First Monday, Volume 15, Number 1 - 4 January 2010

first monday

PEER-REVIEWED JOURNAL ON THE INTERNET

Detecting spam in a Twitter network

by Sarita Yardi, Daniel Romero, Grant Schoenebeck, and danah boyd

Abstract

Spam becomes a problem as soon as an online communication medium becomes popular. Twitter's behavioral and structural properties make it a fertile breeding ground for spammers to proliferate. In this article we examine spam around a one–time Twitter meme — "robotpickuplines". We show the existence of structural network differences between spam accounts and legitimate users. We conclude by highlighting challenges in disambiguating spammers from legitimate users.

Contents

Introduction

Spam is a curious game of cat and mouse — spammers try to mask themselves as legitimate users, and legitimate users want to avoid spammers. Spam has plagued almost every popular online medium — Clay Shirky (2004) remarked that a communication channel "isn't worth its salt" until the spammers descend.

Spam is a moving target and difficult to measure. Statistics on the amount of spam in e–mail vary. In 2009, Microsoft reported that 97 percent of all e–mail messages sent over the Web were unwanted, Google reported that spam hovered between 90–95 percent, and Symantec reported that spam accounted for 90.4 percent of all e–mail (Swidler, 2009; Symantec, 2009; Waters, 2009). Most e–mail spam is filtered by e–mail clients and goes unnoticed, yet spammers persist, in sophisticated and creative ways. The economics of spam is in the scam — how much money can spammers make off of Internet users who click on their links? The point–of–sale is the scam that extracts money or value from Web users (Anderson, *et al.*, 2007).

E–mail, blog, and social network spam are familiar and often vexing parts of online communication experiences. Twitter spam has joined the ranks, but differs from other kinds of spam. In this article, we measure and analyze spam behavior on Twitter. Similar to e–mail spam, Twitter spam varies in style and tone; some approaches are well–worn and transparent and others are deceptively sophisticated and adaptable. On e–mail, the driving question for spammers is whom should they target and how? In contrast, on Twitter, the question is what to target and when — what trending topic should spammers latch on to, and how long can they stay there before drowning out real content and effectively killing the meme? At its best, Twitter spam is annoying and clogs up a user's tweetstream; at worst, it can compromise users' security or privacy.

To explore spam behavior on Twitter, we tracked an endogenous Twitter meme — the

#robotpickuplines hashtag — through its entire lifecycle, from inception and growth through decay. By endogenous meme, we mean one whose lifecycle begins and ends within Twitter. This is different than #MichaelJackson or #iranelections, which were the result of external network influences from news and mass media.

We describe the evolution of the #robotpickuplines meme and the subsequent infiltration of spammers. We show when and where spam occurs on Twitter and what strategies spammers use to fly under the radar. This goal of this research is to characterize behavioral patterns and to help protect from current and future spam intrusions.

Related work

Spam is not a new problem. To evade content–based filters, spammers have adopted techniques such as image spam and e–mail messages (Swidler, 2009). These techniques attempt to mislead filters that are designed to detect text patterns. Spammers also evade IP–based blacklists, which must be continually updated to keep up with spammers. There are generally two approaches to spam filtering: (1) examining the content of the message and (2) tracking the IP address of the sender (Feamster, 2008). However, there are weaknesses in both of these approaches. First, spammers can easily change message content, and they can vary messages by spam recipient, making them harder to track and identify. Second, they can change the sending IP address dynamically so that multiple hosts can exhibit the same spamming behavior from different IP addresses over time (Feamster, 2008).

There is a growing body of research exploring alternative approaches to spam detection. Ramachandran, *et al.* (2007) propose a new class of techniques called behavioral blacklisting, which identify spammers based on their network–level behavior. Rather than attempting to blacklist individual spam messages based on what the message contains, behavioral blacklisting classifies a message on spatial and temporal traffic patterns of how the e–mail itself was sent (Feamster, 2008; Ramachandran, *et al.*, 2007). These kinds of patterns are more difficult to detect, but are also more difficult for spammers to game.

Twitter introduces new kinds of spam behavior. First, spammers may exhibit unusually high following to friend ratios because spammers auto–follow many users but users may not reciprocate. Second, spammers may retweet and change legitimate links to illegitimate ones, the process of which is obfuscated by URL shorteners. Last, temporal patterns of tweeting may vary with frequency, volume, and distribution over time. These multi–pronged approaches can all be part of a conspired effort to go undetected.

Twitter

Twitter is a microblogging service where users can post 140 character messages called tweets. Unlike Facebook and MySpace, Twitter is directed, meaning that a user can follow another user, but the second user is not required to follow back. Most accounts are public and can be followed without requiring the owner’s approval. With this structure, spammers can easily follow legitimate users as well as other spammers.

Hashtags are indicated by a # symbol and are combined with keywords to indicate a topic of interest. Hashtags become popular when many people use the hashtag. Popular topics, known as “trending” topics, appear on the main Twitter page and can significantly increase the number of tweets containing that topic. Another convention is the use of URL shorteners to share links of interest within the 140 character constraint. Some popular URL shorteners, such as bit.ly, enable users to track the number of clicks their link receives.

