Five incidents, one theme: Twitter spam as a weapon to drown voices of protest

John-Paul Verkamp Minaxi Gupta

School of Informatics and Computing, Indiana University
{verkampj, minaxi}@cs.indiana.edu

Abstract

Social networking sites, such as Twitter and Facebook, have become an impressive force in the modern world with user bases larger than many individual countries. With such influence, they have become important in the process of worldwide politics. Those seeking to be elected often use social networking accounts to promote their agendas while those opposing them may seek to either counter those views or drown them in a sea of noise. Building on previous work that analyzed a Russian event where Twitter spam was used as a vehicle to suppress political speech, we inspect five political events from 2011 and 2012: two related to China and one each from Syria, Russia, and Mexico. Each of these events revolved around popular Twitter hashtags which were inundated with spam tweets intended to overwhelm the original content.

We find that the nature of spam varies sufficiently across incidents such that generalizations are hard to draw. Also, spammers are evolving to mimic human activity closely. However, a common theme across all incidents was that the accounts used to send spam were registered in blocks and had automatically generated usernames. Our findings can be used to guide defense mechanisms to counter political spam on social networks.

1 Introduction

Social networks, such as Facebook, Twitter, and Google+, are an increasingly important part of the daily lives of billions of users. With politicians and other important figures increasingly reaching out to social networks to communicate, it is only to be expected that those with malicious intent would follow. Indeed, multiple sources such as [2, 5, 10, 12, 15] have shown how Twitter can be used to spread spam and malicious content. Others have shown how both legitimate and compromised accounts on social networks are manipulated to make spam campaigns more effective [1, 4, 11].

More to the topic of this paper, researchers have recently studied the use of Twitter spam as a vehicle to spread propaganda or to suppress political expression [7, 13]. Several examples of the latter have appeared in the news over the last couple of years. This includes attempts to suppress protests against the disputed Russian parliamentary elections [6]; suppression of information regarding the arrest of the Chinese artist Ai Weiwei [3]; attempts to deter Twitter users from learning about Tibet [9]; inundation of pro-revolution tweets from Arab Spring incidents [16]; and dilution of protests against the Mexican presidential candidate, Enrique Peña Nieto [8].

Recent work by Thomas et al. studied how twenty five thousand fraudulent Twitter accounts were marshaled to send hundreds of thousands of spam tweets in an attempt to disrupt political conversations following the announcement of Russia's parliamentary election results [13]. These accounts were drawn from a pool of over one million fraudulent accounts serving the *spam-as-a-service* market place. The authors found that fraudulent accounts were created using machines located all over the world. They logged on to Twitter from geographically diverse locations as well. In contrast, more than half of the legitimate accounts logged in only from Russia. Also, the IP addresses of almost 40% of machines posting spam appeared in blacklists, suggesting that they were already known to be compromised.

Given that political speech on Twitter has been suppressed on multiple other occasions, this paper is motivated by the desire to identify characteristics of diverse events from various countries. Doing so will enable us to judge whether it is possible to filter politically-motivated spam. This question is important because Thomas et al.'s work reported that only about half of the spam in the Russian incident was filtered by Twitter's existing spam filtering mechanisms.

Toward our goal, we analyze five different incidents spread over 14 months where political speech on Twitter was suppressed via spam. Two of these events are related to China and one each is related to Syria, Russia, and Mexico. Each of these events revolved around popular Twitter hashtags which were inundated with spam tweets intended to dilute their content. Of these, Thomas et al. studied the Russian incident. Overall, we confirm a few previously known behaviors and identify a few new ones. To our dismay, we find that spammer behaviors vary sufficiently across incidents such that generalizations are hard to draw. Further, we also find that spammers are evolving to become indistinguishable from legitimate users. These observations in turn imply that previous approaches–such as training supervised machine learning classifiers–are unlikely to be directly applicable and further research is needed to address the problem of politically motivated spam.

The key observations from comparing various incidents are the following:

- Spam tweets in three incidents follow a distinct spiking pattern. Spam in the two other incidents is either sustained or dwarfed by non-spam.
- Two of the incidents exhibit strong signs of scheduled activity. However, spammers in these incidents took care to mimic diurnal patterns typical of human activity, perhaps in order to escape detection.
- Non-spam tweets use more conjunctions and prepositions compared to spam tweets. However, this analysis is challenging for Chinese language tweets because of the lack of word breaks.
- In two incidents, URLs in tweets are less common than the baseline while in one other incident they are significantly more common.
- In two incidents, spammers target users directly using @ mentions. In the others, spammers rely primarily on hashtag popularity.
- Spam accounts are registered in blocks in each incident and the usernames used are automatically generated.
- Spammers are increasingly customizing account profiles in newer incidents while older incidents relied heavily on default profiles.

