Adapting Social Spam Infrastructure for Political Censorship

Kurt Thomas[*], *Chris Grier*[*], *Vern Paxson*[*†]
[*]*University of California, Berkeley* [†]*International Computer Science Institute*
{*kthomas, grier, vern*}@*cs.berkeley.edu*

Abstract

As social networks emerge as an important tool for political engagement and dissent, services including Twitter and Facebook have become regular targets of censorship. In the past, nation states have exerted their control over Internet access to outright block connections to social media during times of political upheaval. Parties without such capabilities may however still desire to control political expression. A striking example of such manipulation recently occurred on Twitter when an unknown attacker leveraged 25,860 fraudulent accounts to send 440,793 tweets in an attempt to disrupt political conversations following the announcement of Russia's parliamentary election results.

In this paper, we undertake an in-depth analysis of the infrastructure and accounts that facilitated the attack. We find that miscreants leveraged the spam-as-a-service market to acquire thousands of fraudulent accounts which they used in conjunction with compromised hosts located around the globe to flood out political messages. Our findings demonstrate how malicious parties can adapt the services and techniques traditionally used by spammers to other forms of attack, including censorship. Despite the complexity of the attack, we show how Twitter's relevance-based search helped mitigate the attack's impact on users searching for information regarding the Russian election.

1 Introduction

In recent years social networks have emerged as a significant tool for both political discussion and dissent. Salient examples include the use of Twitter, Facebook, and Google+ as a medium for connecting United States government officials with citizens to drive public discourse [6, 16, 17]. The Arab Spring that swept over the Middle-East also embraced Twitter and Facebook as a tool for organization [20], while Mexicans have adopted social media as a means to communicate about violence at the hands of drug cartels in the absence of official news reports [9]. Yet, the response to the growing importance of social networks in some countries has been chilling, with the United Kingdom threatening to ban users from Facebook and Twitter in response to rioting in London [11] and Egypt blacking out Internet and cell phone coverage during its political upheaval [18].

While nation states can exert their control over Internet access to outright block connections to social media [28], parties without such capabilities may still desire to control political expression. An example of this recently occurred on Twitter during protests tied to Russia's parliamentary elections [5]. The protests began in Moscow's Triumfalnaya Square and quickly moved online as both pro-Kremlin and anti-Kremlin parties posted to Twitter to express their opinions on Russia's election outcome. In response to these discussions, a wave of bots swarmed the hashtags that legitimate users were using to communicate in an attempt to control the conversation and stifle search results related to the election [13]. This attack highlights the possibility of manipulating social networks for partisan goals through the nefarious use of sybil accounts, sidestepping any requirement for controlling Internet access.

In this paper we present an in-depth analysis of how unknown parties attempted to control the political conversations surrounding Russia's disputed election. We examine the accounts and infrastructure the attackers relied upon, as well as the impact of their efforts on Twitter users searching for information pertaining to the election and protests. While previous researchers have explored the potential of using posts from sybil accounts to skew product ratings and to generate fake content [12, 15, 27], we show that the attackers specifically adapted spam infrastructure to manipulate political speech. These events demonstrate that malicious parties are now using the *spam-as-a-service* marketplace that has emerged for social networks [23] for multiple ends beyond spam.

The attack consisted of 25,860 fraudulent Twitter accounts used to inject 440,793 tweets into legitimate conversations about the election. We find evidence that these accounts originated from a pool of 975,283 fraudulent accounts, 80% of which remain dormant in preparation for use in future spam campaigns. We contrast the geolocation of logins for legitimate users and those of bots, finding that 56% of logins tied to users discussing the Russian election were located in Russia, compared to just 1% of spam accounts. Equally striking, the attack relied on machines distributed across the globe, 39% of which appear in IP blacklists, a strong indicator that the miscreants involved relied on compromised hosts.

Despite the volume of traffic generated by the attack, its impact was partially mitigated by relevance rankings integrated in search results that aim to filter out spam tweets. On average, search results that used relevance metrics returned 53% fewer bot-generated tweets. These techniques highlight how personalized search results can defend against censorship-based attacks, even in the presence of thousands of fake accounts.

In summary, we frame our contribution as follows:

- We present an in-depth analysis of the profiles, tweets, login behavior, and social graph of accounts attempting to censor political discussion.
- We explore the infrastructure required to carry out such an attack, finding that spam services were repurposed to enable censorship.
- We characterize the impact of the attack on legitimate users searching for information regarding the election and protests.

