ROCKET SHIP OR BLIMP? – IMPLICATIONS OF MALICIOUS ACCOUNTS REMOVAL ON TWITTER

Agnieszka Onuchowska
University of South Florida, aonuchowska@mail.usf.edu

Donald J. Berndt
University of South Florida, dberndt@usf.edu

Sagar Samtani
University of South Florida, ssamtani@usf.edu

Follow this and additional works at: https://aisel.aisnet.org/ecis2019_rip

Recommended Citation
ROCKET SHIP OR BLIMP? – IMPLICATIONS OF MALICIOUS ACCOUNTS REMOVAL ON TWITTER

Research in Progress

Onuchowska, Agnieszka, University of South Florida, Tampa, USA, aonuchowska@mail.usf.edu
Berndt, Donald J., University of South Florida, Tampa, USA, dberndt@usf.edu
Samtani, Sagar, University of South Florida, Tampa, USA, ssamtani@usf.edu

Abstract

In this study we investigate how the removal of malicious accounts that follow legitimate accounts owned by popular people impacts the popularity of tweets posted by celebrities and politicians. Using retweet counts, we analyze to what extent malicious accounts contribute to amplification of tweets across the network. We organize tweets into three broad categories (Rocket Ship, Jet or Blimp) and investigate how the distribution of tweets is influenced by a cleanup of malicious accounts. To understand how the suspension of malicious accounts impacts the propagation of messages on Twitter, we conduct a descriptive statistical analysis of retweets of a total of 464 Donald Trump tweets. We find a statistically significant difference in the mean count of retweets and favorites before and after the malicious account removal. Preliminary results of our analysis show that the implications of Twitter’s cleanup initiatives, which targeted malicious accounts, are visible in the narrowing amplitudes of retweet values. However, the distribution of tweet categories based on the number of retweets remains unchanged.

Keywords: Twitter, Malicious Accounts, Tweet Categorization, Donald Trump
1. Introduction and Motivation

A major challenge introduced by malicious accounts on social media is the potential for false impressions to be generated or amplified by such accounts. Malicious accounts are usually driven by bots, automated programs that spread malicious content, propagate false information or exert influence on social media users (Ferrara et al., 2016; Stringhini et al., 2015; Cao et al., 2016). Bots are used to propagate malware, spam or illegitimate links (Alarifi et al., 2016); they can also contribute to attempts at polarizing political discussions or even influencing elections (Ferrara et al., 2016) by inflating trending topics related to a given political candidate.

Among the biggest challenges related to the management of social media outlets is developing policies that minimize the impact of malicious accounts (Lapowsky, 2018). Social media companies have been actively pursuing malicious accounts, with the goal of reducing their overall number as well as the impact of such accounts’ bad actions. For example, before the US midterm elections in November 2018, Facebook and Instagram blocked accounts showing “coordinated inauthentic behavior” that seemed designed to interfere in the elections’ outcome (Gleicher, 2018). In November 2017, Twitter (2017) updated the Automation Rules policy, which prohibits scripting the Twitter website to generate spam or otherwise bother Twitter users. At the same time, Twitter publicly communicated that automated tweets which seek to manipulate trending topics would be filtered out and any originating accounts would be suspended. Consequently, to protect Twitter users from manipulation and abuse, the company decided to remove roughly 70 million suspicious accounts from the network in June and July 2018 (Timberg & Dwoskin, 2018). Targeted accounts had been actively used to propagate and amplify news of questionable provenance (Fung, 2018). Twitter reported that as a result of the cleanup, some Twitter accounts owned by politicians (e.g., Barack Obama or Donald Trump) or celebrities (such as Katy Perry, Justin Bieber, Oprah Winfrey or Ellen DeGeneres) lost significant numbers of followers (Fung, 2018, Moore, 2018).