A number of researchers have examined Twitter networks recently. Krishnamurthy, *et al.’s* (2008) early analysis of Twitter characterized users and their behavior, geographic growth patterns, and current size of the network. Java, *et al.* (2007) examined the follower network on Twitter, including over 1.3 million tweets from over 76,000 users. Their study reported high degree correlation and reciprocity in the follower network and revealed there is great variety in users’ intentions and usages on Twitter. Huberman, *et al.* (2009) demonstrated that Twitter users only interact with a small subset of their social connections. However, the role of Twitter spam in these results has not been explored extensively.

Researchers have also investigated reasons why people use Twitter, such as finding common ground and connectedness, as well as benefits for informal communication at work (Zhao and Rosson, 2009). Honeycutt and Herring (2008) described conversational practices on Twitter based on the @ reply that is used to refer to others and to direct messages to others. boyd, *et al.* (2009) examined conversational practices in Twitter based on retweeting and the ways that authorship, attribution, and communicative fidelity are negotiated.

We build on this previous work in our analysis of spam behavior from both a social and network analysis perspective. In this paper we address the following questions:

> *RQ1: Does age of account differ between spammers and legitimate users?*
>
> *RQ2: Do spammers tweet more frequently than legitimate users?*
>
> *RQ3: Do spammers have more friends than followers?*
>
> *RQ4: Are spammers clustered?*
>
> *RQ5: Can spammers be located based on network structure?*

Methods

The #robotpickuplines hashtag was started by @grantimahara, a robot builder, modelmaker and television host on *Mythbusters* (http://dsc.discovery.com/fansites/mythbusters/mythbusters.html). As a well–known figure, he had over 20,000 followers as of June 2009. On the morning of 5 June, along with @cybernortis and others, he tweeted a series of robot jokes, leading to:

> RT @khylsty it would bring a whole new meaning to
> "nice rack"
> :) - LOL

Users retweeted his tweet and added their own new tweets. Within an hour, at 11:18am PST, he tweeted:

> Hah! We're trending, now baby!
>
> Really, this is crazy! #robotpickuplines started only about an hour ago tailing off a discussion about robot bodies…

Traffic spiked quickly and contained a mix of retweets and original posts, mostly sexual jokes of varying quality (*e.g.*, references to floppy drives and hard drives). The most commonly retweeted tweet was:

> Hi my name's Vista. Can I crash at your place tonight?

Most activity occurred in the first 24 hours of the hashtag's lifecycle and then trailed off over the next few days. After about two hours, the hashtag was placed on Twitter's top 10 trend list and stayed there for three–four hours, at which point it was replaced by other topics.

Using whitelisted Twitter accounts, we tracked the hashtag over the next four days and captured 17,803 tweets from 8,616 unique users. User participation followed a power law distribution where 6,021 users tweeted one time, 2,595 tweeted two or more times, and a dedicated 205 tweeted 10 or more times using the #robotpickuplines.

Spammers started to use the hashtag when it became a trending topic, and the spam lifecycle mirrored the meme lifecycle, with a slight lag. We subsequently collected the entire 1st degree network of the 8,616 users in the dataset, which contained all their followers and friends. Followers are users who follow a given user (indegree), and friends are users who the given user follows (outdegree). The 1st degree network contained 3,048,360 directed edges, 631,416 unique followers, and 715,198 unique friends. We used Network Workbench (Network Workbench Team, 2006) and GUESS (Adar, 2006), to analyze network properties and create the visualizations.

Identifying spammers

Some spam tweets can be easily identified because they include links to an external URL. In some cases, the links are transparent (*e.g.*, http://www.nomoredatingpigs.com); in others, they are masked by URL shorteners (*e.g.*, http://www.bit.ly/KLYbHo), and the burden is placed on host sites like http://www.bitly.com to help detect illegitimate links. Other patterns we have observed spammers exhibit are more than one hashtag on disparate topics, letter+number patterns in usernames, and suggestive keywords (*e.g.*, “naked”, “girls”, “webcam”).

We used a simple algorithm to detect spam in #robotpickuplines based on these properties: (1) searches for URLs; (2) username pattern matches; and, (3) keyword detection. We manually coded 300 randomly sampled tweets from #robotpickuplines as spam or not spam and ran this algorithm on the set. Our algorithm matched 91 percent of the time, with 27 missed spam tweets and 12 false positives. We use this baseline algorithm to mark all the tweets in #robotpickuplines.

Results

We examined behavioral patterns among spam accounts. Figure 1a shows traffic over the first 24 hours of use of the #robotpickuplines hashtag. The meme started at 11am CST and spiked around 3pm when it became a trending topic. It dropped around 4am and picked up again, although less heavily, the next day.