2 Background

In this section, we outline how an attacker can attempt to censor Twitter through the use of fraudulent accounts. We also discuss existing defenses for detecting abusive accounts, as well as previous research into how attackers can purchase tweets or accounts through spam-as-a-service marketplaces. While spammers traditionally use these underground markets to acquire resources for spam campaigns, malicious parties can easily adapt these services for other forms of attacks.

2.1 Diluting Hashtag Content

Hashtags have emerged on Twitter as a mechanism for organizing conversations around topics that occur outside a user's social graph. As a result, users can view global and local trends that capture popular conversations or use Twitter's search functionality for more nuanced queries. Because any user can embed a hashtag in their tweets, conversations are susceptible to a specific attack we term *message dilution*. In the attack, automated sybil accounts post conflicting or incomprehensible content with hashtags used by legitimate users, effectively hijacking the previous conversation. Similar scenarios arise outside of Twitter, such as when fraudulent accounts generate fake reviews to skew rating systems [12], marketers masquerade as fans of a product to sway public opinion [8], or political parties engage in "astroturfing." In the absence of a curator or relevance-based search, sybil accounts can simply outproduce content compared to legitimate users in order to reshape a hashtag's meaning or to bury relevant information.

2.2 Detecting Social Network Abuse

The challenge of identifying sybil accounts that participate in a message dilution attack is akin to detecting spam and abusive behavior. A wide variety of strategies to this end have appeared, which include analyzing the social graph of sybil accounts [4, 29], characterizing the arrival rate and distribution of posts [7], analyzing statistical properties of account profiles [1, 21], and detecting spam URLs posted by accounts [22].

For the purposes of this paper, we rely on Twitter's internal spam detection algorithm. While the implementation of this algorithm is not published, the system's rules target the frequent formation of relationships, posting duplicate content, frequently posting to multiple hashtags, and posting irrelevant or misleading content to trending topics [25]. While the system is imperfect, we discuss our technique for correcting the possibility of false negatives in the context of our study in Section 3.

2.3 Spam-as-a-Service

In response to the challenge of generating fraudulent accounts and reviews, a number of semi-legitimate and illegitimate services have appeared that provide content generation, favorable reviews, and user registrations as a service [15, 27]. Similarly, an underground marketplace has emerged that facilitates purchasing social network accounts and URL advertisements [23]. These *spam-as-a-service* marketplaces allow parties to sell and trade their resources, which include lists of email addresses, fraudulent accounts, network proxies, and compromised hosts. As we will show, these services are not just limited to spammers; the attack carried out on Twitter relied on access to thousands of compromised machines and accounts likely purchased in bulk from spam-as-a-service markets.

3 Methodology

Before analyzing the attack, we discuss our technique for identifying automated accounts that posted to twenty distinct topics pertaining to the Russian election between December 5–6, 2011. In total, 46,846 accounts participated in discussions of the disputed election results, 25,860 of which we identify as bots. Although these accounts do not fit with traditional views of spam where an account advertises a product or scam, we refer to these accounts as *spam accounts* in this paper. The other accounts were legitimate users on both sides of the political spectrum.

Hashtag	Translation	Accounts
чп	Catastrophe	23,301
6дек	December 6th	18,174
5дек	December 5th	15,943
выборы	Election	15,082
митинг	Rally	13,479
триумфальная	Triumphal	10,816
победазанами	Victory will be ours	10,380
5dec	December 5th	8,743
навальный	Alexey Navalny [1]	8,256
ridus	Ridus [2]	6,116

Table 1: Top 10 hashtags related to the Russian election used between December 5–6.

Statistic	Spam	Nonspam
Accounts	25,860	20,986
Tweets (Dec 5–6, 2011)	440,793	876,774
Tweets (May, 2011–Jan, 2012)	2,445,382	- -

Table 2: Summary of accounts who participated in hashtags pertaining to the Russian election (December 5–6) and the activities of spam accounts outside of the election period.

3.1 Attacked Hashtags

To characterize the attack, we begin by identifying all of the accounts that posted a tweet containing the hashtag #триумфальная between December 5–6, 2011. This hashtag corresponds with the protests at Moscow's Triumfalnaya Square and was previously reported to be a target of spam accounts [24]. Due to the possibility that the attack targeted multiple hashtags, we take the set of users who tweeted #триумфальная and aggregate all the other hashtags they tweeted with during the attack window, filtering out hashtags with fewer than 1,500 participants. In total, we identify twenty hashtags posted by 46,846 accounts. Table 1 shows the top ten of these hashtags and their translation.