2 Methodology and data overview

We analyzed five different political events–two from China, one from Russia (previously analyzed by Thomas et al.

Incident	Dates	Primary hashtag	Interpretation	Expanded set of hashtags
Syria	1-13 April 2011	#syria	Syria	#syria, #bahrain, #egypt, #libya, #syria, #jan25 (Egypt), #feb14, #tahrir (Egypt), #yemen, #feb17 (Libya), #kuwait,
China'11	4-6 April 2011	#aiweiwei	Chinese artist, Ai Weiwei	#aiww, #aiweiwei, #cn417 (Jasmine), #5mao (5 May), #freeaiww, #freeaiweiwei, #cn424 (Jasmine), #tateaww, #cnjasmine
Russia	5-6 December 2011	#триумфальная	Triumphal Square in Moscow	#чп (abbr of Чрезвычайное Происшествие, extrodinary incident), #6дек (Dec 6),#5дек (Dec 5), #выборы (elections), #митинг (meeting), #триумфальная (Triumphal Square), #победазанами (victory is ours), #5dec, #навальный (surname, likely Navalny), #ridus
China'12	12-15 March 2012	#freetibet	Free Tibet	#tibet, #freetibet, #china, #monday, #西藏 (Tibet),
China'12	12-15 March 2012	#freetibet	Free Tibet	#tibet, #freetibet, #china, #monday, #西藏 (Tibet), #beijing, #shanghai, #india, #apple, #hongkong
Mexico	19-20 May 2012	#marchaAntiEPN	March against EPN (initials of presidential) candidate)	#marchaantiepn, #marchaantipeña, #marchamundialantiepn, #marchayosoy132 (I am 132nd to march), #votomatacopete (vote for another), #epn, #epnveracruznotequiere (no more EPN), #pr, #amlocomp (initials of competitor), #yosoy132,

Table 1: Dates and hashtags of interest for each of the five Twitter spam incidents considered (non-English hashtags translated where possible)

[13]), one from Syria, and one from Mexico–where Twitter spam is thought to have played a significant role in suppressing event-related tweets. In order to collect data for these incidents, we used a portion of Twitter's firehose data, which gave us a statistically valid sampling of an estimated one in ten tweets[1]. Although we have access to the full 10% for each time period, we begin by filtering out all but a single hashtag for each incident, gathered from news stories [3, 6, 8, 9, 16]. We refer to these hashtags as "seed" hashtags. The primary seed hashtag for each incident is shown in Table 1, along with the dates of each incident.

The seed hashtags are good starting points but do not paint a complete picture of the magnitude of each incident. Therefore, for each incident, we started by initializing a set seed hashtags $\mathcal{S}$ and collected all available tweets $\mathcal{T}$ involving any hashtag in $\mathcal{S}$. We then updated $\mathcal{S}$ to contain the top n most common hashtags in $\mathcal{T}$. We chose $n = 10$ hashtags to focus on the key hashtags related to each incident and also to avoid noise in our data set arising from irrelevant hashtags. We repeated the process of hashtag expansion until $\mathcal{S}$ stabilized for each individual incident. In each case, the algorithm took no more than three iterations. The final sets of hashtags are also shown in Table 1.

The spam tweets in the Syrian event started in the beginning of April 2011 and continued till the 13th of the month. The Chinese instances took place in early April 2011, and mid-March 2012. The Russian attack revolved around the election on December 5-6, 2011, while the Mexican event peaked on May 19-20 2012. In each case, we collected three weeks before and one after the incident in order to paint a more complete picture. The date expansion gives us a chance to determine traits such as how active both spam and non-spam accounts were in the time leading up to the attack, as first noted by Thomas et al. [13].

Once we collected the tweets involved in each incident, it was necessary to identify which tweets were legitimate and which were involved in spam campaigns. In order to do so, we used Twitter's built in spam detection facilities. Since we were looking at these events after they have occurred, we were able to query each user account and determine if it had been identified as a spam account or not. Further, we made the assumption that all tweets from spam accounts were spam and vice versa. While this does not account for the possibility of compromised accounts taking part in the attacks, we have yet to find any compromised account being used in any incident based on a manual inspection of spam accounts.

One final aspect that we investigated was account activity for both spam and non-spam accounts that did not directly involve one of the hashtags in our lists. We found that spam accounts in general had very few other tweets while legitimate accounts maintained a steady flow of other activity, peaking slightly during each incident. Since the behavior in each incident is the same when considering all tweets or only hashtag-related tweets, we use the former for our analysis.

