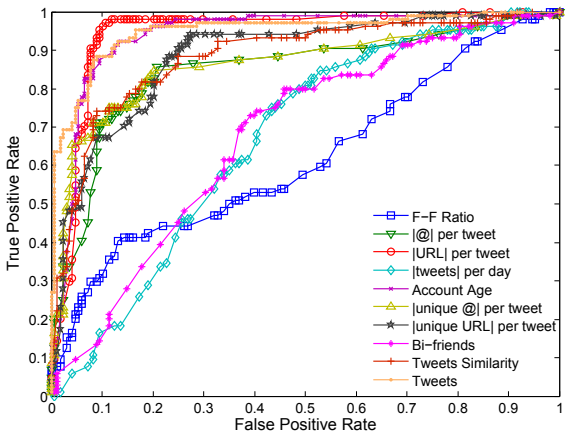


Figure 3: Twitter – Feature Comparison

- The ratio of the number of @usernames in the 20 most recently posted tweets to the number of tweets ($|@username|/|tweets|$).
- The ratio of the number of *unique* @usernames in the 20 most recently posted tweets to the number of tweets ($|unique\ @username|/|tweets|$).

Additionally, we measure the average content similarity over all pairs of tweets posted by a user:

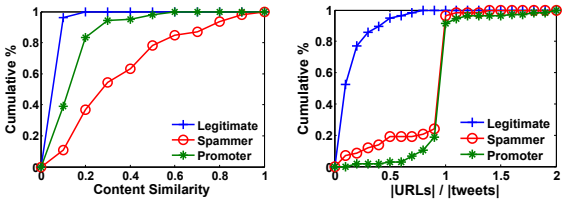
$$\sum_{a,b \in \text{set of pairs in tweets}} \frac{similarity(a,b)}{|\text{set of pairs in tweets}|}$$

where the content similarity is computed using the standard cosine similarity over the bag-of-words vector representation $\overrightarrow{V}(a)$ of the tweet content:

$$similarity(a,b) = \frac{\overrightarrow{V}(a) \cdot \overrightarrow{V}(b)}{|\overrightarrow{V}(a)||\overrightarrow{V}(b)|}$$

We finally considered some text-based features to model the content in the tweets. Since tweets are extremely short (140 characters or less), we consider a bag-of-words model and a sparse bigrams model [9]. We remove punctuation, make all letters lowercase, tokenize each word in the bag-of-words model and tokenize a pair of words in the sparse bigrams model. The sparse bigrams model generates a pair of words separated by no more than k words. We assigned k = 3 in our system, while k = 0 yields ordinary bigrams. If a tweet is "check adult page view models", the sparse bigrams will generate the features "check adult", "check page", "check view", "adult page", "adult view", "adult models", "page view", "page models", "view models". We weighted terms and bigrams using tf-idf weighting as in the previous MySpace classification.

In order to know how much discrimination power each feature has for spammers, promoters and legitimate users, we generated ROC curves of the proposed features using Decorate in Figure 3. Average posted tweets per day ($|tweets|\ per\ day$), percentage of bidirectional friends (*bi-friends*), and ratio of number of following and number of followers (*F-F Ratio*) have low discrimination powers relatively, while ratio of number of unique URLs in recently posted top 20 tweets and number of the tweets ($|unique\ URL|\ per\ tweet$), ratio of number of @username in recently posted top 20 tweets and number of the tweets ($|@|\ per\ tweet$), ratio of number of unique @username in recently posted top 20 tweets and number of the tweets ($|unique\ @|\ per\ tweet$), and av-



(a) Content Similarity (b) Avg URLs per Tweet

Figure 4: Cumulative Distribution of Features Extracted from Users

Table 3: Performance results of top 10 classifiers

Weka Classifier	Accuracy	F_1	FP
Decorate	**88.98%**	**0.888**	**5.7%**
LogitBoost	87.86%	0.877	6.2%
HyperPipes	85.29%	0.846	8.1%
Bagging	84.56%	0.844	7.5%
RandomSubSpace	84.19%	0.837	8.3%
BFTree	83.82%	0.84	7.2%
FT	83.46%	0.832	8.3%
SimpleLogistic	83.46%	0.832	8.5%
LibSVM (SVM)	83.09%	0.825	10.2%
ClassificationViaRegression	82.72%	0.823	9.1%

erage content similarity between a user's tweets (*tweets similarity*) have good discrimination powers. Account age, text-based features extracted from tweets, and ratio of number of URLs in recently posted top 20 tweets and number of the tweets ($|URL|$ per $tweet$) have the best discrimination power. Overall, the proposed all features have positive discrimination power.

