Detecting Spam on Social Networking Sites: Related Work

Antonio Lupher, Cliff Engle, Reynold Xin

{alupher, cengle, rxin}@cs.berkeley.edu

1. RELATED WORK

The rise of social media has made Social Networking Services (SNSs) more attractive targets for spam and fraud, leading to increasingly sophisticated attacks. This trend is reflected in recent research, as papers have focused on identifying and classifying the various types of social media spam. Many of these studies employ techniques previously used to combat conventional email and web spam. SNSs also provide opportunities to take advantage of user reputation and other social graph-dependent features to improve classification. Nevertheless, most research has been carried out on publicly-available data from SNSs, making it difficult up until now to measure the effect of private user data on algorithms for detecting site misuse.

1.1 Social Spam Features

Heymann et al. [9] survey the field of spam on SNSs, identifying several common approaches. Identification-based approaches identify spam to train classifiers based on labels submitted by users or trusted moderators. Rank-based approaches demote visibility of questionable content, while interface-based approaches apply policies to prevent unwanted behavior. This work groups classification-based approaches with detection, although classifiers can be used in conjuction with user information to prevent spam before it happens.

A number of researchers have focused on collecting, identifying features and classifying various genres of spam on social networks. Zinman and Donath [29] extract bundles of profile-based and comment-based features from MySpace profiles, but the relatively poor performance of their classifier highlights the difficulties in manual classification social network spam. Several studies take the approach of baiting spammers with social "honeypots", profiles created with the sole intent of attracting spam.[22, 15] They then use the data collected to train classifiers with features including friend request rate and ratios of URLs to text. Webb et al. [25] use the honeypot approach as well and provide examples of various types of spammers, the typical demographics of their profiles as well as the web pages that they tend to advertise.

Gao et al. [6] look at Facebook wall posts, analyzing temporal properties, URL characteristics, post ratios and other features of malicious accounts. They also pinpoint various spam "campaigns" based on products advertised in a given time frame. They note that spam on Facebook often exhibits burstiness and is mainly sent from compromised accounts. Benevenuto et al. [1] identify attributes of spam on video SNSs and use a Support Vector Machine (SVM) for classification.

Not all undesirable content on SNSs is necessarily spam or a scam. SNSs and online communities witness inappropriate user behavior, where users post offensive and harassing content. Yin et al. [27] combine sentiment analysis and profanity word lists with contextual features to identify harassment on datasets from Slashdot and MySpace. Other work looks at SNSs as platforms to collect data about users in order to aid direct attacks on the user's computers or to compromise a large number of accounts. [20, 10]

1.2 Social Spam Detection Systems

SocialSpamGuard [13] is a social media spam detection system that analyzes text and image features of social media posts. The demo system uses GAD clustering [12] for sampling spam and ham posts, then trains a classifier with text and image features. However, the system is built on top of Facebook features that are publicly accessible and thus cannot make use of sensitive user data (*e.g.,* IP addresses) to increase its effectiveness.

De Wang et al. [24] propose a cross-site spam detection framework to share spam data across all social networking sites, building classifiers to identify spam in profiles, messages and web pages. This multi-pronged approach lends itself to associative classification, in which, for example, a message would be classified as spam if it contained a link to a web page that had a high probability of being spam. Unfortunately, the differing characterestics of various social networks *e.g.,* the length of messages in Facebook vs. Twitter, can reduce the benefits of sharing spam corpora across diverse sites.

Facebook [21] provides an overview of their "immune system" defences against phishing, fraud and spam. The system is composed of classifier services, an ML-derived Fea-

ture Extraction Language (FXL), feature loops to aggregate and prepare features for classification and a policy engine to take action on suspected misuse. While the discussion remains high-level and includes few implemention particulars, it does include significant detail on the various types and characterists of undesirable activity on the site, including fake profiles, harassment, compromised accounts, malware and spam.

