Devils, Angels, and Robots: Tempting Destructive Users in Social Media

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

Abstract

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, content polluters, and malware disseminators. In an effort to preserve community value and ensure long-term success, we present a prototype system for automatically detecting and profiling *destructive users* in social media. We described the architecture of the system – inspired by the "broken windows" theory embraced by law enforcement – the results and insights gained from a preliminary study conducted to determine the efficacy of our approach, and a discussion of our ongoing research.

Introduction

Social media sites like Facebook, Twitter, and Digg display many characteristics of traditional offline communities. Prominent examples include the rise and fall of trendy destinations (c.f. Friendster versus Facebook) and the income and social status divide that separates groups of online participants (e.g., MySpace versus Facebook (boyd 2009)). In our ongoing research, we are studying the impact of *policing* on the quality and continued use and adoption of social media sites. Social media is already being inundated with 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 neighborhoods, we are studying a suite of novel approaches for protecting social media sites. Our current work is inspired by the popular Broken Windows theory.

The Broken Window theory to law enforcement was first proposed by James Wilson and George Kelling in 1982 (Wilson and Kelling 1982). The inspiration for the theory was a series of experiments pertaining to the link between anonymity and destruction conducted by Stanford sociologist Philip Zimbardo. In one of the experiments a 1959 Oldsmobile automobile was abandoned in the Bronx neighborhood of New York City; a second car was abandoned in Palo Alto (Zimbardo 1969). The license plates were removed from both cars and the hoods were left opened. Useful parts were stripped from the automobile abandoned in the Bronx after only ten minutes. Three days later, the car was a "battered, useless hulk of metal." In contrast, the car abandoned in Palo Alto was untouched. Zimbardo believed vandalism needed to be "primed" because it did not frequently occur in Palo Alto. Hence, Zimbardo primed the situation by taking a sledgehammer to the car; within days, the Palo Alto car was destroyed by passersby and flipped over. Utilizing Zimbardo's observations Wilson and Kelling suggest that if a window is broken in a building and not repaired, soon all the windows in the building will be broken, and that "serious street crime flourishes in areas in which disorderly behavior goes unchecked." In essence, *quickly fix the small problems or you will face much larger ones*.

We believe that the Broken Windows theory[1] can inform the design and ongoing maintenance of social media sites. Our hypothesis is that inappropriate behavior even if it does not overtly affect users must be stopped, or else it will lead to behavior that is much more invasive and destructive.

System Architecture

Concretely, we have developed and deployed a prototype system for detecting and profiling *destructive users* on the Twitter micro-blogging service (Krishnamurthy, Gill, and Arlitt 2008) (see Figure 1). Twitter users are able to post 140 character messages (tweets), and follow the posts of other users. Twitter automatically sends an email to inform a user that their postings are being followed. More and more these messages are not an indication of friends or colleagues discovering a user's Twitter feed, but of an unwelcome automated entity. These accounts, which we will call *Robots*, often have an abnormal number of followers, are following an abnormal number of users, and post either a few tweets or an excessive amount. Most users are indifferent to these Robots, and do not actively block them from following their accounts. In our manual inspection of 500 Robots, we identified a mix of strategies including duplicate messages, the targeting of specific users (through the inclu-

[1] Recent studies have questioned the efficacy of a broken windows based approach to preventing crime (Jean 2007). Nevertheless, we still find value in Zimbardo's original observations. Success or failure of policies based on broken windows does not prove nor disprove Zimbardo's conclusions.