To further illustrate, Figure 4(a) presents the cumulative distributions of content similarity in tweets posted by each user class. It shows clear distinction among legitimate users, spammers and promoters. The content similarity in tweets of each spammer is the largest compared to the other classes because some of them post almost the same content or even duplicate tweets. Promoters have a goal of promoting something like online business, marketing and so on; naturally their tweets include common terms like the name of a product. Therefore, the content similarity in their tweets is larger than legitimate users' one because legitimate users post tweets about their news such as what they are doing, where they are and so on. The content similarity in tweets of legitimate users is the smallest. Figure 4(b) shows the cumulative distributions of the average number of URLs in the tweets of each user. Tweets posted by legitimate users include the smallest number of URLs; not surprisingly, the majority of spammers and promoters post tweets with URLs. The curves of spammers and promoters are overlapped near 1 in the X axis, meaning that promotor and spammer behavior is closely coupled in our dataset.

Table 3 shows the performance results for the top 10 classifiers. Each of the top 10 classifiers achieved an accuracy greater than 82.7%, an F_1 measure over 0.82, and a false positive rate less than 10.3%. As in the case with MySpace, the meta classifiers (Decorate, LogitBoost, etc.) produced better performance than tree classifiers (BFTree and FT) and function-based classifiers (SimpleLogistic and LibSVM). The best classifier was Decorate, which obtained an accuracy of 88.98% accuracy, an F_1 measure of 0.888, and a 5.7% false positive rate. As in the MySpace analysis, we additionally considered different training mixtures of spam and legitimate training data (from 10% spam / 90% legitimate to 90% spam / 10% legitimate); we find that the classification metrics are robust across these changes in training data.

Table 4: Statistics of MySpace dataset

Public Profiles	Private Profiles	Total Profiles	Size
1,576,684	274,988	1,851,672	150GB

Table 5: Statistics of Twitter dataset

User Profiles	Tweets	Following	Followers	Size
215,345	4,040,415	51,650,754	65,904,253	11.3GB

4.4 Summary

Based on our empirical study over both MySpace and Twitter, we find strong evidence that social honeypots can attract spam behaviors that are strongly correlated with observable features of the spammer's profiles and their activity in the network (e.g., tweet frequency). These results hold across two fundamentally different communities and confirm the hypothesis that spammers engage in behavior that is correlated with observable features that distinguish them from legitimate users. In addition, we find that some of these signals may be difficult for spammers to obscure (e.g., content containing a sales pitch or deceptive content), so that the results are encouraging for ongoing effective spam detection.

5. RC3: LARGE-SCALE SOCIAL SPAM CLASSIFICATION IN THE WILD

So far, we have seen that the deployed social honeypots can collect evidence of spam behavior, and that these behaviors are correlated with spam signals which can support automatic spam classification. In this final study, we explore whether these classifiers can be effectively deployed over large collections of unknown profiles (for which we have no assurances of the degree of spam or legitimate users). Concretely, we apply the developed classifiers for both MySpace and Twitter over datasets "in-the-wild" to better understand the promise of social honeypots in defending against new and emerging spam and zero-day spam attacks.

5.1 Data and Setup

For this final study, we considered two large datasets.

MySpace Dataset: The first dataset is a crawl over MySpace, including about 1.5 million of public profiles collected in 2006 and 2007. A full description of this dataset and its characteristics is available in [8]. Table 4 summarizes statistics of this dataset.

Twitter Dataset: We also collected a large dataset from Twitter for the period September 2 to September 9, 2009. We sampled the public timeline of Twitter (which publishes a random selection of new tweets every minute), collected usernames, and then used the Twitter API to collect each user's recently posted top 20 tweets, plus the user's following (friends) and followers' information. Table 5 presents statistics of this Twitter dataset. It consists of 215,345 user profiles, 4,040,415 tweets.