In contrast to research that focuses on dynamically detecting spam based on user activity, Irani et al. [11] show that static features associated with user signups on MySpace are enough to train an effective social spam classifier. They note that C4.5 decision tree algorithms provide better performance than naive Bayes in this case. As in other works, this only examines publicly available profile information collected by social honeypots. Private data collected on users including browser features, IP addresses and geographic location would conceivably improve classifier performance substantially.

Bosma et al. [4] explore user-generated spam reports as a tool for building an unsupervised spam detection framework for SNSs. Their approach counts the number of spam reports against a suspected spammer and adds weight to reports based on user reputation. Determining reputation and trustworthiness of users in social networks has been well studied [2, 7, 28] and appears to be a promising addition to social spam classification. The framework uses a Bayesian classifier and links messages with similar content, but does not take into account other features. Nevertheless, this is one of the few studies to test its framework on non-public data, inlcuding private messages, spam reports and user profiles from a large Dutch social networking site.

1.3 Spam Email & Web

Much work has been done on protecting traditional email systems from spam. Blanzieri [3] offers a comprehensive overview of machine learning techniques that can be applied to email filtering. Hao et al. [8] describe a reputation engine based on lightweight features such as geographic distance between sender and receiver, geolocation anomalies and diurnal patterns. While the target was conventional spam, these and similar features are applicable to spam on SNSs as well.

Whittaker et al. [26] describe a scalable phishing machine learning classifier and blacklisting system with high accuracy. Since a considerable amount of social media spam includes links to phishing sites, being able to detect them is critical. Along similar lines, Monarch [23] is a system that provides scalable real-time detection of URLs that point to spam web pages as determined by URL features, page content and hosting properties of the target domain.

Blog comment spam have also attracted considerable attention from researchers who have applied machine learning [14, 19] and NLP [18] techniques to the problem. Likewise, Markines et al. [17] apply similar techniques to spam on social bookmarking sites.

1.4 Machine Learning and Data Mining

Many of the data mining algorithms used to detect spam and patterns of misuse on SNSs are designed with the assumption that the data and the classifier are independent. However, in the case of spam, fraud and other malicious content, users will often modify their behavior to evade detection, leading to degraded classifier performance and the need to re-train classifiers frequently. Several researchers tackle this adversarial problem. Dalvi et al. [5] offer a modified Naive Bayes classifier to detect and reclassify data taking into account the optimal modification strategy that an adversary could choose. Lowd and Meek [16] provide a framework for reverse engineering a classifier to determine whether an adversary can efficiently learn enough about a classifier to effectively defeat it.

2. REFERENCES

[1] F. Benevenuto, T. Rodrigues, V. A. F. Almeida, J. M. Almeida, and M. A. Gonçalves. Detecting spammers and content promoters in online video social networks. In *SIGIR*, pages 620–627, 2009.

[2] J. Bian, Y. Liu, D. Zhou, E. Agichtein, and H. Zha. Learning to recognize reliable users and content in social media with coupled mutual reinforcement. In *Proceedings of the 18th international conference on World wide web*, WWW '09, pages 51–60, New York, NY, USA, 2009. ACM.

[3] E. Blanzieri and A. Bryl. A survey of learning-based techniques of email spam filtering. *Artif. Intell. Rev.*, 29:63–92, March 2008.

[4] M. Bosma, E. Meij, and W. Weerkamp. A framework for unsupervised spam detection in social networking sites. In *ECIR 2012: 34th European Conference on Information Retrieval*, Barcelona, 2012.

[5] N. N. Dalvi, P. Domingos, Mausam, S. K. Sanghai, and D. Verma. Adversarial classification. In *KDD*, pages 99–108, 2004.

[6] H. Gao, J. Hu, C. Wilson, Z. Li, Y. Chen, and B. Y. Zhao. Detecting and characterizing social spam campaigns. In *Proceedings of the 10th annual conference on Internet measurement*, IMC '10, pages 35–47, New York, NY, USA, 2010. ACM.

[7] R. Guha, R. Kumar, P. Raghavan, and A. Tomkins. Propagation of trust and distrust. In *Proceedings of the 13th international conference on World Wide Web*, WWW '04, pages 403–412, New York, NY, USA, 2004. ACM.