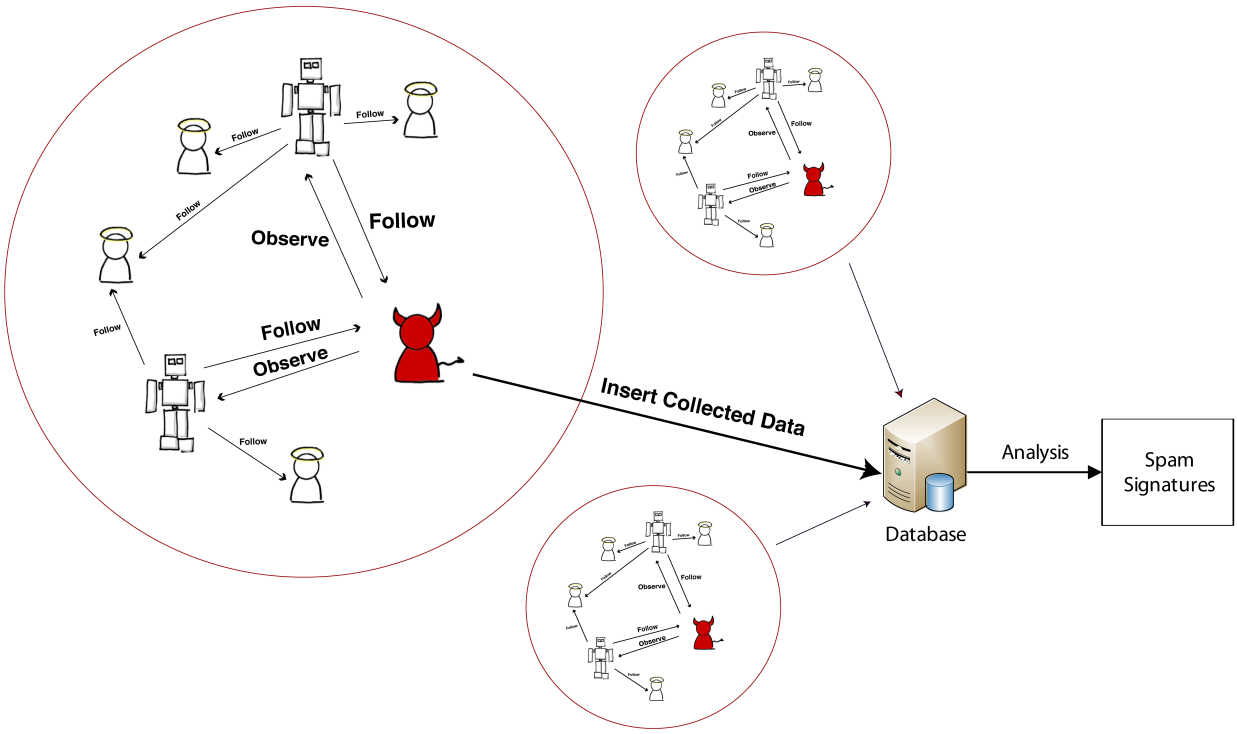


Figure 1: System Architecture

sion of $@username$), spam link dissemination, and phishing URL dissemination.

Due to the automated nature of the entities we are trying to stop, our system is designed to minimize human involvement in Robot detection. Instead of seeking out Robots, our approach is designed to induce them to find us through the deployment of special entities (and some associated control scripts), akin to the use of honeypots (Spitzner 2003)(Webb, Caverlee, and Pu 2008). Because of the tempting nature of our entities we refer to them as *Devils*. In keeping with the metaphor we refer to regular contributing Twitter users as *Angels*.

Each of our Devils poses as a legitimate Twitter user. We are able to manipulate how often a day they post (as a random number distributed over a certain range), the type of messages they post, and the topology of their following/follower network. There are four types of tweets our Devil can post. The first, is a normal textual tweet. The second, is a tweet which is an '@' reply to one of the other Devils. The third, is a tweet that contains a link. The fourth, is a tweet that contains one of Twitter's current Top 10 trending topics, which are n-grams that are frequently used on the Twitter system.

The Devils post at various times. Their goal is to tempt a Robot to follow them. Any user which follows one of the Devils is assumed to be a Robot and not an Angel. Once a Robot begins to follow one of the Devils, a separate process observes and records the behavior of the Robot. The content of the Robot's tweets, how many followers the Robot has, how many users the Robot is following, and how these metrics change over time since our Devil tempted the robot are all noted.

