Do Bots impact Twitter activity?

Zafar Gilani
Computer Laboratory
University of Cambridge
szuhg2@cam.ac.uk

Reza Farahbakhsh
Institut Mines Telecom
CNRS Lab UMR5157
reza.farahbakhsh@it-sudparis.eu

Jon Crowcroft
Computer Laboratory
University of Cambridge
jac22@cam.ac.uk

ABSTRACT

The WWW has seen massive growth in population of automated programs (bots) for a variety of exploits on online social networks (OSNs). In this paper we extend on our previous work to study the affects of bots on Twitter. By setting up a bot account on Twitter and conducting analysis on a click logs dataset from our web server, we show that despite bots being in smaller numbers, they exercise a profound impact on content popularity and activity on Twitter.

Keywords

information propagation; bot characterisation; bot activity analysis

1. INTRODUCTION

Bots, or automated entities, exist in vast quantity on online social networks (OSNs). They are created for a number of different purposes, such as news, marketing, spamming, spreading malicious content, and more recently political campaigning. According to an estimate 51.8% of all Web traffic is generated by bots.[1] Similarly, OSNs such as Twitter have seen a massive surge in bot population as Twitter itself reported in 2014 that 13.5 million (then 5% of the total Twitter population) are either fake or spam accounts.[2] Twitter insists these numbers do not include accounts that use third-party scheduling tools or social media management apps. The rise of bots on Twitter is further evident from a number of studies that analyse this phenomenon [2, 3] as well as a number of articles and blogs discussing bots.[3]

The existence of bots on Twitter is owed to three main reasons: (*i*) registering a Twitter account only needs passing a soft inspection: an email address, CAPTCHA recognition, and recently a mobile phone number; (*ii*) Twitter APIs that let programmers automate actions on Twitter; and (*iii*) organisations and individuals exploiting automation for their purposes and agenda.

In this study, we quantify the impact of bots on content popularity and activity on Twitter and the Web.

[1]Bot traffic report 2016 – `http://bit.ly/2kzZ6Nn`
[2]Twitter's 2014 Q2 SEC filing – `http://bit.ly/1kBx4M8`
[3]Bots in press and blogs – `http://bit.ly/2dBAIbB`

WWW 2017 Companion, April 3–7, 2017, Perth, Australia.
ACM 978-1-4503-4914-7/17/04.
http://dx.doi.org/10.1145/3041021.3054255

2. METHODOLOGY AND RESULTS

This section briefly describes the (*i*) methodology of setting up a bot account (as part of ***Stweeler***[4]), (*ii*) how the bot operates, (*iii*) click logs dataset, and (*iv*) bot characterisation using this dataset.

2.1 Setting up a bot account

We extend our previous work [4] and collect click logs dataset from our web server powered by our Twitter bot. The bot operates as follows: (*i*) The bot fetches a popular 'job' related tweet from the Twitter Streaming API. It then disassembles the text and URL in the tweet. (*ii*) The URL is then fetched into our web server (WS). The WS runs a shortener module that shortens the URL into our domain name. The shortener is needed to enable redirecting click traffic to our web server in order for us to collect click logs. (*iii*) The bot reassembles the tweet using the text and shortened URL. (*iv*) The tweet is posted to our Twitter account. In essence, our Twitter bot and WS performs a simple 'tweet manipulation' to avoid retweeting, which would otherwise prevent us from obtaining click logs dataset. (*v*) Finally, whenever a user (Twitter user or from the Web) clicks on our tweet(s) or URL(s), our WS records the click. Table 1 shows the type of information that is collected. Note that in order to respect the ethical boundaries of social media research, we **only** collect publicly available data about users and hash sensitive information such as IP addresses.

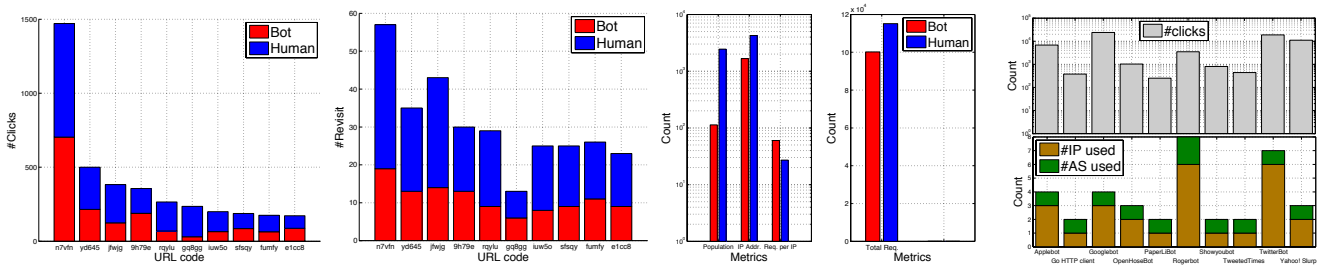
Table 1: Data collected through click logging.

Data attribute	Description
Click timestamp	Date and time of click, local to our web server.
Tweet ID	Tweet ID which received a click.
Hashed IP address	Hashed IP address of the machine that clicked the URL in the tweet identified by Tweet ID.
AS number	Obtained using the IP addresses from CAIDA.
User agent string	This records the `HTTP_USER_AGENT` string of the user clicking the URL in the tweet identified by Tweet ID.

