A Postmortem of Suspended Twitter Accounts in the 2016 U.S. Presidential Election

Huyen Le∗ G.R. Boynton‡ Zubair Shafiq∗ Padmini Srinivasan∗

∗Department of Computer Science, The University of Iowa, Iowa City, IA, USA
‡Department of Political Science, The University of Iowa, Iowa City, IA, USA

{huyen-t-le,bob-boynton,zubair-shafiq,padmini-srinivasan}@uiowa.edu

Abstract—Social media sites such as Twitter have faced significant pressure to mitigate spam and abuse on their platform in the aftermath of congressional investigations into Russian interference in the 2016 U.S. presidential election. Twitter publicly acknowledged the exploitation of their platform and has since conducted aggressive cleanups to suspend the involved accounts. To shed light on Twitter’s countermeasures, we conduct a postmortem analysis of about one million Twitter accounts who engaged in the 2016 U.S. presidential election but were later suspended by Twitter. To systematically analyze coordinated activities of these suspended accounts, we group them into communities based on their retweet/mention network and analyze different characteristics such as popular tweeters, domains, and hashtags. The results show that suspended and regular communities exhibit significant differences in terms of popular tweeter and hashtags. Our qualitative analysis also shows that suspended communities are heterogeneous in terms of their characteristics. We further find that accounts suspended by Twitter’s new countermeasures are tightly connected to the original suspended communities.

I. INTRODUCTION

Social media usage in elections. Social media sites such as Twitter and Facebook have played a significant role in elections across the world (e.g., U.S. [1, 2], Australia [3], Britain [4], India [5]). Social media sites are used by people to discuss election campaigns as well as by politicians to directly reach out to the electorate. A recent survey by the Pew Research Center shows that two-thirds of U.S. social media users discuss political issues on these sites [6]. Moreover, about a quarter of U.S. adults directly relied on social media sites to keep up with the 2016 presidential election campaigns of Donald Trump and Hillary Clinton [7].

Social media manipulation and countermeasures during elections. Unfortunately, the open nature of social media platforms also makes them susceptible to manipulation. Prior research has extensively reported on the widespread nature of spam and other types of abuses in popular online social networks [8], [9], [10], [11]. It is perhaps unsurprising that social media sites have been targeted during elections beyond “vanilla” spam. There were reports of social media misinformation campaigns targeting various political candidates going as far back as the 2010 U.S. midterm election [12]. There have also been numerous reports of widespread misinformation campaigns during the 2016 U.S. presidential election [13], [14], including reports by the Office of the Director of National Intelligence [15] and the Senate Intelligence Committee [16] concluding Russian state-sponsored misinformation campaigns. Popular social media sites such as Twitter [17] and Facebook [18] publicly acknowledged the exploitation of their platforms during the 2016 U.S. presidential election by state-sponsored attackers. They have since announced a number of “cleanups” [19], [17], [20], [21], which have resulted in suspension of millions of accounts [22], [23].

Analysis of social media manipulation during elections. There is significant interest in understanding the countermeasures deployed by social media sites to counter spam and misinformation campaigns specifically targeting elections. One line of research has specifically focused on characterizing state-sponsored misinformation campaigns during the 2016 U.S. presidential election. For example, researchers showed that RU-IRA (Russian Internet Research Agency) Twitter accounts systematically manipulated political discourse [13], were able to reach a substantial number of Twitter users [24], and produced content that had a mostly conservative agenda [25]. Another line of research has more broadly focused on measuring the impact of countermeasures deployed by Twitter and Facebook. For example, researchers found that popular Twitter accounts in aggregate lost over half a billion followers due to a recent Twitter “purge,” in which former president Obama lost about 2 million followers and president Trump lost about a half million followers [26].

Gaps in prior work. There are two main gaps in prior research that we hope to address in this paper. First, prior research studying social media manipulation during the 2016 U.S. presidential election is mostly limited to analyzing a few thousand Russian or Iranian state-sponsored accounts publicly disclosed by social media operators [24], [25], [27], [28]. We argue that social media manipulation during the 2016 U.S. presidential election was likely more diverse and at a much bigger scale. Second, while prior research has...
studied the impact of new countermeasures deployed by social media platforms, there is a dearth of research on understanding the inner-working of these new countermeasures. We argue that better understanding the newly deployed countermeasures can shed light into their potential blind spots and lead to development of more effective solutions.

