Computational
Propaganda
Research Project

Measuring Traffic Manipulation on Twitter

Ben Nimmo, ***Atlantic Council***

UNIVERSITY OF
OXFORD

OXFORD
INTERNET
INSTITUTE

Table of Contents

Executive summary 4

Introduction 5

Defining "Twitter traffic manipulation" 6

Tools and techniques of manipulation 7

- Bots 7
- Botnets 7
- Cyborgs 8

Comparing manipulated and organic content 8

Indicators of manipulation 8

Methodology 9

Indicators and results 10

- Average Posts per User 10
  - Covfefe 10
  - Davos 11
  - #4thofJuly 11
  - Wednesday, Thursday, and Friday 11
  - Organic Traffic – Characteristics 11
  - #StopAstroturfing 12
  - #DigDoug 13
  - Qadhafi of the Gulf and Tamim the Glorious 13
  - Pro-Le Pen hashtags 13
  - Posts Per user – Organic vs Manipulated Traffic 14
- Proportion of Retweets 14
- Proportion of Traffic from the 50 Most Active Accounts 16

The Coefficient of Traffic Manipulation 17

Conclusion 19

Appendix 1: Case studies 20

- Control series 20
  - Covfefe 20
  - Davos 20
  - Wednesday, Thursday, and Friday 20
  - #4thofJuly 21
- Manipulated series 21
  - #StopAstroturfing 21
  - Gulf Crisis 23
  - #DigDoug 24
  - French Election 26

Appendix 2: Archived posts 30

- Polish study 30
- US study 30
- French study 30

BIBLIOGRAPHY 34

Table of Tables

Table 1 Scan Covefefe 10
Table 2 Scan Davos 11
Table 3 Scan #4thofJuly 11
Table 4 Scan weekdays 11
Table 5 Characteristics of organic traffic 12
Table 6 #StopAstroturfing 12
Table 7 Scan #DigDoug 13
Table 8 Scan Qadhafi of the Gulf and Tamim the Glorious 13
Table 9 Scan Pro-Le Pen hashtags 14
Table 10 Posts per user 14
Table 11 Control samples posts per user 14
Table 12 Known manipulated samples posts per user 15
Table 13 Proportion of retweets – organic vs manipulated traffic 15
Table 14 Control sample posts from top 50 most active accounts 16
Table 15 Manipulated samples posts from top 50 most active accounts 16
Table 16 Posts from 50 most active accounts – organic vs manipulated traffic 17
Table 17 Coefficient of traffic manipulation by scan 17
Table 18 Coefficient of traffic manipulation by scan without extreme cases 18

Table of Figures

Figure 1 Traffic on #StopAstroturfing and #StopNGOSoros, July 22, 2017 21
Figure 2 Tweets and retweets by @annawitme, posting the identical wording 22
Figure 3 Traffic on "Qadhafi of the Gulf," August 27, 2017 23
Figure 4 Traffic on "Tamim the Glorious," September 19, 2017 23
Figure 5 Retweets of the same post made at the same second by an apparently Turkish botnet 24
Figure 6 Traffic on #DigDoug, November 15–16, 2017, from machine scan 25
Figure 7 The number of times each of the most active accounts posted #DigDoug 26
Figure 8 Tweet by @messsmer, archived at http://archive.is/G1Hjj. 26
Figure 9 Different memes in the #LaFranceVoteMarine shared by cyborg @ImRiyrr. 28
Figure 10 Near-simultaneous posts on #LaFranceVoteMarine on hashtag launch 29
Figure 11 Traffic on #Marine2017, February 14, 2017, from machine scan 29

EXECUTIVE SUMMARY

Twitter is a major platform for political communication and campaigning, used extensively by political parties, lobbying groups, and activists in many democratic debates around the world.

It has proven to be vulnerable to traffic manipulation by small but coordinated user groups, and those who control automated accounts, known as "bots". These have demonstrated an ability to materially distort Twitter traffic, forcing chosen phrases and hashtags into the "trending" lists, and generating very high volumes of traffic from a very small base of human users.

This paper proposes a computational method to calculate the extent to which a given flow of Twitter traffic has been subject to manipulation by such groups. It examines three indicators:

- the average number of tweets (including retweets) per user;
- the percentage of retweets as a proportion of total traffic;
- the proportion of traffic generated by those fifty accounts which used the given term most often.

These three factors indicate whether Twitter traffic was generated organically by a large number of users or pushed by a small one; whether it was driven by a high proportion of original posts, or by large-scale retweeting; and whether it was driven by a small user group, or a broader movement.

We combine the three factors into a single metric, the Coefficient of Traffic Manipulation (CTM). This is a relative measure, rather than an absolute one. It allows us to compare different Twitter traffic flows against measurable criteria and assess which of those movements appear to have been most subject, or least subject, to manipulation.

As such, it can serve as an early-warning system for researchers, indicating Twitter traffic flows which do appear to have been manipulated. These can then be subject to further study, to determine what manipulation measures were used.

INTRODUCTION

Twitter has served as a primary medium for political campaigning at least since the election of then-US President Barack Obama in 2012 (Rutledge 2013).However, the platform is vulnerable to manipulation by small groups of users. Malicious actors have, for example, used manipulative techniques to spread political disinformation (Agarwal et al. 2017), game Twitter's "trending" algorithm (Yubao Zhang et al. 2017), and propagate low-quality or "fake" news (Shao et al. 2017). Such manipulated traffic can reach a significant number of internet users. In January 2018, 11% of users in the United Kingdom, United States and ten other countries said they had used Twitter for news in the previous week; 23% had used it for any purpose (Newman et al. 2018).

Ironically, the relative ease with which Twitter traffic can be manipulated stems from the platform's own high level of data transparency, combined with the low level of transparency demanded of Twitter users. Twitter's willingness to grant developers access to its Application Programming Interface (API) means that many commercial and amateur programmers have been able to create apps which allow users to automate tweets, likes, and follows (Ferrara et al. 2016; Davis et al. 2016).

The relative lack of transparency demanded of users at the moment of account creation has made it possible for individuals to create hundreds or thousands of separate accounts (Cox 2017). Taken together, these factors have allowed users to create large numbers of automated accounts, known as "bots", and form them into synchronised networks of remote-controlled accounts, called "botnets".Such botnets allow the controller to generate tens of thousands of tweets and retweets over a short period, thus creating the misleading appearance of a substantial organic movement, involving many thousands of people (Kollanyi, Howard, and Woolley 2016; DiResta et al. 2017).

While widely reported, bots and botnets are only one form of online manipulation: users have developed a complex range of techniques to amplify their messages. These include the simultaneous posting of content at a set time, either by automation or by coordination, sometimes referred to as a "thunderclap"; the coordinated posting of high volumes of content by a low number of highly active users, typically pre-planned on a platform which is not visible to the public, such as in a direct-messaging group; and, at the most labour-intensive end of the spectrum, the employment of voluntary or paid "troll armies" to manage multiple accounts (Gallagher 2016).

These complex efforts vary in detail, but share a common goal: to manipulate online debate by generating a very large volume of content from a small group of users, while camouflaging the actual size of the group in question, thereby creating the spurious impression of a large, spontaneous, and popular movement. Sometimes, the effort is directly aimed at Twitter's "trending" algorithm and is designed to bring the chosen message to the attention of more users by forcing it to trend (Yubao Zhang et al. 2017); at other times, it appears more diffuse, aimed at generating a large volume of traffic for its own sake.

Both Twitter and independent users have sought to crack down on automated manipulation. In 2016, Davis et al. launched their "Botornot" online service to help users identify bots. In January 2018, Twitter reported that its automated systems prevented over half a million suspicious logins daily "for being generated through automation" (Twitter Public Policy 2018). On February 21, 2018, Twitter announced changes to the rules governing the use of its API by developers, aimed at banning large-scale duplication of identical messages across accounts (Roth 2018). At the time of this writing, it was unclear how effective these changes would be.

Despite these efforts, identifying and tackling complex traffic manipulation, which uses both automated and human accounts, remains a challenge. Fortunately, Twitter's unusual degree of data transparency also enables researchers to conduct large-scale analyses of traffic flows, identifying the key characteristics of organic and manipulated traffic. Based on such analyses, this paper proposes a computational method for identifying how much any given Twitter traffic flow has been subjected to deliberate manipulation. We term this method the Coefficient of Traffic Manipulation (CTM).

The purpose of the CTM is to identify traffic flows within Twitter which appear to have been subjected to a range of manipulative behaviour patterns, and to distinguish them from genuinely spontaneous, organic movements. Such manipulative behaviour could include the intervention of a large number of bots, a small number of hyperactive human users, or any combination of the two. The CTM is not intended to distinguish between these types of behaviour, or to sort between bots and humans. Its intent is to provide an initial analysis of any given flow, based on data which are readily available from Twitter's public APIs, to identify those which appear to have been manipulated, so that they can be studied in more depth.

The CTM is relative, rather than absolute. Our goal is to situate any given traffic flow between Twitter movements which are known to have been organic, and movements which are known to have been manipulated.

Flows which appear to have been manipulated can then be subjected to more detailed analysis, to determine the extent of any manipulation, expose how it was carried out, and propose remedial action. As such, the CTM method can serve as an early-warning system for attempts to manipulate Twitter traffic, regardless of whether the attempt was conducted using bots, humans, or a combination of methods.

DEFINING "TWITTER TRAFFIC MANIPULATION"

We define "Twitter traffic manipulation" as "an attempt by a small group of users to generate a large flow of Twitter traffic, disproportionate to the number of users involved."[1] Our definition thus hinges on the question of "proportionate" traffic; that is, a level of traffic activity which can be considered "normal" when compared with known organic Twitter traffic flows. We derive the parameters of "normal", organic traffic from Twitter flows which can be

[1] Traffic manipulation on Twitter is often referred to as "astroturfing", in the sense of an attempt to create the semblance of a grassroots movement. We recognise the term, but eschew it in this paper, as "astroturfing" can also refer to various physical manifestations of manipulation, such as paying people to attend rallies.

proven, or safely assumed, to have been free of large-scale manipulation attempts.[2] These control traffic flows are described in detail in Appendix 1.

TOOLS AND TECHNIQUES OF MANIPULATION

Twitter traffic flows are manipulated using a range of tools. Our Coefficient of Traffic Manipulation does not distinguish between such tools; however, it is important to understand them, as they form the basis of our calculations. Traffic manipulation on Twitter typically uses a combination of the following assets:

1. Automated bot accounts, often coordinated in a large network (botnet);
2. Partially-automated accounts ("cyborgs"), which show some human behavioural patterns, but post with machine-like frequency (Zi Chu et al. 2012);
3. Human-run accounts.

Bots

Bots can be defined as "social media accounts that automate interaction with other users" (Kollanyi, Howard, and Woolley 2016). Not all bots are harmful: some have benign uses, such as news aggregators and automated reply services (Ferrara et al. 2016). Other bots share poetry and photography.[3] Such bots tend to be open about their nature.