Table 2 shows a summary of the numbers of spam and legitimate tweets and accounts involved in each incident. As this table shows, the five incidents varied widely. The Syrian incident had the most overall tweets but the percentage of spam tweets was significantly lower than for any other attack. In contrast, percentage of spam tweets in Russia, China'12, and Mexico were much higher, between 62-73%. When comparing spam accounts, we find that while both the percentage of spam tweets in China'12 and Mexico was high, the percentage of spam accounts was low. This implies that individual spam accounts in these incidents were far more active than legitimate accounts. In fact, we found that ten of the spam accounts in the China'12 incident each produced over 5,000 tweets before being shut down. In Mexico, 50 spam accounts produced a sustained 1,000 tweets per day throughout the incident. In both cases, automated detection of such events would do well to focus on finding and stopping prolific accounts. On the other hand, Russia employed the highest number of spam accounts but with relatively fewer tweets per account. In this instance, it would be necessary to detect individual tweets, since finding accounts will have only a marginal effect at best.

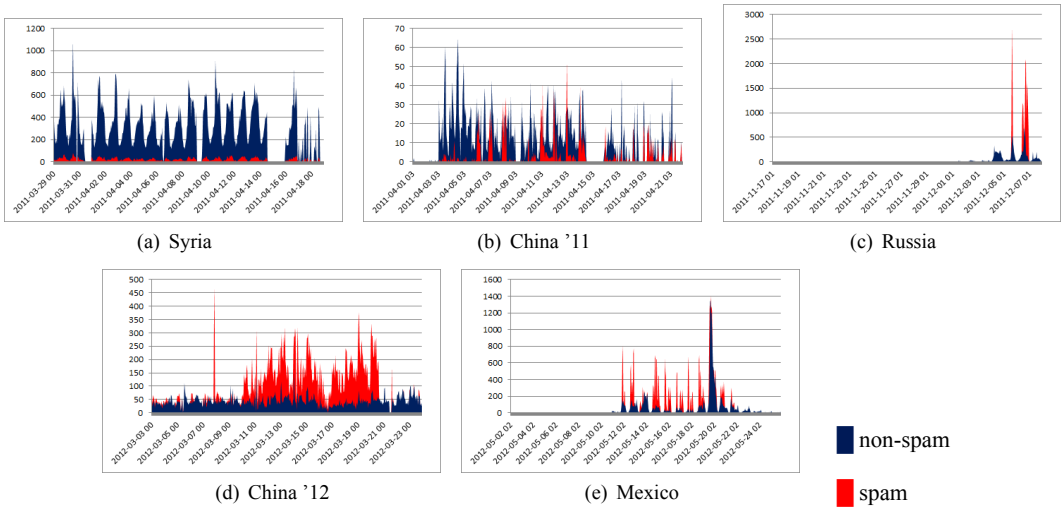
3 Analysis of tweets

Here, we analyze the various aspects of tweets from each of the five incidents and compare them along various dimensions. We also compare our findings to those in [13], where the authors conducted a thorough analysis of the Russian incident.

[1] Based on measured sample size versus that studied in [13]

Incident	Non-spam Tweets		Non-spam Accounts		Spam Tweets		Spam Accounts		Comments
Syria	**1,540,000**	(94%)	157,000	(98%)	107,000	(**6%**)	3,000	(2%)	Most overall tweets, smallest % spam tweets
China '11	58,000	(80%)	3,950	(88%)	15,000	(20%)	550	(12%)	Smallest attack, relatively low % spam
Russia	151,000	(31%)	12,000	(32%)	338,000	(**69%**)	**25,000**	(68%)	High % spam, highest number of spam accounts
China '12	227,000	(27%)	10,300	(86%)	600,000	(**73%**)	1,700	(14%)	Highest % spam, fewer high volume spam accounts
Mexico	306,000	(38%)	28,800	(90%)	498,000	(**62%**)	3,200	(10%)	High % spam, fewer high volume spam accounts

Table 2: Tweet and account statistics



(a) Syria (b) China '11 (c) Russia

(d) China '12 (e) Mexico

Figure 1: Volume of daily spam and non-spam tweets before, during, and after each incident

3.1 Daily tweet volume

We begin by analyzing daily tweet volume. Figure 1 shows the volume of tweets for each incident[2]. We plot the volume for a time period before, during, and after each incident date reported in the news articles. The first thing to note is that each incident possesses unique characteristics both in terms of relative volume of spam versus non-spam, as well as the duration of persistence of spam. Specifically, spam in the case of Russia, China'12, and Mexico has distinct peaks that dwarf non-spam. These are newer incidents, suggesting that censorship-related spam is getting more voluminous. In contrast, non-spam is much higher in volume with respect to spam in the context of Syria, suggesting that it was not successful in masking non-spam tweets. The China'11 incident is somewhat in between these two extremes, with spam and non-spam dominating on different occasions. We add that the Russian incident was the shortest lived of all, with spam drowning non-spam with 4-6 times the tweet volume. Our data on tweet volume for this event agrees with that reported by Thomas et al. [13]. We also note that in all but the Syrian incident, spam was longer lived than noted in the news articles. This highlights the potential inaccuracies in news reporting when matched with technical realities.

