SOCIAL SPAMBOTS

Richard Bernache*

CITE AS: 4 GEO. L. TECH. REV. 307 (2019)

TABLE OF CONTENTS

I. INTRODUCTION: WHAT IS A SPAMBOT?

Spam, unsolicited commercial content spread at a mass scale, is everywhere on the Internet, in our email and on the sites we visit.[1] At a minimum, spam is a nuisance, just another email or post online to delete or ignore. At its worst, falling victim to a spam attack risks infecting a system with malware or compromising an individual's identity. This proliferation is due in large part to spambots, Internet robots designed to spread spam autonomously. Spambots can be programmed in any number of modern coding languages and are created for a variety of purposes including the mass spreading of advertising content, proliferation of credible or false information,

* Georgetown University Law Center, J.D. Candidate 2021. Saint Michael's College, B.A. in Theatre 2016. First and foremost, I would like to thank the staff and editors of the Georgetown Law Technology Review for their work and patience during the writing process. Thanks also to my friends and family for their unyielding support while I undertake this new part of my professional life..

[1] Pedram Hayati et al., *Characterisation of Web Spambots Using Self Organising Maps*, 2 INT'L J. COMPUTER SCI. & ENGINEERING 87, 87–88 (2011). [https://perma.cc/83HH-7267].

or the spread of malware and other harmful software.[2] This explainer focuses on two types of spambots that ordinary human users are likely to encounter: email spambots, which scrape the web looking for email addresses and then auto-send mail with malware, and social media spambots, which pose as human users and autogenerate content containing spam. Web developers have spent significant resources into developing spam prevention mechanisms. However, the most relied upon spambot prevention tool, the CAPTCHA, has not been able to keep up with the evolution of current spambots, leaving our email and social media accounts vulnerable.

II. TYPES OF SPAMBOTS

A. Email Spambots

Email was among the earliest targets of online spam campaigns. The dispersion of spam through email quickly proliferated, and now the majority of all email sent contains spam.[3] The most basic types of email spam are not context specific, meaning they are not connected to the human target's personal history. These emails seek to either convince recipients to follow a link containing malware or to steal a human user's identity.[4] In order to maximize their success, many email spammers target massive numbers of human users in the hope that even a relatively low number of recipients open and follow the malicious link.[5] Deploying spambots to scrape websites for email addresses and generating and auto-sending spam emails is inexpensive.[6] This allows spammers to target more users.

B. Social Media Spambots

While email spambots primarily spread malware, social media bots perform a number of different tasks. For example, Twitter has become a focus of the public conversation around spambots, in large part due to efforts by Russian intelligence agencies to spread misinformation during the 2016

---

[2] *Id.* at 88; Rob Dubbin, *The Rise of Twitter Bots*, NEW YORKER (Nov. 14, 2013), https://www.newyorker.com/tech/annals-of-technology/the-rise-of-twitter-bots [https://perma.cc/X5KF-FQ92]; Chencheng Shao et al., *The Spread of Low-credibility Content by Social Bots*, 9 NATURE COMMS. 1, 5 (2018), https://www.nature.com/articles/s41467-018-06930-7.pdf [https://perma.cc/VW5Z-CWK9].

[3] Cristian Lumezanu & Nick Feamster, *Observing Common Spam in Tweets and Emails*, INTERNET MEASUREMENT CONF. PROC. (2012), http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.297.7011&rep=rep1&type=pdf [https://perma.cc/9RK7-BX58].

[4] *Id;* Dubbin, *supra* note 2.

[5] Lumezanu & Feamster, *supra* note 3.

[6] Hayati et al., *supra* note 1, at 88.

United States presidential election.[7] In total, Twitter identified 50,258 bot accounts linked to the Internet Research Agency (IRA), a Russian propaganda organization.[8] These bots were used to spread false or misleading information regarding the U.S. election and ultimately approximately 1.4 million human users interacted with an IRA bot by liking, retweeting, replying to, or following the bot.[9] These adversarial bot accounts were able to flourish on Twitter due to the website's open application program interface (API).[10] Twitter allows companies and individuals to apply for access to Twitter's various APIs.[11] APIs connect a developer's computer program with Twitter's "endpoints" which are the various types of information (users, tweets, direct messages, and ads) that a developer is able to access via use of Twitter APIs.[12] Accessing Twitter's tweet and reply APIs allows a developer to collect tweets of a certain topic and program bots to post tweets via the API connection.[13] Exploitation of Twitter's API by spammers caused Twitter to add additional steps to the API application process and the site has begun to more aggressively police violations of its spam policy.[14] However, the effectiveness of these new measures has yet to be analyzed.

