Bots and Automation over Twitter during the U.S. Election

COMPROP DATA MEMO 2016.4 / 17 NOV 2016

Bence Kollanyi
Corvinus University
kollanyi@gmail.com
@bencekollanyi

Philip N. Howard
Oxford University
philip.howard@oii.ox.ac.uk
@pnhoward

Samuel C. Woolley
University of Washington
samwooll@uw.edu
@samuelwoolley

ABSTRACT

Bots are social media accounts that automate interaction with other users, and political bots have been particularly active on public policy issues, political crises, and elections. We collected data on bot activity using the major hashtags related to the U.S. Presidential Election. We find that that political bot activity reached an all-time high for the 2016 campaign. (1) Not only did the pace of highly automated pro-Trump activity increase over time, but the gap between highly automated pro-Trump and pro-Clinton activity widened from 4:1 during the first debate to 5:1 by election day. (2) The use of automated accounts was deliberate and strategic throughout the election, most clearly with pro-Trump campaigners and programmers who carefully adjusted the timing of content production during the debates, strategically colonized pro-Clinton hashtags, and then disabled activities after Election Day.

WHAT ARE POLITICAL BOTS?

A growing number of political actors and governments worldwide are employing both people and bots to shape political conversation. [1], [2] Bots can perform legitimate tasks like delivering news and information, or undertake malicious activities like spamming, harassment and hate speech. Whatever their uses, bots on social media platforms are able to rapidly deploy messages, replicate themselves, and pass as human users.

Networks of such bots are called "botnets," a term combining "robot" with "networks" and a term that is generally used to describe a collection of connected computers with programs that communicate across multiple devices to perform some task. There are legitimate botnets, like the Carna botnet, which gave us our first real census of device networks, and there are malicious botnets, like those that are created to launch spam and distributed denial-of-service (DDoS) attacks and to engineer theft of confidential information, click fraud, cyber-sabotage, and cyberwarfare. [3], [4] Over social media, botnets are interconnected automated accounts built to follow and re-message one another. These social botnets, often comprised of hundreds of unique accounts, can be controlled by one user operating from a single computer.

Social bots are particularly prevalent on Twitter, but they are found on many different platforms that increasingly form part of the system of political communication in many countries. [5] Highly automated accounts post, tweet, or message of their own accord. The most rudimentary bot profiles lack basic account information such as coherent screen names or profile pictures. Such accounts have become known as "Twitter eggs" because the default profile picture on that social media site is of an egg. While social media users get access from front-end websites, bots get access directly through a code-to-code connection, mainly through the site's wide-open application programming interface (API) that enables real-time posting and parsing of information.

Bots are versatile, cheap to produce, and ever evolving. Unscrupulous Internet users now deploy bots beyond mundane commercial tasks like spamming. Bots are the primary applications used in carrying out DDoS and virus attacks, email harvesting, and content theft. A subset of social bots are given overtly political tasks and the use of political bots varies from country to country. Political actors and governments worldwide have begun using bots to manipulate public opinion, choke off debate, and muddy political issues. Political bots tend to be developed and deployed in sensitive political moments when public opinion is polarized. How were highly automated accounts used around Election Day in the United States?

SAMPLING AND METHOD

This data set contains approximately 19.4m tweets collected November 1-9, using a combination of hashtags associated with the primary Presidential candidates. Since our purpose is to discern how bots are being used to amplify political communication, the analysis focuses upon the 18.9m tweets captured.

Twitter provides free access to a sample of the public tweets posted on the platform. The platform's precise sampling method is not known, but the company itself reports that the data available through the Streaming API is at most one percent of the overall global public communication on Twitter any given time. [6] In order to get the most complete and relevant data set, the tweets were collected by following particular hashtags identified by the team as being actively used during the debate. A few additional tags were added in the week before the election as they rose to prominence. The programming of the data collection and most of the analysis were done by using the statistics package R.

Selecting tweets on the basis of hashtags has the advantage of capturing the content most likely to be about this important political event. The streaming API yields (1) tweets which contain the keyword or the hashtag; (2) tweets with a link to a web source, such as a news article, where the URL or the title of the web source includes the keyword or hashtag; (3) retweets that contain a message's original text, wherein the keyword or hashtag is used either in the retweet or in the original tweet; and (4) quote tweets where the original text is not included but Twitter uses a URL to refer to the original tweet.