Our primary motivation in this study is to understand how the removal of malicious accounts that follow influential people impacts the popularity of tweets posted by such celebrities and politicians. First, we intend to investigate to what extent malicious accounts contribute to the amplification of tweets over the network. In this study we analyze the propagation of Donald Trump’s tweets before and after malicious accounts were cleaned up in June and July 2018. In order to establish whether the suspension activities that Twitter undertook have any impact on the number of (1) retweets and (2) favorites we apply a sign paired t-test. We establish that there is a statistically significant difference in the daily mean number of retweets and favorites before and after the malicious accounts’ cleanup. Second, we attempt to analyze the lifespan of Donald Trump’s tweets to better understand the nature and characteristics of propagation over time. To do that, we focus on Trump’s tweets published in November 1-13, 2018 as we expect a higher user responsiveness given Midterm elections which took place on November 6. In order to understand the types of tweets generated by Donald Trump, we categorize his tweets (Rocket Ship, Blimp or other categories), depending on the total number of retweets. For the purpose of this study we primarily focus on retweets as previous research has proven that retweets are more effective than favorites in estimating informativeness of a tweet (Kwak et al., 2011). In order to assess the effectiveness of Twitter’s current malicious account cleanup attempts based on the Automation Rules, we extracted a tweet dataset and analyzed the descriptive statistics.

The core factor, which motivates us to focus on Donald Trump’s account, is firstly related to the interest his tweets generate. Secondly, a Gallup study published in May 2018 indicated that out of the 52 million accounts following Trump’s Twitter account, 29% are likely to be fake (Dodds, 2018), which further motivated us to investigate the problem. Lastly, our analysis is prompted in part by Donald Trump’s complaint related to the cleanup of bot accounts and overall reduction in the number of his followers, which he shared on Twitter in October 2018: ”A few weeks ago it was a Rocket Ship, now it is a Blimp! Total Bias!” (Trump, 2018).

This paper is structured as follows. First, we discuss previous research related to the analysis of malicious accounts on Twitter. Then, we review past research, which investigated the problem of malicious accounts on Twitter from different perspectives. Next, we present our study’s methods and discuss preliminary findings. Finally, we discuss the implications of the study and possible next steps for future research.
2. Background

Past research related to the existence of malicious or bot accounts on the Twitter networks is quite diverse. For example, significant attention in past publications was dedicated to the explanation of how fake content is populated within the Twitter network (Gupta et al., 2013) or how bots are used to amplify messages on social media (Savage et al., 2016, Vosoughi et al., 2018, Mønsted et al., 2017, Salge & Karahanna, 2018). Also, many studies have been conducted to identify and describe the types of bot accounts, which influence the Twitter network (Ferrara et al., 2016, Forelle et al., 2017). Ethical issues related to bot behavior have also been investigated (Salge & Berente, 2017). Multiple studies have focused on the investigation of common features of malicious accounts. For example, Chu et al. (2012) found that bot-generated tweets often contain spam and show regular posting patterns as opposed to human-generated tweets, which exhibit more complex timing behaviors. Significant portions of these studies have focused on effective ways to detect malicious accounts on Twitter. For example, Cao et al. (2014) proposed cluster analysis to detect large groups of malicious users. Alarifi et al. (2016) generated a classifier that uses machine learning techniques to detect “Sybils” (machine-controlled accounts). Cresci et al. (2015) used existing fake account detection methods as a baseline to deploy crawling cost analysis and design for the most optimal fake account classifiers. Miller et al. (2012) defined algorithms that cluster legitimate Twitter users and treat outliers as fake accounts, whereas Cao et al. (2012) deployed social graphs to establish perceived likelihood for an account being a fake (Sybil) account. To our knowledge the research about the extent to which the suspension of malicious accounts impact message propagation on social media networks has not been widely investigated in the past.

3. Method and Testbed

To understand how the suspension of malicious accounts impacts the propagation of messages on Twitter, in the first part of the study we conducted a descriptive statistical analysis of retweets of 323 Donald Trump tweets collected between June 15 and July 30, 2018. This is followed by a sign paired t-test, which helped us to establish whether the difference of means of daily retweets and favorites collected before and after malicious accounts’ suspension is statistically significant.

