Seven Months with the Devils: A Long-Term Study of Content Polluters on Twitter

Kyumin Lee and **Brian David Eoff** and **James Caverlee**
Texas A&M University
College Station, TX 77843
{kyumin, bde, caverlee} @cse.tamu.edu

Abstract

The rise in popularity of social networking sites such as Twitter and Facebook has been paralleled by the rise of unwanted, disruptive entities on these networks—including spammers, malware disseminators, and other content polluters. Inspired by sociologists working to ensure the success of commons and criminologists focused on deterring vandalism and preventing crime, we present the first long-term study of social honeypots for tempting, profiling, and filtering content polluters in social media. Concretely, we report on our experiences via a seven-month deployment of 60 honeypots on Twitter that resulted in the harvesting of 36,000 candidate content polluters. As part of our study, we (i) examine the harvested Twitter users, including an analysis of link payloads, user behavior over time, and followers/following network dynamics and (ii) evaluate a wide range of features to investigate the effectiveness of automatic content polluter identification.

Introduction

Social networking services such as Twitter, Facebook, Digg, and MySpace are similar in nature to a public commons. They provide a forum for participants to engage, share, and interact, leading to great community value and ancillary services like search and advertising. These services must balance encouraging participation—which without renders the resource worthless—while discouraging abuse—which if left unchecked, will quickly destroy the value of the resource. In our ongoing research, we are studying the impact of *policing* on the quality and continued use and adoption of social media sites in the presence of spam, malware, and other instances of "vandalism" (Benevenuto et al. 2009).

In analogue to how law enforcement observes criminal behavior, enforces laws and community standards, and deters bad behavior in offline communities, we present a long-term study of protecting social media sites via social honeypots (Webb, Caverlee, and Pu 2008; Lee, Caverlee, and Webb 2010; Lee, Eoff, and Caverlee 2010). Similar in spirit to traditional honeypots for luring and monitoring network-level attacks, social honeypots target community-based online activities, typically through the deployment of a honeypot profile (e.g., a Twitter account), a related bot for monitoring the profile and its interactions with other users in the system, and an incrementally updated classification component for identifying and filtering accounts "in-the-wild" (e.g., that have not necessarily contacted one of the social honeypots directly). Compared to traditional spam detection methods in online communities (which often rely on user referral systems which can be gamed by spammers or by costly human-in-the-loop inspection of training data for building classifiers, which can be made quickly outdated by adaptive strategies), social honeypots have the advantages of (1) automatically collecting evidence of content polluters; (2) no interference or intrusion on the activities of legitimate users in the system; and (3) robustness of ongoing polluter identification and filtering, since new evidence of polluter behavior and strategy can be easily incorporated into content polluter models.

Specifically, this paper presents the first long-term study of social honeypots via a seven-month deployment of 60 honeypots on Twitter that resulted in the harvesting of 36,000 candidate content polluters. We provide a detailed examination of the harvested Twitter users, including an analysis of link payloads, user behavior over time, and followers/following network dynamics. We experimentally evaluate a wide range of features – including user demographics, properties of the Twitter follower/following social graph, Tweet content, and temporal aspects of user behavior – to investigate the effectiveness of automatic content polluter identification, even in the presence of strategic polluter obfuscation. Finally, we empirically validate the social honeypot-derived classification framework on an alternative Twitter spam dataset, which shows the flexibility and effectiveness of the proposed approach.

Related Work

To detect spam, researchers have proposed several methods, for example, via link analysis to detect link farms (Becchetti et al. 2006; Benczur, Csalogany, and Sarlos 2006). Others are spam email analysis based on data compression algorithms (Bratko et al. 2006), machine learning (Goodman, Heckerman, and Rounthwaite 2005; Sahami et al. 1998) or statistics (Fetterly, Manasse, and Najork 2004; Ntoulas et al. 2006; Yoshida et al. 2004).

Spammers have extended their targets to social networking sites because of the popularity of the sites and easy access to user information like name, gender, and age. Re-

cently, researchers have shown how many users are vulnerable to context-aware attack emails, and described aspects of Facebook that made such attacks possible (Brown et al. 2008; Felt and Evans 2008). Another work described how social networks could be maliciously used for social phishing (Jagatic et al. 2007). Other researchers have studied the privacy threats related to public information revelation in social networking sites (Acquisti and Gross 2006; Backstrom, Dwork, and Kleinberg 2007; Boyd 2007; Gross, Acquisti, and Heinz 2005).