Our method counted tweets with selected hashtags in a simple manner. Each tweet was coded and counted if it contained one of the specific hashtags that were being followed. If the same hashtag was used multiple times in a tweet, this method still counted that tweet only once. If a tweet contained more than one selected hashtag, it was credited to all the relevant hashtag categories.

Unfortunately, not enough users geotag their profiles to allow analysis of the distribution of this support around the world or within the United States. Furthermore, analyzing sentiment on social media such as Twitter is difficult. [7], [8] Contributions using none of these hashtags were not captured in this data set. It is also possible that users who used one or more of these hashtags, but were not discussing the election, had their tweet captured. Moreover, if people tweeted about the election, but did not use one of these hashtags or identify a candidate account, their contributions were not analyzed here. Any comparison with previous data memos should consider that they were are based on shorter sample periods taken during the presidential debates, taken on different days of the week, and use a larger number of relevant hashtags.

FINDINGS AND ANALYSIS

This sample allows us to draw some clear conclusions about the character and process of political conversation over Twitter during the election. Specifically, we are able to both parse out the amount of social media content related to the two major candidates and investigate how much of this content is driven by highly automated accounts. We can parse the volume of tweets by perspective, assess the level of automation behind the different perspectives, and evaluate the particular contribution of bots to the traffic on this issue.

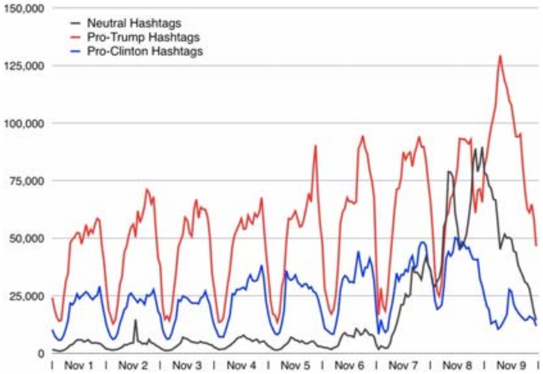
Comparing the Candidates on Twitter. Table 1 reveals that 18.90m tweets used some combinations of these hashtags. This table reveals that the overall volume of pro-Twitter Trump traffic (55.1 percent), was much greater than the volume of tweets containing only hashtags associated with the Clinton camp (19.1 percent). The overall volume of neutral election-related traffic (15.2 percent) was also significantly smaller than the pro-Trump traffic.

Table 1: Twitter Activity around Voting Day, 2016

	All Tweets in Sample	
	N	%
Pro-Trump	10,426,547	55.1
Pro-Clinton	3,618,778	19.1
Neutral	2,879,084	15.2
Trump-Neutral	434,897	2.3
Clinton-Neutral	217,509	1.2
Trump-Clinton	1,233,872	6.5
Trump-Clinton-Neutral	99,563	0.5
Total	18,910,250	100.0

Source: Authors' calculations from data sampled 1-9/11/16.
Note: Pro-Trump hashtags include #AmericaFirst, #benghazi, #CrookedHillary, #DrainTheSwamp, #lockherup, #maga3x, #MAGA, #MakeAmericaGreatAgain, #NeverHillary, #PodestaEmails, #projectveritas, #riggedelection, #tcot, #Trump2016, #Trump, #TrumpPence16, #TrumpTrain, #VoterFraud, #votetrump, #wakeupamerica; pro-Clinton hashtags include #Clinton, #ClintonKaine16, #democrats, #dems, #dnc, #dumptrump, #factcheck, #hillary2016, #Hillary, #HillaryClinton, #hillarysupporter, #hrc, #ImWithHer, #LastTimeTrumpPaidTaxes, #NeverTrump, #OHHillYes, #p2, #strongertogether, #trumptape, #uniteblue; neutral hashtags include *#Election2016, #Elections2016, #uselections, #uselection, #earlyvote, #iVoted, #Potus.*

Figure 1: Hourly Twitter Traffic, by Candidate Camp



Source: Authors' calculations from data sampled 1-9/11/16.
Note: This figure is based on the hashtags used in the tweets

Much smaller proportions of the tweets were categorized for mixes of hashtags. As human users made up their minds about whom to vote for and began expressing their preferences over Twitter, the proportion of clearly pro-Trump and pro-Clinton content using hashtags from each camp rose to 74.2 percent.