In order to identify which accounts were bots, we rely on Twitter's internal spam detection algorithm that monitors abusive behavior including accounts that excessively post to multiple hashtags. At the time of our analysis, the algorithm had suspended 24,203 of the accounts that posted to at least one of the twenty hashtags. In addition to suspended accounts, we include 1,657 accounts that were not suspended but exhibit patterns akin to the bots including similar-looking automatically generated email addresses and creation times correlated with a burst in spam account creation. We discuss the details of how we generate these criteria further in Section 4.

3.2 Dataset

After identifying all of the accounts that tweeted with hashtags pertaining to the Russian election and labeling them as spam or nonspam, we aggregate all of the tweets sent by both spam and legitimate accounts during December 5–6, 2011. Furthermore, we aggregate all of the tweets sent by spam accounts between May, 2011 when attackers registered their first account until January, 2012 when we started our analysis. In summary, our dataset consists of over 2.4 million spam tweets, 440,793 of which were posted during the attack, as shown in Table 2.

In addition to the posting activity of accounts, we build our analysis on registration data and periodic logging tied to each account. This data includes an account's email address, the IP address used to register the account, and all subsequent IP addresses used to access the account between November, 2011–January, 2012. This information allows us to analyze how attackers registered thousands of accounts and the hosts they used to access Twitter.

Finally, in order to gauge the impact of the attack on users searching for information related to the Russian election, we aggregate all of the tweets returned by search queries conducted between December 5–6 that correspond with one of the twenty hashtags attacked. We subsequently identify all of the tweets tied to bots and label them as spam, allowing us to measure the attack's success at diluting search results. We provide a more detailed summary of the search queries performed in Section 5.

4 Analysis

We deconstruct the attack into three components: the tweets sent prior to and during the attack; the registration data tied to the accounts involved; and the IP addresses that attackers used to access Twitter. We find that the accounts used in the attack generated politically-motivated tweets long before December 5–6, 2011. In order to spread these messages, the attackers acquired thousands of accounts from spam-as-a-service markets that controlled nearly a million fraudulent accounts. Similarly, the attack relied on tens of thousands of compromised machines located around the globe, 39% of which were blacklisted for email spam and malware distribution.

4.1 Tweets

In order to control the information users' found when they accessed hashtags pertaining to the Russian election, the attackers posted 440,793 tweets that targeted 20 hashtags organically adopted by users. At its height, the attack generated 1,846 tweets per minute, as shown in Figure 1. The entire attack consisted of a short burst in traffic on the first day, followed by a sustained flow of incomprehensible tweets interspersed with partisan jeers

[1] Prominent blogger arrested during protest in Moscow

[2] Russian media outlet

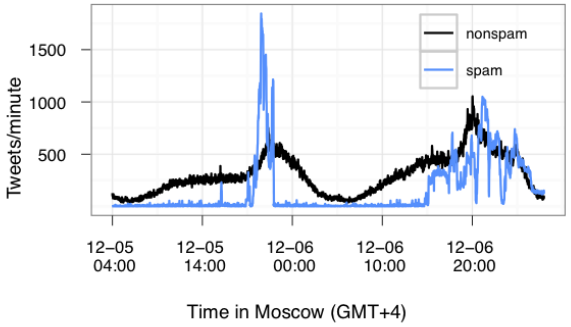


Figure 1: Number of tweets sent per minute during the attack on December 5–6. Tweets generated by bots appear in two large spikes beginning around 8PM the first day and 3PM the second day.

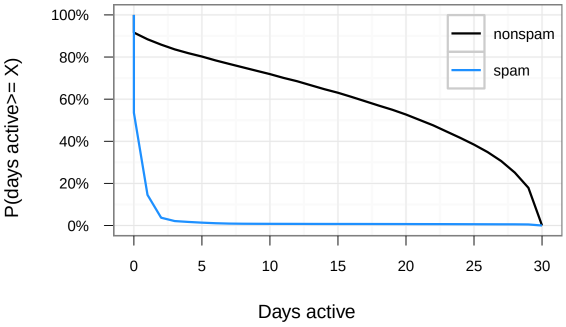


Figure 2: Total number of days in November, just prior to the attack, that an account tweets at least once. Nonspam users were frequently active, while spam accounts remained dormant.

the following day, effectively diluting the content available to Twitter users who were following the election discussions.