Additionally, due to advances in publicly available source codes for social media bots and advanced artificial intelligences (AIs) capable of conversational text, creating social media bots now requires less individual programming knowledge.[15] These source codes are readily available on various tech blogs, allowing an unsophisticated human spammer to use spambots to spread malicious content.[16] Simple spambots may post the same pre-scripted content written by the human spammer according to a programmed schedule.[17] More sophisticated social media bots may integrate

---

[7] *Update on Twitter's Review of the 2016 US Election*, TWITTER BLOG, (last updated Jan. 31, 2018), https://blog.twitter.com/official/en_us/topics/company/2018/2016-election-update.html [https://perma.cc/XST3-WMRU].
[8] *Id.*
[9] *Id.*
[10] Dubbin, *supra* note 2.
[11] *About Twitter's API*, TWITTER, https://help.twitter.com/en/rules-and-policies/twitter-api (accessed Oct. 31, 2019) [https://perma.cc/74WV-XP65].
[12] *Id.*
[13] *Id.*
[14] Yoel Roth & Rob Johnson, *New Developer Requirements to Protect Our Platform*, TWITTER BLOG (July 24, 2018), https://blog.twitter.com/developer/en_us/topics/tools/2018/new-developer-requirements-to-protect-our-platform.html [https://perma.cc/XUP6-QSL7].
[15] *See generally* Emilio Ferrara, Univ. S. Cal. Info. Sci. Inst., *Bots, Elections, and Social Media: A Brief Overview*, (arXiv:1910.01720v1, Oct. 3, 2019), https://arxiv.org/pdf/1910.01720.pdf [https://perma.cc/QRX5-JC8V].
[16] *Id.*
[17] Emilio Ferrara, *The History of Digital Spam*, 62 COMMS. ACM 82, 88 (Aug. 14, 2019), https://arxiv.org/pdf/1908.06173.pdf [https://perma.cc/YX6X-AWJF].

text generating AI such as those provided by companies like ChatBots.io in order to auto-generate conversational text and engage with and respond to real human users.[18] These AIs allow social media bots to better mimic human user behavior, increasing the effectiveness of spam campaigns.[19]

While many Twitter bots are used for nefarious purposes, many early Twitter bots used Twitter's API solely for innocuous, often times comedic, purposes. One comedic bot, Pentametron, was programmed to roam Twitter searching for rhyming couplets, and retweets the rhyme to the bot's twenty-six thousand followers.[20] Other bots seek to spread malware or harvest and steal user identities just as email spambots do.[21] Others can be purchased in order to boost a user's follower count to project a greater following than the individual actually possess. In the process, these bots follow large numbers of other human users in order to mask the fact that they are not actually human and to protect the identity of whoever purchased the follower boost.[22]

C. Cross Platform Spam Attacks

The effectiveness of spambots increases when the spam is spread across multiple platforms.[23] Such cross-platform spam attacks can be characterized as "context aware spam."[24] By using bots to web-scrape social media and other online profiles of human users, spambots are able to generate content which is harder to identify as spam. For example, a spambot may scrape sufficient information from a user's Facebook profile to determine the user's friends, date of birth, and email address.[25] The bot is then able to use this information to create a spam email with a malware link that is context specific to the individual target. This could be synthesized into an email from a user's Facebook friend with a link to a birthday e-card.[26] By capitalizing on the user's personal data, the spammer is able to increase the number of users who follow the malicious link.[27]

---

[18] Ferrara, *supra* note 15, at 3.
[19] Ferrara, *supra* note 17, at 84.
[20] Dubbin, *supra* note 2.
[21] Alexis Madrigal, *Here's Why 9,000 Porny Spambots Descended on a High Schooler's Twitter Account*, ATLANTIC (Nov. 25, 2013), https://www.theatlantic.com/technology/archive/2013/11/why-did-9-000-porny-spambots-descend-on-this-san-diego-high-schooler/281773 [https://perma.cc/D8XQ-E4YY].
[22] *Id.*
[23] Lumezanu & Feamster, *supra* note 3.
[24] Garrett Brown et al., *Social Networks and Context-Aware Spam*, *in* COMPUTER SUPPORTED COOPERATIVE WORK CONF. 403 (2008), http://www-personal.umich.edu/~kborders/p403-brown.pdf [https://perma.cc/3JL9-TD6U].
[25] *Id.* at 407.
[26] *Id.*
[27] *Id.*

III. EVOLUTION OF SPAMBOTS

Early spambots were typically easily identified by human users because bots shared many common behavioral patterns which did not mirror human to human online communication. For example, early email spambots frequently misspelled or inserted random extra characters to commonly filtered words in order to avoid detection by spam filters.[28] Early social media spambots often posted the same advertisement or link incessantly and featured nonsensical usernames.[29]