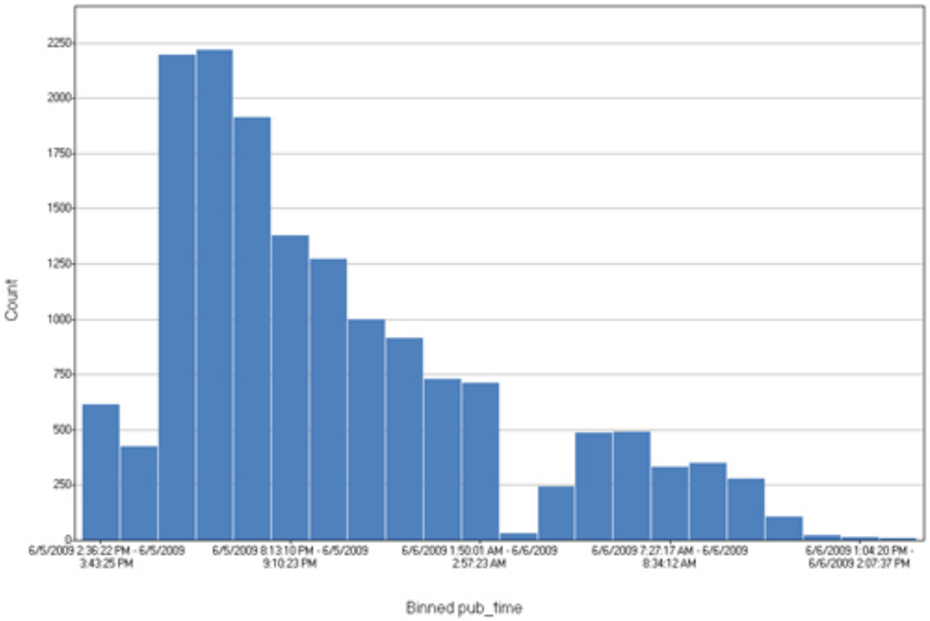

Figure 1a: 24–hour lifecycle of #robotpickuplines. Note: y–axis not to scale.

Figure 1b shows the spam tweets using #robotpickuplines. We find that 14 percent of tweets are spam, and spam lags slightly behind the meme trend. Spam picks up about five hours after @grantimahara’s first post and decays quickly. Figure 1b has been scaled up to compare to 1a, the raw number of spam tweets is actually 14 percent of legitimate accounts for this hashtag. The uptake slightly lags the meme itself. The drop–spike–drop pattern from 2–4am is unexpected; larger and more case studies may reveal broader trends.

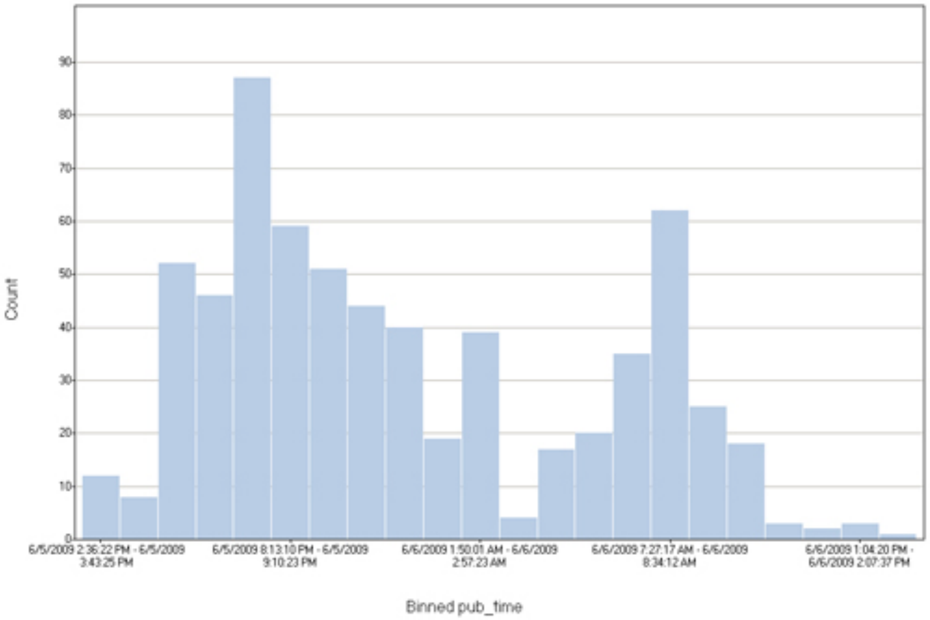

Figure 1b: 24–hour lifecycle of #robotpickuplines spammers. Note: y–axis not to scale.

Q1: Does age of account differ between spammers and legitimate users?

We anticipated that spam accounts would be newer than legitimate accounts because spammers can easily drop used accounts and create new ones; however, we found that they were not significantly different (legitimate: mean=258 days, stddev=170 days; spam: mean=269 days, stddev=128). The distribution of both spammers and legitimate users by age of accounts showed that most accounts were created 100–200 days ago, which maps to a date created between February–April 2009 (when Twitter was heavily publicized by celebrities and the media) (see Figures 2a and 2b).