[8] S. Hao, N. A. Syed, N. Feamster, A. G. Gray, and S. Krasser. Detecting spammers with snare: Spatio-temporal network-level automatic reputation engine. In *USENIX Security Symposium*, pages 101–118, 2009.

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

[10] M. Huber, M. Mulazzani, E. Weippl, G. Kitzler, and S. Goluch. Friend-in-the-middle attacks: Exploiting social networking sites for spam. *IEEE Internet Computing*, 15:28–34, May 2011.

[11] D. Irani, S. Webb, and C. Pu. Study of static classification of social spam profiles in myspace, 2010.

[12] X. Jin, S. Kim, J. Han, L. Cao, and Z. Yin. A general

framework for efficient clustering of large datasets based on activity detection. *Stat. Anal. Data Min.*, 4:11–29, February 2011.

[13] X. Jin, C. X. Lin, J. Luo, and J. Han. Socialspamguard: A data mining-based spam detection system for social media networks. *PVLDB*, 4(12):1458–1461, 2011.

[14] P. Kolari, A. Java, T. Finin, T. Oates, and A. Joshi. Detecting spam blogs: A machine learning approach. In *AAAI*, 2006.

[15] K. Lee, J. Caverlee, and S. Webb. Uncovering social spammers: social honeypots + machine learning. In *SIGIR*, pages 435–442, 2010.

[16] D. Lowd and C. Meek. Adversarial learning. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, KDD '05, pages 641–647, New York, NY, USA, 2005. ACM.

[17] B. Markines, C. Cattuto, and F. Menczer. Social spam detection. In *AIRWeb*, pages 41–48, 2009.

[18] G. Mishne, D. Carmel, and R. Lempel. Blocking blog spam with language model disagreement. In *AIRWeb*, pages 1–6, 2005.

[19] D. Nagamalai, B. C. Dhinakaran, and J.-K. Lee. Bayesian based comment spam defending tool. *CoRR*, abs/1011.3279, 2010.

[20] C. Patsakis, A. Asthenidis, and A. Chatzidimitriou. Social networks as an attack platform: Facebook case study. In *Proceedings of the 2009 Eighth International Conference on Networks*, pages 245–247, Washington, DC, USA, 2009. IEEE Computer Society.

[21] T. Stein, E. Chen, and K. Mangla. Facebook immune system. In *Proceedings of the 4th Workshop on Social Network Systems*, SNS '11, pages 8:1–8:8, New York, NY, USA, 2011. ACM.

[22] G. Stringhini, C. Kruegel, and G. Vigna. Detecting spammers on social networks. In *Proceedings of the 26th Annual Computer Security Applications Conference*, ACSAC '10, pages 1–9, New York, NY, USA, 2010. ACM.

[23] K. Thomas, C. Grier, J. Ma, V. Paxson, and D. Song. Design and evaluation of a real-time url spam filtering service. In *IEEE Symposium on Security and Privacy*, pages 447–462, 2011.

[24] D. Wang, D. Irani, and C. Pu. A social-spam detection framework. In *Proceedings of the 8th Annual Collaboration, Electronic messaging, Anti-Abuse and Spam Conference*, CEAS '11, pages 46–54, New York, NY, USA, 2011. ACM.

[25] S. Webb, J. Caverlee, and C. Pu. Social honeypots: Making friends with a spammer near you. In *CEAS*, 2008.

[26] C. Whittaker, B. Ryner, and M. Nazif. Large-scale automatic classification of phishing pages. In *NDSS*, 2010.

[27] D. Yin, B. D. Davison, Z. Xue, L. Hong, A. Kontostathis, and L. Edwards. Detection of Harassment on Web 2.0, 2009.

[28] J. Zhang, J. Tang, and J.-Z. Li. Expert finding in a social network. In *DASFAA*, pages 1066–1069, 2007.

[29] A. Zinman and J. S. Donath. Is britney spears spam? In *CEAS*, 2007.