**Proposed research.** To address the first limitation, we propose to retrospectively analyze the activities of suspended Twitter accounts that engaged in political discourse during the 2016 U.S. presidential election. We believe that a postmortem analysis of targeted cleanups (identifying using suspended accounts) provides a valuable ground truth that can be leveraged to study a wide variety of social media manipulation during the 2016 U.S. presidential election at scale. To address the second limitation, we propose to utilize two sets of suspended accounts (identified about a year apart) before and after Twitter announced new countermeasures against spam [19], [29] to analyze Twitter’s new countermeasures. We believe that an examination of how the newly suspended accounts connect to the older suspended accounts can shed light on the inner-working of Twitter’s countermeasures.

In this paper, we identify nearly a million suspended Twitter accounts that engaged with the presidential election campaigns of Donald Trump and Hillary Clinton over the duration of four months leading up to the 2016 U.S. presidential election. Then, to systematically analyze the coordinated behavior of a million suspended accounts, we group them into different communities based on their retweet and mention activities. Next, in order to examine the characteristics of suspended communities, we transfer measures for individual-level features into community-level features along five dimensions: dominant poster (responsible for posting most tweets), dominant retweet content producer (responsible for producing content for other users to retweet), burstiness (temporal bumps of tweets), dominant domain (which was tweeted most frequently), and dominant hashtag (which was tweeted most frequently).

**Key findings.** We summarize our key observations as follows:

- By systematically comparing characteristics of suspended account communities versus regular (not suspended) communities, we find differences across all five defined dimensions, but most significantly for the dominant poster and dominant hashtag dimensions. Through community-level analysis along different dimensions, we hope to provide insights to social media platforms as well as the broader research community for developing effective methods of detecting malicious accounts based on their group-level activities.

- By qualitatively analyzing suspended communities along different dimensions, we find that each of the five proposed dimensions is useful in identifying heterogeneous themes across suspended communities. We find communities that contain a large number accounts engaged in state-sponsored propaganda, those engaged in selling of political merchandise, as well as pornographic materials. This demonstrates that our analysis of targeted cleanups by Twitter through suspended accounts provided a general ground truth to study political spam and misinformation campaigns during the 2016 U.S. presidential election.

- By analyzing suspended accounts before and after Twitter’s deployment of new countermeasures, we find that more than 90% of the newly suspended accounts have direct connections to the communities of suspended accounts that were detected earlier. Moreover, a large fraction of the newly suspended account retweet or mention the top retweet content producers of old suspended communities. These findings suggest that Twitter’s new countermeasures are targeting accounts that are linked to previously suspended accounts. It also suggests that our community based methodology can be used to identify more users that are possibly eligible for suspension.

**II. RELATED WORK**

Our work uses Twitter collection related to the 2016 U.S. presidential election. There are a variety of works related to this election on multiple topics, such as policy discussion [30], [31], [2], political disinformation [13], [32], [33], Russian trolls [24], [25], [27] and bot activities [14], [25], [34]. Overall, they have shown that the 2016 U.S. presidential election has been manipulated by state-sponsored propaganda as well as distorted by rumors and social bots.

In this work, we focus on the topic of suspended accounts. Although this topic is limited to a very small dataset (e.g., Russian trolls) in election-related analysis, it has received large attention from research communities in terms of spam and fake account detection [9], [10], [35], [36], [37]. One of the most similar works to ours is by Thomas et. al. [10], who analyzed Twitter suspended accounts as spammers. The authors examined these accounts as a whole on a number of properties such as active duration, tweet rates, relationships, domain usage and compared to non-spam accounts in some cases. However, unlike their work, since our dataset is within four months before the 2016 U.S. presidential election day and directly related to two main candidates Donald Trump and Hillary Clinton, we suspect there can be different themes causing user suspension other than spam. In fact, the suspended percentage in our data collection is 9.5%, nearly triple theirs (3.3%). Moreover, we group suspended users into communities and analyze their activity at the community level rather than as a whole.

More recently, Volkova et al. [37] built classifiers to distinguish deleted and suspended accounts from active ones in three different languages. The authors found that neural network models trained on text and network features produce the highest performance for most of tasks, despite the fact that the network features they used are very simple (e.g. number of mentions). Thus, although our goal is not to predict suspended accounts, by analyzing suspended communities we can provide more insights for researchers to develop a more effective method of predicting suspended accounts based on their community activities.
III. METHODOLOGY

A. Data Collection

During the 2016 U.S. presidential election, as a part of our previous studies on political discourse [2], [38], we collected tweets around two major presidential candidates Clinton and Trump. Specifically, we used Twitter’s streaming API with filter keywords as full names of candidates (e.g. “hillary clinton” for Clinton) to collect tweets for each candidate. Since this API caps the tweets at 1% of all public tweets and there are more than 500 million tweets posted per day on Twitter [39], we are set to capture up to five million tweets per day for each candidate. However, since in our data collection the highest daily tweet count (for Trump) was less than two million, we were still able to capture the vast majority of tweets for both candidates.