Aside from privacy risks, researchers have also identified attacks that are directed at these sites (Heymann, Koutrika, and Garcia-Molina 2007). Researchers also showed that social networking sites are susceptible to two broad classes of attacks: traditional attacks that have been adapted to these sites (e.g., malware propagation) and new attacks that have emerged from within the sites (e.g., deceptive spam profiles) (Webb, Caverlee, and Pu 2008). Researchers have also begun proposing solutions to solve emerging security threats in social networking sites. Heymann et al. presented three anti-spam strategies such as identification-based strategy (detection), rank-based strategy and limit-based strategy (prevention) (Heymann, Koutrika, and Garcia-Molina 2007). Zinman and Donath attempted to detect fake profiles using learning algorithms (Zinman and Donath 2007). Benevenuto et al. presented two methods to detect spammers and content promoters in a video social networking site (Benevenuto et al. 2009). In the security aspect, Grier et al. (Grier et al. 2010) collected tweets containing URLs in Twitter, and analyzed what kind of spam pages the URLs link to and studied the limits of using blacklists to detect tweets containing spam links. Recently, researchers have begun studies of trending topic spam on Twitter (Irani et al. 2010; Benevenuto et al. 2010).

Tempting Content Polluters

As the first step toward detecting content polluters in Twitter, we present in this section the design of our Twitter-based social honeypots. Concretely, we created and deployed 60 social honeypot accounts on Twitter whose purpose is to pose as Twitter users, and report back what accounts follow or otherwise interact with them. We manipulate how often the honeypot accounts post, the content and type of their postings and their social network structure. Our Twitter-based social honeypots can post four types of tweets: (1) a normal textual tweet; (2) an "@" reply to one of the other social honeypots; (3) a tweet containing a link; (4) a tweet containing one of Twitter's current Top 10 trending topics, which are popular n-grams.

To seed the pool of tweets that the social honeypot accounts would post we crawled the Twitter public timeline and collected 120,000 sample tweets (30,000 for each of our four types). The social honeypot accounts are intentionally designed to avoid interfering with the activities of legitimate users. They only send @ reply messages to each other, and they will only follow other social honeypot accounts.

Once a Twitter user makes contact with one of the social honeypots via following or messaging the honeypot, the information is passed to the Observation system. The Observation system keeps track of all the users discovered by the system. Initially, all information about each user's account and all the user's past tweets are collected. Every hour the Observation system checks each user's status to determine if more tweets have been posted, the number of other accounts that the user is following, the number of other Twitter accounts following the user and if the account is still active.

The system ran from December 30, 2009 to August 2, 2010. During that time the social honeypots tempted 36,043 Twitter users, 5,773 (24%) of which followed more than one honeypot. One user was tempted by twenty-seven different honeypots. After removing users who followed more than one honeypot, 23,869 users remained. Figure 1 shows the number of polluters tempted per day.

Who are the Harvested Twitter Users?

Our overall goal is to automatically attract content polluters via our social honeypots so that we can provide ongoing and dependable policing of the online community. Of course, a user identified by the social honeypot system is not necessarily a content polluter. Our intuition, however, is that given the behavior of the social honeypots there is no reason for a user who is not in violation of Twitter's rules to be tempted to message or follow them. Since social honeypot accounts post random messages and engage in none of the activities of legitimate users, it seems reasonable that the likelihood of a legitimate user being tempted to be similar, if not less, than the likelihood an error would be made in hand-labeling the type of users.

Users Detected via Social Honeypots vs. Official Twitter Spammers. To support this intuition, we first investigated the 23,869 polluters the honeypots lured to see if any were considered as official violators of Twitter's terms of service (Twitter 2010). We found that Twitter eventually suspended the accounts of 5,562 (or 23% of the total polluters identified by the social honeypots). We observe that of the 5,562, the average time between the honeypot tempting the polluter and the account being suspended was 18 days. In one case, the honeypot snared a polluter 204 days before Twitter terminated the account. In other words, the social honeypots identified polluters much earlier than through traditional Twitter spam detection methods (again, on average by 18 days). But what of the remaining 77% (18,307) of the polluters that were caught but not suspended by Twitter? Are these merely legitimate accounts that have been erroneously labeled as polluters?

Cluster Analysis of Harvested Users. To better understand who these harvested Twitter users are, we manually investigated them via cluster analysis. We used the Expectation-Maximization (EM) algorithm (Dempster et al. 1977) and a set of features for representing each harvested Twitter user (described more fully in the following section) to find groups of harvested users with similar appearances/behaviors. EM is a well-known clustering algorithm, and finds the best number of clusters, assigning a probability distribution about the clusters to each instance (each harvested user account). EM discovered nine clusters. We in-