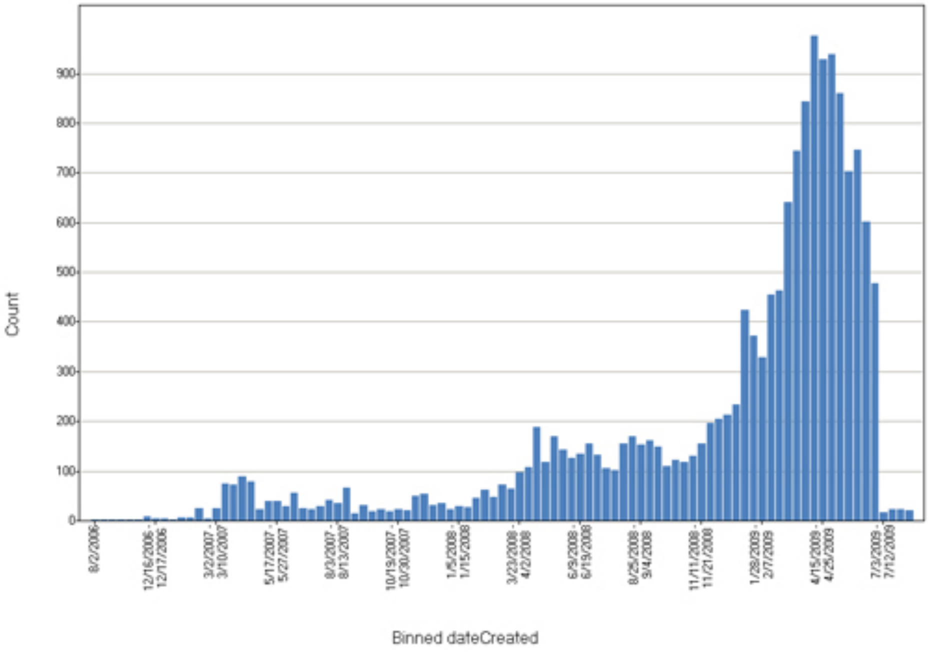

Figure 2a: Date that legitimate accounts were created.

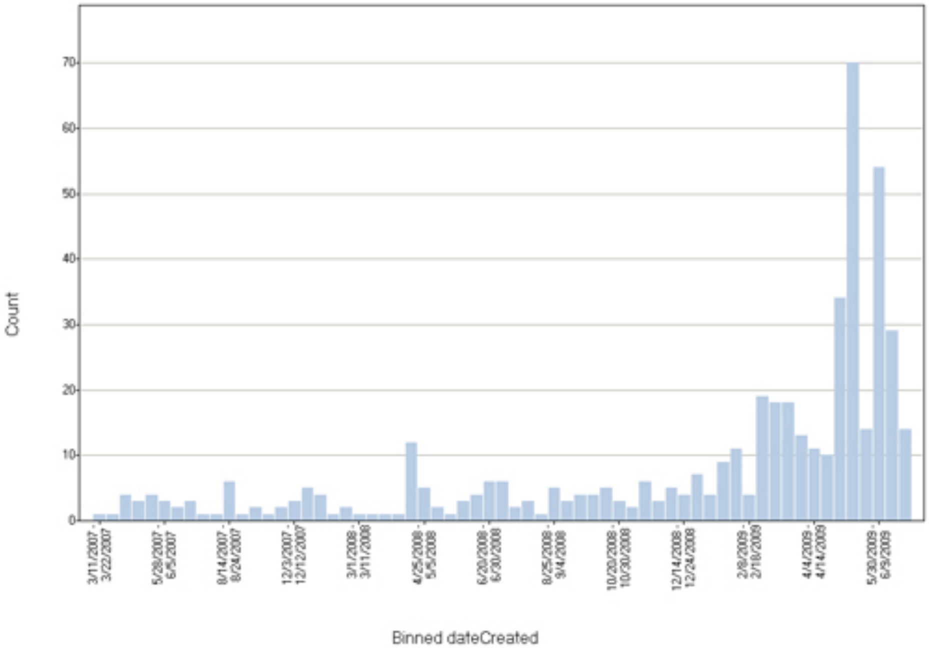

Figure 2b: Date that spam accounts were created.

Q2: Do spammers tweet more frequently than legitimate users?

We calculated the average distribution of tweets for each user by subtracting the date of first tweet from the date of last tweet and divided by total number of tweets. The average number of tweets per day was higher among spam accounts than legitimate accounts (legitimate: mean=6.7; spam: mean=8.66).

We also compared retweet and @reply behavior from legitimate accounts versus spammers. Results showed that 19 percent of legitimate tweets and 21 percent of spam tweets contained other hashtags within the set of #robotpickuplines tweets. Use of @replies was slightly higher with 26 percent legitimate uses and 24.8 percent spam uses. A chi–square test was performed and there was no significant difference in retweets or replies between spammers and legitimate users, but there was significant different in number of tweets $\chi^2(1,n=300)=4.464$, $\rho<0.05$, and use of hashtags $\chi^2(1,n=300)=3.847$, $\rho<0.05$.

Q3: Do spammers have more friends than followers?

We then examined structural properties of the network. We hypothesized that follower–to–friend ratio would be higher for legitimate accounts than for spammers because spambots may auto–follow Twitter users *en masse*. We calculated the ratio to be not significantly different (1.38 for legitimate, 1.12 for spammers). However, the total number of followers and friends for spammers was three times that of legitimate users (see Figures 3 and 4).