However, spambots have rapidly improved their behavior such that it has become significantly more difficult for human users to detect bot accounts on their own. Modern social media spambots have seen dramatic improvements in their ability to mimic human behavior. Spambots attempt to hide their presence by following or friending large numbers of real human users.[30] This tactic is made more effective when these real human users follow the spambot back, deceiving detection mechanisms which rely on targeting bots' tendency to have limited followers while generating a large amount of content.[31] This technique has been improved by reducing the amount of content a bot produces each day in order to more closely approximate the typical human user's Twitter usage.[32] Bots also focus on trending topics, such as popular songs and YouTube videos and generate tweets of their own which appear to follow online trends.[33]

Identifying these more advanced bots typically requires the ability to observe bots as a group, which is often beyond the capabilities of a normal human user. For example, an examination of bots used in the 2014 mayoral election in Rome revealed that these bots were unlikely to be detected without the ability to track the common behaviors of a large group of bots.[34] These bots followed a large number of ordinary users, tweeted at a relatively infrequent pace to mimic common human user patterns, and normally tweeted

---

[28] Li Zhuang et al., *Characterizing Botnets from Email Spam Records*, *in* FIRST USENIX WORKSHOP ON LARGE-SCALE EXPLOITS & EMERGENT THREATS 3 (2008), http://static.usenix.org/events/leet08/tech/full_papers/zhuang/zhuang.pdf [https://perma.cc/ZU6N-7P28].

[29] Dubbin, *supra* note 2.

[30] Stefano Cresci et al., *On the Capability of Evolved Spambots to Evade Detection via Genetic Engineering*, 9 ONLINE SOC. NETWORKS & MEDIA 1, 4 (2019), https://www.sciencedirect.com/science/article/pii/S246869641830065X [https://perma.cc/8QYK-RLQY].

[31] Xia Hu et al., *Online Social Spammer Detection*, *in* 28TH AAAI CONF. ARTIFICIAL INTELLIGENCE 59, 59 (2014), https://www.aaai.org/ocs/index.php/AAAI/AAAI14/paper/viewFile/8467/8399 [https://perma.cc/F673-9A47].

[32] Cresci et al., *supra* note 30 at 4.

[33] *Id.*

[34] *Id*.

about popular songs and media. However, when the candidate these bots were designed to promote tweeted, all bots retweeted the candidate's content within minutes of each other.[35] The ordinary user would be unable to determine this synchronized behavior unless they were able to observe the tweets of several bots. In fact, some human users actually replied to bot generated content during Rome's mayoral election.[36] In order to protect users from spam attacks, sites have long since used spam detection and prevention mechanisms, such as CAPTCHA, to stop the proliferation of spam to the site's users.[37] As spambots become increasingly more adept at deceiving the ordinary human users, these prevention mechanisms are even more crucial as tools to stop bots from reaching human users. However, even the most frequently used spam prevention tool, the CAPTCHA test, has become increasingly ineffective at preventing spambots from gaining access to sites and human users.

IV. SPAMBOT DETECTION AND PREVENTION—CAPTCHA TESTS

CAPTCHA, or Completely Automated Public Turing test to tell Computers and Humans Apart, is the most popular and well recognized anti-spambot tool.[38] CAPTCHAs are tests that a user must solve before accessing a webpage. CAPTCHAs come in different forms but all must be easily solvable by humans and easily generated, but must not be easily solved by bots.[39] Early CAPTCHAs were primarily text based and typically featured an image of a word or words, which were distorted through some combination of text warping or color overlay.[40] These images are decipherable by human users, who are able to read the text through the distortions, but unreadable by bots that rely optical character recognition (OCR) software to convert images to text.[41] The OCR software would be unable to accurately decipher the text obscured by the CAPTCHA distortions. In 2014, Google announced its own

---

[35] *Id.*

[36] *Id.*

[37] Suphannee Sivakorn, *I'm Not a Human: Breaking the Google reCAPTCHA*, BLACK HAT (2016), https://www.blackhat.com/docs/asia-16/materials/asia-16-Sivakorn-Im-Not-a-Human-Breaking-the-Google-reCAPTCHA-wp.pdf [https://perma.cc/L85W-E55P].

[38] Hayati, *supra* note 1, at 89.

[39] Marti Motoyama et al., *Re: CAPTCHAs–Understanding CAPTCHA-Solving Services in an Economic Context*, *in* PROC. 19TH USENIX SEC. SYMPOSIUM 2 (2010), http://cseweb.ucsd.edu/~savage/papers/UsenixSec10.pdf [https://perma.cc/Z7E5-PTSL].

[40] Jeff Yan and Ahmad Salah El Ahmad, *Usability of CAPTCHAs or Usability Issues in CAPTCHA Designs*, *in* PROC. 4TH SYMPOSIUM ON USABLE PRIVACY & SEC. (2008), https://prof-jeffyan.github.io/soups08.pdf [https://perma.cc/9CUJ-R9LM].