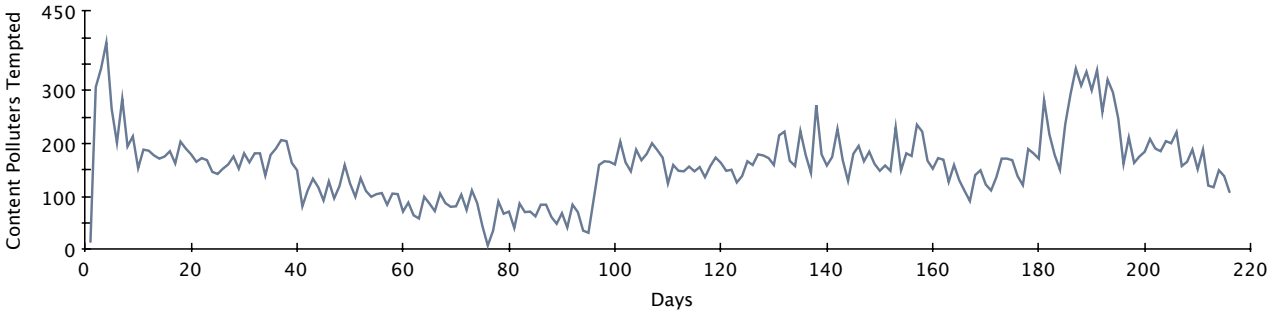

Figure 1: A chart of the number of content polluters tempted per day. On the fourth day of the study the honeypots were able to tempt a total of 391 content polluters, the most in a single day. The third highest single-day temptation was 339, which occurred on day 191.

vestigated each of the clusters, focusing on major clusters which included a large number of harvested users. Based on our observations, we grouped these users into four categories of content polluters (illustrated in Table 1):

- Duplicate Spammers: These content polluters post nearly identical tweets with or without links.
- Duplicate @ Spammers: These content polluters are similar to the Duplicate Spammers, in that they post tweets with a nearly identical content payload, but they also abuse Twitter's @username mechanism by randomly inserting a legitimate user's @username. In this way, a content polluter's tweet will be delivered to a legitimate user, even though the legitimate user does not follow the content polluter.
- Malicious Promoters[1]: These content polluters post tweets about online business, marketing, finance and so on. They have a lot of following and followers. Their posting approach is more sophisticated than other content polluters because they post legitimate tweets (e.g., greetings or expressing appreciation) between promoting tweets.
- Friend Infiltrators: Their profiles and tweets are seemingly legitimate, but they abuse the reciprocity in following relationships on Twitter. For example, if user A follows user B, then user B typically will follow user A as a courtesy. Previous literature (Mislove et al. 2007; Weng et al. 2010) has shown that reciprocity is prevalent in social networking web sites including Twitter. After they have a large number of followers, friend infiltrators begin engaging in spam activities (e.g., posting tweets containing commercial or pornographic content).

What we see is that although not suspended by Twitter, these accounts are engaging in aggressive promotion and negative behaviors, e.g., following a large number of users, and shortly dropping them, exclusively posting promotional material, posting pornographic material, and so on.

[1]While product promotion is allowed by Twitter, accounts of this nature often are guilty of violating Twitter's definition of spam which includes if the account's updates consist mainly of links, and if the account repeatedly follow and unfollow other users or promotes third-party sites that claim to get you more followers.

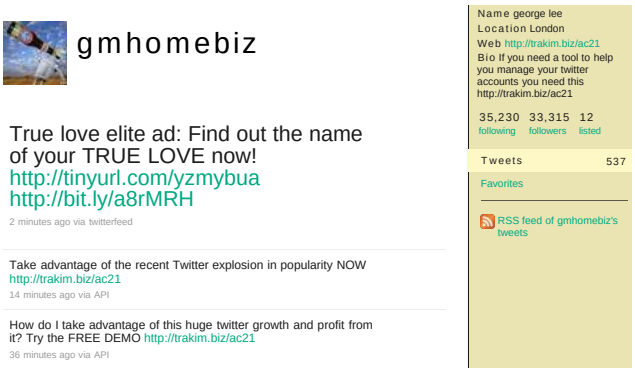

Figure 2: The Twitter homepage of gmhomebiz, a user tempted by our social honeypots.

Followers and Following. We next investigated the properties of the collected content polluters, to explore what behaviors and properties these users displayed. First, we found that on average they followed 2,123 accounts, and the average number of followers they had was 2,163. These numbers are higher than most legitimate users which only have between 100 and 1,000 followers and following counts (Krishnamurthy, Gill, and Arlitt 2008; Weng et al. 2010). Figure 2 shows the account homepage of a content polluter the social honeypots tempted. It appears to be a legitimate user; the profile information has been fully completed, and the appearance of the page has been customized. However, this account is following 35,230 users, and has a following of 33,315. Those counts are drastically different from most legitimate users who typically follow fewer than 1,000 users.

