The Implications of Twitterbot Generated Data Traffic on Networked Systems

Zafar Gilani, Jon Crowcroft
University of Cambridge
(szuhg2,jac22)@cam.ac.uk

Reza Farahbakhsh
Institut Mines Telecom - Sud-Paris
reza.farahbakhsh@it-sudparis.eu

Gareth Tyson
Queen Mary University of London
g.tyson@qmul.ac.uk

ABSTRACT

The explosion of bots on the Web brings an unprecedented increase in traffic from *non-human* sources. This work studies bot traffic on Twitter, finding that almost 50% of traffic is generated and propagated by a rapidly growing bot population — a major concern for networked systems in the future.

CCS CONCEPTS

• **Information systems** → **Social networks**; • **Networks** → *Network measurement*;

KEYWORDS

information propagation, bot network traffic, bot generated content

ACM Reference format:
Zafar Gilani, Jon Crowcroft, Reza Farahbakhsh, and Gareth Tyson. 2017. The Implications of Twitterbot Generated Data Traffic on Networked Systems. In *Proceedings of SIGCOMM Posters and Demos '17, Los Angeles, CA, USA, August 22–24, 2017,* 3 pages.
https://doi.org/10.1145/3123878.3131983

1 INTRODUCTION

Automated agents, *bots*, exist in a vast quantity on online social networks (OSNs) such as Twitter. Their purpose defines their intent, *e.g.,* news, marketing, spamming, spreading malicious content, and more recently political campaigning. 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.[1] 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 [1, 2, 5] as well as many articles and blogs discussing bots.[2]

Hence, the combined popularity of social media and online bots may mean that a significant portion of network traffic can be attributed to bots. This conjecture is not without support: according to one estimate, 51.8% of all Web traffic is generated by bots.[3] Such a radical shift from traditional views on web traffic brings about both new research questions and engineering opportunities. For example, can we model the amount of traffic produced by bots? Can we predict their behaviour? Can we adapt our network and content delivery infrastructure to better meet their needs, and mitigate overheads? The latter is of particular importance, as the above preliminary evidence seems to suggest that much of our network utilisation is due to (low priority) bots.

To explore the above questions, we have focused on Twitter, which is reputed to contain bots and, fortuitously, is easy to collect data for. In this initial study, we seek to discover: (*i*) the amount of data traffic bots generate on Twitter, and (*ii*) the nature of this traffic in terms of media type, *i.e.,* URL, photo (JPG/JPEG), animated image (GIF), and video (MP4). We also shed light on the possibilities of how this ever-increasing bot traffic might affect networked systems and their properties. Finally, we propose that automated identification of bot traffic should be used within traffic shaping and engineering policies, such that it can be de-prioritised.

Table 1: Types of bot traffic uploaded by Twitter users.

Type	Description
URL & schemes	URL hosts and URI schemes (4,849 http and 289,074 https instances). These are extracted from the `[text]` tweet attribute. 162,492 URLs by bots and 131,431 by humans.
photos (JPG/JPEG)	A photos is extracted from the URL in `[media_url_https]` attribute. In total 23.31 GB of photo data is uploaded by 3,062 bots and humans in one month.
animated images (GIF)	Though these are animated photos, Twitter saves the first image in the sequence as a photo, and the animated sequence as a video under the `[video_info]` attribute. In total 2.92 GB of animated image data is uploaded.
videos (MP4)	Video files accompany a photo which is extracted by Twitter from one of the frames of the video. A video is pointed to by the URL in `[video_info][url]` attribute. In total 16.08 GB of video data is uploaded.