For the month prior to the attack, the majority of spam accounts that existed at the time remained dormant, in contrast to legitimate users, per Figure 2. However, over the entire course of May, 2011 up until the attack, the bots generated nearly 1.8 million tweets during sporadic periods of activity. The first salvo of coordinated tweets appeared in May, when 4,215 accounts tweeted for the first time with content deriding a prominent Russian anti-corruption blogger. Similar examples occur throughout the dataset when thousands of accounts activate to promote one-sided political opinions interspersed with unrelated news headlines, as determined by Google Translate. Yet, with no legitimate followers or hashtags tied to the early tweets, there was no one to see the content. The uniformity in the types of spam tweets sent, even months before the December 5–6 attack, implies that the accounts were under the sole control of miscreants rather than leased at different intervals. Otherwise, we would expect to observe mismatching messages from competing spammers renting access to the accounts.

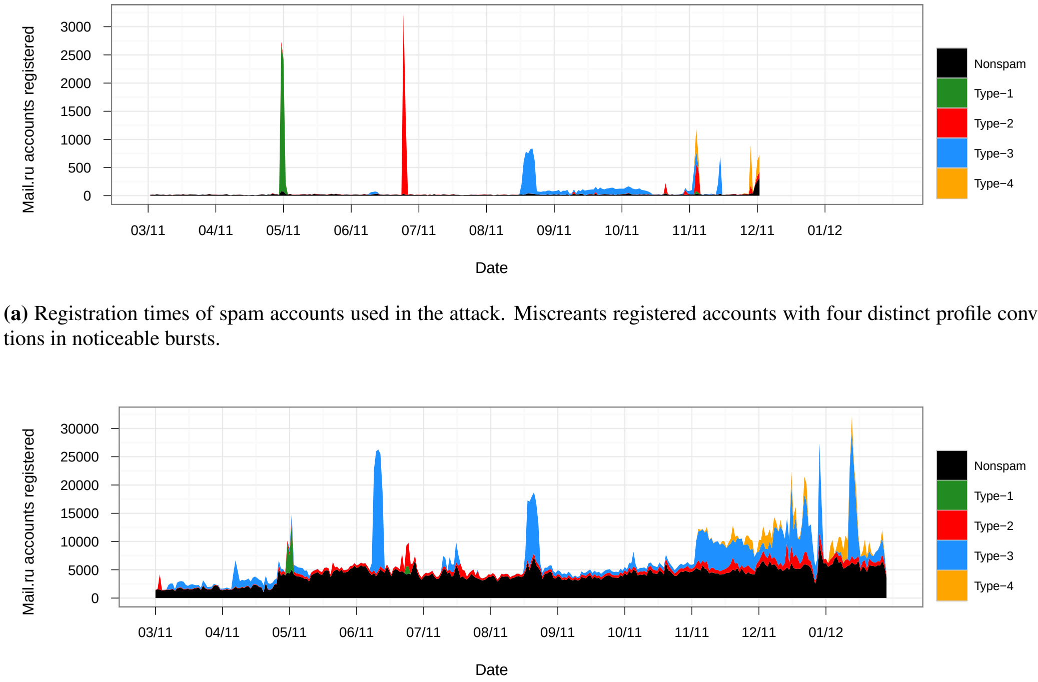
4.2 Accounts

Registration & Profile. Manipulating Twitter search results using bots requires the acquisition of thousands of accounts that are either fraudulent or compromised. In order to understand where the bot accounts originated from, we begin by analyzing the profiles and registration data tied to each suspended account. We find that 99.5% of the accounts were registered with a distinct mail.ru email address. 95% of these mail.ru email accounts were valid and belonged to the attacker, as indicated by the account's controller clicking on a URL sent to the email addresses after registration. The remaining 5% of mail.ru email addresses were awaiting verification before the account tied to the email was suspended.

Looking into account registration further, we examine the naming conventions used for the screennames, real names, and email addresses of each spam account. We identify a number of patterns tied to bot accounts with mail.ru emails that regularly repeat, but are absent from legitimate users with mail.ru email addresses. Due to the adversarial nature of identifying spam accounts, we do not reveal the patterns, but discuss their accuracy and importance. In total, we identify four distinct patterns of account registrations which we codify into regular expressions which we denote Type-1 through Type-4.