Table 1: The settings of the five Devils

Devil #	#1	#2	#3	#4	#5
Posting Frequency (per day)	0-10	10-20	10-20	20-30	20-30
% of Normal Tweets	25	40	20	20	20
% of @ Tweets	25	20	40	20	20
% of Link Tweets	25	20	20	40	20
% of Trending Topic Tweets	25	20	20	20	40

Exploratory Study and Results

A preliminary study was conducted to determine if our Devils would be able to tempt Robots into following them. We created five Devils, each having a different posting frequency and tweet ratio as shown in Table 1. The study ran for two weeks from November 24 to December 8, 2009. During this time our Devils were able to tempt 131 Robots to begin following them. These Robots had created 158,847 tweets. Table 2 presents detailed results. The higher posting frequency was, the larger number of Robots a Devil tempted. For example, Devil #4 and #5 tempted larger number of Robots than Devil #2 and #3. Devil #1 tempted the smallest number of Robots. Devils that tweeted messages containing trending topics or plain tweets were able to tempt more

Table 2: Breakdown of the Robots following the five Devils

Devil #	#1	#2	#3	#4	#5
# of Robots	3	22	19	40	47
Avg # of followings of the Robots	320	2,808	3,315	3,946	3,108
Avg # of followers of the Robots	242	2,431	3,053	3,457	2,966
Avg # of tweets of the Robots	210	1,558	2,263	1,431	505

Robots than tweets that contained links or @ replies.

Table 3 presents details on four of the Robots the Devils were able to tempt. None of these Devils had their accounts suspended by Twitter during the period of our study. Each Robot had significant fluctuations in the number of users they were following and the number of users that were following them. In three of the four examples the Robot unfollowed a number of users so as to give the appearance of a similar count of users following and users being followed by the them.

Of the 131 Robots our Devils tempted Twitter detected and suspended the account of only 10 for violating the terms of service (Twitter 2009), meaning that over 92% of the Robots are still alive in spite of engaging in negative behavior (e.g., following a large number of users, and shortly dropping them, posting promotional material, posting pornagraphic material). Of the 10, the average time between our Devil discovering a Robot and the account being suspended was 49 hours. In one case, the Devils identified the Robot five days earlier than Twitter.

Future Work and Conclusion

Through our research we hope to answer four questions:

- What behavior/profile of Twitter user is most tempting to Robots?
- What is the structure of Robots' following/followers network?
- How does this network structure change over time?
- Given the information we've collected, can we build a realtime system for identifying destructive Twitter users?

To address these questions we are designing a more comprehensive study, which will be over the course of a much longer time period, and with a larger number of more sophisticated Devils. Given the success of our preliminary study we feel confident that we can achieve answers to the above questions.

In 1968, Garrett Hardin (Hardin 1968) published the groundbreaking article "The Tragedy of the Commons." Hardin states that there will always be a financial benefit to a person abusing a common resource. The irony is that this abuse destroys the resource. To ensure that social media sites such as Twitter remain useful, destructive users must be quickly discovered and removed, even if their behavior is currently just a minor annoyance.

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.

References

Benevenuto, F.; Rodrigues, T.; Almeida, V.; Almeida, J.; and Gonçalves, M. 2009. Detecting spammers and content promoters in online video social networks. In *SIGIR '09: Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, 620–627. New York, NY, USA: ACM.

boyd, d. 2009. Feature implications of user choice: the cultural logic of "myspace or facebook?". *interactions* 16(6):33–36.

Hardin, G. 1968. The tragedy of the commons. *Science* 162(3859):1243–1248.

Jean, P. K. B. S. 2007. *Pockets of Crime: Broken Windows, Collective Efficacy, and the Criminal Point of View*. The University of Chicago Press.

Krishnamurthy, B.; Gill, P.; and Arlitt, M. 2008. A few chirps about twitter. In *WOSP '08: Proceedings of the first workshop on Online social networks*, 19–24. New York, NY, USA: ACM.

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

Twitter. 2009. The twitter rules.