3.2 Timing of tweets

Next, we analyze if spam tweets show evidence of automation. We do so in two ways. First, in Figure 2, we show the volume of tweets per minute totaled over the course of the entire incident, such that the value for the X-axis label, 0:05-0:06, is the sum of all tweets occurring more than 5 but less than 6 minutes past any hour throughout the period of the incident. While the trend is less obvious in three cases, the Russian incident shows definite spikes at 5 and 15 minutes past the hour. Similarly, the Mexican event shows spikes occurring every fifteen minutes starting at the hour. Thus, both these incidents exhibit definite evidence of scripted behavior, most likely running at a specific time based on a cron job or the like. Automation in the context of censorship was also noted in [14], where Winter et al. described periodic scans of Tor relays.

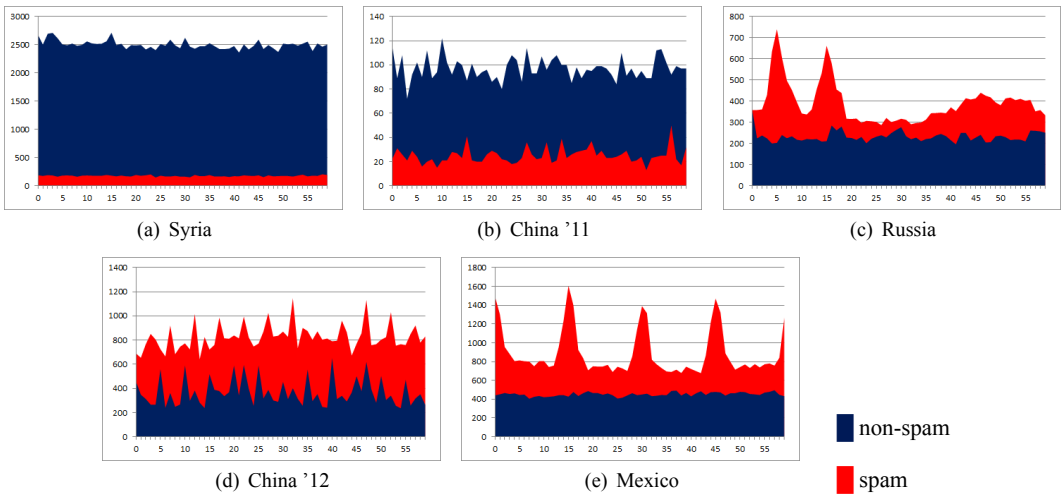
Next, we look at the hourly view, where we plot the volume of tweets per hour totaled over the course of the entire incident (Figure 3). We find that both spam tweets in Russian and Mexican events peak at the same time as regular traffic. Given that both these events show evidence of automation, this suggests that spammers took special care to mimic the diurnal patterns exhibited by human activity. None of the other events had a noteworthy correlation.

3.3 Tweet content

Here, we analyze the content of the tweets. There are four components of tweet text: words, hashtags (start with a #), URLs, and mentions (@ and then a username). Treating the combination of all of these as a bag of words, we first look at the top-10 most popular items in spam and non-spam tweets. Unlike the common practice in natural language processing, we kept stop words for this analysis because a lack of stop words indicates that these tweets are not a normal conversation.

The analysis of top-10 most popular items revealed interesting aspects (Table 3). First, retweeting, for which many Twitter clients insert an *rt* in the tweets, is more popular in non-spam tweets than spam for Russian, China'12, and Mexican incidents (which would almost certainly not be using typical clients but rather automated tools). However, this is not true for Syrian and China'11, where both spam and