However, some bots are malicious, and masquerade as human users for the purposes of deception. Classification of such malicious bot falls into three categories: spambots, which spread spam on various topics; paybots, designed to attract users to pay-per-click advertising sites; and influence bots, which are intended to influence the online conversation on a given topic by amplifying selected content (Subrahmanian et al. 2016).

Influence bots can be used for various purposes. The goal can be to force a chosen hashtag into the "Trending" lists (Yubao Zhang et al. 2017), or to silence or intimidate users (Alfonso 2012). In especially charged political moments, it may be to swamp online discussion with unrelated content (Alexander 2015). Some influence bots appear to be operated individually, achieving their impact by posting very high volumes of content – sometimes in excess of 600 posts per day.[4] However, such rapid posting risks detection by Twitter's algorithm. One solution is to harness a large number of bots together in a botnet.

Botnets

Botnets are networks of automated accounts which are programmed to behave in the same way without human intervention. Typically, influence botnets achieve their effect by mass-automating retweets, likes and follows, or by mass-posting specific texts (Barojan 2018).nIn

[2] The prevalence of automated systems on Twitter has historically been such that we consider all Twitter traffic flows, except the smallest, to be subject to a degree of incidental manipulation. Our purpose is to identify attempts at manipulation which are large enough to substantially distort traffic.

[3] For example, https://twitter.com/PoetryBot01 and https://twitter.com/QUEEN_PHOTObot.

[4] We use the term "posts" to refer to tweets (i.e. apparently original content) and retweets.

general, networked bots post at a lower rate than solo bots; the impact on traffic stems from the aggregate number of accounts involved. Individual botnets counting tens of thousands of accounts have been observed (Alexander 2015).

Botnets allow individual users to generate tens of thousands of posts over a short period, creating the misleading appearance of a substantial organic movement (Kollanyi, Howard, and Woolley 2016; DiResta et al. 2017). However, running a botnet is a violation of Twitter's terms of service. Botnets which Twitter detects can be suspended *en masse*; bot herders have therefore adopted a range of techniques to mask their activities from Twitter's detection algorithms (Agarwal et al. 2017).

Cyborgs

One of the most common masking techniques is to subject an account to automation, but for a human user to post periodic comments and replies from it, so that its behaviour pattern appears less machine-like. These semi-automated accounts are known as "cyborgs". Cyborgs typically behave in a manner akin to high-volume bots, posting rapidly (often in the order of 100–250 times per day) and with a high proportion of retweets (typically around 75–-90%), but interspersing these with a substantial minority of authored posts.

The intervention of human activity makes cyborgs harder to detect by algorithm. It also makes it harder for Twitter to justify suspending the account, since it is not purely automated. The requirement for human intervention also means that cyborgs are harder to deploy on an industrial scale. They can therefore supplement bot activity, but are unlikely, in the short term, to replace it entirely.

COMPARING MANIPULATED AND ORGANIC CONTENT

The precise combination of human, bot, and cyborg accounts used in each manipulation campaign varies. However, most campaigns have the same, ultimate goal: to generate a high volume of posts from a low number of human users, creating the false impression of a large-scale organic movement. This being so, there is only a limited number of ways in which the manipulation can be achieved:

1. Each account tweets a high number of authored posts;
2. Each account tweets a small number of authored posts, and a very high number of retweets;
3. A large number of bots or cyborgs retweets authored posts;
4. A large number of bots or cyborgs posts identical content.

In practice, most manipulation attempts which we have observed use a combination of methods 1, 2, and 3. Twitter's own algorithm appears broadly effective in preventing the large-scale simultaneous posting of identical content (method 4).

INDICATORS OF MANIPULATION

Based on this understanding of the methodology of manipulation, we can identify three numerical factors which are likely to indicate an attempt at traffic manipulation.
These indicators are:

A. The average number of posts per account;
B. The proportion of retweets in the total traffic flow;
C. The number of posts from the fifty most active accounts.

These indicators correspond to the main methods which malicious actors use to create the impression of large-scale traffic. A small user group can programme a botnet to retweet their content on an industrial scale (B) or post very actively themselves (C). Since the whole point is to generate a disproportionately large number of posts from a relatively small group, either activity is likely to result in a disproportionately high average number of posts per account (A).

By combining these indicators in a composite Coefficient of Traffic Manipulation, we can generate an initial assessment of any given traffic flow, compared with known organic flows. The CTM can highlight subject flows which were generated by a disproportionately small user base, as would be the case if a small group of individuals had tried to "game" the algorithm by each posting many times in a short period; traffic featuring a disproportionate number of retweets, as would be the case if it had been amplified by retweet bots; traffic featuring an overall number of users disproportionate to the volume of traffic, as would be the case in a planned and coordinated posting campaign by a small group; and any combination of the above.

The purpose of the CTM is not to distinguish between those methods in the first instance, but to identify traffic flows which merit further study, so that the exact nature of the traffic can be established.

METHODOLOGY

To test our hypothesis, we conducted a series of machine scans of Twitter traffic. Data were collected by searching for a given term (a hashtag or keyword) over a set time period, using Twitter's public APIs. Each search recorded all mentions of the selected search term, to an upper limit of 100,000 posts, and laid them out in chronological order. Each individual post was identified by the content, the account from which it was posted, date, time, and status as a tweet or retweet.

We divided our samples into a control series and a series of manipulated flows, with seven scans in each series. In our control series, we were confident, based on contextual evidence, that the traffic had not been subjected to large-scale manipulation. We drew our series of manipulated flows from earlier studies, in which we had demonstrated the presence of large-scale attempts at manipulation by both bot and human users (Nimmo 2017a, 2017b, 2017c). The detailed analysis of our control and manipulated flows, together with the reason for including each term, is set out in Appendix 1. Archive links to the posts we reference are provided in Appendix 2.

The search terms for our control series were:

- The word "covfefe", collected on May 31, 2017;

- The word "Davos", collected in two scans in January 2018, before and during the World Economic Forum in Davos;
- The hashtag #4thofJuly, collected on July 4, 2017;
- The word "Wednesday", collected on September 11, 2018;
- The word "Thursday", collected on September 12, 2018;
- The word "Friday", collected on September 21, 2018.

The search terms for our manipulated samples were:

- The hashtags #Marine2017, #LePionMacron, and #LaFranceVoteMarine, collected in February 2017;
- The hashtag #StopAstroturfing, collected in June 2017;
- The hashtags #قذافي_الخليج ("Qadhafi of the Gulf") and #تميم_المجد("Tamim the Glorious"), collected in September 2017;
- The hashtag #DigDoug, collected in December 2017.

Having collected the raw data, we established the total number of posts (tweets + retweets) in the traffic flow, the number of unique accounts from which those accounts had come, and the number of tweets compared with the number of retweets. This allowed us to calculate the average number of posts per user, the percentage of posts generated by the fifty most active accounts, and the proportion of retweets to total traffic.

Based on these indicators, we calculated the overall Coefficient of Traffic Manipulation for each traffic flow. We found that the Coefficient of Traffic Manipulation for our control samples was substantially lower than any of the results for our manipulated series. The variation between manipulated samples was far higher than the variation between control samples. We therefore concluded that the Coefficient of Traffic Manipulation can give a reliable indication of the degree to which a given traffic flow has been subject to some form of attempted manipulation, separating out organic flows from those which have been subjected to some, or extreme, manipulation.

INDICATORS AND RESULTS

Average Posts per User

Our first measure was the average number of posts per user. We considered this to be the most reliable single indicator of attempted manipulation, as it expresses the overall intensity of posting within the Twitter flow, regardless of whether the posts come from humans, bots, or cyborgs.

Covfefe

On "covfefe", our scan gathered a total of 90,000 posts, representing 41 minutes of traffic (the number of posts was defined by search quota limitations at the time). We found that these posts were generated by 42,415 users, for an average of 2.14 posts per user:

Table 1 Scan Covefefe

Phrase	Number of posts	Number of users	Ave posts / user
Covfefe	90,737	42,415	2.14

Davos

On "Davos", we conducted two scans of 100,000 posts each, covering the week prior to the World Economic Forum 2018, and the opening of the forum itself. Our first scan ran from 16 to 22 January, 2018, during the preparatory phase; the second ran from 22 to 23 January, 2018. Our first scan collected 100,000 posts from 65,120 users, for an average of 1.54 posts per user. The second collected the same number of posts from 63,037 users, for an average of 1.59 posts per user.

Table 2 Scan Davos

Phrase	Number of posts	Number of users	Ave posts / user
Davos, scan 1	100,000	65,120	1.54
Davos, scan 2	100,000	63,037	1.59

#4thofJuly

Our next control scan, #4thofJuly, was particularly important. Given the popularity of the Fourth of July celebration, we expected the hashtag to be subjected to a range of incidental manipulation by users adopting it for unrelated purposes, such as spambots, paybots, and marketing accounts. We therefore took #4thofJuly as a benchmark of the extent to which such incidental amplifiers distort traffic. In fact, our scan of 99,866 posts was generated by 85,575 users, for an average of 1.17 posts per user. This was lower than we had expected, and the lowest measure for any of our scans.

Table 3 Scan #4thofJuly

Phrase	Number of posts	Number of users	Ave posts / user
#4thofJuly	99,866	85,575	1.17

Wednesday, Thursday, and Friday

Our final scans considered the words "Wednesday", "Thursday", and "Friday" on the appropriate days of the week in September 2018. These words are so generic that we considered them to represent a further baseline measure of online activity, with incidental, but not targeted, manipulation. Each scan was set to gather 100,000 tweets; in each case, the posts came from a little over 80,000 users. This gave an average of between 1.2 and 1.25 posts per user, falling between #4thofJuly and the Davos scans in our range.

Table 4 Scan weekdays

Phrase	Number of posts	Number of users	Ave posts / user
Wednesday	99,871	80,518	1.24
Thursday	99,909	82,626	1.21
Friday	100,000	82,226	1.22

Organic Traffic – Characteristics

These control samples suggest that the typical ratio of posts per user for organic traffic, even on politically charged phrases, lies in a range between 1.1 and 2.2, taking covfefe as our upper boundary, and #4thofJuly as our lower one. Our decision to keep our control samples to a similar size meant that they were generated over widely differing timespans. The 90,000 tweets on covfefe were generated in just 41 minutes. Our first scan on Davos covered a week.

To verify whether the rate of posting remained constant over different periods, we corrected all our scans to cover a 41-minute range. (For the slowest traffic, we chose 41-minute ranges when traffic was close to its peak rate, to generate a statistically meaningful sample.)

Table 5 Characteristics of organic traffic

Phrase	Number of tweets in 41 minutes	Number of users	Ave tweets / user
Covfefe	90,737	42,415	2.14
#4thofJuly[5]	20,767	18,599	1.17
Davos, scan 1[6]	1,558	1,436	1.08
Davos, scan 2[7]	6,112	4,853	1.26
Wednesday[8]	5,717	4,890	1.17
Thursday[9]	4,393	3,860	1.14
Friday[10]	17,898	15,884	1.13

In fact, correcting for time gave only a minor variation in range, from 1.08 to 2.14 posts per user on average. Again, covfefe marked the upper limit of the range; this time, our first Davos scan marked the lower limit. This suggests that an average number of tweets per user in the range of 1.1 to 2.2 can be considered normal for organic Twitter traffic on political or current events, regardless of the time period involved. With this "normal" range of averages established, we calculated the average for our manipulated samples.