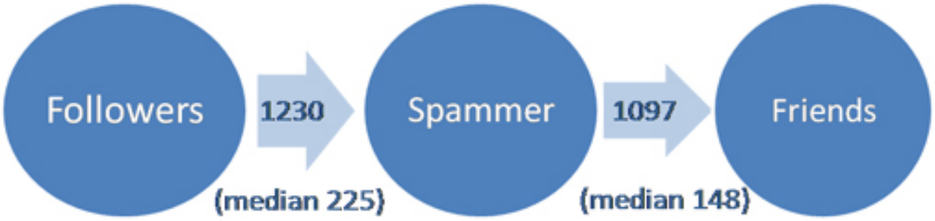

Figure 3: Average number of spam follower and friends.

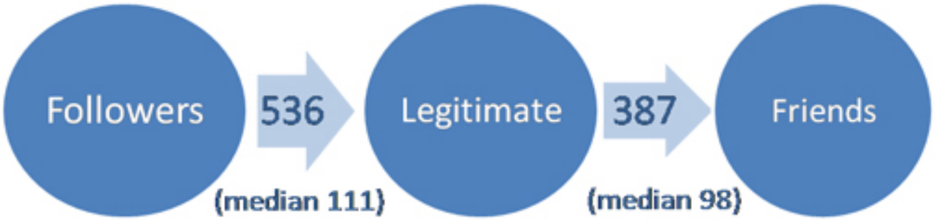

Figure 4: Number of legitimate follower and friends.

Q4: Are spammers clustered?

We examined the 2nd degree network of #robotpickuplines users for overlapping links. We examined two questions: 1) are there local clusters of users who might have learned about the meme from one another?; and, 2) are there local clusters of spammers? (*e.g.*, do spammers follow one another to boost their follower count?).

For both spammers and legitimate users, we find some overlap in follower and friend ties; 3,236 accounts (out of 8,616) followed one or more other people who tweeted about #robotpickuplines. These accounts had an average outdegree of 2.28. The outdegree distribution revealed that 1,718 followed only one other user who tweeted #robotpickuplines and 10 users had an outdegree of greater than 10. There was little variation between spammers and legitimate users in the 1st degree network distribution. However, the spam graph is relatively small; larger sample sizes might reveal differences in network structures.

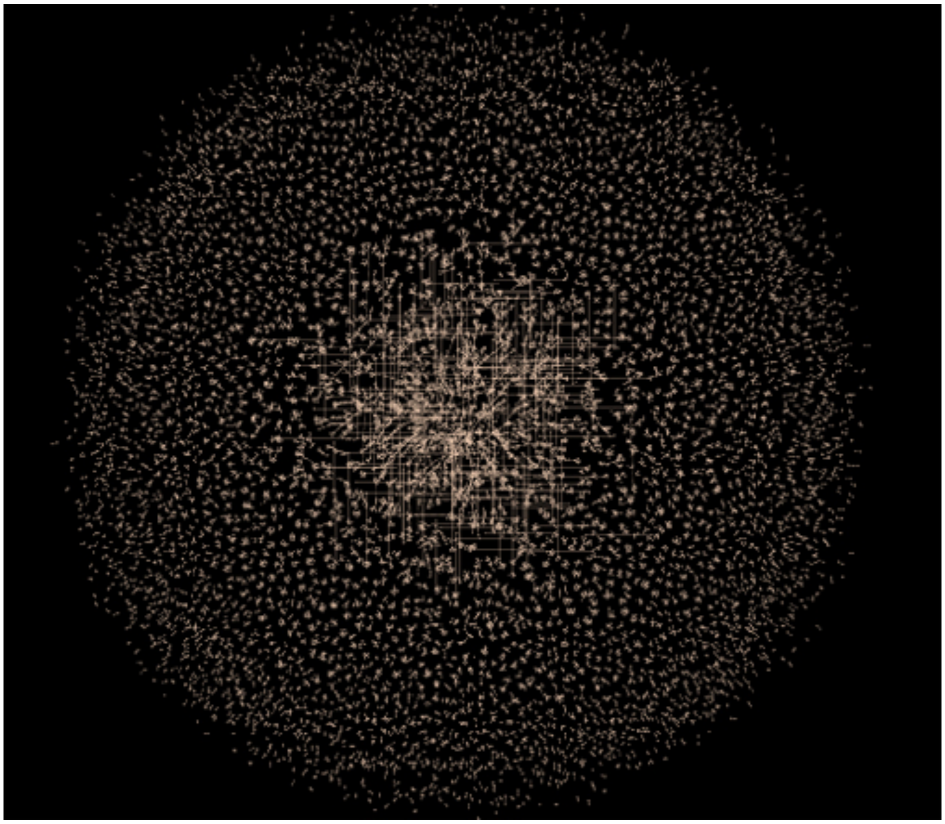

Figure 5a: #robotpickuplines network.