2.2 Bot detection

For the purposes of this study we implemented a simple bot detection mechanism using two most relevant features: click frequency, and User agent strings. From 21-11-2015 to 08-01-2017, our Twitter bot account received more than 223000 clicks, out of which more than 44.91% have been by some sort of automated agent or a bot. Firstly, we employ time series analysis that takes

[4]***Stweeler***- `https://github.com/zafargilani/stcs`



(a) clicks (b) revisits (c) IP addresses and requests (d) IPs and ASs by bots

Figure 1: Click logs dataset.

Table 2: Click logs dataset – statistics.

Fact	Figures
Timeframe	From 21-11-2015 to 08-01-2017
Total clicks	223062
Clicks by bots	100194 **(>44.91%)**
Unique visitors	2563
Unique recurring bots	113 (**4.08%**)

into account the frequency of clicks by a single Twitter user account. As shown in [1] higher tweet frequency is indicative of automated behaviour. We then perform User agent string analysis, which reveals properties such as description or a URL containing description of the tool responsible for initiating clicks on our URLs. Note that tools accessing our tweets must have a Twitter account as guests are only allowed a sample of tweets by Twitter. Moreover, we find that there are a total of 2563 unique visitors, out of which only 113 are unique bots that have a recurring presence. These facts are summarised in Table 2.

2.3 Characterisation

We highlight important behavioural properties of bots and humans such as: click activity, revisiting a previously visited URL, and the use of IP addresses and Autonomous Systems (AS) to launch requests to our web server.

Surprisingly, only 4.08% of the visitors to our tweets or URLs are bots but are responsible for almost half of the clicks (44.91%). Bots account for a large chunk of the traffic produced on and contributed to the Twitter CDN and the Web. This finding points to interesting implications since bots not only access these URLs on the Web, but may also repost or retweet these tweets on their Twitter page or elsewhere using the website or platform-specific APIs. This is evident from Figure 1.

Figure 1(a) shows the number of clicks received by top 10 most popular URLs that our bot posted on its Twitter page. The URL code is the shortened suffix that replaces the original URL. The most popular URL for bots (*n7vfn*) advertises a UI/UX job in Sunnyvale CA, and the least popular URL for bots (*gq8gg*) advertises a job in Nairobi. The top 10 list would change by at least 3 URLs if bots had not existed, thus clearly showing that bots cause the rise in URL popularity.

Revisits are more typical for humans than bots, as observed in Figure 1(b). This is because these bots usually follow tweet streams which always flow forwards, thus requiring additional functionality for fetching historic profile. Moreover, some of the bots in our click logs dataset are actually content crawlers that maintain databases to avoid performing repeated activity.

Figure 1(c) shows the distribution of IP addresses used by bots vs IP addresses used by humans. 113 bots use 1667 unique IP addresses to generate a total of 100194 requests. On the other hand 2450 humans use 4258 unique IP addresses to generate a total of 115137 requests. Human activity per IP address is considerably lower (27 requests per IP) than bots (60 requests per IP).

Lastly, Figure 1(d) shows the distribution of number of unique IP addresses and Autonomous Systems (AS) used by the top 10 most active bots (rank based on User agent string analysis), along with their click activity. The top most active bots detected from our click logs dataset tend to be Twitter bots that make use of the Twitter API to perform actions (Twitterbot = 18828 clicks), web crawlers and indexers (Googlebot = 15790, Yahoo! Slurp = 11022, Applebot = 6755), and content curators and publishers (PaperLiBot = 249, TweetedTimes = 437). There is a possibility that Twitter might also inject its own bots for account profiling, spam detection, monitoring and reporting, by using its BotMaker software.

Typically, the top most active bots use multiple static IP addresses from within a single AS, possibly to parallelise tasks. Interestingly, this possibility is further supported by the fact that all except one AS (25 of 26) are designated as type 'Content' (content hosting and distribution system), while only one is designated as type 'Transit/Access' (connecting networks through itself). Furthermore, in our dataset for the top 10 most active bots, there was one exception of an unusually aggressive (but benign) bot called Rogerbot, a web crawler for a marketing firm, that used 6 IPs from 2 ASes to register 3485 clicks.

3. CONCLUSION

Extending our past work, we perform a characterisation of bot activity on Twitter. We show bots play a significant role in boosting URL popularity, demonstrate differences in URL revisiting behaviour, and exercise increased usage of IP addresses and ASes to launch requests. Future work will include reliable classification and methodical characterisation of bots and humans on Twitter.

Acknowledgement: This work is partially funded by EU Metrics project (Grant EC607728) and Celtic plus CONVINcE. We thank Mario Almeida (UPC) who was initially involved in development of ***Stweeler***.

4. REFERENCES

[1] Chu, Zi, Steven Gianvecchio, Haining Wang, and Sushil Jajodia. "Who is tweeting on Twitter: human, bot, or cyborg?." In 26th ACSAC, pp. 21-30. 2010.

[2] Lee, Kyumin, Brian David Eoff, and James Caverlee. "Seven Months with the Devils: A Long-Term Study of Content Polluters on Twitter." In 5th ICWSM. 2011.

[3] Edwards, Chad, Autumn Edwards, Patric R. Spence, and Ashleigh K. Shelton. "Is that a bot running the social media feed? Testing the differences in perceptions of communication quality for a human agent and a bot agent on Twitter." Computers in Human Behavior 33 (2014): 372-376.

[4] Gilani, Zafar, Liang Wang, Jon Crowcroft, Mario Almeida, and Reza Farahbakhsh. "Stweeler: A Framework for Twitter Bot Analysis." In 25th WWW, pp. 37-38. 2016.