2 METHODOLOGY AND RESULTS

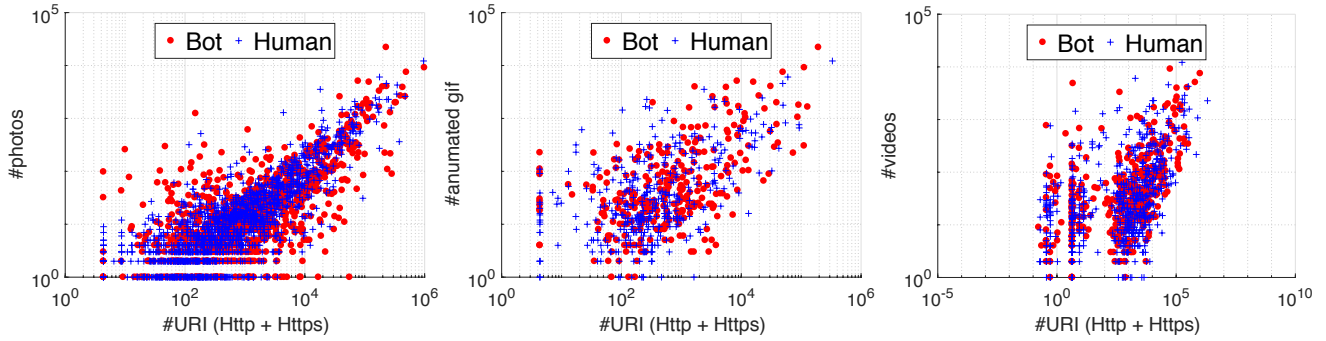
2.1 Data Collection

We focus on Twitter as a core platform serving bots. We use the Twitter Streaming API to collect a sampled set of Tweets. Following this, we use our previous work, ***Stweeler***[4] [4], to classify accounts as either bots or humans. The data consists of 722,109 tweets generated by 3,062 accounts in one month: 42.58% are bots and 57.42% are humans. For each tweet created, we extract the media and URLs. Importantly, Twitter automatically creates different resolutions of photos and videos, as well as generating images from animated

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

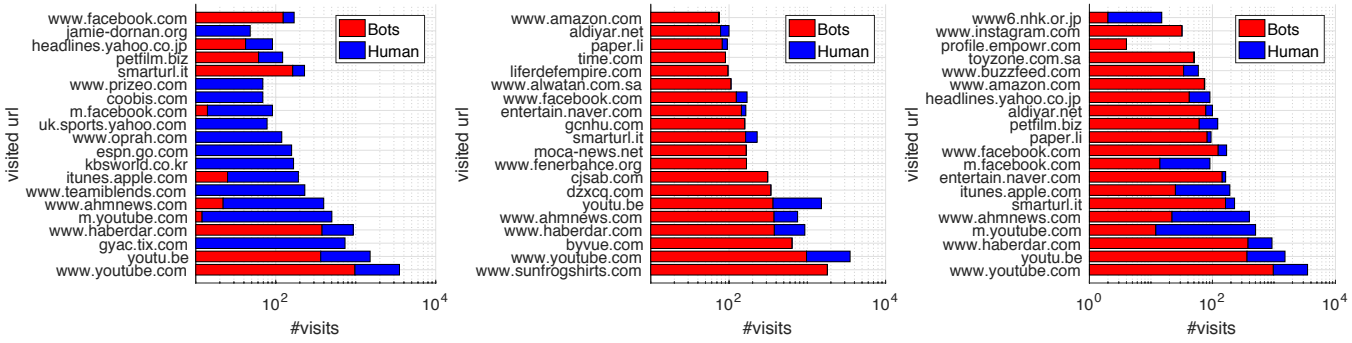
SIGCOMM Posters and Demos '17,
August 22–24, 2017, Los Angeles, CA, USA

ACM ISBN 978-1-4503-5057-0/17/08.
https://doi.org/10.1145/3123878.3131983



(a) No. of photos (JPG/JPEG) uploaded by bots & humans per URI (http + https).
(b) No. of animated images (GIF) uploaded by bots & humans per URI (http + https).
(c) No. of videos (MP4) uploaded by bots & humans per URI (http + https).

Figure 1: Media (photos, animated images, videos) uploaded by bots and humans on Twitter.