#StopAstroturfing

We encountered the hashtag #StopAstroturfing in Poland (Nimmo 2017b). This Twitter flow was marked by a particularly blatant use of bots and cyborgs. We therefore considered it our most primitive case of attempted manipulation. The total number of tweets was lower than in our control samples, at just under 16,000; however, these came from just 2,408 users, for an average of 6.61 tweets per user. This is three times the upper limit of our control range, a clearly suspicious figure.

Table 6 #StopAstroturfing

Phrase	Number of tweets	Number of users	Ave tweets / user
#StopAstroturfing	15,915	2,408	6.61

[5] Scan limited from 12:49 to 13:30 on July 4, 2017.

[6] Scan limited from 16:00 to 16:41 on January 19, 2018.

[7] Scan limited from 13:52 to 14:33 on January 23, 2018.

[8] Scan limited from 06:00 to 06:41 on September 12, 2018.

[9] Scan limited from 08:00 to 08:41 on September 13, 2018.

[10] Scan limited from 11:00 to 11:4 on September 21, 2018.

#DigDoug

We encountered our American sample, #DigDoug, in a study of Twitter traffic around the Alabama senatorial election (Nimmo 2017c). The traffic was primarily driven by human users and apparent cyborgs, rather than large-scale bot use; it was fuelled by a very significant degree of coordination within a small user group. Our scan returned 42,939 posts from 8,982 users, for an average of 4.78 posts per user. While below the level of the Polish example, this was still more than twice the upper level of our "normal" range.

Table 7 Scan #DigDoug

Phrase	Number of tweets	Number of users	Ave tweets / user
#DigDoug	42,939	8,982	4.78

Qadhafi of the Gulf and Tamim the Glorious

We encountered the phrases "Qadhafi of the Gulf" and "Tamim the Glorious" in a study of Twitter traffic around the dispute between Saudi Arabia and Qatar in August to September 2017. Our analysis showed a combination of bot amplification and apparently coordinated human use. Traffic on "Qadhafi of the Gulf" generated 33,976 posts from 11,548 users, for an average of 2.94 posts per user. Traffic on "Tamim the Glorious" was substantially higher, generating 93,429 posts from 25,582 users, for an average of 3.65 posts per user. Both these results lay well above the upper boundary of our "normal" range, but below our earlier manipulated samples.

Table 8 Scan Qadhafi of the Gulf and Tamim the Glorious

Phrase	Number of posts	Number of users	Ave posts / user
Qadhafi of the Gulf	33,976	11,548	2.94
Tamim the Glorious	93,429	25,582	3.65

Pro-Le Pen hashtags

We studied the hashtags #Marine2017, #LePionMacron, and #LaFranceVoteMarine during the build-up to the French presidential election (Nimmo 2017a). These hashtags were driven by the most sophisticated manipulation operation we have studied, combining bots, cyborgs, coordinated human posting, and a large stock of prepared content.

On each occasion, these campaigns managed to make their hashtag the top trending phrase on French Twitter, marking them out as extremely effective. The earliest flow, #Marine2017, generated 24,001 posts from 2,723 users, for an average of 8.81 posts per user. #LePionMacron, one week later, raised 30,673 posts from 5,166 users, for an average of 5.94 posts per user. Finally, #LaFranceVoteMarine generated 47,075 posts from 8,860 users, for an average of 5.31 posts per user.

All of these results lay well above the upper limit of our "normal" range, reflecting the very considerable degree of manipulation by amplifier bots and cyborgs. However, the average diminished over time, with #LaFranceVoteMarine generating twice as many posts as #Marine2017 from almost four times as many accounts. This suggests that the online operation successfully managed to attract more users (whether human or automated) to its campaigns, as they evolved.

Table 9 Scan Pro-Le Pen hashtags

Phrase	Number of posts	Number of users	Ave posts / user
#Marine2017	24,001	2,723	8.81
#LePionMacron	30,673	5,166	5.94
#LaFranceVoteMarine	47,075	8,860	5.31

Posts Per user – Organic vs Manipulated Traffic

When set out in order, these scans showed a consistent difference between organic and manipulated flows. All our control flows returned lower rates of posting per user than any of our manipulated samples. The lowest average number of posts per user for a manipulated flow (Qadhafi of the Gulf) was one-third higher than the highest average number of posts for an organic one (covfefe).

Table 10 Posts per user

Phrase	Number of posts	Number of users	Ave posts / user
#4thofJuly	99,866	85,575	1.17
Thursday	99,909	82,626	1.21
Friday	100,000	82,226	1.22
Wednesday	99,871	80,518	1.24
Davos, scan 1	100,000	65,120	1.54
Davos, scan 2	100,000	63,037	1.59
Covfefe	90,737	42,415	2.14
Qadhafi of the Gulf	33,976	11,548	2.94
Tamim the Glorious	93,429	25,582	3.65
#DigDoug	42,939	8,982	4.78
#LaFranceVoteMarine	47,075	8,860	5.31
#LePionMacron	30,673	5,166	5.94
#StopAstroturfing	15,915	2,408	6.61
#Marine2017	24,001	2,723	8.81

We therefore conclude that the average number of posts per user is a reliable indicator of one particular method of manipulating Twitter traffic.

Proportion of Retweets

Our second measure was the percentage of retweets within the total posts. Given the role of retweet bots in many manipulation campaigns, we considered that this measure would indicate how much traffic was driven by users posting comments, or by otherwise silent retweeters, whether human or bot. Our control samples returned figures of between 49.66% (for "Friday") and 75.12% (for our second Davos scan).

Table 11 Control samples posts per user

Phrase	Total posts	Retweets	Proportion of retweets
Friday	100,000	49,659	49.66%

Covfefe	90,737	50,457	55.61%
#4thofJuly	99,866	58,267	58.34%
Thursday	99,909	63,902	63.96%
Davos 1	100,000	66,167	66.17%
Wednesday	99,871	69,222	69.31%
Davos 2	100,000	75,124	75.12%

Our manipulated samples returned figures of between 66.78% (for #DigDoug) and 94.95% (for "Qadhafi of the Gulf").

Table 12 Known manipulated samples posts per user

Phrase	Total posts	Retweets	Proportion of retweets
#Marine2017	24,001	20,892	87.05%
#LePionMacron	30,673	27,385	89.28%
#LaFranceVoteMarine	47,075	39,939	84.84%
Qadhafi of the Gulf	33,976	32,286	94.95%
Tamim the Glorious	93,429	72,315	77.40%
#StopAstroturfing	15,915	13,889	87.27%
#DigDoug	42,939	28,673	66.78%

These results did not reveal as clear a separation between our control and manipulated samples as the average number of posts per user: the manipulated #DigDoug recorded lower levels than "Wednesday" or our second Davos scan, with "Tamim the Glorious" only a little higher. However, none of our control samples returned a retweet proportion of over 80%; by contrast, only two of the manipulated samples showed a proportion of below 80%, while the "Qadhafi of the Gulf" sample was over 90%.

This last fact is particularly important, because "Qadhafi of the Gulf" returned the lowest average number of posts per user of any of our manipulated samples. Taken together, the two indicators suggest that the traffic in this sample was chiefly driven by an unusually large number of retweet amplifiers (bot or human). Our detailed analysis tended to confirm this hypothesis. This underlines the importance of factoring in multiple indicators, to take into account different manipulation techniques.

Table 13 Proportion of retweets – organic vs manipulated traffic

Phrase	Total posts	Retweets	Proportion of retweets
Friday	100,000	49,659	49.66%
Covfefe	90,737	50,457	55.61%
#4thofJuly	99,866	58,267	58.34%
Thursday	99,909	63,902	63.96%
Davos 1	100,000	66,167	66.17%
#DigDoug	42,939	28,673	66.78%
Wednesday	99,871	69,222	69.31%
Davos 2	100,000	75,124	75.12%
Tamim the Glorious	93,429	72,315	77.40%

#LaFranceVoteMarine	47,075	39,939	84.84%
#Marine2017	24,001	20,892	87.05%
#StopAstroturfing	15,915	13,889	87.27%
#LePionMacron	30,673	27,385	89.28%
Qadhafi of the Gulf	33,976	32,286	94.95%

Proportion of Traffic from the 50 Most Active Accounts

Our final indicator was the proportion of posts generated by the fifty accounts which mentioned the search term most often during the scan (we will refer to these as the "most active accounts"). Since it is a proportion, this indicator depends on the total number of posts in the traffic flow, as well as the number of posts from the top fifty; larger flows will, by definition, tend to return lower proportions than smaller flows.

We opted for this approach to measure the *impact* of the most active accounts on total traffic, rather than their *effort*: our concern is whether a specific group of users managed to materially distort a specific traffic flow, not whether they tried to. All our control samples returned a proportion of traffic of between 1.5% and 3%.

Table 14 Control sample posts from top 50 most active accounts

Phrase	Total posts	Posts from top 50	Percentage of traffic from top 50
Covfefe	90,737	2,065	2.27%
#4thofJuly	99,866	1,778	1.78%
Davos 1	100,000	2,849	2.85%
Davos 2	100,000	2,819	2.82%
Wednesday	99,871	2,503	2.50%
Thursday	99,909	1,630	1.63%
Friday	100,000	1,477	1.48%

Our manipulated samples returned proportions of between 2.4% and 36%.

Table 15 Manipulated samples posts from top 50 most active accounts

Phrase	Total tweets	Tweets from top 50	Percentage of traffic from top 50
#Marine2017	24,001	8,640	35.83%
#LePionMacron	30,673	8,532	27.82%
#LaFranceVoteMarine	47,075	11,408	24.23%
Qadhafi of the Gulf	33,976	1,397	4.11%
Tamim the Glorious	93,429	2,266	2.43%
#StopAstroturfing	15,915	5,568	34.99%
#DigDoug	42,939	8,494	19.78%

This separated the flows into two distinct bands, with one outlier. All our control samples, and "Tamim the Glorious", showed results of under 3%. Our manipulated cases from the United States, Poland, and France all showed results close to, or exceeding, 20%. "Qadhafi of the Gulf" stood apart, on 4.11% – significantly above our control samples, but significantly below our other manipulated ones. This, again, suggests that the manipulation of this traffic flow was achieved by a high number of accounts which posted a few retweets each, rather than a smaller group of more active accounts.