In order to evaluate the accuracy of our expressions at identifying spam accounts, we apply each regex to the 46,846 accounts in our December 5–6 dataset. While this classification approach is simple, our expressions identify an additional 1,657 spam accounts posting to election-based hashtags that were uncaught by Twitter's suspension algorithm. We manually validate the labels for 150 of the newly labeled spam accounts and find only 2% are false positives. Even more impressive, when we apply the expressions to all mail.ru registrations in the past year, we identify *975,283 spam accounts*, only 20% of which Twitter's algorithm had suspended at the time of our analysis. Furthermore, 80% of these accounts have no friends, followers, or tweets despite existing for months. We repeat our manual validation for 150 of the flagged mail.ru accounts and find only 4% are false positives. Due to the false positive rate, the number of accounts we identify should be treated as a rough estimate of the number of spam accounts registered with mail.ru emails that mirror the accounts used in the attack.

We further validate our classification approach both on accounts within the attack and for all accounts tied to mail.ru email addresses. Figure 3a shows the regis-



(a) Registration times of spam accounts used in the attack. Miscreants registered accounts with four distinct profile conventions in noticeable bursts.

(b) Registration times of all mail.ru accounts (note that the scale is 10x of that in the previous plot). Due to a lack of diversity in account profiles, we can readily detect other fraudulent accounts based on the same account signatures as those used in the attack.

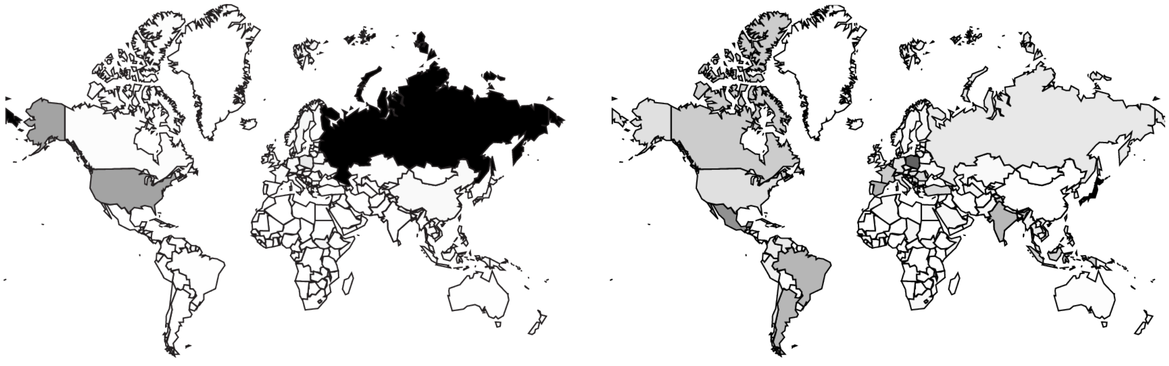
Figure 3: Pattern of registrations for accounts used in the attack and other accounts registered by the same spam-as-a-service programs where the attackers purchased accounts from.

tration dates of bots from March 2011 up until the date of the attack. Miscreants registered accounts in bulk, with account types rarely overlapping during the same period. In contrast, legitimate account registrations are uniformly distributed during the entire period. The registration times for the accounts used in the attack overlap with an abnormal volume of Twitter accounts registered to mail.ru email addresses, shown in Figure 3b. The registration spikes in June, August, and January are labeled exclusively as spam, while legitimate registrations remain roughly stable throughout the entire period.

In total, the accounts used in the attack represent only 3% of all the mail.ru accounts that our expressions flag as Type 1–4. This indicates the accounts were likely purchased from a spam-as-a-service marketplace that registers and sells accounts in bulk, such as `buyaccs.com`. These markets have an incredible negative impact on Twitter. For instance, the software registering these Type 1–4 accounts is responsible for *over 80% of fraudulent accounts* tied to mail.ru email addresses suspended by Twitter within the last year. With accounts readily available to any party willing to pay, spam-as-a-service shops simplify the re-purposing of spam infrastructure to whatever end, be it traditional scams or politically motivated censorship.

Social Graph. While most automatically generated accounts rarely engage in forming relationships with other Twitter accounts [23], the spam accounts involved in the attack attempted to simulate a social graph. A median account had 121 following—or outbound relationships—76% of which terminated at other bots. Similarly, a median account had 122 followers—or inbound relationships—85% of which originated from accounts involved in the attack. Even though the attackers acquired accounts registered across multiple months, all of the accounts were used to form a complete sybil network. As a result, all spam accounts involved in the attack that were not singletons were reachable via only spam relationships in an average of 3 hops. The motivation for an attacker to interconnect spam accounts is unclear, but may be a result of assumptions that the presence of social connections will make accounts less susceptible to suspension or improve the relevance ranking of content posted by spam accounts, discussed in Section 5.