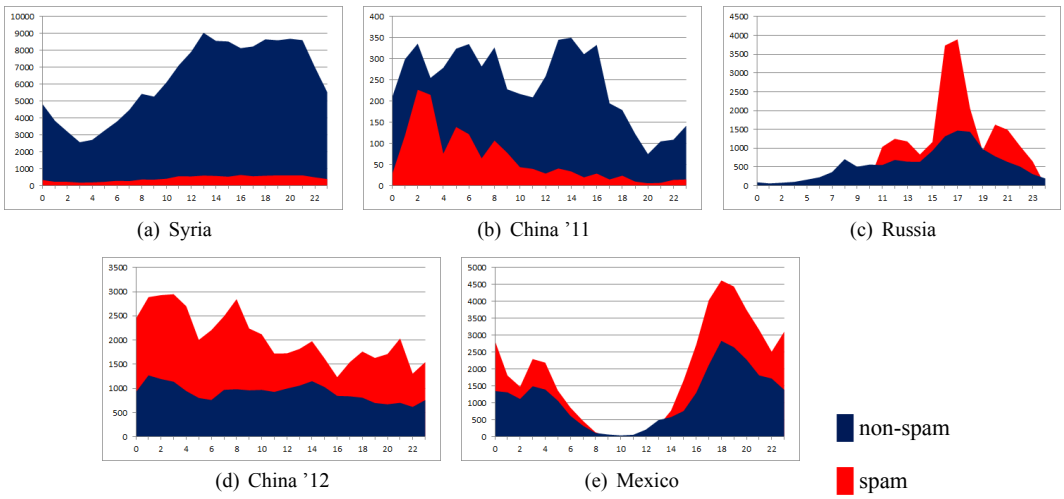
[2]There is an interruption in data collection during the Syria and China '11 incidents.



(a) Syria (b) China ’11 (c) Russia

(d) China ’12 (e) Mexico

Figure 2: Tweet volume per minute



(a) Syria (b) China ’11 (c) Russia

(d) China ’12 (e) Mexico

Figure 3: Tweet volume per hour (all times in UTC)

non-spam have retweets in the top-10. Hashtags are popular both in spam and non-spam tweets across all incidents and they are relevant to individual incidents in all cases. However, for Russia, the hashtags around which the data was collected are not important enough to qualify for the top ten positions, particularly in the non-spam cases.

Spam and non-spam in the Syrian incident are similar, in that both have the same top-10 items, albeit with minor shuffling in order. Also interesting is the presence of two URLs in the top-10 list for China’11. These are for URL shortening services. Upon following the chains of redirects, each URL leads to a product placement web page. Given that there is a sustained increase in spam traffic for this incident, exceeding 90% of the entire traffic at its peak, and that spam seems to last longer than the activity related to the incident, this suggests that a portion of spam in this case may simply be an artifact of spammers latching on to a popular hashtag. However, the other popular items in tweets suggest that the rest of the spam is related to the incident.

The China’11 and China’12 incidents had a couple of peculiarities with respect to the other incidents. For one, both spam and non-spam in these incidents made very little use of propositions and conjunctions (the aforementioned stop words). This is unfortunately a matter of the difficulty in parsing Chinese language tweets. Without spaces to act as delimiters, it is non-trivial to pull apart words other than where hashtags are present (which require spaces). In the other cases though, stop words commanded many of the top-10 spots for spam as well as non-spam for all other incidents. Further, only in the two Chinese incidents did *mentions* (of another account or person, tagged by an ’@’) feature in the top-10. No other incident had any mentions in top-10 spots. However, while three mentions featured in China’12 for spam, one featured in China’11 for non-spam, suggesting that just like other features, mentions can also not be used to distinguish such spam from non-spam.

	URLs		Mentions		Retweets	
Incident	Spam	Non	Spam	Non	Spam	Non
Syria	**41.0%**	**96.4%**	59.1%	60.4%	44.2%	45.2%
China ’11	**58.8%**	**36.2%**	69.7%	68.3%	**3.3%**	**29.8%**
Russia	**2.8%**	**36.8%**	**4.2%**	**54.6%**	**3.1%**	**35.8%**
China ’12	60.6%	64.5%	**81.3%**	**36.4%**	**0.2%**	**13.7%**
Mexico	**1.0%**	**32.8%**	**1.9%**	**80.7%**	**1.6%**	**68.9%**

Table 4: Percentage of tweets with URLs, mentions, and retweets

Syria	*Spam*:	rt, #bahrain, #egypt, #libya, the, in, #syria, to, في (in), of
	Non-spam:	rt, #egypt, #bahrain, #libya, the, in, #syria, في (in), to, من (of)
China '11	*Spam*:	#aiww, rt, #5mao (May 5), #cn417, 艾未未的童话涉嫌抄袭 (headline about Ai Weiwei), url_1, #cn424, url_2, #aiweiwei, #china
	Non-spam:	rt, #aiww, #aiweiwei, #cn417, ai, @aiww, #freeaiww, #5mao, the, #freeaiweiwei
Russia	*Spam*:	на (on), #победазанами (victory is ours), не (no), #чп, и (and), #выборы (elections), в (in), #6дек (Dec. 6), я (I), площади (areas)
	Non-spam:	#выборы, rt, в, на, #чп, и, не (not), за (for), с (with), #митинг (meeting)
China '12	*Spam*:	#tibet, #freetibet, @degewa, @tibet, #西藏 (#tibet), #degewa, #china, and, @sfchoi8964, #315
	Non-spam:	#china, #tibet, rt, in, #beijing, #shanghai, the, to, #hongkong, #freetibet
Mexico	*Spam*:	#marchaantiepn, marcha (march), la (the), de (of), anti, epn (initials), i, rt, #marchaantipeña, marchaantiepn
	Non-spam:	#marchaantiepn, la, rt, de, a, en (in), no, el (the), que (that), y (and)