Figure 1 displays the rhythm of this traffic over the sample period. It reveals that, in contrast with the findings from our analysis of the debates, most tweets contained either pro-Trump or pro-Clinton hashtags. The use of neutral hashtags diminished by Election Day. Large dips in traffic coincide with night time in the United States. Figure 1 includes a total of 18.9m tweets from 3.7m users who tweeted using the sampled hashtags, but not the candidate's user names because the @ mentions reveal little about the political affinity of the user. During the election itself, the amount of candidate-committed traffic outstripped the volume of neutral traffic.

Automated Political Traffic. A fairly consistent proportion of the traffic on these hashtags was generated by highly automated accounts. These accounts are often bots that are either irregularly curated by people or actively maintained by people who employ scheduling algorithms and other applications for automating social media communication. We define a high level of automation as accounts that post at least 50 times a day using one of these election related hashtags, meaning 450 or more tweets on at least one of these hashtags during the data collection period.

Extremely active human users might achieve this pace of social activity, especially if they are simply retweeting the content they find in their social media feed. And some bots may be relatively dormant, waiting to be activated and tweeting only occasionally. But this metric captures accounts generating significant amounts of issue-specific traffic wherein high levels of automation probable. Finally, self-disclosed bots were identified by searching for the term "bot" in either the tag or account description. While this is a small proportion of the overall accounts, we expect the actual number of bots to be much higher—many bots, after all, are built to avoid obvious methods of identification. Future research will involve a more detailed analysis of the disclosed and hidden bots and searching for a wider range of terms referring to bots in the account name and description data.

Table 2 reveals the different levels of automation behind the traffic associated with clusters of hashtags. To track the activity of political bots around election time, we have clustered the hashtags by their candidate associations. To evaluate the role of automation, we organize these clusters of opinion based on hashtag use. After this, we create a subcategory of accounts that use high levels of automation. Table 2 indicates the level of traffic, by political camp and associated hashtags. This table distinguishes between the messages that exclusively used a hashtag known to be associated with a perspective and then the combinations of mixed tagging that are possible. When comparing the highly automated accounts tweeting for Trump versus those messaging for Clinton, it appears that the pro-Trump tweets out-numbered pro-Clinton tweets 5:1 during this period.

Table 2 also reveals that automation is used at several different levels by accounts taking different perspectives in the election. The accounts using exclusively neutral hashtags are rarely automated (only about 4 percent reveal a high level of automation). However, one-third of all the tweets using a mixture of all hashtags are generated by accounts that use high level of automation.

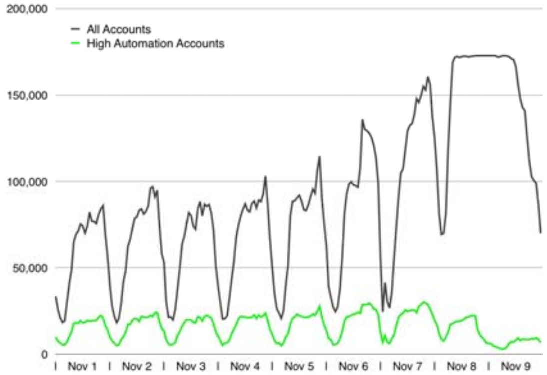
Figure 2 reveals the relative flow of traffic overall alongside traffic from accounts with high levels of automation. As with many Twitter-based conversations surrounding political events, the most active accounts here are either obvious bots or users with such high levels of automation that they are essentially bot-driven accounts—most likely making use of software applications to automate their Twitter presence and thus dominate conversation. During waking hours, highly automated accounts were generating between 20 and 25 percent of the traffic about the election during the days leading up to the vote. On Election Day, the server was recording 170K tweets per hour and we reached the cap set by Twitter for capturing data—again, one percent of global traffic captured in real time. The pace of automated political campaigning dropped off after Election Day—a reminder that campaigners and programmers behind bot accounts often disable their purpose-built automation on victory.

Table 2: Twitter Content, By Hashtag and Level of Automation

	Low %	High %	All N	%
	Exclusive Hashtag Clusters			
Pro-Trump	77.1	22.9	10,426,547	100
Pro-Clinton	86.4	13.6	3,618,778	100
Neutral	96.4	3.6	2,879,084	100
	Mixed Hashtag Clusters			
Trump-Neutral	83.0	17.0	434,897	100
Clinton-Neutral	92.3	7.7	217,509	100
Trump-Clinton	75.5	24.5	1,233,872	100
Trump-Clinton-Neutral	86.7	13.3	99,563	100
Sum	82.1	17.9	18,910,250	100

Source: Authors' calculations from data sampled 1-9/11/16.
Note: Low volume users are average human users, high volume accounts post more than 50 times per day on average.