[41] Haley Tsukayama, *The Surprising Tool Bots Use to Get Around Those Pesky CAPTCHAs*, WASH. POST (June 9, 2016), https://www.washingtonpost.com/news/the-switch/wp/2016/06/09/the-surprising-tool-bots-use-to-get-around-those-pesky-captchas/ [https://perma.cc/LJ7M-CPHT].

spambot prevention tool, "No CAPTCHA reCAPTCHA" (reCAPTCHA), in part because of the advances made by AI algorithms which were solving text-based CAPTCHAs with 99.8% accuracy.[42]

Google's reCAPTCHA has become the most popular CAPTCHA service.[43] reCAPTCHA functions by presenting a widget on a webpage that asks the user to certify "I'm not a robot" by clicking a box.[44] When the box is clicked, the user's past usage behavior is automatically analyzed by a risk analysis algorithm to determine level of confidence that the user is human.[45] If this level of confidence is high, meaning the user is very likely to be human, the user is given permission to access the webpage without completing a CAPTCHA test.[46] While much of how this confidence level is estimated is proprietary information, research has shown that browsing history as determined through Google's tracking cookie plays a critical role in determining the confidence level.[47]

If the algorithm cannot determine with high confidence that the user is human, then the user is given either an image CAPTCHA or a text CAPTCHA. Image CAPTCHAs ask users to match images to a prompt.[48] For example, a user may be asked to "select all images which include wine."[49] While human users are able to select the images that match the prompt, spambots are unable to accurately differentiate between the images generated by the program. Image CAPTCHAs are reCAPTCHA's preferred form of test, and text CAPTCHAs have been gradually phased out.[50]

However, Google's reCAPTCHA has not been entirely effective at preventing bots from accessing sites. reCAPTCHA initially reused images frequently, allowing human users to collect and create databases of "tags" which can be used to program spambots to correctly select the prompted images of a reCAPTCHA.[51] Image annotation services, such as Google Reverse Image Search, can generate lists of tags which describe the image, making the process of generating tags more efficient.[52] Afterwards, the bots can be programmed to match the prompt of the reCAPTCHA (for example,

---

[42] *Are You a Robot? Introducing "No CAPTCHA reCAPTCHA,"* GOOGLE SECURITY BLOG (Dec. 3, 2014), https://security.googleblog.com/2014/12/are-you-robot-introducing-no-captcha.html [https://perma.cc/EA5S-3VH7].
[43] Sivakorn, *supra* note 37.
[44] *Are You a Robot? Introducing "No CAPTCHA reCAPTCHA,"* supra note 41.
[45] *Id.*
[46] Sivakorn, *supra* note 37.
[47] *Id.*
[48] Are You a Robot? Introducing "No CAPTCHA reCAPTCHA", supra note 41.
[49] Sivakorn, *supra* note 37.
[50] *Id.*
[51] *Id.*
[52] *Id.*

select all images which include wine) tags listed for each reused image and correctly select the matching images.[53] One such experiment, conducted in 2016, was able to use bots to solve image CAPTCHAs with seventy percent accuracy.[54]

Even when spammers are unable to generate large image and tag databases on their own, they have been able to navigate around CAPTCHAs by employing human workers to solve text and image CAPTCHAs. These bad actors hire CAPTCHA solving services that employ workers to solve CAPTCHAs for the entirety of their workday.[55] These human workers and spambots are connected through a plug-in distributed by the CAPTCHA solving service.[56] When the bot encounters a CAPTCHA, the test is sent to a human user who solves it. The solution is then communicated back to the bot, which enters the solution and gains access to the site.[57] Integrating human solvers into spambots in this way thwarts the fundamental purpose of CAPTCHAs, which were designed to allow humans, but not the spam spreading bots, to access the sites.[58]

V. CONCLUSION

Spambots are rapidly evolving and remain ever present on online platforms. As these bots evolve to outpace prevention mechanisms, the sheer amount of spam content online is likely to increase. Additionally, as these spam techniques become more effective in deceiving ordinary human users, more and more users may fall victim to malware attacks, identify theft, or disinformation campaigns. In order to keep pace with developments in spambot technology, prevention mechanisms must continue to evolve.

---

[53] *Id.*
[54] *Id.*
[55] Christopher Mims, *How Spammers Use Low-cost Labor to Solve CAPTCHAs*, MIT TECH. REV. (Aug. 11, 2010), https://www.technologyreview.com/s/420200/how-spammers-use-low-cost-labor-to-solve-captchas/ [https://perma.cc/9CWS-XQ5P].
[56] Motoyama et al., *supra* note 39, at 6**.**
[57] *Id.*
[58] *Id.* at 5.