Tweeting Activity. The discovered content polluters posted on average only four tweets per day. We assume the controllers of these accounts are aware that if they post a large number of tweets per day, they will be easily detected by Twitter and their accounts will be suspended. Instead, they post a few tweets per day attempting to mimic the pattern of a legitimate user. However, they cannot hide the large number of users they follow and the large number of users following them since their goal is to promote to a vast audience.

Behavior Over Time. This observation and intuition led us

Table 1: Content Polluter Examples

Content Polluters	Tweets
Duplicate Spammer	T1: OFFICIAL PRESS RELEASE Limited To 10,000 "Platinum Founders" Reseller Licenses http://tinyurl.com/yd75xyy T2: OFFICIAL PRESS RELEASE Limited To 10,000 "Platinum Founders" Reseller Licenses http://tinyurl.com/yd75xyy
Duplicate @ Spammer	T1: #Follow @_anhran @PinkySparky @RestaurantsATL @combi31 @BBoomsma @TexMexAtl @Daniel-StoicaTax T2: #Follow @DeniseLescano @IsabelTrent @kxtramoney @PhoinixROTC44 @ATL_Events @HoldemTalkRadio
Malicious Promoter	T1: The Secret To Getting Lots Of Followers On Twitter http://bit.ly/6BiLk3 T2: Have Fun With Twitter - Twitter Marketing Software http://bit.ly/6ns0sc
Friend Infiltrator	T1: Thank you for the follows, from a newbie T2: @EstherK Yes I do and and thatnks for the follow

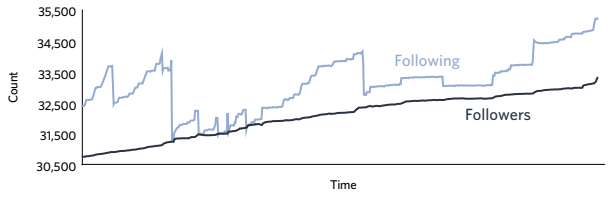

Figure 3: The graph shows the changing number of users following the *gmhomebiz* account and the number of users followed.

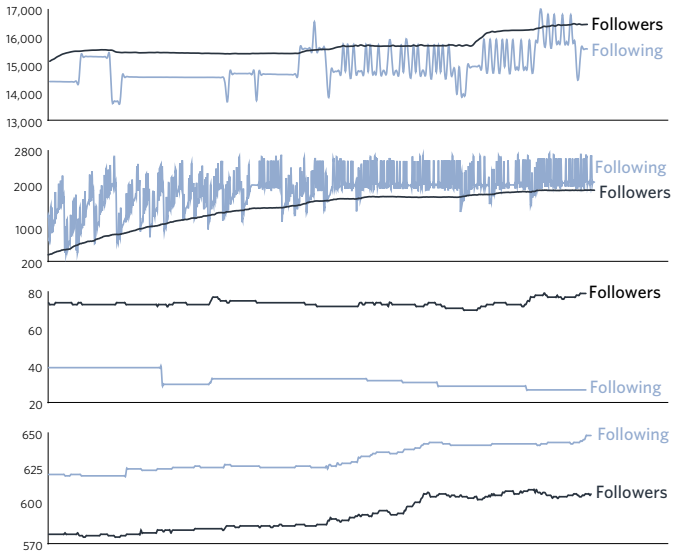

Figure 4: The top two graphs are of content polluter accounts. The bottom two are legitimate users. The accounts in the top two graphs are engaging in the act of "follower churn" (Twitter 2010).

to investigate their temporal and historical profile information which includes the number of following and followers collected by our system once per hour, since they were tempted. The number of users the content polluters were following fluctuated significantly over time. Figure 3 presents a portion of the temporal information of the content polluter shown in Figure 2. This polluter manipulated the number of

Table 2: Top five URLs posted by Content Polluters

Freq.	URL	Linked Page
2,719	www.thetweettank.com	twitter bot software
2,348	shop.cooliohigh.com	sunglasses seller
2,227	friendfeed.com	social networking site
1,919	www.tweetsbot.com	twitter bot software
771	wefollow.com	twitter 3rd party site

accounts it was following in order to achieve a balance between the number of following and followers, presumably to maintain a balance so that Twitter will not investigate and possibly suspend the account. To further illustrate, Figure 4 shows the change in the number of following and followers for two content polluters and two legitimate users.

Link Payloads. Twitter users often add an URL to the text of a tweet; thus allowing them to circumvent Twitter's 140 character limitation. Table 2 shows the five most frequently posted URLs, where we have converted shortened URLs (e.g., http://bit.ly/6BiLk3) to their original long form for easier understanding. Most linked to disreputable pages such as automatic promotion/bot software and phishing sites, with some links being inserted into hundreds of tweets in a clear attempt at link promotion.