Table 16 Posts from 50 most active accounts – organic vs manipulated traffic

Phrase	Total tweets	Tweets from top 50	Percentage of traffic from top 50
Friday	100,000	1,477	1.48%
Thursday	99,909	1,630	1.63%
#4thofJuly	99,866	1,778	1.78%
Covfefe	90,737	2,065	2.27%
Tamim the Glorious	93,429	2,266	2.43%
Wednesday	99,871	2,503	2.50%
Davos 2	100,000	2,819	2.82%
Davos 1	100,000	2,849	2.85%
Qadhafi of the Gulf	33,976	1,397	4.11%
#DigDoug	42,939	8,494	19.78%
#LaFranceVoteMarine	47,075	11,408	24.23%
#LePionMacron	30,673	8,532	27.82%
#StopAstroturfing	15,915	5,568	34.99%
#Marine2017	24,001	8,640	35.83%

THE COEFFICIENT OF TRAFFIC MANIPULATION

Each of these indicators was designed to assess one of the ways in which malicious actors can attempt to manipulate Twitter flows. By combining all three indicators, we believe that a reliable indication of the relative extent to which a given flow has been manipulated can be calculated.

This calculation should hold true, regardless of whether the manipulation was achieved using bots, cyborgs, human posters, or a combination of all of them, and should therefore be more resilient against attempts to evade detection by changing the mixture of tactics deployed. To give equivalent weight to each indicator in our final calculation, we divided the percentage of retweets by 10.

Our Coefficient of Traffic Manipulation is expressed as follows, where *C* is the coefficient of traffic manipulation, *R* is the percentage of retweets, *F* is the percentage of traffic from the top fifty users, and *U* is the average number of tweets per user:

$$C = \frac{R}{10} + F + U$$

This gives a coefficient of traffic manipulation for each scan, as follows:

Table 17 Coefficient of traffic manipulation by scan

Phrase	R/10	F	U	Coefficient
Friday	4.966	1.48	1.22	7.666
#4thofJuly	5.834	1.78	1.17	8.784
Thursday	6.396	1.63	1.21	9.236
Covfefe	5.561	2.27	2.14	9.971
Wednesday	6.931	2.50	1.24	10.671
Davos 1	6.617	2.85	1.54	11.007
Davos 2	7.512	2.82	1.59	11.922
Tamim the Glorious	7.74	2.43	3.65	13.82
Qadhafi of the Gulf	9.495	4.11	2.94	16.545
#DigDoug	6.678	19.78	4.78	31.238
#LaFranceVoteMarine	8.484	24.23	5.31	38.024
#LePionMacron	8.928	27.82	5.94	42.688
#StopAstroturfing	8.727	34.99	6.61	50.327
#Marine2017	8.705	35.83	8.81	53.345

Taken together, our scans fall into three distinct groups: the control experiments, all of which scored a coefficient of under 12; the two Gulf samples, which scored between 13 and 17; and the five US, French and Polish examples, which all scored over 30. These five latter samples showed the clearest indications of large-scale manipulation. The most influential factor in their CTM scores was the very high impact of the fifty most active accounts; this is the most direct indication of a small group of users working disproportionately hard to give the appearance of a large-scale movement. These five traffic movements were relatively small-scale – under 50,000 tweets – and conducted over a short but intense span of a few hours. Filtering out these extreme cases, however, the results remain indicative:

Table 18 Coefficient of traffic manipulation by scan without extreme cases

Phrase	R/10	F	U	Coefficient
Friday	4.966	1.48	1.22	7.666
#4thofJuly	5.834	1.78	1.17	8.784
Thursday	6.396	1.63	1.21	9.236
Covfefe	5.561	2.27	2.14	9.971
Wednesday	6.931	2.50	1.24	10.671
Davos 1	6.617	2.85	1.54	11.007
Davos 2	7.512	2.82	1.59	11.922
Tamim the Glorious	7.74	2.43	3.65	13.82
Qadhafi of the Gulf	9.495	4.11	2.94	16.545

The lower group of seven control samples clusters together, with a variation of just over four points between the lowest and highest. The upper group, consisting of the two manipulated Gulf samples, begins 1.9 points higher. This suggests that, even in larger and less obviously manipulated cases, the CTM can give a reliable indication of the degree of manipulation to which a traffic flow has been subjected.

CONCLUSION

Conducted by skilled operators – the Le Pen supporters, in particular, were strikingly successful in getting their hashtags to trend – Twitter traffic can be manipulated and distorted by a combination of high-volume human users and high-volume bots. Working together, these can give a small group of users the appearance of a large and organically trending movement. Such activities leave traces on the Twitter flow. If a small group attempts to manipulate traffic by posting at very high volumes, this will show up as an abnormally high rate of tweets per user, or an abnormally high level of activity from the fifty most active accounts ("normal" being defined with reference to our control series). If manipulators use a large number of retweet bots, this will distort the proportion of retweets as a percentage of total traffic.

By combining these factors, we can assess how much a given traffic flow appears to have been manipulated by a small user group. Small variations can be expected in every flow, but our initial findings suggest that the combination of three factors – the average number of posts per user, the proportion of traffic generated by the fifty most active accounts, and the proportion of traffic from retweets – does allow us to isolate those Twitter traffic flows which are most likely to have been subject to significant manipulation. It is important to note that our CTM is relative, not absolute. We do not believe that there is a single numerical cut-off point, above which any traffic flow can be considered to have been manipulated – although we note that all of our control samples scored under 12.

However, with that caveat, we believe that the CTM does provide a useful first warning of potentially manipulated Twitter traffic. If automated, the CTM could serve as an early-warning system, flagging up suspicious traffic for more detailed analysis. As such, it could provide a first step in tackling the ongoing problem of traffic manipulation.

APPENDIX 1: CASE STUDIES

Control series

Covfefe

Our first case study was the word "covfefe," an apparent typographical error tweeted by US President Donald Trump on May 31, 2017. As soon as it was posted, the word triggered an explosive traffic flow from supporters and defenders of the president, and many unaffiliated commentators.

According to contemporaneous reporting, Trump's original tweet generated 126,000 retweets within five hours, before being deleted; by June 5, "covfefe" had been mentioned 2.6 million time across Twitter, Facebook and Instagram (Terry 2017).

We assessed that initial traffic on "covfefe" was not subject to planned or large-scale manipulation. President Trump's error cannot have been predicted by outsiders and appears to have taken the White House itself by surprise. The traffic was on such a large scale, and came from so many varied sources, that we considered it highly unlikely that any one group could have manipulated the traffic on a scale large enough to materially distort it.

We therefore considered "covfefe" to have been a reliable example of organic large-scale Twitter traffic.

Davos

We conducted two scans on the word "Davos" in the build-up to, and during, the annual meeting of the World Economic Forum in Davos, Switzerland, in January 2018.

"Davos", which has strong political connotations, generated high volumes of Twitter traffic immediately before, and during, the meeting. In 2018, traffic was particularly driven by news that President Trump was to attend.

However, while the Forum generated intense political interest, this interest was very widely spread. Our scans showed that "Davos" was used intensively by media and commentators around the world in widely varying contexts (we saw particularly vigorous engagement from India, Canada, Pakistan, the United States, and from celebrity and corporate business accounts internationally).

We therefore concluded that it would have been difficult for any one group, in any one country, to post such high volumes that they materially distorted the traffic.

Wednesday, Thursday, and Friday

To study organic traffic on non-political topics, we scanned mentions of the words "Wednesday", "Thursday" and "Friday" on a Wednesday, Thursday, and Friday in September 2018.

We considered that searches for the names of the days, without hashtags or other slogans added, would return data from such a wide variety of sources that it would be difficult for any one group to generate so many posts that they materially distorted it.

We also considered that the non-political nature of the words meant that political manipulators would have little incentive to try and influence them. This, in turn, meant that we could expect automated traffic from non-political bots, such as spambots, but not from large-scale political ones, providing a valid snapshot of traffic which has not been subject to deliberate, as opposed to incidental, automation.

#4thofJuly

As a final control sample on a less politically charged topic, we conducted a machine scan of 100,000 tweets on the hashtag #4thofJuly, on July 4, 2017.

This hashtag was predictable and time-limited, and an obvious candidate to make it into the trending topics. It was therefore likely to be targeted by groups which routinely try to take advantage of trends to amplify their messaging, such as spambots and marketing accounts.

Nevertheless, we assessed that the term was of sufficiently broad interest, and sufficiently popular, that it would take a major effort to materially distort the traffic.

As such, we considered that #4thofJuly represented a valuable control sample, as it provided a realistic flow in which spambots and paybots were likely to have been present in some volume, but the overall flow was too wide and varied to have been materially manipulated by influence accounts.

Manipulated series

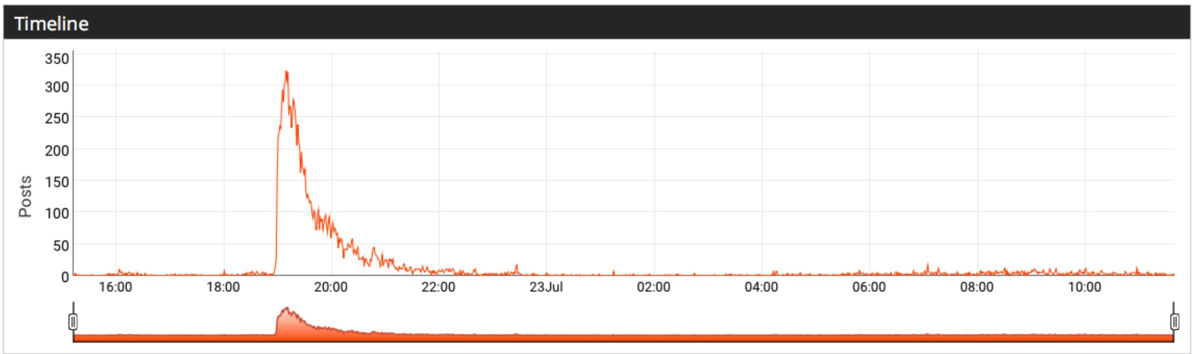
We identified six case studies of Twitter flows in which we were able to conclude with a high degree of confidence that they were deliberately manipulated by relatively small user groups in order to make the hashtags trend.

#StopAstroturfing

We observed the simplest case in Poland in July 2017. This featured a flow of almost 16,000 tweets using the hashtags #StopAstroturfing and #StopNGOSoros, in response to protests against reforms to the nation's Supreme Court.

This traffic was unusual for two reasons. Firstly, over 11,000 of the tweets posted identical wording as tweets or retweets, suggesting that their activity was either pre-planned or driven by bots. Secondly, a machine scan of the traffic showed that the number of posts jumped seven-fold in just two minutes at exactly 19:00 GMT, from 29 posts a minute to 217 posts a minute. It then remained at that rate for less than half an hour, before falling back to the original level, almost as abruptly (see Figure 1).