In this work, we are interested in analyzing Twitter suspended accounts during the 2016 U.S. presidential election. Thus, in February 2018, using Tweepy - a Python library to access the Twitter API - we were able to examine how many accounts were suspended in our previous tweet collection around Clinton and Trump. Specifically, Twitter API returns the request of the user status with the response code of 63 if the user was suspended. Table I shows the statistics of our data during nearly four months, from June 01 to November 08, 2016 (except for most of August and some individual days due to the crash in our collecting process). In total, there were 912,979 accounts suspended out of 9,572,020 which made the suspended percentage to be 9.5% on average.

<table>
<thead>
<tr>
<th>TABLE I</th>
<th>STATISTICS OF SUSPENDED USERS AS OF FEBRUARY 2018 IN OUR TWEET COLLECTION AROUND CLINTON AND TRUMP DURING NEARLY FOUR MONTHS, FROM JUNE 01 TO AUGUST 01 &amp; FROM SEPTEMBER 09 TO NOVEMBER 08, 2016 (EXCEPT SEPTEMBER 18-20, OCTOBER 31, AND NOVEMBER 1).</th>
</tr>
</thead>
<tbody>
<tr>
<td>Users</td>
<td>unique 9,572,020 suspended 912,979</td>
</tr>
<tr>
<td>Tweets</td>
<td>unique 9,025,222 suspended 9,218,751</td>
</tr>
</tbody>
</table>

B. Suspended Communities

Motivated by the lack of using network activities in understanding user suspension, we decide to analyze suspended accounts based on their retweeting and mentioning activities. We decide to use retweet and mention activities because of their representativeness for tweeting interactions. Specifically, retweet and mention activities among accounts demonstrates explicit engagement, especially since users can retweet/mention another user who they do not follow. In the case of follower network the level of engagement may be regarded as weaker and more passive. Besides, we are also unable to have their follower network since these are suspended users. To this end, we first build the retweet and mention network starting from these suspended users. Table II shows the statistics of this network’s directed graph, with the frequency weighted edge coming from the suspended user who retweeted/mentioned to the user who was retweeted/mentioned. The total of nodes or users is nearly one million while the total of edges is more than 14 million.

<table>
<thead>
<tr>
<th>TABLE II</th>
<th>STATISTICS OF RETWEET AND MENTION NETWORK FROM SUSPENDED AND REGULAR USERS.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Retweet + Mention Network</td>
</tr>
<tr>
<td>Nodes</td>
<td>955,083</td>
</tr>
<tr>
<td>Edges</td>
<td>14,321,863</td>
</tr>
<tr>
<td>Biggest Community’s Size</td>
<td>13,809</td>
</tr>
<tr>
<td>Number of Communities</td>
<td>72,864</td>
</tr>
<tr>
<td>Number of Communities (Size ≥ 10)</td>
<td>9,554</td>
</tr>
</tbody>
</table>

Next, suspecting that suspended users tend to act together (a.k.a synchronized or coordinated behaviors) or more strongly connected to each other, which was shown for spammers and malicious accounts in previous work [40], [41], [11], we want to analyze suspended accounts at the group level for better interpretation. To this end, we apply Louvain community detection algorithm [42] on the retweet and mention network to identify communities for these suspended users, so called suspended communities. In general, this algorithm tries to build communities so that connections are much more inside but less outside a community. Specifically, it optimizes modularity value which measures the density of edges inside communities to edges outside communities. Small communities are found at first by optimizing modularity value locally, then each small community is considered as one node and the process is repeated. In this study, we aim to have a manageable size of a community for better analysis so we constrain size of the biggest community at most 15K. To this end, the algorithm results in 72,864 communities with the size of the biggest one as 13,809 (suspended users). Moreover, we do not focus on very small communities (size < 10) so we finally have 9,554 suspended communities to analyze.