Profiling Content Polluters

In this section, we aim to automatically "profile" Twitter-based content polluters by developing automatic classifiers for distinguishing between content polluters and legitimate users. Do we find that content polluters engage in particular behaviors that make them clearly identifiable? Or do they strategically engage in behaviors (e.g., posting frequency, history of friends in the network) that make them "invisible" to automated detection methods? For example, we have seen that our harvested content polluters post ∼four tweets a day, which seems well within "normal" behavior (in contrast to email spammers who issue millions of spam emails).

Classification Approach and Metrics

To profile content polluters on Twitter, we follow a classification framework where the goal is to predict whether a

candidate Twitter user u is a content polluter or a legitimate user. To build a classifier c

$$c : u \rightarrow \{polluter, legitimate\ user\}$$

we used the Weka machine learning toolkit (Witten and Frank 2005) to test 30 classification algorithms, such as naive bayes, logistic regression, support vector machine (SVM) and tree-based algorithms, all with default values for all parameters using 10-fold cross-validation. 10-fold cross-validation involves dividing the original sample (data) into 10 equally-sized sub-samples, and performing 10 training and validation steps. In each step, 9 sub-samples are used as the training set and the remaining sub-sample is used as the validation set. Each sub-sample is used as the validation set once.

Table 3: Dataset

Class	User Profiles	Tweets
Polluters	22,223	2,380,059
Legit Users	19,276	3,263,238

For training, we relied on a dataset[2] (summarized in Table 3) of content polluters extracted by the social honeypots and legitimate users sampled from Twitter.

Content Polluters: We filtered the original 23,869 polluters collected by the social honeypots to exclude those that were (nearly) immediately identified and suspended by Twitter. The reason why we dropped these short-lived polluters is that Twitter already has their own solution for the short-lived polluters, and our target is content polluters that are alive for a long time (at least two hours, since our system tempted them). For the remaining 22,223 polluters, we collected their 200 most recent tweets, their following and follower graph, and their temporal and historical profile information including the number of following and followers collected by our system once per hour since they were tempted by a honeypot.

Legitimate users: To gather a set of legitimate users, we randomly sampled 19,297 Twitter users. Since we have no guarantees that these sampled users are indeed legitimate users (and not polluters) and hand labeling is both time consuming and error-prone, we monitored the accounts for three months to see if they were still active and not suspended by Twitter. After three months, we found that 19,276 users were still active and so we labeled them as legitimate users. Even though there is chance of a false positive in the legitimate user set, the results of our classifier study should give us at worst a lower bound on accuracy since the introduction of possible noise in the training set would only degrade our results.

We compute precision, recall, F-measure, accuracy, area under the ROC curve (AUC), false negatives (FNs) and false positives (FPs) as metrics to evaluate our classifier. In the confusion matrix, Table 4, a represents the number of correctly classified polluters, b (called FNs) represents the number of polluters misclassified as legitimate users, c (called FPs) represents the number of legitimate users misclassified as polluters, and d represents the number of correctly classified legitimate users. The precision (P) of the polluter class is $a/(a+c)$ in the table. The recall (R) of the polluter class is $a/(a+b)$. F_1 measure of the polluter class is $2PR/(P+R)$. The accuracy means the fraction of correct classifications and is $(a+d)/(a+b+c+d)$. AUC is a measure showing classification performance. The higher AUC is, the better classification performance is. 1 AUC value means a perfect performance.

[2] Available at http://infolab.tamu.edu/data

Table 4: Confusion matrix

		Predicted	
		Polluter	Legitimate
Actual	Polluter	a	b
	Legit User	c	d

Features

The quality of a classifier is dependent on the discriminative power of the features. Based on our previous observations, we created a wide variety of features belonging to one of four groups: User Demographics (**UD**): features extracted from descriptive information about a user and his account; User Friendship Networks (**UFN**): features extracted from friendship information such as the number of following and followers; User Content (**UC**): features extracted from posted tweets; and User History (**UH**): features extracted from a user's temporal and historical profile information.

The specific features for each feature group are:

UD the length of the screen name, and the length of description

UD the longevity of the account

UFN the number of following, and the number of followers

UFN the ratio of the number of following and followers

UFN the percentage of bidirectional friends:

$$\frac{|following \cap followers|}{|following|}\ and\ \frac{|following \cap followers|}{|followers|}$$

UFN the standard deviation of unique numerical IDs of following

UFN the standard deviation of unique numerical IDs of followers

UC the number of posted tweets

UC the number of posted tweets per day

UC $|links|\ in\ tweets\ /\ |tweets|$

UC $|unique\ links|\ in\ tweets\ /\ |tweets|$

UC $|@username|\ in\ tweets\ /\ |tweets|$

UC $|unique\ @username|\ in\ tweets\ /\ |tweets|$

@username features can detect a content polluter posting tweets with various @usernames.