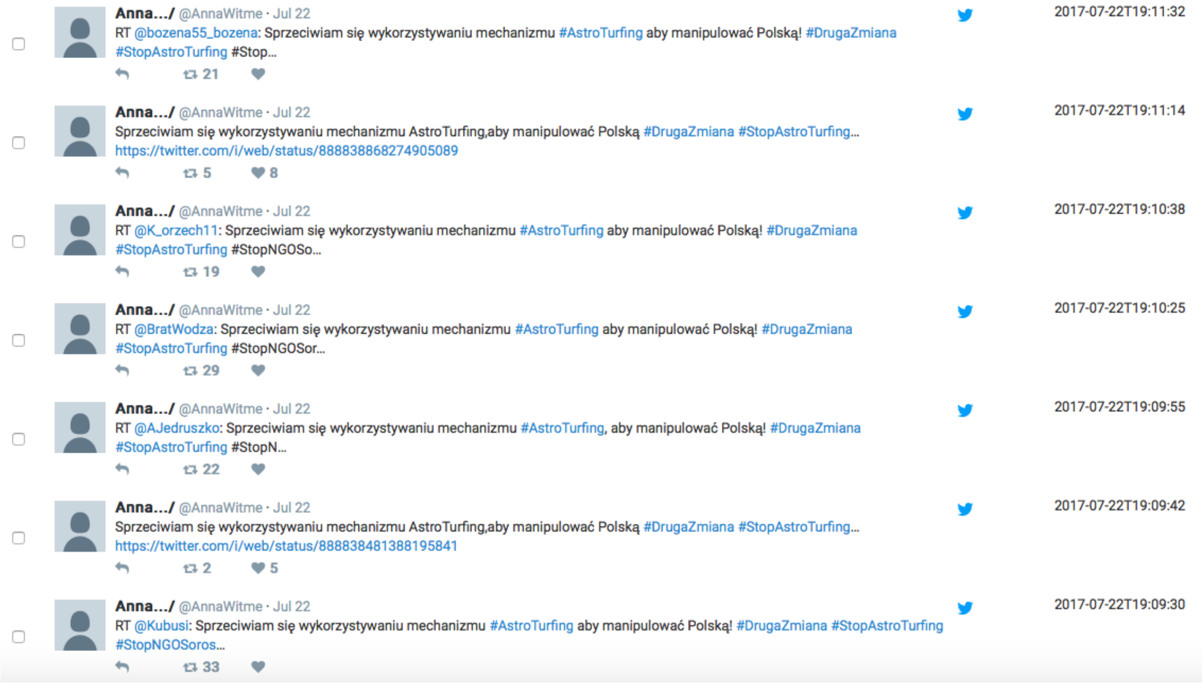
Figure 1 Traffic on #StopAstroturfing and #StopNGOSoros, July 22, 2017



Our detailed scan identified a significant number of confirmed bots which were heavily involved in driving the traffic, posting dozens or hundreds of times each. These included, as an example, @annawitme, which was created on December 8, 2016.[11] By July 23, 2017, it had posted 67,666 tweets and 123,204 likes, for an average rate of 833.5 engagements per day: a clearly botlike figure.

During the course of the scan, @annawitme posted the same text 199 times in less than two hours, either as individual posts or retweets (see Figure 2). Several dozen other accounts behaved similarly.

Figure 2 Tweets and retweets by @annawitme, posting the identical wording



☐ **Anna.../** @AnnaWitme · Jul 22 — 2017-07-22T19:11:32
RT @bozena55_bozena: Sprzeciwiam się wykorzystywaniu mechanizmu #AstroTurfing aby manipulować Polską! #DrugaZmiana #StopAstroTurfing #Stop...
⇄ 21 ♥

☐ **Anna.../** @AnnaWitme · Jul 22 — 2017-07-22T19:11:14
Sprzeciwiam się wykorzystywaniu mechanizmu AstroTurfing,aby manipulować Polską #DrugaZmiana #StopAstroTurfing... https://twitter.com/i/web/status/888838868274905089
⇄ 5 ♥ 8

☐ **Anna.../** @AnnaWitme · Jul 22 — 2017-07-22T19:10:38
RT @K_orzech11: Sprzeciwiam się wykorzystywaniu mechanizmu #AstroTurfing aby manipulować Polską! #DrugaZmiana #StopAstroTurfing #StopNGOSo...
⇄ 19 ♥

☐ **Anna.../** @AnnaWitme · Jul 22 — 2017-07-22T19:10:25
RT @BratWodza: Sprzeciwiam się wykorzystywaniu mechanizmu #AstroTurfing aby manipulować Polską! #DrugaZmiana #StopAstroTurfing #StopNGOSor...
⇄ 29 ♥

☐ **Anna.../** @AnnaWitme · Jul 22 — 2017-07-22T19:09:55
RT @AJedruszko: Sprzeciwiam się wykorzystywaniu mechanizmu #AstroTurfing, aby manipulować Polską! #DrugaZmiana #StopAstroTurfing #StopN...
⇄ 22 ♥

☐ **Anna.../** @AnnaWitme · Jul 22 — 2017-07-22T19:09:42
Sprzeciwiam się wykorzystywaniu mechanizmu AstroTurfing,aby manipulować Polską #DrugaZmiana #StopAstroTurfing... https://twitter.com/i/web/status/888838481388195841
⇄ 2 ♥ 5

☐ **Anna.../** @AnnaWitme · Jul 22 — 2017-07-22T19:09:30
RT @Kubusi: Sprzeciwiam się wykorzystywaniu mechanizmu #AstroTurfing aby manipulować Polską! #DrugaZmiana #StopAstroTurfing #StopNGOSoros...
⇄ 33 ♥

[11] The original account has been deleted, but is archived at http://archive.is/cUCDK.

We therefore concluded that this Twitter flow was subject to targeted manipulation by a combination of bots, cyborgs, and hyperactive human users.

Gulf Crisis

Two case studies concerned Twitter traffic during the diplomatic dispute between Saudi Arabia and Qatar in 2017. In a detailed analysis of this traffic, we found numerous cases of botnet intervention on both sides, together with strong indications that the traffic had been coordinated offline.

Our analyses included the hashtags #قذافي_الخليج ("Qadhafi of the Gulf", sample from August 27, 2017), and #تميم_المجد("Tamim the Glorious", sample from September 19, 2017).[12]

Automated scans of these hashtags showed repeated spikes, in which traffic more than doubled in volume for the space of a few minutes, or even seconds (see Figures 3 and 4).

Figure 3 Traffic on "Qadhafi of the Gulf," August 27, 2017

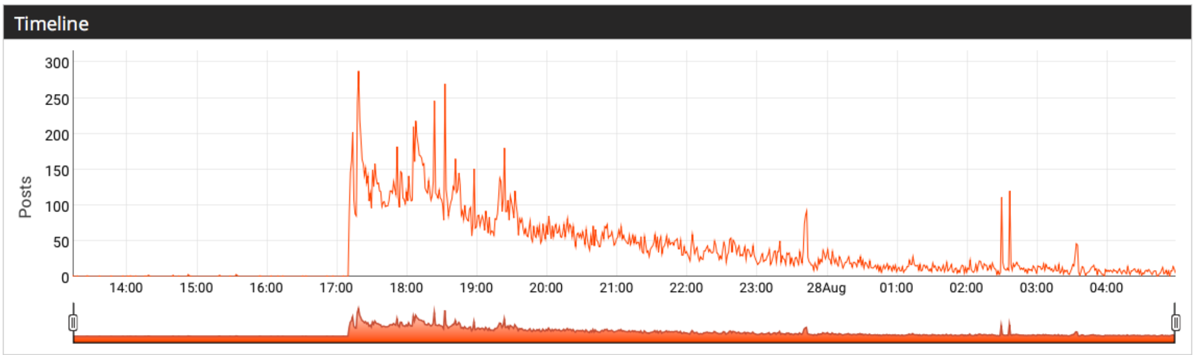
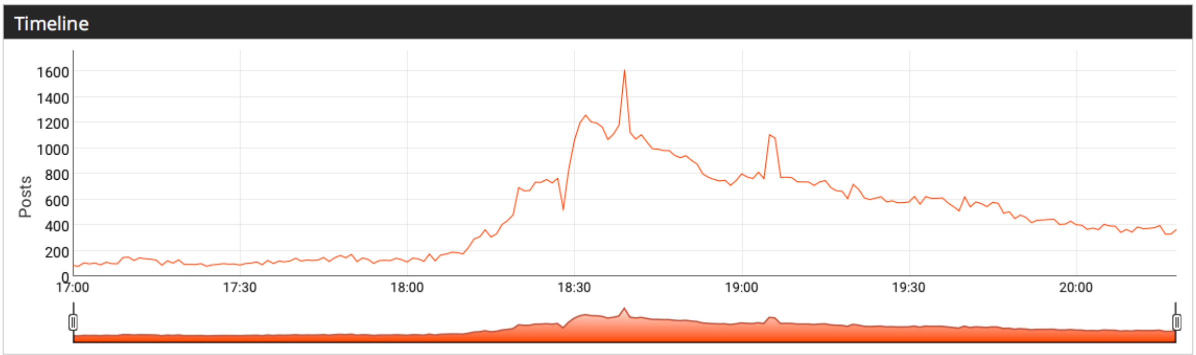


Figure 4 Traffic on "Tamim the Glorious," September 19, 2017



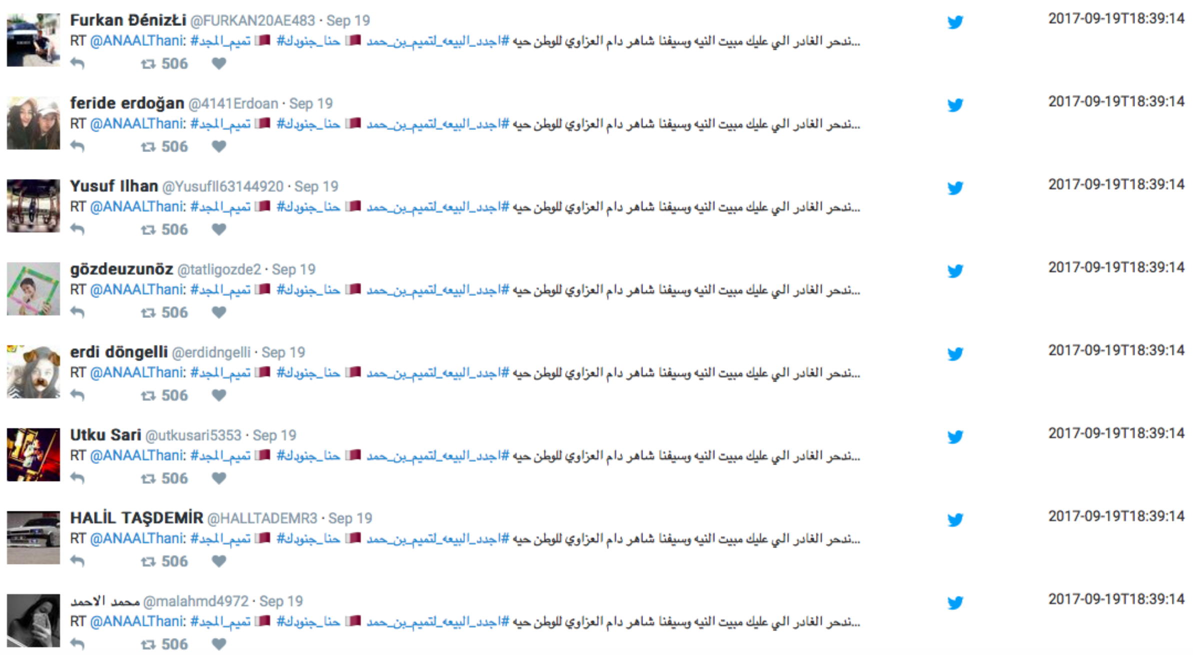
On "Qadhafi of the Gulf," the scans also showed that initial traffic jumped from zero to over 200 tweets per minute in just three minutes, reached its highest rate within nine minutes, and then fell back precipitously, before recording a number of discernible spikes.