In both cases, the collected profiles are unseen to our system, meaning that we do not know ground truth as to whether a profile is spam or legitimate. As a result, the traditional classification metrics presented in the previous section would be infeasible to apply in this case. Rather than hand label millions of profiles, we adopted the spam precision metric to evaluate the quality of spam predictions. For spam precision, we evaluate only the predicted spammers (i.e., the profiles that the classifier labels as spam). Spam precision is defined as:

$$SpamPrecision = \frac{true\ positives}{true\ positives + false\ positives}$$

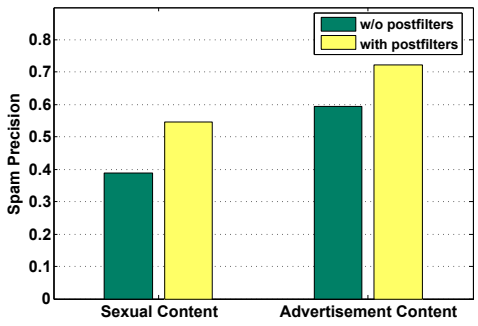


Figure 5: MySpace – Spam Precision

5.2 Finding Unknown Spam on MySpace

As in the previous section, we trained a classifier over a training set consisting of 388 legitimate profiles (labeled by us) and 627 deceptive spam profiles collected from social honeypots. In the interests of efficiency, we used the LibSVM classifier – an implementation of support vector machines – which is a widely popular classifier and classifies a large dataset quickly with high accuracy. Its classification time is faster than meta classifiers that proved successful in the previous experiments. We sampled from the 1.5 million public profiles a smaller test set of 43,000 profiles. We repeated this sampling procedure four times so we had four different test sets.

As we classified each of four test sets, a human inspector verified whether the newly found predicted spam was actually spam, added the new instances to the training set, and the process continued. In each test set, LibSVM classifier predicted about 30 users as spammers. In each subsequent iteration, we found that the spam precision increased.

Figure 5 shows the evaluation results of the fourth test set. The left two bars of the figure present spam precision based on sexual content. If an unseen profile is classified to a deceptive spam profile by LibSVM, and its "About Me" content includes sexual content, it will be considered as a real deceptive spam profile. The right two bars of the figure present spam precision based on advertisement content. If predicted spam profile's "About Me" content includes advertisement content, it will be considered as a real deceptive spam profile. Note that there are two results: with postfilter and without postfilter. We found that LibSVM incorrectly predicted spam labels for profiles containing special profile layout links, e.g., "click here to get a theme for your myspace" or "click here for myspace backgrounds", which are similar to spammer techniques for inserting links into spam profiles. These types of profile layout links are common on MySpace, which allows users to adjust their profile layouts. To correct these errors, we inserted a "postfilter" to check for these standard links and remove their spam labels.

As we can see, using postfilters improved about 40% spam precision in sexual content and about 21% in advertisement content. Detecting spammers whose profiles include advertisement content is easier than detecting spammers whose profile include sexual content. Even with the fairly good results (70% spam precision), the results are significantly worse than what we observed in the previous section over the controlled dataset. We attribute this decline in performance to the time-based mismatch between the harvested social honeypots and the large MySpace dataset. The honeypots were deployed in 2007, but the large MySpace data was crawled in 2006. As a result, the spam signatures developed from the honeypots have difficulty identifying spam in an earlier dataset when those spam signature may have not been in use at all. Even with these challenges, the results are fairly good.

Table 6: Example of "About Me" content in new deceptive spam profiles

"About Me" content
I moved to san diego 3 months ago with my boyfriend, well, ex-boyfriend now . . . one thing i did find is this webcam site. it pays pretty decent and the best part is that its really fun, too . . . , click here to visit me.

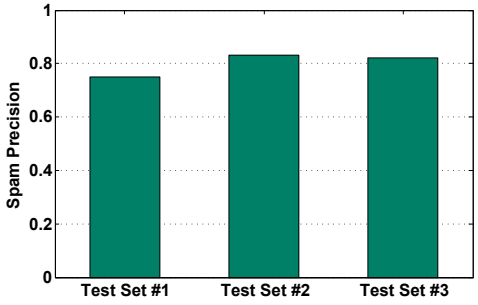


Figure 6: Twitter – Spam Precision

As an example, Table 6 illustrates a legitimate-appearing profile in the first part of the "About Me" content, but then inserts a URL which links to an external site (usually porn or sexual sites) in the middle part or the last part of the "About Me" content.