UC the average content similarity over all pairs of tweets posted by a user

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

Table 5: Top 10 features

Feature	χ^2 value	Avg of Polluters	Avg of Legitimate Users
standard deviation of following	26,708	35,620,487	19,368,858
the change rate of $\|following\|$	23,299	29.6	1.5
standard deviation of followers	22,491	35,330,087	22,047,831
$\|following\|$	15,673	2,212	327
longevity of the account	15,467	279	506
ratio of the number of following and followers	12,115	11.1	1.5
$\|links\|$ per tweet	11,827	0.65	0.21
$\|@username\|$ *in tweets* / $\|tweets\|$	9,039	0.2	0.51
$\|unique\ @username\|$ *in tweets* / $\|tweets\|$	8,859	0.12	0.17
$\|unique\ links\|$ per tweet	7,685	0.48	0.18

UC the ZIP compression ratio of posted tweets:

$$\frac{uncompressed\ size\ of\ tweets}{compressed\ size\ of\ tweets}$$

The compression ratio can detect a content polluter posting nearly identical tweets because when we compress its tweets, their compressed size is significantly decreased.

UH the change rate of number of following obtained by a user's temporal and historical information:

$$\sqrt{\frac{1}{n-1}\sum_{i=1}^{n-1}(f_{i+1}-f_i)}$$

where n is the total number of recorded temporal and historical information, and f_i means the number of following of the user extracted in ith temporal and historical information.

We computed the χ^2 value (Yang and Pedersen 1997) of each of the features to determine their discriminative power. The larger the χ^2 value is, the higher discriminative power the corresponding feature has. The results showed all of our features had positive discrimination power, though with different relative strengths. Table 5 shows the top-10 features. The standard deviation of numerical IDs of following returned the highest χ^2 value because polluters follow users randomly. Content polluters' following standard deviation is much higher than legitimate users' standard deviation. The change rate of number of following outperforms other features because polluters increased the number of following in order to contact a larger number of users and promote to them, and then decreased the number of following in order to maintain a balance between the number of following and followers to avoid being suspended by Twitter. The average change rate of polluters was 29.6, while the average change rate of legitimate users was 1.5. Like the standard deviation of following, polluters' follower standard deviation is much higher than legitimate users' follower standard deviation. Polluters had larger followings than legitimate users, and shorter longevity than legitimate users.

Classification Results

Using the classification setup described above and these feature groups, we tested 30 classification algorithms using the Weka machine learning toolkit (Witten and Frank 2005). Across most classifiers (25 of the 30 tested), we find that the results are consistent, with accuracy ranging from 95% to 98%, indicating that the strength of classification lies mainly in the choice of features and is relatively stable across choice of particular classifier. For the other 5 of the 30 tested, accuracy ranges from 89% to under 95%. Tree-based classifiers showed the highest accuracy results. In particular, Random Forest produced the highest accuracy as shown in Table 6. Its accuracy was 98.42% and 0.984 F_1 measure.

Table 6: The performance result of Random Forest

Classifier	Accuracy	F_1	AUC	FNs	FPs
Random Forest	98.42%	0.984	0.998	301	354

We additionally considered different training mixtures of polluters and legitimate users, ranging from 1% polluter and 99% legitimate to 99% polluter and 1% legitimate. We find that the classification quality is robust across these training mixtures.

Table 7: Boosting and bagging of the Random Forest classifier

Classifier	Accuracy	F_1	AUC	FNs	FPs
Boosting	**98.62%**	**0.986**	0.995	287	**287**
Bagging	98.57%	**0.986**	**0.999**	**248**	345

In order to improve the Random Forest classifier, we additionally applied standard boosting (Freund and Schapire 1997) and bagging (Breiman 1996) techniques. Both create multiple classifiers and combine their results by voting to form a composite classifier. Table 7 shows the results. Both outperformed the original Random Forest. Boosting of Random Forest classifier produced the best result, 98.62% accuracy and 0.986 F_1 measure. These results provide strong evidence that social honeypots attract polluter behaviors that are strongly correlated with observable features of their profiles and their activity in the network.

Handling Strategic Polluter Obfuscation

As time passes, the designers of content polluters may discover which features are signaling to our system that their polluters are not legitimate Twitter users, and so these features may lose their power to effectively profile and detect polluters. Thus, we tested the robustness of the polluter-based classifier by constraining the classifier to have access

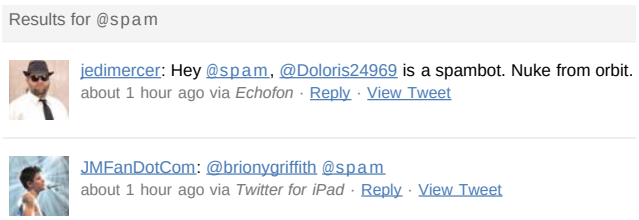

Figure 5: "@spam" search results

only to certain feature groups, mimicking the scenario in which content polluters strategically target our system (and, hence, entire feature groups lose their ability to distinguish polluters). We also considered scenarios in which one, two or even three entire feature groups lose their effectiveness.