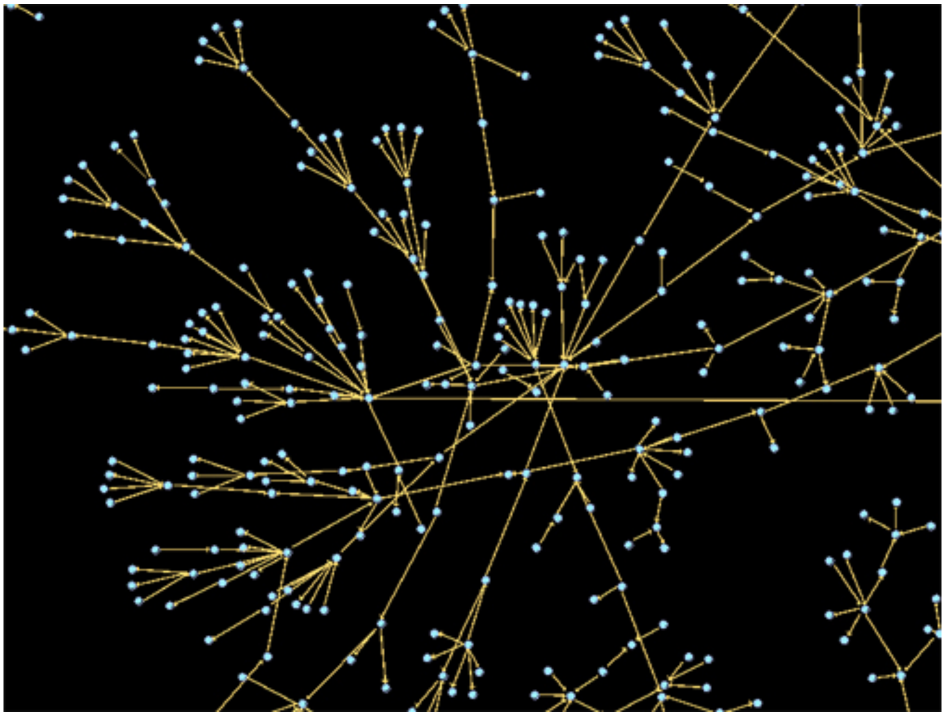

Figure 5b: More connected accounts.

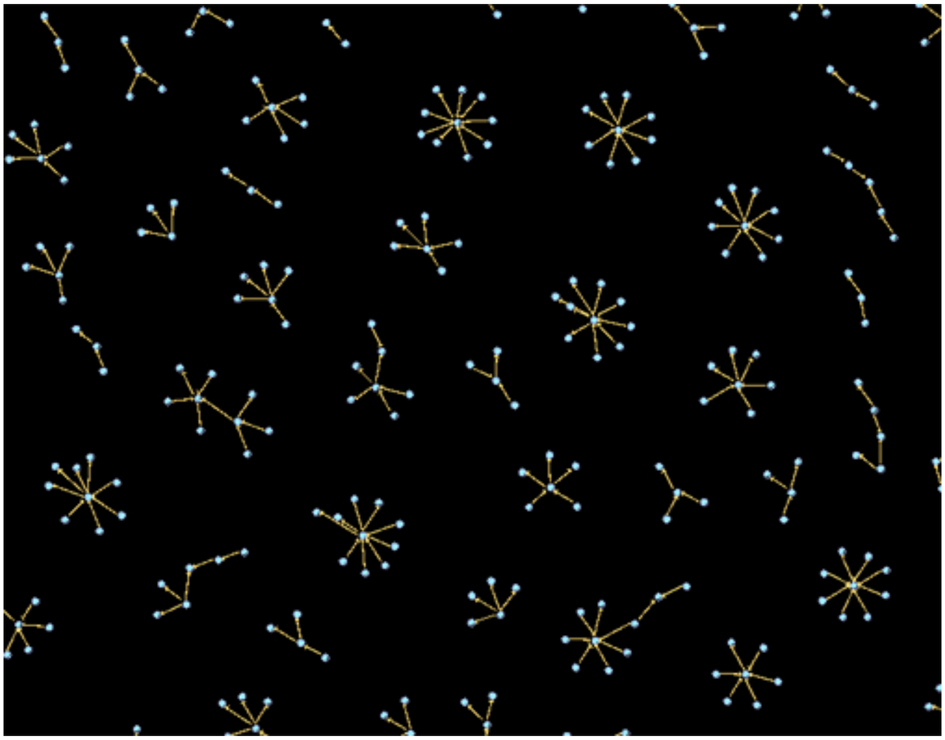

Figure 5c: Less connected accounts.

H5: Spammers have distinct structural properties compared to legitimate users

Last, we hypothesized that spammers would follow each other and legitimate "regular" users, and regular users would follow each other and celebrities (see Figure 5). To test this, we selected 100 random user IDs off the public timeline and took five backward hops (*e.g.*, picking a random node and clicking on one of her followers). After five hops, we reached a spammer 63 out of 100 times.

The corollary is that clicking on a random node's "friend" (a node she is following) will lead to a high–profile user, such as a celebrity, athlete, or politician. In other words, popularity and legitimacy are indicated by high indegree and spam is indicated by high outdegree. This link structure is similar to PageRank and is susceptible to many kinds of spam attack.

