The Social Honeypot Project: Protecting Online Communities from Spammers*

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

We present the conceptual framework of the Social Honeypot Project for uncovering social spammers who target online communities and initial empirical results from Twitter and MySpace. Two of the key components of the Social Honeypot Project 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.

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. OVERALL FRAMEWORK

Spammers are increasingly targeting Web-based social systems (like Facebook, MySpace, YouTube, etc.) as part of phishing attacks, to disseminate malware and commercial spam messages, and to promote affiliate websites. 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.

In our ongoing research, we are developing approaches for uncovering and investigating social spammers through a prototype system called the Social Honeypot Project. Concretely, the Social Honeypot Project is designed to (i) automatically harvest spam profiles from social networking communities; (ii) develop robust statistical user models for distinguishing between social spammers and legitimate users; and (iii) actively filter out unknown spammers based on these user models. Drawing inspiration from security researchers who have used honeypots to observe and analyze malicious activity (e.g., [1]), the Social Honeypot Project deploys and maintains *social honeypots* for trapping evidence of spam profile behavior. 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 to pages on the Web-at-large). What entails *suspicious user behavior* can be optimized for the particular community and updated based on new observations of spammer activity.

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 the Social Honeypot Project to support ongoing and active *automatic* detection of new and emerging spammers (See Figure 1). 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 particular classifier used (e.g., Naive Bayes, SVM), the spam classifier can be optimized over the known spam and legitimate profiles. In our design of the overall architecture of the Social Honeypot Project we include human inspectors in-the-loop for validating the quality of these extracted spam candidates.

2. SOCIAL SPAM DETECTION RESULTS

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 support public access to their profiles, so all data collection can rely on purely public data capture.

MySpace Social Honeypot Deployment: 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 (fully described in [2]). Based on a four month evaluation period (October 2007 to January 2008), we collected 1,570 profiles that sent unsolicited friend requests to the honeypots.

*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.

WWW 2010, April 26–30, 2010, Raleigh, North Carolina, USA.
ACM 978-1-60558-799-8/10/04.

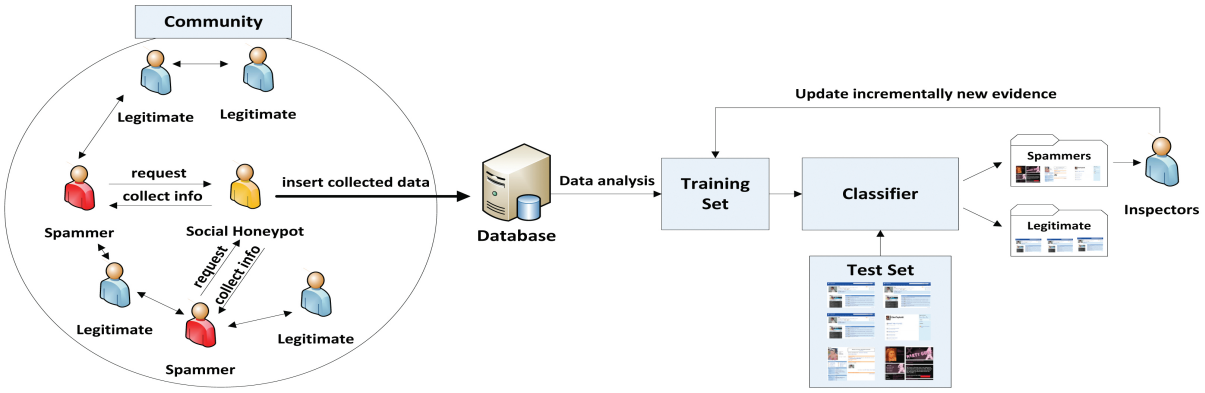

Figure 1: The Social Honeypot Project: Overall Framework

Twitter Social Honeypot Deployment: Similarly, we created and deployed a mix of honeypots within the Twitter community to track unsolicited "followers." From August 2009 to September 2009, these social honeypots collected 500 users' data.

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, 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. 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. The classification experiments were performed in the Weka [3] using 10-fold cross-validation to improve the reliability of classifier evaluations and evaluated using standard metrics such as precision, recall, accuracy, the F_1 measure, false positive and true positive.

Table 1: Spam Classification Results

	Accuracy	F_1	FP
MySpace	99.21%	0.992	0.7%
Twitter	88.98%	0.888	5.7%

In Table 1, we report the results for spam classification over both MySpace and Twitter.[1] 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. We additionally find that some features are stronger spam predictors than others (see Figure 2 for ROC curves for the Twitter dataset); in this case features like tweets per day and account age are not strong spam signals, whereas number of URLs per tweet and inter-tweet similarity are strong signals.

[1]Additional experimental details available from `http://infolab.tamu.edu/projects/social_honeypots/`

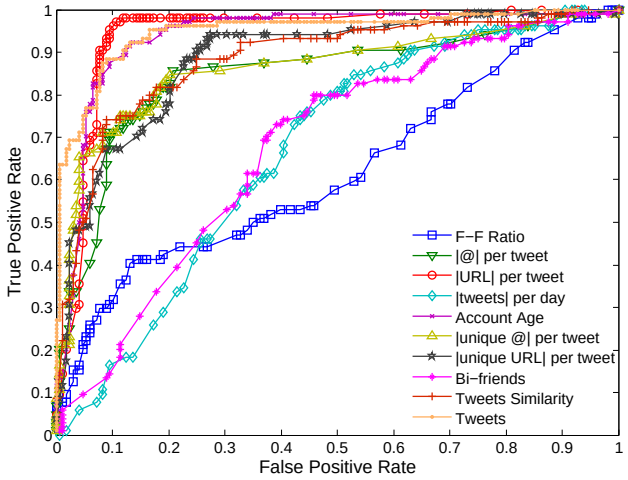

Figure 2: Twitter – Feature Comparison

3. CONCLUSION

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.

4. REFERENCES

[1] L. Spitzner. *Honeypots: Tracking Hackers.* Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 2002.

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

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