The presence of a sybil graph is interesting for two reasons. First, it indicates that a single party controlled all of the accounts used in the attack. Relationships between the accounts were formed as far back as May, 2011, requiring coordination between the accounts long before



(a) Nonspam

(b) Spam

Figure 4: Geolocation of user logins. Higher density regions are shown in black. Over 56% of logins tied to legitimate users originate from Russia, compared to only 1% of logins for spam accounts.

the attack. Furthermore, 80% of the nearly 1 million fraudulent mail.ru accounts we identify have 0 friends and followers and remain dormant. It does not appear that building social relationships is the responsibility of the account creator, providing further evidence that control of the accounts changed hands at some point. We conclude that the miscreants who launched the attack adapted the accounts to their needs and generated social connections, while the party registering the accounts provided them without tweets or relationships.

4.3 IP Addresses

Diversity and Lifetime. In addition to acquiring thousands of spam accounts, the attack relied on a diverse body of IP addresses to circumvent Twitter's IP-based restrictions. We find that miscreants registered 84% of the bots with unique IP addresses. After sign up, this diversity decreases; only 49% of 110,189 IP addresses used to access spam accounts between November, 2011–January, 2012 were unique across accounts.

To translate this into the number of machines under the attacker's control, we first examine the lifetime of IP addresses used to access accounts. We find that 80% of the IP addresses tied to the spam accounts were present in our logs from November–January for only a single day. This same phenomenon is true for the 20,986 legitimate accounts, where 84% of IP addresses used to access the accounts persist for one day. We performed a reverse DNS lookup on all the IPs tied to the bots and find that each of the IP addresses belongs to ISP address pools. The hosts tied to these IP addresses are likely residential, as indicated by the presence of `dsl`, `cable`, `dynamic` and a number of other heuristics in the reverse lookup's naming convention.

Due to heavy churn in IP addresses over time, it is difficult to estimate the number of unique hosts ever used by attackers. Instead, we limit our analysis to a single day. At the height of the attack on December 6th, 11,356 unique IP addresses were used to access spam accounts. If we assume that IP addresses are stable for at least a day, then tens of thousands of hosts were available to the attackers.

Geolocation and Origin. In order to understand where the hosts controlled by the attackers originate from and how they compare to legitimate users, we examine the geolocation of IP addresses used by both types of parties. To start, we generate a list of the unique IP addresses used to access each of the 46,846 accounts between November, 2011–January, 2012. We then map these to their country of origin using the MaxMind database [14] and aggregate the totals across spam and legitimate accounts.

Figure 4 shows our results. 56% of all legitimate logins originate from Russia, compared to only 1% of logins tied to spam accounts. The IPs used by the attack are located around the globe, with Japan accounting for the largest set of logins (14%). These results imply that the bots are using compromised machines or proxy services to access Twitter.

Blacklist Membership. If attackers relied on compromised machines, there is a possibility that the hosts used to access Twitter were also used by other parties—or the same party—for other malicious behavior such as distributing email spam [2]. To this end, we used a list of 47 million suspicious IP addresses taken from the CBL blacklist [3], which contains IPs flagged for email spam and spreading malware. We then tested whether an IP addresses used to access any of the 25,860 accounts employed in the attack ever appears in the blacklist between October 2011–January 2012. When we perform our analysis, we ignore the timestamp for when an IP address is listed and unlisted to account for any delay between an attacker using an IP address and its subsequent blacklisting. However, this approach may also overestimate the number of malicious hosts.

We find that the CBL blacklist contains 39% of the IP addresses tied to bots. This indicates that hosts used to attack Twitter are also used by a malicious party to generate spam or distribute malware. However, in order to judge the accuracy of the blacklists, we repeat the same experiment using a list of IP addresses used to access legitimate accounts. We find that 21% of benign IP addresses are also listed. While we cannot definitively determine why blacklists are flagging IPs tied to legitimate users, it may result from blacklists biasing their classification of Russian IPs, or arise due to DHCP churn causing aliasing with other infected hosts. The imprecision we detect in blacklists reiterates previous research on the limitations of blacklists [19], especially in the context of social networks [10]. Nevertheless, because IPs used in the attack are more likely to be listed, we can infer that some of the attack's hosting infrastructure was simultaneously used for more traditional spam/malware activities.