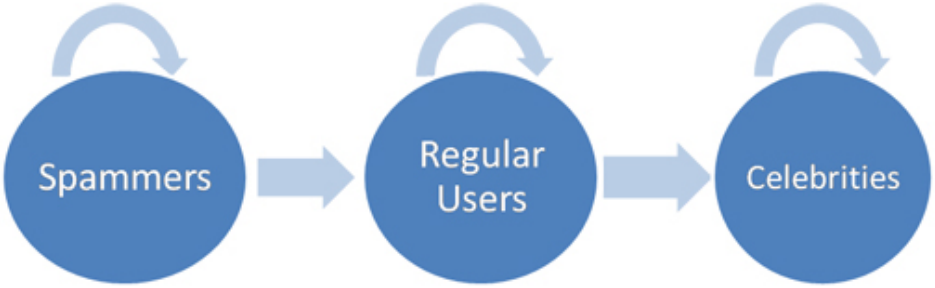

Figure 6: Network structure of spammers on Twitter.

Discussion

This paper is the first to examine Twitter memes and spam based on network and temporal properties. Twitter shares some properties with earlier communication mediums, but introduces new structural properties. We find that spam accounts are not significantly newer, they exhibit slightly higher retweet and @reply properties, they have a higher number of followers and friends, and indegree and outdegree may be associated with legitimacy.

These results have implications for detecting spam in Twitter as well as understanding behavioral properties of spammers in a directed network graph. First, spam accounts were not significantly newer than legitimate accounts for this dataset. This suggests that some spammers have been able to go undetected by Twitter for a long period of time. They can do this with spam campaigns that involve sending a few pieces of spam from any given account over a short period of time.

Spammers tweeted and replied slightly more frequently than legitimate users in our dataset. This may be explained by their strategies for flying under the radar by limiting the amount of content that is sent from any given account. Future work could examine frequency of tweeting behavior to explore if patterns are different between spammers and legitimate users. However, hashtag use was significantly higher among spammers. This result is not surprising; an informal view of trending topics on Twitter reveals spammers often use multiple unrelated hashtags as well as URLs in their tweets. More hashtags means their tweet will show up in more search results than it otherwise would, and therefore more people will see it and potentially click on a link.

While the ratio of friends to followers is not significantly different between legitimate users and spammers, the number of both is three times higher on average among spammers. This suggests that spammers invest a lot of time in following other users (and hoping other users follow them back). This again gives spammers an incentive to appear legitimate; users are unlikely to follow an account back if it looks like a spammer.

Finally, spammers are likely to be found at the edges of the Twitter graph rather than the center. They can follow many accounts and there is little downside in doing so because they are not actually reading the tweetstream that comes in from all the users they have followed. They are not likely to be at the center of the graph where celebrities like Oprah or Shaq, who have millions of followers, are located.

Limitations

Each of these characteristics of spammers suggests that there are structural properties that can be leveraged to detect and limit spam in Twitter. However, there are a number of challenges in designing spam detection applications. First, it is difficult to disambiguate spam from legitimate tweets without having a number of false positives. In many cases, users send tweets with multiple hashtags that are legitimate. Many users, such as marketers, tread a fine line between reasonable and spam–like behavior in their efforts to gather followers and attention. Second, marketers often have a high number of friends and followers, making it difficult to isolate spammers from this category of users.

Corporate accounts also exhibit this behavior. For example, in December 2009, Zappos.com (http://twitter.com/Zappos) had almost 1.6 million followers and followed back 400,000 of these users. Finally, this is a case study and the algorithm we used to detect spam specifically fitted the #robotpickuplines hashtag. In many cases, people share links and post sexual language legitimately, and URL detection alone will return a significant number of false positives. While the goal of this research is to produce broader patterns in network behavior of spammers on Twitter, more and larger studies are needed. Additionally, more work is needed to understand why people tweet and what their motivations are.

Modeling human behavior

Spammers can plan attacks by distributing tweets, managing frequency of tweets, leveraging hashtags, and managing following to follower ratios. All of this is not so different from what real Twitter users have to negotiate with their Twitter audiences on an ongoing basis. Are they talking too much? Are their tweets interesting enough? Are they following the right people? The problem of spam detection becomes a fundamentally human problem. Trust, reputation, and reciprocity are paramount.

Indeed, the number of followers has more social value than number of following, and we can impute some amount of status, importance, and attention from a Twitter user who has a high number of followers. Spammers may try to game the system by auto–following users then unfollowing them to invert their followers/friend ratio.

Behavioral and structural network approaches may offer more robust and resilient approaches to spam detection.

Future work

There are a number of open challenges in disambiguating legitimate human behavior from spam behavior on Twitter. Spammers can take a legitimate link and redirect it to their own pages, but legitimate users do this too. We have observed that users will change a bit.ly link to a particular reputable news article to their own bit.ly link to the same article to track how many people are clicking on their links.

Reciprocity, trust, and reputation can be manipulated by deceptive spammers. Phishing is a social engineering approach in which criminals masquerade as legitimate users by capturing access to their accounts (Dhamija, *et al.*, 2006). Phishes can systematically and temporarily cycle through user accounts on Twitter without staying long enough for the user to notice their account has been compromised.