Table 3: Top-10 items in spam and non-spam tweet text (meaning of non-English words is in brackets)

The next aspect we consider is the meta information contained in each tweet. This includes mentioning other users by their username; using hashtags; using URLs linking to external content; and using retweets or replies. Since the incidents are characterized by hashtags, the presence of a hashtag is not an interesting feature; however, all the other features are. Our findings are summarized in Table 4. We find that significantly fewer spam tweets in Russia and Mexico use mentions than non-spam tweets. In contrast, more spam tweets use mentions than non-spam tweets in China'12, contradicting the trend. The other two incidents varied little in this regard.

The URLs are a similar story. More non-spam tweets in Syria, Russia, and Mexico contain a URL compared to spam tweets. In both Russia and Mexico, the URLs often eventually land on news articles, although interestingly they were not related to the elections in either case. This possibly implies that the spammers were using said sources to create semi-legitamate looking Tweets. However, the opposite is true for China'11, where more spam tweets contain a URL–although this is likely a case of product placement as previously mentioned. Spam and non-spam tweets in China'12 varied little. Perhaps a strong trend is exhibited in retweets, where all but the Syrian incident revealed that non-spam tweets were more likely to use this feature compared to spam. Replies were too infrequently used by both spam and non-spam tweets to have significant differences.

3.4 Tweet recipients

Here, we compare the tweet recipients for spam and non-spam accounts for each incident. In general, we note that both spam and non-spam accounts tend to average very few followers, with the majority of followers concentrated in only a few accounts. As such, spammers cannot rely on followers in order to target their campaigns and must rely instead on either mentions or popular hashtags.

For the case of mentions, any user who is mentioned in a tweet (by an @ then their username) will be notified as such. This allows for direct targeting of users and has been successfully used for spam campaigns in the past. However, this does not appear to uniformly be the case among the incidents studied. In Syria, China'11, and China'12, this is the case where 60-80% of spam tweets contain a mention (as shown in Table 4), however Table 5 shows that in Syria and China'12 this may be at least partially related to creating a reasonable fake identity. This table shows how often the spam accounts mention other spam accounts, building a sort of network. However, the interesting case is that of China'11, where a high percentage of those mentions (77.5%) target 'other' accounts, ones which do not use any hashtags we are studying throughout the event (and are thus not otherwise a part of our data set). This at least partially supports the conclusion that some of the traffic captured in China'11 is opportunistic spam, attempting to sell products rather than merely flooding the hashtags.

	@non-spam	@spam	neither
Syria	4.7%	78.3%	17.0%
China '11	1.1%	**21.5%**	**77.5%**
Russia	**10.7%**	63.8%	25.4%
China '12	0.7%	75.0%	24.3%
Mexico	4.8%	51.6%	43.6%

Table 5: Social graph implied by mentions

Conversely, however, both in Russia and Mexico, spam accounts use very few mentions (4.2% and 1.9% respectively; see Table 4). This supports the idea that in neither incident are the spam accounts attempting to spam in a targeted manner, but are rather attempting to flood the entire hashtag. Nevertheless, when spam accounts do mention other users, the highest percentage is other spam accounts.

4 Analysis of accounts

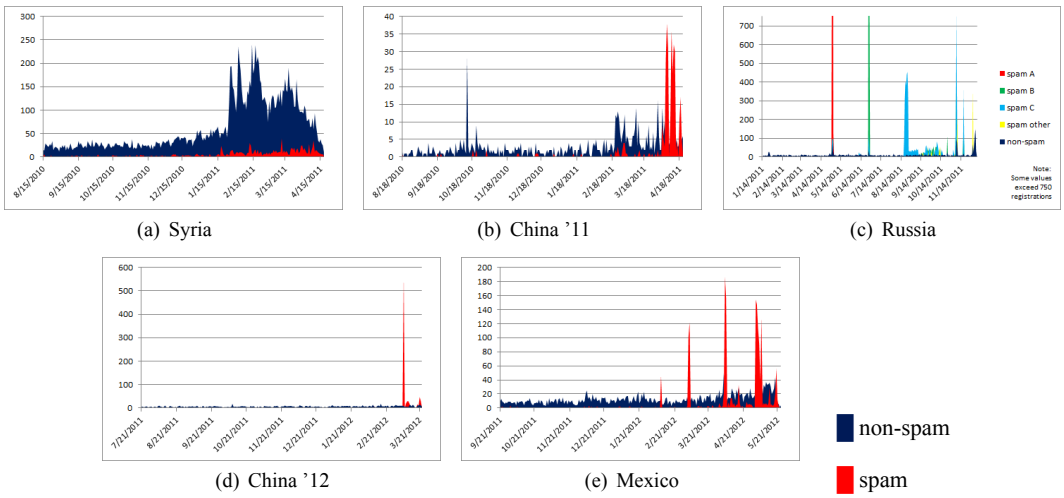
Now we analyze the various aspects of accounts used in each of the five incidents and compare them along various dimensions. We again compare our findings to those in [13], where the authors conducted a thorough analysis of the Russian incident.