Table 8: The performance results of various feature group combinations

Feature Set	Accuracy	F_1	AUC	FNs	FPs
UD	76.17%	0.762	0.839	6,007	3,882
UH	85.34%	0.854	0.899	4,130	1,950
UC	86.39%	0.864	0.932	2,811	2,837
UFN	96.46%	0.965	0.992	510	958
UD+UC	88.61%	0.886	0.953	2,469	2,256
UD+UH	92.45%	0.925	0.967	1,743	1,389
UC+UH	94.38%	0.944	0.979	1,111	1,221
UFN+UH	97.11%	0.971	0.994	496	702
UD+UFN	97.50%	0.975	0.995	437	597
UFN+UC	97.92%	0.979	0.996	413	448
UD+UC+UH	95.42%	0.954	0.985	878	1,022
UD+UFN+UH	97.78%	0.978	0.996	395	524
UFN+UC+UH	98.13%	0.981	**0.997**	361	413
UD+UFN+UC	**98.26%**	**0.983**	**0.997**	**333**	**388**

Following the classification setup described above, we trained the content polluter classifier based on different feature group combinations, resulting in the effectiveness measures shown in Table 8. First, in the extreme case in which only a single feature group is available (either User Demographics, User Friendship Network, User Content, or User History), we see results ranging from 76.17% accuracy to 96.46% accuracy. Considering pairs of features, we can see that the classification accuracy ranges from 88.61% to 97.92%. Even in the case when the content polluters obfuscate the single most distinguishing signal (User Friendship Network), the UD+UC+UH case resulted in 95.42% accuracy and nearly equaled the performance across the other measures. Together, these results indicate that the signals distinguishing content polluters from legitimate users are not tied to a single "super-feature", but are a composite of multiple inter-related features. This gives us confidence going forward that content polluters cannot trivially change a single behavior (e.g., by manipulating their follower-following ratio) and become invisible. Rather they must become more like legitimate users, which necessarily decreases the effectiveness and impact of their pollution attempts (e.g., by reducing the number of links per tweet, reducing the number of @*username* per tweet, and so on).

Validation with Twitter @spam

Complicating the effective profiling of content polluters is the potential mismatch between models built on polluters harvested by social honeypots and for content polluters in Twitter-at-large. We have seen that the framework presented in this paper is effective at tempting large amounts of content polluters and at discriminating between polluters and legitimate users. However, it could be argued that there is inherent bias in the testing framework we have employed since the capacity of the classification framework to distinguish polluters (even with 10-fold cross validation) is linked to the collection method via social honeypots or that we are guilty of over-fitting the models. Thus, we also evaluated the quality of our approach by applying the learned content polluter models to a test set that was collected *entirely separately* from our Twitter-based social honeypots.

To collect a content polluter dataset orthogonally from the honeypot framework, we monitored Twitter's spam reporting channel over a four month period using the Twitter search API. Twitter supports user-based spam reporting via a special @spam Twitter account which allows users to report suspicious or malicious users to @spam. Figure 5 illustrates a sample @spam search result. In the first result, *jedimercer* reported *Doloris24969* as a suspicious account to @spam account. Twitter investigates the reported user and if Twitter itself determines that the user has engaged in harmful activities, only then is the account suspended. If we find that our approach can effectively identify spam accounts from this alternative source, we will have confidence in the robustness and wide applicability of our results. Accounts suspended in this way may behave differently from the ones detected by our honeypots (e.g., they may never follow another account as our honeypots require).

Concretely, we constructed our orthogonal test set by searching for tweets containing "@spam" once per 20 minutes, and extracted the user account names (@*username*) listed in the tweets. When each user was reported to @spam, we collected their descriptive profile information, friendship information, tweets, and temporal and historical profile information. We continued to monitor these reported candidate spam accounts to identify only those that were actually suspended by Twitter (in effect, throwing away all of the false positives reported by users but not subsequently found to be spammers by Twitter itself). The four month observation period led to a total of 2,833 suspended accounts.

Following the classifier setup described in the previous section, we trained a Random Forest classifier using the content polluter data collected by our social honeypots and the set of legitimate users (recall Table 3). The trained classifier predicted class labels (content polluter or legitimate user) for the 2,833 suspended users in the separate @spam testing set.

We find that our approach leads to 96.75% accuracy and 0.983 F_1 measure over these @spam profiles. When we applied bagging to the Random Forest classifier, we achieved an even better result, 98.37% accuracy and 0.992 F_1 measure. We did not compute AUC because the test set does not include legitimate users. These results indicate that there is a strong capacity of our approach to detect harmful users on Twitter, even if they have not been directly discovered by

our social honeypots.