5.3 Finding Unknown Spam on Twitter

Unlike the MySpace data mismatch, the social honeypots deployed on Twitter pre-date the large Twitter dataset collection. Hence, we are interested in this final experiment to discover if these honeypots can effectively detect new and emerging spam. For Twitter classification, we again relied on the training set consisting of 104 legitimate users' data, 61 spammers' data and 107 promoters' data which were used in Section 4. For prediction, we considered two cases: legitimate or spam+promoter. We randomly selected a test set of 1,000 users' data from the total dataset of about 210,000 users. We repeated this process three times, resulting in three test sets. The feature generator produces the same features used in the previous section. We selected Decorate as a classifier because it showed the best performance in the previous Twitter study. Human inspectors view spam data predicted by the classifier, and then decide whether or not they are real spam data. They will add newly found spam data with labels (spam or non-spam) to the training set in order to iteratively improve the classifier's accuracy.

Figure 6 presents spam precision results obtained from the three test sets. In each test set, the Decorate classifier predicted about 20 users as spammers. We assessed whether the predicted spammers were real spammers. In the first iteration, spam precision was 0.75, nearly matching the performance of the controlled classifier reported in the previous section. By the third iteration, the spam precision was 0.82. We see in this experiment how the social honeypots provide strong ability to discover unknown spam; and as these newly discovered spammers are added to the training set, the classifier becomes more robust (resulting in the improvement from the first to the third iteration).

As an example, Figure 7 shows an example of newly found spammer. The spammer's tweets include URLs which link to sex search tool pages. It is interesting that the spammer has 205 followers, meaning that this spammer has successfully inserted himself into the social network without detection. We additionally found that

Figure 7: An example of newly found spammer on Twitter

about 20% of the users predicted to be spammers were *bots* that post tweets automatically using the Twitter API.

Based on this large-scale evaluation of spam "in-the-wild", we can see how social honeypots can enable effective social spam detection of previously unknown spam instances. Since spammers are constantly adapting, these initial results provide positive evidence of the robustness of the proposed approach.

6. CONCLUSIONS AND NEXT STEPS

In this paper, we have presented the design and real-world evaluation of a novel social honeypot-based approach to social spam detection. Our overall research goal is to investigate techniques and develop effective tools for automatically detecting and filtering spammers who target social systems. By focusing on two different communities, we have seen how the general principles of (i) social honeypot deployment, (ii) robust spam profile generation, and (iii) adaptive and ongoing spam detection can effectively harvest spam profiles and support the automatic generation of spam signatures for detecting new and unknown spam. Our empirical evaluation over both MySpace and Twitter has demonstrated the effectiveness and adaptability of the honeypot-based approach to social spam detection.

In our continuing research, we are interested to explore a number of extensions. First, to what degree can traditional email and web spam approaches be incorporated into the social honeypot framework? For example, email spam and phishing approaches relying on data compression algorithms [6], machine learning [15, 21] and statistics [25] could inform the further refinement of the proposed approach. Similarly, web spam approaches have extensively studied the link structure of the web (e.g., [2, 3]); adapting these link-based approaches to the inherent social connectivity of online communities could further improve social spam effectiveness.

Second, we are interested to explore how social honeypots can be augmented by other recent approaches to deal with spam in social systems, including Heymann et al. [13] and Benevenuto et al. [4]. These prior approaches have focused on particular communities (e.g., social tagging systems, online video sharing sites); in what ways can their domain-specific techniques be incorporated into the social honeypot approach?

Finally, we are interested to expand and refine the social honeypots. For example, it may be worthwhile to both scale up the number of social honeypots (say, to the 1000s) and to consider more variation in the demographics and behaviors of the social honeypot profiles (say, by constructing clique-based social honeypots to measure whether honeypots that are more "connected" induce more spammer activity than "loner" honeypots.). Similarly, we are interested to diversify the domains in which we deploy the honeypots (while respecting the terms of service of each community). Do we find that spammers engage in similar behaviors across domains? And if so, perhaps we can use this cross-domain spammer correlation to further improve the effectiveness of social spam detection.

7. ACKNOWLEDGMENTS

This work is partially supported by a Google Research Award and by faculty startup funds from Texas A&M University and the Texas Engineering Experiment Station.

8. REFERENCES

[1] L. A. Adamic, J. Zhang, E. Bakshy, and M. S. Ackerman. Knowledge sharing and yahoo answers: everyone knows something. In *WWW*, 2008.