5 Impact

The attack on Twitter is a compelling example of how miscreants can adapt spam infrastructure to censor legitimate access to relevant information surrounding controversial events. However, even with control of thousands of fraudulent accounts and compromised machines, the attack was partly mitigated by Twitter's relevance ranking of tweets, which personalizes search results and emphasizes popular content. We provide a brief overview on the different search mechanisms available to Twitter users before evaluating the fraction of search results that were affected by politically-slanted spam.

5.1 Search: Relevance vs. Real-time

Twitter search offers two modes of operation: *real-time* and *relevance* mode. Real-time mode returns tweets in order of most recently posted first. This type of indexing is susceptible to message dilution attacks as spammers merely need to outproduce legitimate content that users post to hashtags. In contrast, the relevance search mode incorporates signals that capture the popularity of a tweet while at the same time surfacing content from accounts whose social graph and interests overlap those of the account submitting a search query [26]. As a result, the algorithm ranks content by its importance, reducing the impact of mass producing tweets on a single topic. However, to add some dynamism to search results to prevent popular content from being locked at the top, the freshness of a tweet is also considered in the ordering of the most relevant tweets. By default, Twitter returns relevance-ranked searches.

Search Mode	Tweets Returned
Real-time	2,923,022
Relevance	17,276,281
Relevance (top 5 most recent)	3,743,919

Table 3: Number of tweets returned to users searching for hashtags related to the Russian election.

5.2 Search Pollution

To measure the frequency of spam in search results, we aggregate all of the tweets returned by queries performed between December 5–6, 2011 related to one of the attacked hashtags. If a bot posted a tweet that appears in the search results, we assume that tweet was spam. We assume all other accounts and tweets are legitimate. On average, searches return 15 tweets per query. We consider all of these tweets in our analysis even though users may only view a fraction when searching. Consequently, our analysis may overestimate a user's perception of spam in search results.

Table 3 shows a summary of the data used in our analysis. Twitter users generated over 233,000 real-time search queries related to the election. In aggregate, these search results contained 2.9 million tweets. Users relied on the default relevance-ranked search far more frequently. In total, users performed 1.1 million relevance searches which returned 17 million tweets. Analyzing the fraction of spam in each search query, we find that relevance-based searches returned 53% fewer spam tweets compared to real-time searches, a testament to the volume of tweets produced by bots in the real-time feed. If we restrict our analysis to the five most recent relevance-ranked tweets that were returned in searches—those that appear at the top of the page and are most likely to be seen—we find that relevance-mode returned 64% fewer spam tweets. These results highlight that integrating a user's social graph and interests into the tweets returned by searches can to a degree mitigate the impact of message dilution attacks.

6 Conclusion

We have analyzed how attackers adapted fraudulent accounts and compromised hosts—traditionally tied to spamming—in order to control political speech on Twitter. In particular, we examined an attack launched by unknown miscreants that leveraged 25,860 accounts to send 440,793 tweets in order to disrupt conversations about the Russian election, protests, and purported fraud. We showed that the accounts used in the attack were likely purchased from a spam-as-a-service program that controlled at least 975,283 Twitter accounts and mail.ru email addresses. In contrast to legitimate Russian users participating in discussions of the election, only 1% of the IP addresses used by attackers originated in Russia. Instead, the attackers controlled hosts located around the globe, over 39% of which were blacklisted for involvement in more classic spam activities. Despite the large volume of malicious tweets, Twitter's search relevance algorithm, which personalizes search results and weights popular content, eliminated 53% of the tweets sent during the attack compared to the real-time search results with no protections, offering a promising approach for defending against future censorship attacks.

7 Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. 0433702 and 0905631. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. This work is partially supported by the Office of Naval Research under MURI Grant No. N000140911081.

References

[1] F. Benevenuto, G. Magno, T. Rodrigues, and V. Almeida. Detecting Spammers on Twitter. In *Proceedings of the Conference on Email and Anti-Spam (CEAS)*, 2010.

[2] J. Caballero, C. Grier, C. Kreibich, and V. Paxson. Measuring Pay-Per-Install: The Commoditization of Malware Distribution. In *Proceedings of the USENIX Security Symposium*, 2011.

[3] CBL. Composite Blocking List. `http://cbl.abuseat.org/`, 2012.