Figure 2: Total Hourly Twitter Traffic around Voting Day, 2016, by Level of Automation



Source: Authors' calculations from data sampled 1-9/11/16.
Note: We define heavily automated accounts as tweeting 50 times or more per day on election topics.

Additional Observations on Automation. To understand the distribution of content production across these users, we then look at segments of the total population of contributors to these hashtags. There is a noticeable difference between the usage patterns of typical human users and accounts that are bots or otherwise highly automated. For example, the top 20 accounts, which were mostly bots and highly automated accounts, averaged over 1,300 tweets a day and they generated more than 234,000 tweets during this short period. The top 100 accounts, most of which still used high levels of automation, generated around 450,000 tweets at an average rate of 500 tweets per day. In contrast, the average account in the whole sample generated only one tweet every second day. While heavily automated accounts are usually the most active, there is a long tail of human users with only occasional Twitter activity.

Highly automated accounts—the accounts that tweeted 450 or more times with a related hashtag and user mention during the data collection period—generated close to 18 percent of all Twitter traffic about the Presidential election. Interestingly, Figure 2 also shows that automated postings significantly decreased the day after election whereas, in the days immediately before the election, highly automated accounts generated between as much 25 percent of all the Twitter activity on these political hashtags. That volume is significant, considering that this number of posts was generated by only 4,160 highly automated accounts in a sample of more than 3.7m users. It is very difficult for human users to maintain this rapid pace of social media activity without some level of account automation, though it is likely that not all of these are bot accounts.

CONCLUSIONS

Across the first three debates and the election (See Data Memos 2016.1, 2016.2, and 2016.3) we find that the proportion of highly automated twitter activity changed over time, increasing during the debates from 23 to 27 percent, and then dropping to 18 percent during the lead up to the election. The pace of highly automated pro-Trump social media activity grew from the first debate to the election. During the first debate, highly automated accounts generated four pro-Trump tweets for every pro-Clinton tweet. But by Election Day, the highly automated accounts generated five pro-Trump tweets for every pro-Clinton tweet.

Table 3 summarizes the important trends across the major events of the 2016 campaigns season. Pro-Trump traffic was many times higher than pro-Clinton traffic. Moreover, pro-Trump hashtags were inserted into more and more combinations of neutral and pro-Clinton hashtags, such that by the time of the election fully 81.9 percent of the highly automated content involved some pro-Trump messaging. In many kinds of Twitter conversations, this means that the pro-Trump accounts were moving into the political conversations that had previously involved neutral or pro-Clinton hashtags. The proportion of the overall sample generated by automation increased over the debates. This proportion appears to have diminished during the election—to 17.9 percent—but this reflects the longer sample period and the fact that many of the highly automated accounts were disabled after Election Day.

In the last debate 30.8 percent of the traffic about the debates was using relatively neutral hashtags, but this proportion was halved by Election Day. In the lead up to Election Day, only 15.2 percent of the traffic was using neutral hashtags, pro-Trump traffic grew from 46.7 to 55.1 percent and pro-Clinton traffic grew from 10.4 to 19.1 percent. As voters made up their minds about who to vote for, their expression of commitment solidified in the use of clear candidate specific hashtags. In the first debate 52.7 percent of the content was associated with a defined camp, but by the election 74.2 percent of the content was associated with one candidate or the other.

Table 3: Summary of Highly Automated Activity

	First Debate	Second Debate	Third Debate	Election
For each pro-Clinton tweet from a highly automated account, the number of pro-Trump tweets	4.4	4.2	6.9	4.9
Percent of pro-Trump content from highly automated accounts that either used pro-Trump hashtags or mixed with the pro-Clinton or Neutral hashtags	67.2	66.6	67.2	81.9
Proportion of hashtag sample generated by highly automated accounts	23.3	26.1	27.2	17.9

Source: Authors' calculations from data sampled during the first debate (26-29/09), second debate (9-12/10), third debate (19-22/10), and election (1-9/11).

Note: We define heavily automated accounts as tweeting 50 times or more per day on election topics.