[2] L. Becchetti, C. Castillo, D. Donato, S. Leonardi, and R. Baeza-Yates. Link-based characterization and detection of web spam. In *SIGIR Workshop on Adversarial Information Retrieval on the Web*, 2006.

[3] A. A. Benczur, K. Csalogany, and T. Sarlos. Link-based similarity search to fight web spam. In *SIGIR Workshop on Adversarial Information Retrieval on the Web*, 2006.

[4] F. Benevenuto, T. Rodrigues, V. Almeida, J. Almeida, and M. Gonçalves. Detecting spammers and content promoters in online video social networks. In *SIGIR*, 2009.

[5] D. Boyd and J. Heer. Profiles as conversation: Networked identity performance on friendster. In *HICSS*, 2006.

[6] A. Bratko, B. Filipič, G. V. Cormack, T. R. Lynam, and B. Zupan. Spam filtering using statistical data compression models. *J. Mach. Learn. Res.*, 7:2673–2698, 2006.

[7] G. Brown, T. Howe, M. Ihbe, A. Prakash, and K. Borders. Social networks and context-aware spam. In *CSCW*, 2008.

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

[9] G. V. Cormack. Email spam filtering: A systematic review. *Found. Trends Inf. Retr.*, 1(4):335–455, 2007.

[10] D. J. Crandall, L. Backstrom, D. Huttenlocher, and J. Kleinberg. Mapping the world's photos. In *WWW*, 2009.

[11] N. Dalvi, P. Domingos, Mausam, S. Sanghai, and D. Verma. Adversarial classification. In *SIGKDD*, 2004.

[12] A. Felt and D. Evans. Privacy protection for social networking platforms. In *Workshop on Web 2.0 Security and Privacy*, 2008.

[13] P. Heymann, G. Koutrika, and H. Garcia-Molina. Fighting spam on social web sites: A survey of approaches and future challenges. *IEEE Internet Computing*, 11(6):36–45, 2007.

[14] T. N. Jagatic, N. A. Johnson, M. Jakobsson, and F. Menczer. Social phishing. *Commun. ACM*, 50(10):94–100, 2007.

[15] D. H. Joshua Goodman and R. Rounthwaite. Stopping spam. *Scientific American*, 292(4):42–88, April 2005.

[16] C. Kreibich and J. Crowcroft. Honeycomb: creating intrusion detection signatures using honeypots. *SIGCOMM Comput. Commun. Rev.*, 34(1):51–56, 2004.

[17] Y.-R. Lin, H. Sundaram, Y. Chi, J. Tatemura, and B. L. Tseng. Splog detection using self-similarity analysis on blog temporal dynamics. In *WWW Workshop on Adversarial Information Retrieval on the Web*, 2007.

[18] A. Nazir, S. Raza, and C.-N. Chuah. Unveiling facebook: a measurement study of social network based applications. In *SIGCOMM*, 2008.

[19] M. F. Porter. An algorithm for suffix stripping. *Program*, 14(3):130–137, 1980.

[20] M. B. Prince, B. M. Dahl, L. Holloway, A. M. Keller, and E. Langheinrich. Understanding how spammers steal your e-mail address: An analysis of the first six months of data from project honey pot. In *the Conference on Email and Anti-Spam (CEAS)*, 2005.

[21] M. Sahami, S. Dumais, D. Heckerman, and E. Horvitz. A bayesian approach to filtering junk E-mail. In *ICML Workshop on Learning for Text Categorization*, 1998.

[22] L. Spitzner. The honeynet project: Trapping the hackers. *IEEE Security and Privacy*, 1(2):15–23, 2003.

[23] S. Webb, J. Caverlee, and C. Pu. Social honeypots: Making friends with a spammer near you. In *the Conference on Email and Anti-Spam (CEAS)*, 2008.

[24] I. H. Witten and E. Frank. *Data Mining: Practical Machine Learning Tools and Techniques (Second Edition)*. Morgan Kaufmann, June 2005.

[25] K. Yoshida, F. Adachi, T. Washio, H. Motoda, T. Homma, A. Nakashima, H. Fujikawa, and K. Yamazaki. Density-based spam detector. In *SIGKDD*, 2004.

[26] A. Zinman and J. S. Donath. Is britney spears spam? In *the Conference on Email and Anti-Spam (CEAS)*, 2007.