(a) Human popular URLs
(b) Bot popular URLs
(c) Combined popular URLs

Figure 2: Visiting trends to popular URLs by bots and humans.

sequences or videos to accompany static display with each dynamic media. Note that we are *only* considering the media originally uploaded by users. This is pointed to by `[sizes][large]`. We do not consider media created or uploaded by Twitter. Complete details of the dataset can be found in [3].

2.2 Data Analysis

Our data reveals a significant presence of content generated by bots (Figure 1). In total, bots account for 55.35% (12.90 GB) of the total photo traffic uploaded on Twitter; 53.58% (1.56 GB) of the total animated image traffic uploaded; and 40.32% (6.48 GB) of the total video traffic uploaded on Twitter. This is despite the fact that they only constitute 42.58% of the accounts under study and generate 53.90% of the tweets. When combined, bots account for a total of 49.52% (20.95 GB) traffic uploaded on Twitter, which is as much as expected from their proportion in the dataset.

It is also worth noting that many bot accounts post URLs. In fact, 55.28% of all URLs are posted by bots. This is important because these have the potential to trigger further traffic generated amongst the accounts that view the tweets. To explore this, Figure 2 presents the most popular domains posted by bots and humans. Significant differences can be observed. For example, whereas humans tend to post mobile sites (*e.g.,* `m.youtube.com`, `m.facebook.com`), bots rather post the desktop version (*e.g.,* `youtube.com`, `facebook.com`). We also see a range of websites exclusively posted by humans, *e.g.,* `espn.com` and `oprah.com`. One can also see a few URLs posted by bots, but never by humans. The most regularly posted URL in our dataset is `sunfrogshirt.com`, which is actually a website for purchasing bespoke t-shirts. This highlights a common purpose of media posting on Twitter: spam and marketing. Note that bots infiltrate human popular URLs more often than humans infiltrate bot popular URLs. This shows that bots can reach further due to their automated ability and can considerably impact network traffic in unusual ways.

3 CONCLUSION AND IMPLICATIONS

Using a large-scale Twitter dataset, we have shown that bots inject significant proportions of network traffic via the uploading of media. Further, by regularly posting links, we posit that they trigger further traffic generation amongst their followers. Overall, bots have a greater propensity to upload material than humans. We therefore argue that Twitter, and similar services, should begin to explicitly factor this within their infrastructural design. Classification mechanisms already allow bots to be detected. Such bots, for example, could be downgraded in terms of Quality of Service priorities, or even have their uploads buffered/delayed until off-peak hours. As bots are automated this seems a sensible strategy, considering the more sensitive nature of user-perceived experience.

To conclude, we argue that bot traffic will impact many aspects of network operations, including traffic engineering, routing, cloud computing, edge computing, content distribution networks, and quality of service. Thus, understanding and addressing these observations is of increasing importance.

Acknowledgment: This work is partially funded by EU Metrics project (Grant EC607728) and EU H2020 ReThink.

REFERENCES

[1] Zi Chu, Steven Gianvecchio, Haining Wang, and Sushil Jajodia. 2010. Who is Tweeting on Twitter: Human, Bot, or Cyborg?. In *Proceedings of the 26th Annual Computer Security Applications Conference (ACSAC '10)*. ACM, New York, NY, USA, 21–30. https://doi.org/10.1145/1920261.1920265

[2] Chad Edwards, Autumn Edwards, Patric R Spence, and Ashleigh K Shelton. 2014. 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.

[3] Zafar Gilani, Reza Farahbakhsh, Gareth Tyson, Liang Wang, and Jon Crowcroft. 2017. Of Bots and Humans (on Twitter). In *Proceedings of the 9th IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM'17)*. https://doi.org/10.1145/3110025.3110090

[4] Zafar Gilani, Liang Wang, Jon Crowcroft, Mario Almeida, and Reza Farahbakhsh. 2016. Stweeler: A Framework for Twitter Bot Analysis. In *Proceedings of the 25th International Conference Companion on World Wide Web (WWW '16 Companion)*. 37–38. https://doi.org/10.1145/2872518.2889360

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