To investigate the cases in which our classifier did not perform well (for the 46 spammers who were misclassified as legitimate users), we manually examined their profiles, friendship networks, and historical behavior. In all cases, the misclassified users have a low standard deviation of numerical IDs of following and followers (which was a strong discriminating feature in our content polluter study). Most of these users were quickly suspended by Twitter after they were first reported, meaning that the historical and temporal profile features were not available to our system. For those users for which we did have sufficient historical and temporal profile information, most engaged in widespread *@username* messages to contact many users rather than directly following users.

Conclusion

Social media sites derive their value by providing a popular and dependable community for participants to engage, share, and interact. This community value and related services like search and advertising are threatened by spammers, malware disseminators, and other content polluters. In an effort to preserve community value and ensure long-term success, we have presented the design and evaluation of a system for automatically detecting and profiling content polluters on Twitter. During our seven-month long study we were able to lure approximately 36,000 abusive Twitter accounts into following our collection of social honeypots. We have seen how these content polluters reveal key distinguishing characteristics in their behavior, leading to the development of robust classifiers.

References

Acquisti, A., and Gross, R. 2006. Imagined communities: Awareness, information sharing, and privacy on the facebook. In *Workshop on Privacy Enhancing Technologies*.

Backstrom, L.; Dwork, C.; and Kleinberg, J. 2007. Wherefore art thou r3579x?: anonymized social networks, hidden patterns, and structural steganography. In *WWW*.

Becchetti, L.; Castillo, C.; Donato, D.; Leonardi, S.; and Baeza-Yates, R. 2006. Link-based characterization and detection of web spam. In *AIRWeb*.

Benczur, A. A.; Csalogany, K.; and Sarlos, T. 2006. Link-based similarity search to fight web spam. In *AIRWeb*.

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

Benevenuto, F.; Magno, G.; Rodrigues, T.; and Almeida, V. 2010. Detecting spammers on twitter. In *Collaboration, Electronic messaging, Anti-Abuse and Spam Conference (CEAS)*.

Boyd, D. 2007. Social network sites: Public, private, or what? In *The Knowledge Tree: An e-Journal of Learning Innovation*.

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

Breiman, L. 1996. Bagging predictors. *Mach. Learn.* 24(2):123–140.

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

Dempster, A.; Laird, N.; Rubin, D.; et al. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)* 39(1):1–38.

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

Fetterly, D.; Manasse, M.; and Najork, M. 2004. Spam, damn spam, and statistics: using statistical analysis to locate spam web pages. In *Workshop on the Web and Databases*.

Freund, Y., and Schapire, R. E. 1997. A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.* 55(1):119–139.

Goodman, J.; Heckerman, D.; and Rounthwaite, R. 2005. Stopping spam. *Scientific American* 292(4):42–88.

Grier, C.; Thomas, K.; Paxson, V.; and Zhang, M. 2010. @spam: the underground on 140 characters or less. In *Computer and communications security (CCS)*.

Gross, R.; Acquisti, A.; and Heinz, III, H. J. 2005. Information revelation and privacy in online social networks. In *Workshop on Privacy in the electronic society*.

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

Irani, D.; Webb, S.; Pu, C.; and Li, K. 2010. Study of trend-stuffing on twitter through text classification. In *Collaboration, Electronic messaging, Anti-Abuse and Spam Conference (CEAS)*.

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

Krishnamurthy, B.; Gill, P.; and Arlitt, M. 2008. A few chirps about twitter. In *Workshop on Online social networks*.

Lee, K.; Caverlee, J.; and Webb, S. 2010. Uncovering social spammers: Social honeypots + machine learning. In *SIGIR*.

Lee, K.; Eoff, B. D.; and Caverlee, J. 2010. Devils, angels, and robots: Tempting destructive users in social media. In *International AAAI Conference on Weblogs and Social Media*.

Mislove, A.; Marcon, M.; Gummadi, K. P.; Druschel, P.; and Bhattacharjee, B. 2007. Measurement and analysis of online social networks. In *SIGCOMM*.

Ntoulas, A.; Najork, M.; Manasse, M.; and Fetterly, D. 2006. Detecting spam web pages through content analysis. In *WWW*.

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

Twitter. 2010. The twitter rules. `http://help.twitter.com/forums/26257/entries/18311`.

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

Weng, J.; Lim, E.-P.; Jiang, J.; and He, Q. 2010. Twitterrank: finding topic-sensitive influential twitterers. In *WSDM*.

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

Yang, Y., and Pedersen, J. O. 1997. A comparative study on feature selection in text categorization. In *ICML*.

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

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