Future work could further examine the relationships between trust, spam, and legitimate human behavior (*e.g.*, Cramer, *et al.*, 2009), and push more conversations about its role in design (*e.g.*, Preece, *et al.*, 2003). Finally, there are broader implications in this work. Behavioral and network approaches to spam detection are designed to detect spammers, but might have other kinds of applications, such as filtering noise from legitimate Twitter users who tweet just a little too much.

About the authors

Sarita Yardi is a PhD candidate in the School of Interactive Computing at Georgia Tech. Her research interest is in social computing with a specialization in social networks and technical capital.
Direct comments to syardi3 [at] gatech [dot] edu

Daniel M. Romero is a graduate student at the Center for Applied Mathematics of Cornell University. His research interest is in the empirical and theoretical analysis of social networks.
Direct comments to dmr239 [at] cornell [dot] edu

Grant Schoenebeck is a PhD candidate in theoretical computer science at University of California, Berkeley. His research interests are in complexity theory and the intersection of economics and computer science.
Direct comments to grant [at] cs [at] berkeley [dot] edu

danah boyd is a Social Media Researcher at Microsoft Research New England and a Fellow at Harvard University’s Berkman Center for Internet and Society. Her research examines social media, youth practices, tensions between public and private, social network sites, and other intersections between technology and society.
Direct comments to danah [at] microsoft [dot] com

Acknowledgements

The research for this article began at Microsoft Research New England. We thank MSR–NE for their support.

References

E. Adar, 2006. “GUESS: A language and interface for graph exploration,” *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Montréal), pp. 791–800; version at http://www.cond.org/guess1.html, accessed 30 December 2009.

D.S. Anderson, C. Fleizach, S. Savage, and G.M. Voelker, 2007. “Spamscatter: Characterizing

Internet scam hosting infrastructure," *Proceedings of the 16th USENIX Security Symposium*, pp. 135–148, and at http://www.usenix.org/events/sec07/tech/anderson.html, accessed 30 December 2009.

d. boyd, S. Golder, and G. Lotan, 2009. "Tweet, tweet, retweet: Conversational aspects of retweeting on Twitter," paper to be presented at the Hawaii International Conference on System Sciences (HICSS–43; 6 January 2010); version at http://research.microsoft.com/apps/pubs/default.aspx?id=102168, accessed 30 December 2009.

H.S.M. Cramer, V. Evers, M.W.v. Someren, and B.J. Wielinga, 2009. "Awareness, training and trust in interaction with adaptive spam filters," *Proceedings of the 27th International Conference on Human Factors in Computing Systems* (Boston), pp. 909–912.

R. Dhamija, J.D. Tygar, and M. Hearst, 2006. "Why phishing works," *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Montréal), pp. 581–590; version at http://people.seas.harvard.edu/~rachna/papers/why_phishing_works.pdf, accessed 30 December 2009.

N. Feamster, 2008. "Fighting spam, phishing, and online scams at the network level," *Proceedings of the 4th Asian Conference on Internet Engineering* (Bangkok), pp. 39–40.

C. Honeycutt and S.C. Herring, 2008. "Beyond microblogging: Conversation and collaboration via Twitter," *Proceedings of the 42nd Hawaii International Conference on System Sciences* (HICSS–42), pp. 1–10; version at http://ella.slis.indiana.edu/~herring/honeycutt.herring.2009.pdf, accessed 30 December 2009.

B.A. Huberman, D.M. Romero, and F. Wu, 2009. "Social networks that matter: Twitter under the microscope," *First Monday*, volume 14, number 1, at http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/2317/2063, accessed 30 December 2009.

A. Java, X. Song, T. Finin, and B. Tseng, 2007. "Why we twitter: Understanding microblogging usage and communities," *Proceedings of the 9th WebKDD and 1st SNA–KDD 2007 Workshop on Web Mining and Social Network Analysis* (San José, Calif.), pp. 56–65; version at http://ebiquity.umbc.edu/paper/html/id/367/Why-We-Twitter-Understanding-Microblogging-Usage-and-Communities, accessed 30 December 2009.

B. Krishnamurthy, P. Gill, and M. Arlitt, 2008. "A few chirps about Twitter," *Proceedings of the First Workshop on Online Social Networks* (Seattle), pp. 19–24.

J. Preece, J. Lazar, E. Churchill, H.d. Graaff, B. Friedman, and J. Konstan, 2003. "Spam, spam, spam, spam: How can we stop it," *Conference on Human Factors in Computing Systems* (CHI '03) extended abstracts, pp. 706–707; version at http://www.ifsm.umbc.edu/~preece/Papers/chi2003_panel_Soam.pdf, accessed 30 December 2009.

A. Ramachandran, N. Feamster, and S. Vempala, 2007. "Filtering spam with behavioral blacklisting," *Proceedings of the 14th ACM Conference on Computer and Communications Security* (Alexandria, Va.), pp. 342–351; version at http://www.ifsm.umbc.edu/~preece/Papers/chi2003_panel_Soam.pdf