[12] Translations courtesy of the BBC Arabic service, with whom these scans were conducted.

This is unusual for large-scale organic traffic, in which a bell curve usually predominates. It is more characteristic of manipulated traffic, in which either a botnet or a coordinated team of users gives an initial boost to traffic, in the hope of catching the attention of unaffiliated users, and then subsequent bot interventions give further impetus.

Visual inspection of the traffic generated during the "spike" moments confirmed the intervention of large-scale, crude botnets – on one occasion, retweeting the same post 449 times in just two seconds (see Figure 5).

Figure 5 Retweets of the same post made at the same second by an apparently Turkish botnet



Furkan ĐénizŁi @FURKAN20AE483 · Sep 19 2017-09-19T18:39:14
RT @ANAALThani: #تميم_المجد 🇶🇦 #حنا_جنودك 🇶🇦 #اجدد_البيعه_لتميم_بن_حمد ندحر الغادر الي عليك مبيت النيه وسيفنا شاهر دام العزاوي للوطن حيه...
506

feride erdoğan @4141Erdoan · Sep 19 2017-09-19T18:39:14
RT @ANAALThani: #تميم_المجد 🇶🇦 #حنا_جنودك 🇶🇦 #اجدد_البيعه_لتميم_بن_حمد ندحر الغادر الي عليك مبيت النيه وسيفنا شاهر دام العزاوي للوطن حيه...
506

Yusuf Ilhan @Yusufll63144920 · Sep 19 2017-09-19T18:39:14
RT @ANAALThani: #تميم_المجد 🇶🇦 #حنا_جنودك 🇶🇦 #اجدد_البيعه_لتميم_بن_حمد ندحر الغادر الي عليك مبيت النيه وسيفنا شاهر دام العزاوي للوطن حيه...
506

gözdeuzunöz @tatligozde2 · Sep 19 2017-09-19T18:39:14
RT @ANAALThani: #تميم_المجد 🇶🇦 #حنا_جنودك 🇶🇦 #اجدد_البيعه_لتميم_بن_حمد ندحر الغادر الي عليك مبيت النيه وسيفنا شاهر دام العزاوي للوطن حيه...
506

erdi döngelli @erdidngelli · Sep 19 2017-09-19T18:39:14
RT @ANAALThani: #تميم_المجد 🇶🇦 #حنا_جنودك 🇶🇦 #اجدد_البيعه_لتميم_بن_حمد ندحر الغادر الي عليك مبيت النيه وسيفنا شاهر دام العزاوي للوطن حيه...
506

Utku Sari @utkusari5353 · Sep 19 2017-09-19T18:39:14
RT @ANAALThani: #تميم_المجد 🇶🇦 #حنا_جنودك 🇶🇦 #اجدد_البيعه_لتميم_بن_حمد ندحر الغادر الي عليك مبيت النيه وسيفنا شاهر دام العزاوي للوطن حيه...
506

HALİL TAŞDEMİR @HALLTADEMR3 · Sep 19 2017-09-19T18:39:14
RT @ANAALThani: #تميم_المجد 🇶🇦 #حنا_جنودك 🇶🇦 #اجدد_البيعه_لتميم_بن_حمد ندحر الغادر الي عليك مبيت النيه وسيفنا شاهر دام العزاوي للوطن حيه...
506

محمد الاحمد @malahmd4972 · Sep 19 2017-09-19T18:39:14
RT @ANAALThani: #تميم_المجد 🇶🇦 #حنا_جنودك 🇶🇦 #اجدد_البيعه_لتميم_بن_حمد ندحر الغادر الي عليك مبيت النيه وسيفنا شاهر دام العزاوي للوطن حيه...
506

We therefore concluded that both flows had been subjected to a combination of bot amplification and the intervention of coordinated human user groups, creating a hybrid structure which materially distorted the traffic.

#DigDoug

Our fourth sample concerned the hashtag #DigDoug, which was deployed in the United States in December 2017 during the Alabama gubernatorial race between Democrat Doug Jones and Republican Roy Moore.

This case study was particularly noteworthy, because the users who attempted to manipulate the traffic did so using open posts.

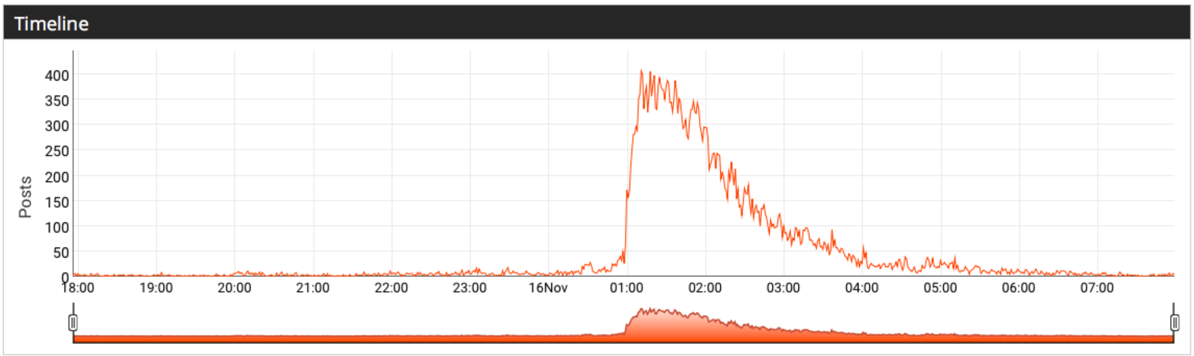
The main driver of traffic was an account called @pmaalai. On November 13, 2017, this account tweeted an invitation to its over 12,000 followers to launch a "Twitter storm" at 8pm Eastern time on November 15. The account asked for retweets and shared a "tweet sheet" on a Google document.[13]

That document, in turn, offered users 48 separate tweet texts, each with a shortened URL which was generated using a service called "click to tweet" (ctt.ec).

The "click to tweet" service allows users to click on the text of their choice. As long as they are already logged in to their Twitter account, they then see the text appear as a prepared tweet and can post it by clicking once more. This is a textbook way of generating disproportionate Twitter traffic from a small but dedicated group by encouraging users to post at a rate usually associated with bots; indeed, the "tweet sheet" advised users to "average about one tweet per minute so Twitter doesn't lock you out".

Sure enough, on November 15 at 8pm Eastern time, Twitter traffic on #DigDoug showed an explosive growth, from 32 tweets per minute just before the hour, to 300 tweets per minute just afterwards (see Figure 6). As in the Gulf case, traffic hit peak flow in the first ten minutes, and then declined steeply, dropping back to pre-spike levels within three hours.

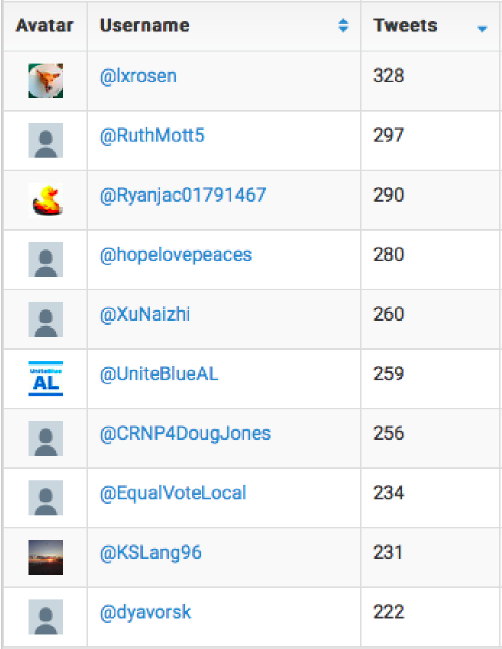
Figure 6 Traffic on #DigDoug, November 15–16, 2017, from machine scan



Much of the traffic was generated by hyperactive accounts which posted the same hashtag dozens or hundreds of times. Over a fourteen-hour period, over thirty accounts each posted the hashtag more than 140 times (see Figure 7).

[13] The original tweet is online at https://twitter.com/Pmaalai/status/930164142912106496, archived at http://archive.is/3GjLw. The Google document is online at https://docs.google.com/spreadsheets/d/1sETXC_u4TtF2n4IRf194d9N9akUSzQx0-WmZO_3T7lk/edit#gid=1955483145.

Figure 7 The number of times each of the most active accounts posted #DigDoug



Avatar	Username	Tweets
	@lxrosen	328
	@RuthMott5	297
	@Ryanjac01791467	290
	@hopelovepeaces	280
	@XuNaizhi	260
AL	@UniteBlueAL	259
	@CRNP4DougJones	256
	@EqualVoteLocal	234
	@KSLang96	231
	@dyavorsk	222

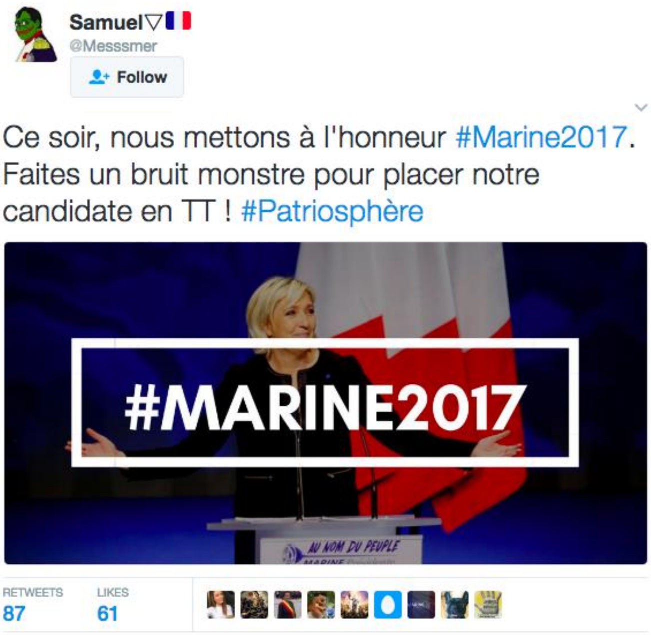
On visual inspection, few of these accounts resembled bots; they interspersed large numbers of retweets with a significant proportion of apparently authored posts, in the manner of cyborgs.

In the light of the considerable preparation which went into this campaign (as witnessed by the "tweet sheet"), and the timed mass-releasing of posts, we concluded that this was an example of attempted large-scale traffic manipulation.

French Election

Our final three samples were generated during the French election campaign in early 2017, by a group of users supporting far-right presidential candidate Marine Le Pen. This group's explicit intention was to manipulate Twitter's traffic so that their hashtags would trend, as evidenced by one of their leaders, @messsmer (see Figure 8).