4.1 Account registrations and usernames

We begin by examining the registration dates for the spam and non-spam accounts. Figure 4 shows the registration dates of all accounts involved in each of the five incidents. We note that in all cases but Syria, spam accounts were registered in blocks while non-spam accounts were not. In fact, for Russia and Mexico there were multiple registration blocks for spam accounts. For Syria, neither type of accounts were registered in blocks.

We note that block registrations are in general not easily correlated if the machines registering the accounts are geographically spread out. In order to check if this is the case, we would need IP addresses of machines doing registrations, as Thomas et al. [13] did. Indeed, they found that non-spam accounts were primarily registered from machines in Russia but spam accounts were registered using machines all over the world. Unfortunately, we cannot investigate how the other four incidents compared with this observation.

In the lack of IP addresses of machines used to register accounts, we look at the account names for different blocks of spam accounts. First, we note that almost all of the accounts in each incident have usernames that appear to be generated in origin. The generating algorithms are different in each case (as shown in Table 6); however they share several common features. In China '12 and Mexico, a vast majority of usernames for spam accounts take the form of *{name}{name}{number}* or *{name}{name}{random}* where



(a) Syria (b) China '11 (c) Russia

(d) China '12 (e) Mexico

Figure 4: Registration dates for spam and non-spam accounts for each incident

the first two parts are common given and family names. Approximately 85% of them are exactly 15 characters in length, the maximum allowed by Twitter. In both cases, usernames that would otherwise be shorter are padded–with both letters and digits in the case of China '12 and with only digits in the case of Mexico.

In Russia, the spam account user names are of the type *{initial}{name}* or *{name}{name}*. Further, unlike Russia and Mexico, where the names are indicative of popular names in the region of interest, names in China'12 are simply western names. The patterns are less obvious for Syria and China'11 though a human eye can tell that they are machine generated. Many end in numbers. Though Thomas et al. reverse engineered account names in the case of Russia, they deemed the algorithm sensitive and did not reveal it. Due to this reason, we can only state that they also found evidence of automatically generated account names for Russia.

Incident	Example usernames
Syria	*Often end in numbers, patterns less common* zuhair77, GC814, walidraafat, ToQiiiZ, GeorgiaKillick0, libyana1702, Bahraini61, ScottsdaleReb, Updates2424
China '11	*Often end in numbers, patterns less common* cnjs2, cnjs5, cnjs10, cnjs11, cnjs12 cxbenben113, dabenben222, huashengdun111, huashengdun203
Russia	*Most are {name}{name} or {initial}{name}* SScheglov, SSchelkachev, SSchelkonogov, SSchelokov, SSchemilov, SScherbakov, SShabalin, SShabarshin,
China '12	*Most are {name}{name}{random/number}, max length* LanelleL4nelle6, LanieSl1dek1103, LarondaGuererro, LatanyaZummoMNS, LatarshaWeed181, LauraHelgerm1nV
Mexico	*Most are {name}{name}{number}, max length* AnaAvil58972814, AnaAvil76571383, AnaLope95971326 AnaRive02382949, AnaSuar79305176, AnaSuar83449134

Table 6: Example usernames for spam accounts used

4.2 Default profile and profile image

Next, we look at the profiles of spam and non-spam profiles. The second column in Table 7 shows the percentage of spam and non-spam profiles that use the default profile for their Twitter accounts. We find that for China'11, Russia, and China'12, a significantly higher percentage of spam accounts use the default profile. The difference is insignificant for Syria but reverse for Mexico, where a higher percentage of spammers change the default profile.

Likewise, there is an interesting case when looking at the percentage of accounts using the default image (not counted towards the default profile). China'11 and Russia, where many spam accounts used the default profile, did not show strong evidence of using the default image. However, a higher percentage of spam accounts in China'12 used the default image along with the default profile. Mexico and Syria did not have any noteworthy trends.