We started collecting the data 12 days before the first malicious account cleanup and finished data collection 12 days after the last cleanup was concluded. As reported by Jacobs (2018), Donald Trump’s account lost a total of over 300,000 followers in three cleanup attempts conducted on June 27, July 12 and July 18. During this time span, the most popular tweet was posted on July 8, got retweeted 114,596 times, received 319,173 favorites and contained the following text: “They just didn't get it but they do now!” accompanied by a 2:29 min movie. The least popular tweet was published with a text “The legendary Gary Player at Turnberry in Scotland!” on June 26 and received 4,883 retweets and 33,218 favorites.

![Figure 1: Donald Trump’s most popular tweet (left) vs. least popular tweet (right) (June 15-July 30 2018)](image-url)
We used an online archive (http://trumptwitterarchive.com/archive) to extract a subset of Trump’s tweets during the timespan of interest, followed by the number of retweets each tweet has received. Simultaneously, we collected the data indicating the number of Donald Trump’s account followers using snapshots of daily changes in Twitter account followers via https://web.archive.org.

In the second part of the study we decided to take a daily snapshot of 141 tweets posted by Donald Trump between November 1 and 13 2018 (around US midterm elections) to understand the dynamics of daily changes in the retweets associated with tweets by Donald Trump. We followed each of the 141 tweets’ retweet counts for seven days. Since we observed that the retweet count plateaued after the fourth day of the tweets being posted, we implemented a 4-day cutoff. The most popular tweet from the timespan was posted on November 2, containing a graphical image with the text “Sanctions are coming November 5”, and received 65,869 retweets and 199,616 favorites by the fourth day after the tweet had been published. The least popular tweet, which Donald Trump posted around the time of the midterm elections, was published on November 1 with the following content: “....His opponent Jared Polis is weak on crime and weak on borders – could never do the job. Get out and VOTE – Walker has my Complete and Total Endorsement!” and received 8,478 retweets and 35,015 favorites after the four days.

The descriptive statistics for the dataset are presented in Table 1, broken out by time periods.

<table>
<thead>
<tr>
<th></th>
<th>June 15 - July 30, 2018</th>
<th>November 1 - 13, 2018</th>
</tr>
</thead>
<tbody>
<tr>
<td>Retweet count</td>
<td>323</td>
<td>141</td>
</tr>
<tr>
<td>Max retweet count</td>
<td>114596</td>
<td>65869</td>
</tr>
<tr>
<td>Min retweet count</td>
<td>4883</td>
<td>8478</td>
</tr>
<tr>
<td>Mean</td>
<td>21322</td>
<td>20779</td>
</tr>
<tr>
<td>Standard deviation</td>
<td>11089</td>
<td>10269</td>
</tr>
<tr>
<td>Median</td>
<td>20197</td>
<td>18255</td>
</tr>
</tbody>
</table>

Table 1: Retweets’ descriptive statistics (June 15-July 30, 2018 vs. November, 1-16, 2018)

4. Results

The initial analysis of the dataset shows that the cleanup of Trump’s Twitter account took place on June 27 (around 100,000 followers were removed), July 12 (around 200,000 followers were removed) and July 18 (around 100,000 were removed).

![Followers tally (June 15 – July 30, 2018)](image)

Figure 2: Donald Trump Twitter account followers tally (June 15 – July 30, 2018)

As a next step, we analyzed the counts of retweets obtained for tweets published between June 15 and July 30, 2018. In order to better understand the trend, we computed the daily average for all tweets published on a given day. We observed that one day after each of the three cleanup events took place, the mean retweet count was significantly lower than earlier. We performed a sign pair t-test to establish whether the difference in the mean count of retweets before and after the cleanup is statistically
significant. We observed that there is a statistically significant difference between the retweet counts before and after the cleanup. We also conducted similar test on mean favorite count for the same dataset capturing Trump’s tweets. We observed that there was a statistically significant difference between the mean favorite counts before and after the cleanup as well. We showcase the results of both tests in Table 2.