Figure 8 Tweet by @messsmer, archived at http://archive.is/G1Hjj.



Samuel▽🇫🇷
@Messsmer

Ce soir, nous mettons à l'honneur #Marine2017. Faites un bruit monstre pour placer notre candidate en TT ! #Patriosphère

RETWEETS 87
LIKES 61

Three of their hashtags were #Marine2017, #LePionMacron ("Macron the pawn", attacking Le Pen's rival, now-President Emmanuel Macron) and #LaFranceVoteMarine ("France votes for Marine").

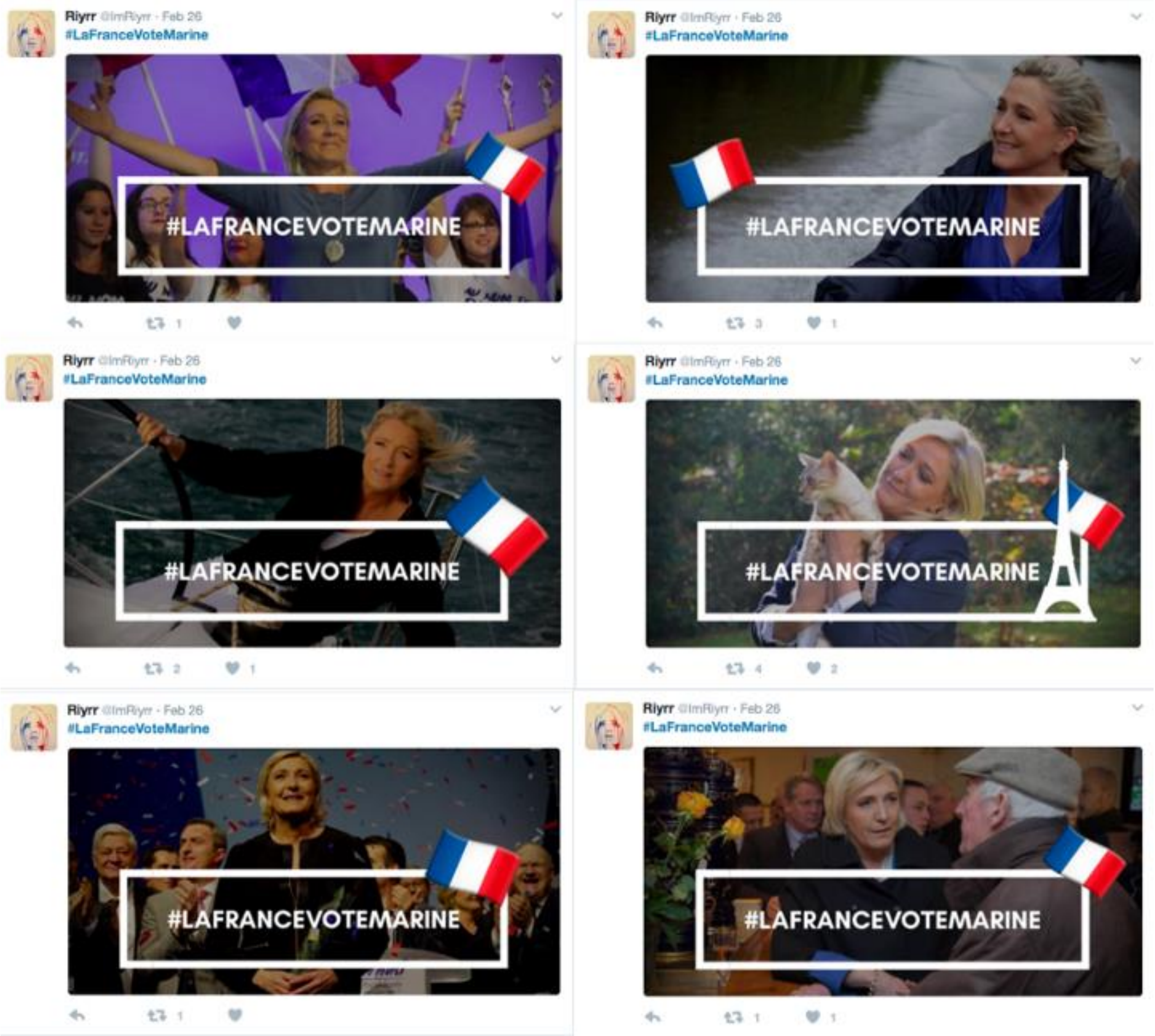
The call to make a hashtag trend should not, in itself, be considered as an attempt at manipulation: manipulation requires an infrastructure which is capable of distorting traffic, such as compromised or fake accounts (Yubao Zhang et al. 2017). It is, however, an early indicator that such attempts might be made.

In the case of the three pro-Le Pen hashtags, our analysis showed that there was such an infrastructure, which combined a large stock of prepared content, a coordinated launch time, and an amplification network of cyborgs and bots. Together, these factors did push the hashtags into the trending topics in France, according to screenshots posted by the lead accounts.[14]

The preparation of content was apparent from the variety of memes shared by the users during the campaign. On #LaFranceVoteMarine, for example, organisers shared over a dozen different memes in the same visual style, as demonstrated in Figure 9.

[14] An example of such a tweet is archived at http://archive.is/Lcll5.

Figure 9 Different memes in the #LaFranceVoteMarine shared by cyborg @ImRiyrr.



Riyrr @ImRiyrr · Feb 26
#LaFranceVoteMarine

Riyrr @ImRiyrr · Feb 26
#LaFranceVoteMarine

Riyrr @ImRiyrr · Feb 26
#LaFranceVoteMarine

Riyrr @ImRiyrr · Feb 26
#LaFranceVoteMarine

Riyrr @ImRiyrr · Feb 26
#LaFranceVoteMarine

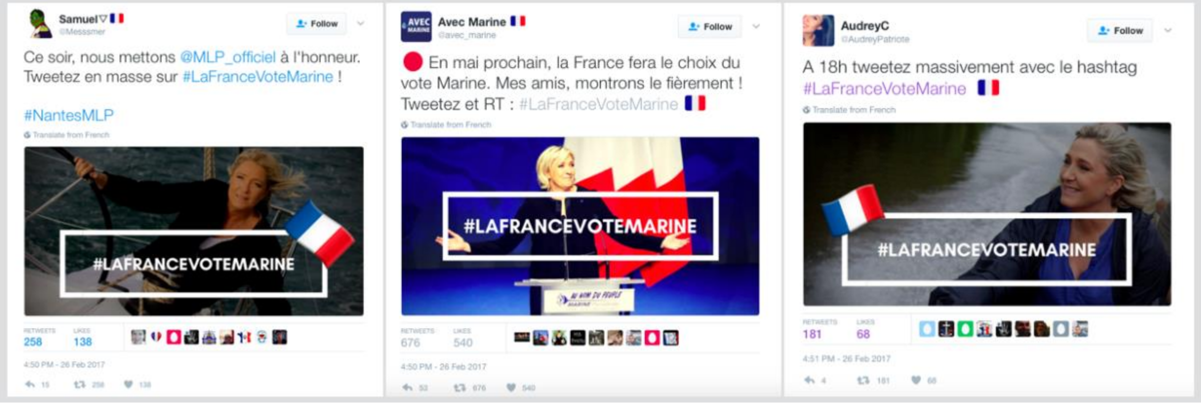
Riyrr @ImRiyrr · Feb 26
#LaFranceVoteMarine

The effort which went into this preparation was clearly considerable: we noted at least 95 different memes accompanying the hashtag #Marine201y.[15] Memes such as these provided a stock of raw material on which users could draw, in a manner akin to the #DigDoug "tweet sheet"; indeed, for #LaFranceVoteMarine, @messsmer tweeted a link to an online repository of memes which supporters could download.[16]

This preparation was followed up by synchronised posting by the leading accounts in the group. On each occasion, half a dozen influential accounts launched their hashtag campaign at exactly 16:50 UTC, 5.50pm French time, with simultaneous tweets using different wording and memes, but the same hashtag, as Figure 10 demonstrates.

[15] A full list of the archived versions, all posted by @ImRiyrr, is provided in Appendix 2.
[16] The tweet is archived at http://archive.is/v76Ri.

Figure 10 Near-simultaneous posts on #LaFranceVoteMarine on hashtag launch



Samuel▽ 🇫🇷
@Messsmer
Follow

Ce soir, nous mettons @MLP_officiel à l'honneur. Tweetez en masse sur #LaFranceVoteMarine !

#NantesMLP

Translate from French

RETWEETS 258 LIKES 138

4:50 PM - 26 Feb 2017

15 258 138

Avec Marine 🇫🇷
@avec_marine
Follow

🔴 En mai prochain, la France fera le choix du vote Marine. Mes amis, montrons le fièrement ! Tweetez et RT : #LaFranceVoteMarine 🇫🇷

Translate from French

RETWEETS 676 LIKES 540

4:50 PM - 26 Feb 2017

52 676 540

AudreyC
@AudreyPatriote
Follow

A 18h tweetez massivement avec le hashtag #LaFranceVoteMarine 🇫🇷

Translate from French

RETWEETS 181 LIKES 68

4:51 PM - 26 Feb 2017

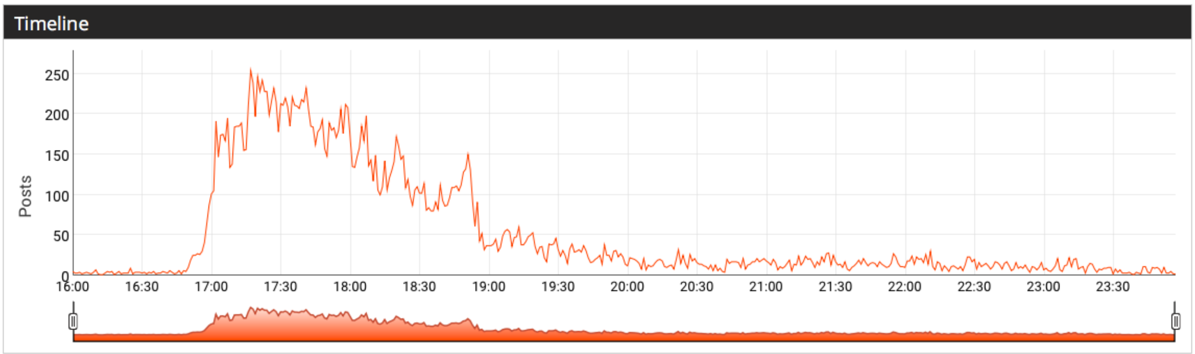
4 181 68

Some of these posts, such as the one by @AudreyPatriote illustrated in Figure 10, called for a timed campaign of amplification from 6pm French time.

These accounts proceeded to post more memes at regular intervals, and to retweet one another. Other apparently human-run accounts shared the same memes and retweeted the leaders, and apparently automated accounts provided amplification on a large scale: @ImRiyrr, for example, posted 95 memes on #Marine2017 in 17 minutes, each with no textual content other than the hashtag.