[4] G. Danezis and P. Mittal. Sybilinfer: Detecting Sybil Nodes Using Social Networks. In *Proceedings of the Network and Distributed System Security Symposium (NDSS)*, 2009.

[5] W. Englund and K. Lally. In Protests, Two Russias Face Off. `http://wapo.st/wiVnV8`, 2011.

[6] J. Epstein. President Obama Google+ Chat Gets Personal. `http://politi.co/zTvgQO`, 2012.

[7] H. Gao, J. Hu, C. Wilson, Z. Li, Y. Chen, and B. Zhao. Detecting and Characterizing Social Spam Campaigns. In *Proceedings of the Internet Measurement Conference (IMC)*, 2010.

[8] P. Gogoi. Wal-Mart's Jim and Laura: The Real Story. `http://buswk.co/wnFI61`, 2006.

[9] J. D. Goodman. In Mexico, Social Media Become a Battleground in the Drug War. `http://nyti.ms/wgWUZb`, 2011.

[10] C. Grier, K. Thomas, V. Paxson, and M. Zhang. @spam: The Underground on 140 Characters or Less. In *Proceedings of the ACM Conference on Computer and Communications Security (CCS)*, 2010.

[11] J. Halliday. David Cameron Considers Banning Suspected Rioters from Social Media. `http://bit.ly/xI8MJs`, 2011.

[12] N. Jindal and B. Liu. Opinion Spam and Analysis. In *Proceedings of the International Conference on Web Search and Web Data Mining*, 2008.

[13] B. Krebs. Twitter Bots Drown Out Anti-Kremlin Tweets. `http://bit.ly/w9Gnaz`, 2011.

[14] MaxMind. Resources for Developers. `http://www.maxmind.com/app/api`, 2010.

[15] M. Motoyama, D. McCoy, K. Levchenko, G. M. Voelker, and S. Savage. Dirty Jobs: The Role of Freelance Labor in Web Service Abuse. In *Proceedings of the USENIX Security Symposium*, San Francisco, CA, August 2011.

[16] Office of the Press Secretary. White House to Host Twitter @TOWNHALL. `http://1.usa.gov/zplVBV`, 2011.

[17] J. Preston. What Does 40 Mean to You? `http://nyti.ms/zfMuQ2`, 2011.

[18] M. Richtel. Egypt Cuts Off Most Internet and Cell Service. `http://nyti.ms/z44cWc`, 2011.

[19] S. Sinha, M. Bailey, and F. Jahanian. Shades of Grey: On the Effectiveness of Reputation-Based "Blacklists". In *3rd International Conference on Malicious and Unwanted Software*, 2008.

[20] socialcapital. Twitter, Facebook and YouTube's Role in Arab Spring. `http://bit.ly/xxBNmo`, 2011.

[21] G. Stringhini, C. Kruegel, and G. Vigna. Detecting Spammers on Social Networks. In *Proceedings of the Annual Computer Security Applications Conference (ACSAC)*, 2010.

[22] K. Thomas, C. Grier, J. Ma, V. Paxson, and D. Song. Design and Evaluation of a Real-Time URL Filtering Service. In *Proceedings of the IEEE Symposium on Security and Privacy*, 2011.

[23] K. Thomas, C. Grier, V. Paxson, and D. Song. Suspended Accounts In Retrospect: An Analysis of Twitter Spam. In *Proceedings of the Internet Measurement Conference*, November 2011.

[24] TrendMicro. The Dark Side of Social Media. `http://http://bit.ly/zn217U`, 2011.

[25] Twitter. The Twitter Rules. `http://support.twitter.com/entries/18311-the-twitter-rules`, 2010.

[26] Twitter Engineering. The Engineering Behind Twitter's New Search Experience. `http://bit.ly/iuRwp8`, 2011.

[27] G. Wang, C. Wilson, X. Zhao, Y. Zhu, M. Mohanlal, H. Zheng, and B. Y. Zhao. Serf and Turf: Crowdturfing for Fun and Profit. In *Proceedings of the International World Wide Web Conference*, 2011.

[28] R. Wauters. China Blocks Access To Twitter, Facebook After Riots. `http://tcrn.ch/yaxKjP`, 2009.

[29] H. Yu, M. Kaminsky, P. Gibbons, and A. Flaxman. Sybilguard: Defending Against Sybil Attacks via Social Networks. In *ACM SIGCOMM Computer Communication Review*, 2006.