<table>
<thead>
<tr>
<th></th>
<th>t-statistic</th>
<th>p-value</th>
<th>Mean of the differences</th>
</tr>
</thead>
<tbody>
<tr>
<td>Retweet count</td>
<td>0.60088</td>
<td>0.5542</td>
<td>955.59</td>
</tr>
<tr>
<td>Favorite (like) count</td>
<td>0.95693</td>
<td>0.349</td>
<td>4766.65</td>
</tr>
</tbody>
</table>

Table 2: Results of a sign pair t-test

However, since the raw time series was very volatile, the next course of action was to apply a smoothing technique based on moving averages. We computed two simple moving averages (one short and one long) to capture possible trends. Figure 3 shows a comparison between the 3-day and 10-day moving averages for retweet tallies. There are roughly six crossing points, where the shorter moving average crosses the longer 10-day moving average, creating seven different episodes (3 breaks to the upside, 4 to the downside). This simple technique does show that Twitter behavior (at least for Donald Trump) tends to be episodic in nature.

Figure 3: Retweet 3-day/10-day moving average (right) (June 15 – July 30, 2018)

4.1 Categorizing Donald Trump's Tweets

In order to categorize the tweets according to the number of retweets they received, we focused on two aspects: 1) the total amount of retweets a certain tweet garners and 2) the number of retweets the tweet gathers on a day-to-day basis. Our initial descriptive statistics showed a highly skewed distribution for retweets. Therefore, a log transformation was performed to help us create reasonable categories. This allowed us to identify Category 1-7 tweets, which we further grouped into three broader categories (“Rocket Ships”, “Jets” and “Blimps”) to get a better understanding of the retweet behavior as shown in Figure 4.
Following Donald Trump’s lead (see the example tweet above), tweets at the high end of the distribution are labeled “Rocket Ships” and the low end “Blimps” (with “Jets” for now making up the middle ground).

- Rocket Ships: Category 4-7 tweets that reached 30,000 retweets or above (marked by a red line in Figure 4).
- Jets: Category 2-3 tweets represent the middle ground for now (marked by a yellow line on Figure 4).
- Blimps: Category 1 tweets with less 10,000 retweets (marked by a blue line on Figure 4).

Out of 141 tweets posted by Donald Trump between November 1 and 13 2018, 11 (8%) were categorized as Blimps, 113 (80%) were categorized as Jets and 17 (12%) as Rocket Ships.

Similar descriptive statistics were computed for a set of 323 of Donald Trump’s tweets which were collected between June 15 and July 30, 2018. As shown in Figure 5, the distribution of tweet categories is similar to the distribution recorded for the tweets collected in November. When comparing percentage values before and after malicious accounts cleanup, Rockets Ship category changed its value from 15% to 13%, Blimp category went from 7% to 8%, whereas Jet changed their percentage value from 78% to 79%.

### 5. Discussion and next steps

Preliminary results of our analysis show that the implications of Twitter’s cleanup initiatives, which targeted malicious accounts, are visible in the narrowing amplitudes of retweet values. The least popular tweet in June-July 2018 (when Twitter cleanup initiatives were conducted) reached 4,883 retweets (on June 26 – before the first cleanup), whereas the most popular tweet in the indicated time was retweeted...
114,596 times (the tweet was posted on July 8 – 16 days after the first cleanup and 4 days before the second cleanup). In contrast, tweets published by Donald Trump around US midterm elections received 8,478 retweets (least popular tweet posted on November 1) and 65,869 retweets (most popular tweet posted on November 2). However, the distribution of tweet categories based on the number of retweets remained unchanged during US midterm elections in November 2018, as compared to retweets of tweets published in June and July 2018.

The contribution of this paper is as follows. First, we assembled a dataset of Donald Trump’s tweets, which we plan to make available in the future. Secondly, our analysis has shown that when it comes to a lifespan of individual tweets posted by Donald Trump, the retweet count stabilizes after the fourth day. What is more, our investigation has shown that Twitter behavior (at least for Donald Trump) tends to be episodic in nature. We then organized tweets into three categories (Rocket Ships, Jets and Blimps) and investigated how the structure of tweets’ distribution is influenced by a cleanup of malicious accounts. Finally, more fine grained categories (varying from Category 1-7 for Donald Trump) were formed based on a log transformation of the retweet amplitude. Overall, our analysis showed that the distribution of tweet categories, before and after malicious accounts’ cleanup did not change significantly. However, there was clearly a narrowing in the amplitude of retweets and favorites (as described above), which was backed by a sign pair t-test we performed to conclude that the difference in the mean count of retweets and favorites before and after the cleanup was statistically significant. This makes intuitive sense if the Twitter account removals reduced the number of bots (serving as amplifiers), as well as somewhat inactive accounts at the other end of the spectrum, thereby narrowing the retweet range.