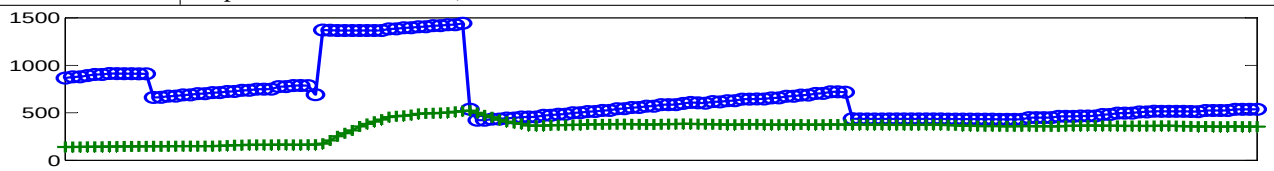
Webb, S.; Caverlee, J.; and Pu, C. 2008. Social Honeypots: Making Friends With A Spammer Near You. In *Proceedings of the Fifth Conference on Email and Anti-Spam (CEAS 2008), Mountain View, CA*.

Wilson, J. Q., and Kelling, G. L. 1982. Broken windows. *The Atlantic*.

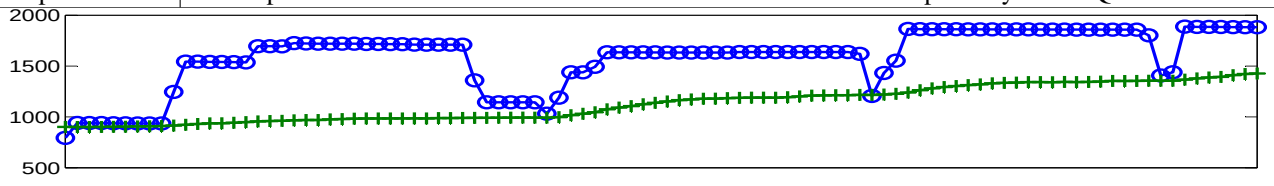
Zimbardo, P. 1969. The human choice: Individuation, reason, and order versus deindividuation, impulse, and chaos. In *Nebraska symposium on motivation*, volume 17, 237–307. University of Nebraska Press.

Table 3: A sample of the various Robots our Devils have tempted. The graphs show the changing number of users following the Robots and users the Robots are following

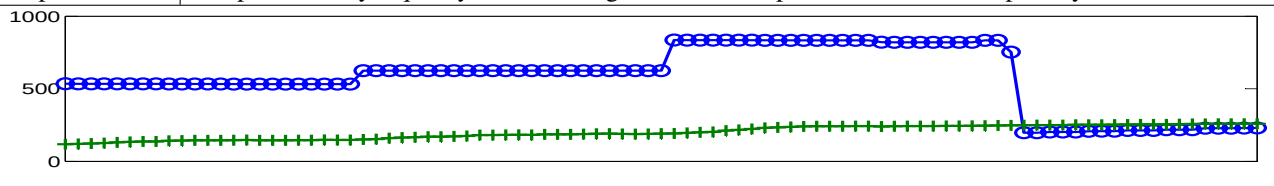
Twitter Account	Tastemystyle
Tempted By	Devil #5
Date Tempted	2009-11-25 01:00:11
Date Suspended	Alive
Sample Tweet 1	Traffic!! http://twitpic.com/qsji4
Sample Tweet 2	watch me do some Exiting Work NO! lol http://twitcam.com/6hax (@tastemystyle live on http://twitcam.com/6hax)



Twitter Account	B2Corporate
Tempted By	Devil #2
Date Tempted	2009-11-27 18:48:50
Date Suspended	Alive
Sample Tweet 1	B2Corporate.com - Windows pay attention...Google Crhome Os is coming... http://bit.ly/8gLMMj
Sample Tweet 2	B2Corporate.com - Human resources and resumes under control: http://bit.ly/5lMcxQ



Twitter Account	localmaplisting
Tempted By	Devil #4
Date Tempted	2009-11-28 01:40:28
Date Suspended	Alive
Sample Tweet 1	Good Evening, please have a visit in our site. http://seomediastar.com
Sample Tweet 2	We provide only a quality service that gets results and price our service comparably.



Twitter Account	thatsbusiness
Tempted By	Devil #5
Date Tempted	2009-11-27 02:56:08
Date Suspended	Alive
Sample Tweet 1	So how do you become the hunted and not be the hunter in Network Marketing? http://bit.ly/SqGyH
Sample Tweet 2	What is it worth to regain your family & freedom & work from home? Simple 2 min testimony video tells all... http://bit.ly/155Oa1