In the first debate we scooped 9.0m tweets from 2.0m users who contributed to using 52 hashtags. For the second we scooped 11.5m tweets from 2.0m users who contributed to 66 hashtags. For the third we scooped 10.0m tweets from 1.6m users who contributed to 72 hashtags. For the election sample, we scooped 19.4m tweets from 3.7m unique users who contributed to 47 hashtags. We distinguish between relatively low activity users who tweet occasionally and highly automated accounts that generate more than 50 tweets day using at least one of these hashtags over the sample period.

Automated accounts tweeting with pro-Clinton hashtags increased their activities from over the course of the campaign period but still never reached the level of automation behind pro-Trump traffic. In this sample the dominance of highly automated pro-Trump tweets increased over automated pro-Clinton tweets to a level of 5:1.

We find that that political bot activity reached an all-time high for the 2016 campaign. Not only did the pace of highly automated pro-Trump activity increase over time, but the gap between highly automated pro-Trump and pro-Clinton activity widened from 4:1 during the first debate to 5:1 by election day. The use of automated accounts was deliberate and strategic throughout the election, most clearly with pro-Trump campaigners and programmers who carefully adjusted the timing of content production during the debates, strategically colonized pro-Clinton hashtags, and then disabled automated activities after Election Day.

ABOUT THE PROJECT

The Project on Computational Propaganda (www.politicalbots.org) involves international, and interdisciplinary, researchers in the investigation of the impact of automated scripts—computational propaganda—on public life. *Data Memos* are designed to present quick snapshots of analysis on current events in a short format. They reflect methodological experience and considered analysis, but have not been peer-reviewed. *Working Papers* present deeper analysis and extended arguments that have been collegially reviewed and that engage with public issues. The Project's articles, book chapters and books are significant manuscripts that have been through peer review and formally published.

ACKNOWLEDGMENTS AND DISCLOSURES

The authors gratefully acknowledge the support of the National Science Foundation, "EAGER CNS: Computational Propaganda and the Production / Detection of Bots," BIGDATA-1450193, 2014-16, Philip N. Howard, Principle Investigator and the European Research Council, "Computational Propaganda: Investigating the Impact of Algorithms and Bots on Political Discourse in Europe," Proposal 648311, 2015-2020, Philip N. Howard, Principal Investigator. Project activities were approved by the University of Washington Human Subjects Committee, approval #48103-EG and the University of Oxford's Research Ethics Committee. 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 or the European Research Council.

REFERENCES

[1] M. C. Forelle, P. N. Howard, A. Monroy-Hernandez, and S. Savage, "Political Bots and the Manipulation of Public Opinion in Venezuela," Project on Computational Propaganda, Oxford, UK, Working Paper 2015.1, Jul. 2015.

[2] P. N. Howard and B. Kollanyi, "Bots, #StrongerIn, and #Brexit: Computational Propaganda during the UK-EU Referendum," *arXiv:1606.06356 [physics]*, Jun. 2016.

[3] "Carna botnet," *Wikipedia*. 24-Nov-2015.

[4] "Denial-of-service attack," *Wikipedia*. 15-Oct-2016.

[5] A. Samuel, "How Bots Took Over Twitter," *Harvard Business Review*, 19-Jun-2015. [Online]. Available: https://hbr.org/2015/06/how-bots-took-over-twitter. [Accessed: 23-Jun-2016].

[6] F. Morstatter, J. Pfeffer, H. Liu, and K. M. Carley, "Is the Sample Good Enough? Comparing Data from Twitter's Streaming API with Twitter's Firehose," *arXiv:1306.5204 [physics]*, Jun. 2013.

[7] Z. Chu, S. Gianvecchio, H. Wang, and S. Jajodia, "Who is tweeting on Twitter: human, bot, or cyborg?," in *Proceedings of the 26th annual computer security applications conference*, 2010, pp. 21–30.

[8] Cook, David, Waugh, Benjamin, Abdinpanah, Maldini, Hashimi, Omid, and Rahman, Shaquille Abdul, "Twitter Deception and Influence: Issues of Identity, Slacktivism, and Puppetry," *Journal of Information Warfare*, vol. 13, no. 1.

[9] P. N. Howard, *Pax Technica: How the Internet of Things May Set Us Free*. New Haven, CT: Yale University Press, 2015.

[10] D. W. Butrymowicz, "Loophole.com: How the Fec's Failure to Fully Regulate the Internet Undermines Campaign Finance Law," *Columbia Law Review*, pp. 1708–1751, 2009.

[11] P. N. Howard, *New Media Campaigns and the Managed Citizen*. New York, NY: Cambridge University Press, 2006.