We see this article as a first step towards a comprehensive and scientifically sound investigation on the implications of malicious account actions on social media outlets. We recognize that manipulation initiatives on social media instigated by malicious accounts have an overwhelming negative effect on societies, cause confusion and distrust and are often used to deepen societal divisions based on nationalistic, racial or religious strains (Ferrara, 2015, Wardle & Derakhshan, 2017). Therefore, as a next step we attempt to extend existing research on behavioral patterns of malicious actors on social media with the use of unsupervised learning. We plan to apply clustering techniques to investigate malicious account behaviors on Twitter based on labeled proprietary data on removed malicious accounts identified and released by Twitter. In October 2018, Twitter started publishing archives of tweets and media which had been identified as malicious and removed by Twitter from the network. According to Twitter, identified accounts had been proliferating potentially state-backed information operations instigated by Russia, Iran, Venezuela and Bangladesh (Twitter, 2018). The data archive presented on the Twitter website contains the information about malicious tweets followed by other media content, such as videos or pictures.

We plan to investigate emergent behavior of malicious accounts that Twitter tagged as connected to state-backed information operations. In the course of preliminary analysis of the dataset related to Bangladesh’s state-backed information operations we were able to extract four different features of malicious account behavior from a dataset released by Twitter as described in Table 3: (1) Account reputation defined as account’s follower count divided by the sum of account’s follower count and following count (Chu et al., 2012), (2) Account tweeting frequency defined as the average number of tweets generated by an account on a daily basis (Dickerson et al., 2014), (3) Age of account defined as total number of days between account creation and removal (Freitas et al., 2015) and (4) Account activity score defined as the number of days, during which an account was tweeting divided by the number of days between account creation and removal.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Account reputation</td>
<td>Follower count/ (follower count + following count) (Chu et al., 2012)</td>
</tr>
<tr>
<td>Tweet frequency</td>
<td>The average number of tweets generated by an account on a daily basis (Dickerson et al., 2014)</td>
</tr>
<tr>
<td>Days tweeted</td>
<td>Number of days when an account tweeted at least one time</td>
</tr>
<tr>
<td>Age of account</td>
<td>Total number of days between account creation and removal (Freitas et al., 2015)</td>
</tr>
<tr>
<td>Activity score (%)</td>
<td>Days tweeting/ days active (%)</td>
</tr>
</tbody>
</table>

Table 3: Description of account features
Due to the high heterogeneity in malicious accounts’ behavior and description, we employed a cluster analysis to group these accounts according to the defined four features. Preliminary research findings show that malicious accounts in the dataset showcase a wide variety of user behavior which we grouped into four clusters. We identified very young accounts with high volatility tweeting behavior (Cluster 1), we were also able to capture a behavior of significantly older accounts which were more consistent in their tweeting activity (Cluster 2), we could extract a group of malicious accounts who did not tweet frequently but still enjoyed quite a high number of followers and behaved like local influencers (Cluster 3) and we also identified simple accounts which were following a lot of other accounts, were not influential but helped in propagation of malicious content (Cluster 4).

![Figure 6: Demonstration of clusters on a two-dimensional plane](image)

One further extension of our work is to analyze the impact of malicious retweets’ and favorites’ counts on the manipulation of conversations on social media. Our final goal is to develop a simulation framework using agent-based modelling. Such a framework would be especially useful in gaining a better understanding of social media network dynamics and help implement and evaluate future policies on such platforms. We believe that constructing an agent-based model of the Twitter environment will help us understand the impacts of malicious accounts on tweet propagation and evaluate how hypothetical Twitter policies could impact tweet propagation on similar social media platforms.
References:


Gupta, A., Lamba, H., & Kumaraguru, P. (2013, September). $1.00 per rt# bostonmarathon# prayforboston: Analyzing Fake Content on Twitter. In eCrime Researchers Summit (eCRS), 2013 (pp. 1-12). IEEE.