The effect was to generate thousands of tweets (between 25,000 and 45,000), with an explosive acceleration in traffic at zero hour, a high volume of posts for roughly two hours, and then a rapid decline, as Figure 11 demonstrates.

Figure 11 Traffic on #Marine2017, February 14, 2017, from machine scan



This group was unusually skilled and sophisticated in its manipulation techniques. On each occasion, it managed to make the chosen hashtag trend by 7pm French time. All the evidence suggested that this sweeping success was due, not to spontaneous organic interest, but to a manipulation campaign which was carefully designed and tightly coordinated.

As such, we concluded that these French examples demonstrated traffic manipulation which went well beyond simple amplification by bots and cyborgs, and spanned the range of methods which online actors can use to distort Twitter flows.

APPENDIX 2: ARCHIVED POSTS

The following links are to archived versions of the accounts and posts which were identified in our manipulated case studies.

Polish study

The tweet which launched the sequence: http://archive.is/hd510
Profile page of bot @annawitme: http://archive.is/cUCDK
Tweets by @annawitme: http://archive.is/wacg1
http://archive.is/GAOnX
http://archive.is/PF2pQ
http://archive.is/DBi9v
http://archive.is/vHdVW
Profile page of bot @ZawszePolska: http://archive.is/V3NbV
Tweets by bot @MichalAlex1975: http://archive.is/cezbs
http://archive.is/SZGDT
http://archive.is/dQURL

US study

Announcement of the "Twitter storm": http://archive.is/3GjLw
The "tweet sheet": http://archive.is/8Vyyx
"Tweet sheet" post by @pmaalai: http://archive.is/1kcpk

French study

Launch of #Marine2017, by @avec_marine: http://archive.is/kO40o
Launch of #Marine2017, by @grimmgrimm84: http://archive.is/JQXg2
Call to make #Marine2017 trend, by @messsmer: http://archive.is/G1Hjj
Memes on #Marine2017, by @messsmer: http://archive.is/5r7rf
Memes on #Marine2017, by @AudreyPatriote: http://archive.is/EtoWQ
Claim that #Marine2017 is trending, by @avec_marine: http://archive.is/Up3bB

Launch of #LePionMacron, by @avec_marine: http://archive.is/F7v6f
Launch of #LePionMacron, by @messsmer: http://archive.is/izvIh
Claim that #LePionMacron is trending: http://archive.is/Lcll5

Launch of #LaFranceVoteMarine, by @messsmer: http://archive.is/lHuKm
Launch of #LaFranceVoteMarine, by @avec_marine: http://archive.is/Zb9m5
Launch of #LaFranceVoteMarine, by @AudreyPatriote: http://archive.is/1htot
Launch of #LaFranceVoteMarine, by @AntreDuPatriote: http://archive.is/3mNpR
Memes on #LaFranceVoteMarine, by @ImRiyrr: http://archive.is/1S3as
Link to memes for #LaFranceVoteMarine, by @messsmer: http://archive.is/v76Ri
Claim that #LaFranceVoteMarine is trending, by @avec_marine: http://archive.is/OT03M

The following links show the 95 memes prepared for the #Marine2017 hashtag campaign, and tweeted by @ImRiyrr.

http://archive.is/d8Kw3
http://archive.is/6xonQ
http://archive.is/XTndW
http://archive.is/AlmPY
http://archive.is/pB1EG
http://archive.is/Ev0SK
http://archive.is/uPkH9
http://archive.is/8jZkS
http://archive.is/apjmg
http://archive.is/ogDzD
http://archive.is/roCBI
http://archive.is/VcA3Q
http://archive.is/xEAFS
http://archive.is/mVfuA
http://archive.is/cbUji
http://archive.is/QJdWI
http://archive.is/udSzr
http://archive.is/kxcoQ
http://archive.is/w28xa
http://archive.is/y8syy
http://archive.is/cC7bh
http://archive.is/fK6dm
http://archive.is/uE5rq
http://archive.is/Jy4Fu
http://archive.is/zSouT
http://archive.is/BXIwh
http://archive.is/renkZ
http://archive.is/RUmK2
http://archive.is/jC1bM
http://archive.is/mK0dR
http://archive.is/BEZrV
http://archive.is/gci5l
http://archive.is/UJCIL
http://archive.is/yehlu
http://archive.is/nuWac
http://archive.is/cLAYU
http://archive.is/CoVog
http://archive.is/EufpE
http://archive.is/Gzzq2
http://archive.is/j4d3L
http://archive.is/8idRM
http://archive.is/9kSSt
http://archive.is/anxTa
http://archive.is/ZEcHS
http://archive.is/3OQKE
http://archive.is/hGaX1
http://archive.is/VaPAK

http://archive.is/yFudt
http://archive.is/0n8Ed
http://archive.is/1qNEU
http://archive.is/W4RHQ
http://archive.is/qSP9Y
http://archive.is/5p9No
http://archive.is/7vtOM
http://archive.is/bF7Ry
http://archive.is/QdruY
http://archive.is/sFq60
http://archive.is/595JJ
http://archive.is/xSKat
http://archive.is/Xv4zP
http://archive.is/O5t4Z
http://archive.is/418jK
http://archive.is/797lP
http://archive.is/afrnd
http://archive.is/NJ5ZW
http://archive.is/D3pPl
http://archive.is/tj4D3
http://archive.is/jDots
http://archive.is/zz2Id
http://archive.is/0f18g
http://archive.is/qV1yj
http://archive.is/FP0Mn
http://archive.is/xbZCt
http://archive.is/XRY2w
http://archive.is/ZXi3U
http://archive.is/qDitX
http://archive.is/5aB7n
http://archive.is/HCBJp
http://archive.is/vQBxq
http://archive.is/8iA9s
http://archive.is/FYuz0
http://archive.is/7G80K
http://archive.is/X0sP9
http://archive.is/Asssb
http://archive.is/Bu7sS
http://archive.is/4gqUj
http://archive.is/HK5w2
http://archive.is/KS4y7
http://archive.is/Bcoow
http://archive.is/eG21f
http://archive.is/e8fTe
http://archive.is/klyWH
http://archive.is/9CdLp
http://archive.is/BkSb9

http://archive.is/tJv2W

BIBLIOGRAPHY

Agarwal, Nitin, et al. 2017. *Examining the use of botnets and their evolution in propaganda dissemination*. Defence Strategic Communications, vol. 2. (March 2017): 87-112. https://issuu.com/natostratcomcoe/docs/full_academic_journal_vol2_issuu_07

Alexander, Lawrence. 2015. *Social network analysis reveals full scale of Kremlin's Twitter bot campaign*. Global Voices. https://globalvoices.org/2015/04/02/analyzing-kremlin-twitter-bots/

Alfonso, Fernando. 2012. *Twitter bots silence critics of Mexico's leading presidential candidate.* The Daily Dot. https://www.dailydot.com/news/pena-nieto-twitter-bots-mexico-election/

Barojan, Donara. 2018. *#BotSpot: Bots target Malaysian elections.* DFRLab. https://medium.com/dfrlab/botspot-bots-target-malaysian-elections-785a3c25645b

Cox, Joseph. 2017. *I bought a Russian bot army for under $100*. The Daily Beast. https://www.thedailybeast.com/i-bought-a-russian-bot-army-for-under-dollar100

Davis, Clayton A., et al. 2016. *Botornot: A system to evaluate social bots.* Proceedings of the 25th International Conference Companion on World Wide Web. International World Wide Web Conferences Steering Committee. https://arxiv.org/pdf/1602.00975.pdf

DiResta, Renee, et al. 2017. *The bots that are changing politics.* Motherboard / Vice News. https://motherboard.vice.com/en_us/article/mb37k4/twitter-facebook-google-bots-misinformation-changing-politics

Ferrara, Emilio, et al. 2016. *The rise of social bots.* Communications of the ACM Volume 59, Number 7 (2016), Pages 96–104.

Gallagher, Erin. 2016. *Manipulating trends & gaming Twitter*. Medium. https://medium.com/@erin_gallagher/manipulating-trends-gaming-twitter-6fd31714c06c

Kollanyi, Bence; Howard, Philip; and Woolley, Samuel. 2016. *Bots and automation over Twitter during the first U.S. presidential debate.* Oxford University Project on Computational Propaganda. http://comprop.oii.ox.ac.uk/wp-content/uploads/sites/89/2016/10/Data-Memo-First-Presidential-Debate.pdf

Newman, Nic, et al. 2018. *Reuters Institute Digital News Report 2018*. Oxford University / Reuters Institute for the Study of Journalism. https://reutersinstitute.politics.ox.ac.uk/sites/default/files/digital-news-report-2018.pdf

Nimmo, Ben. March 2017a. *Le Pen's (small) online army*. DFRLab. https://medium.com/dfrlab/le-pens-small-online-army-c754058630f0

Nimmo, Ben. July 2017b. *Polish astroturfers attack... astroturfing.* DFRlab. https://medium.com/dfrlab/polish-astroturfers-attack-astroturfing-743cf602200

Nimmo, Ben. November 2017c. *Alabama Twitter war.* DFRLab. https://medium.com/dfrlab/electionwatch-alabama-twitter-war-47b34ae89c50

Roth, Yoel. 2018. *Automation and the use of multiple accounts*. Twitter developer blog. https://blog.twitter.com/developer/en_us/topics/tips/2018/automation-and-the-use-of-multiple-accounts.html

Rutledge, Pamela. 2013. *How Obama won the social media battle in the 2012 presidential election*. The National Psychologist, reprinted with permission at http://mprcenter.org/blog/2013/01/how-obama-won-the-social-media-battle-in-the-2012-presidential-campaign/

Shao, Chengcheng, et al. 2017. *The spread of fake news by social bots*. Indiana University. https://arxiv.org/pdf/1707.07592.pdf

Subrahmanian, V. S., et al. 2016. *The DARPA Twitter bot challenge*. Computer 49.6 (2016): 38–46. https://arxiv.org/pdf/1601.05140.pdf

Terry, Kellan. 2017. *Tiger Woods, Covfefe and Sgt Pepper: Inside our PR team's emails*. BrandWatch.com blog post, June 5, 2017. https://www.brandwatch.com/blog/tiger-woods-covfefe-and-sgt-pepper-inside-our-pr-teams-emails/

Twitter Public Policy. 2018. *Update on Twitter's Review of the 2016 U.S. Election*. Twitter official blog, January 19, 2019. https://blog.twitter.com/official/en_us/topics/company/2018/2016-election-update.html

Yubao Zhang et al. 2017. *Twitter Trends Manipulation: A First Look Inside the Security of Twitter Trending*. IEEE Transactions on Information Forensics and Security archive (Volume: 12, Issue: 1, Jan. 2017): 144–156

Zi Chu et al. 2012. *Detecting automation of Twitter accounts: are you a human, bot or cyborg?* IEEE Transactions on Dependable and Secure Computing (Volume: 9, Issue: 6, Nov.–Dec. 2012): 811-824. https://ieeexplore.ieee.org/document/6280553/