	Default profile		Default image	
Incident	Spam	Non-spam	Spam	Non-spam
Syria	46.2%	42.9%	9.4%	6.0%
China '11	**89.4%**	**51.2%**	12.3%	12.6%
Russia	57.8%	34.7%	7.8%	11.1%
China '12	**95.1%**	**47.8%**	**59.0%**	**11.8%**
Mexico	**1.7%**	**27.0%**	0.6%	3.0%

Table 7: Usage of default profile and image

5 Conclusion

We analyzed five political events around which Twitter hashtags related to the events were inundated with spam tweets from politically motivated entities. Unfortunately, things varied sufficiently across incidents that drawing common themes around spam tweets and accounts did not seem promising. This was especially true due to spammer evolution, which seems to be geared toward mimicking human activity closely.

A promising defense avenue could be built around account registrations and usernames, where we found that spam accounts in each incident were registered en masse and in advance and used usernames that could be reverse engineered toward detection purposes. Further work is needed to explore the feasibility of this approach, however.

6 Acknowledgments

We thank Fil Menczer and the Center for Complex Networks and System Research (CNetS) at Indiana University for providing us access to the Twitter streaming API data through their Truthy project. The work in this paper is supported by the National Science Foundation (NSF) under Grant Number OCI-1127406. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

References

[1] Bilge, L., Strufe, T., Balzarotti, B., and Kirda, K. All your contacts are belong to us: Automated identity theft attacks on social networks. In *International Conference on World Wide Web (WWW)* (Apr. 2009).

[2] Cao, Y., Yegneswaran, V., Porras, P., and Chen, Y. PathCutter: severing the self-propagation path of XSS JavaScript worms in social web networks. In *Network and Distributed System Security Symposium (NDSS)* (Feb. 2012).

[3] Chayka, K. Chinese Twitter bots spam aiweiwei hashtag. http://hyperallergic.com/23184/chinese-twitter-spam/, Apr. 2011.

[4] Gao, H., Hu, J., Wilson, C., Li, Z., Chen, Y., and Zhao, B. Y. Detecting and characterizing social spam campaigns. In *ACM SIGCOMM Internet Measurement Conference (IMC)* (Nov. 2010).

[5] Grier, C., Thomas, K., Paxson, V., and Zhang, M. @spam: The underground on 140 characters or less. In *ACM Conference on Computer and Communications Security (CCS)* (Oct. 2010).

[6] Krebs, B. Twitter bots drown out anti-Kremlin tweets. http://krebsonsecurity.com/2011/12/twitter-bots-drown-out-anti-kremlin-tweets/, Dec. 2011.

[7] Lumezanu, C., Feamster, N., and Klein, H. #bias: Measuring the tweeting behavior of propagandists. In *AAAI Conference on Weblogs and Social Media* (June 2012).

[8] Rueda, M. Mexico: Twitterbots sabotage anti-PRI protest. http://univisionnews.tumblr.com/post/23287767289/twitterbots-attack-anti-pri-protest-mexico, May 2012.

[9] Segal, A. China's twitter-spam war against pro-Tibet activists. http://www.theatlantic.com/international/archive/2012/03/chinas-twitter-spam-war-against-pro-tibet-activists/254975/, Mar. 2012.

[10] Sridharan, V., Shankar, V., and Gupta, M. Twitter games: How successful spammers pick targets. In *Annual Computer Security Applications Conference (ACSAC)* (Sept. 2012).

[11] Stringhini, G., Egele, M., Kruegel, C., and Vigna, G. Poultry markets: on the underground economy of twitter followers. In *ACM Workshop on Workshop on Online Social Networks (WOSN)* (Aug. 2012).

[12] Thomas, K., Grier, C., Ma, M., Paxson, V., and Song, D. Design and evaluation of a real-time URL spam filtering service. In *IEEE Symposium on Security and Privacy (SP)* (May 2011).

[13] Thomas, K., Grier, C., and Paxson, V. Adapting social spam infrastructure for political censorship. In *USENIX Workshop on Large-Scale Exploits and Emergent Threats (LEET)* (Aug. 2012).

[14] Winter, P., and Lindskog, S. How China is blocking tor. *USENIX Workshop on Free and Open Communications on the Internet (FOCI)* (Aug. 2012).

[15] Yang, C., Harkreader, R., Zhang, J., Shin, S., and Gu, G. Analyzing spammer's social networks for fun and profit. In *International Conference on World Wide Web (WWW)* (Apr. 2012).

[16] York, J. C. Syria's Twitter spambots. http://www.guardian.co.uk/commentisfree/2011/apr/21/syria-twitter-spambots-pro-revolution, Apr. 2011.