C. Community-Level Features

Previous works [43], [14] have reported the list of signs that could suggest a Twitter account is bot or automated such as very high posting rate, only a few sources or accounts are retweeted, multiple tweets of the same link. While these signs could be missed for an individual account, synchronized behaviors of users in a community could reveal them. Thus, inspired by these previous observations as well as Twitter rules on account suspension [44], we define analogous community-level characteristics to examine evidence of synchronized behaviors at the community level. In essence, we are looking for community-level signals for account suspension. Follows are our five defined dimensions representing distinct community-level features.

1 Note that the retweet and mention network starting from our collection of suspended users also contains other users who were retweeted/mentioned and could be suspended or regular. However, after communities were finally built based on this retweet and mention network, we study communities from members who are suspended users in our collection. Thus, we call communities as suspended communities.
1) Dominant Poster: The highest percentage of tweets in a community that are posted by a single user. The higher this percentage, the more dominant a single user.

2) Dominant Retweet Content Producer: The highest percentage of tweets in a community that are retweeted from a single user (tweets can be posted by either one or multiple users). The higher this percentage, the more dominant a single user.

3) Burstiness: The highest times tweets posted in one hour are higher than the average tweets per hour in a community. The higher this percentage, the more bursty tweets were posted.

4) Dominant Domain: The highest percentage of tweets from a community that contain URLs from a single domain. The higher this percentage, the more dominant a single domain.

5) Dominant Hashtag: The highest percentage of tweets from a community that contain a single hashtag. The higher this percentage, the more dominant a single hashtag.

While the first and second dimensions represent network features of these suspended communities, the third dimension represents temporal features and the two last represent content features. In fact, we also analyze suspended communities on other features such as tweet text and language. However, the findings either mostly overlap with these five dimensions (e.g. tweet text) or are almost meaningless (e.g. language). Thus, in this paper we only present our analysis and findings on these five dimensions.

Note that since these proposed dimensions serve as collective means to measure the characteristics of a community, the higher values in these five dimensions the more distinct and possibly suspicious the community activities. Thus, we are interested in communities which have high values in these dimensions. Specifically, we focus on communities having at least 50% of tweets posted by one user, or having at least 50% of tweets retweeted from one user, or having tweets posted in one hour which is at least 200 times the average tweets per hour, or having at least 50% of tweets containing URLs from one domain, or having at least 50% of tweets containing the same hashtag. In other words, we specify these high thresholds for our further analysis. While as an arbitrary choice, the thresholds as 50% and 200 times serve well as moving these values higher will not change the meaning of our analysis and findings. Further details can be seen in next section Results.

IV. RESULTS

A. Suspended and Regular Communities

To serve as a baseline for observing special characteristics of suspended accounts’ activities, we compare the activities of suspended accounts to those of regular accounts. Here regular accounts are users who Twitter API returns the request of user status as “regular”. In total, our data collection have 7,740,693 regular users who posted nearly 48 million tweets. We apply our previous approach of building suspended communities for these regular users. Specifically, we first build the retweet and mention network starting from these regular users. Table III shows the statistics of this network’s directed graph, with the frequency weighted edge coming from the regular user who retweeted/mentioned to the user who was retweeted/mentioned. The total of nodes or users is nearly 6.5 million while the total of edges is nearly 39 million. We then use the Louvain community detection algorithm with the same constraint for the size of the biggest community (at most 15K). The algorithm results in 690,337 communities with the size of the biggest one as 13,537 (regular users). Finally, we have 67,734 regular communities with size $\geq 10$.

We then analyze both suspended and regular communities along our proposed community-level features. Figure 1 plots CDFs of suspended and regular communities for these five dimensions. We further use the Kolmogorov-Smirnov (KS) test to investigate whether the distributions of suspended and regular communities in each dimension are significantly different. We are able to reject the null hypothesis that both distributions are the same at the 0.02 significance level for the dominant hashtag dimension and at the 0.0025 significance level for the dominant poster dimension. Thus, among five investigated dimensions we find that suspended and regular communities exhibit significant difference in terms of hashtag dominance in posting content and very significant difference in terms of user dominance in posting behavior.

Moreover, since we are interested in communities which have high values in these five dimensions, we focus on communities contributing to the CDFs from the vertical line to the right in each plot of Figure II. Due to the big differences between the total number of suspended communities (9,554) and regular communities (67,734), we only compare the level of suspended and regular communities having high values in these five dimensions in terms of percentage for a fair comparison. Specifically:

- The feature displaying the largest difference in percentages between suspended and regular communities is the dominant poster dimension. Figure 1(a) shows that 38.3% suspended communities have at least 50% of tweets posted by a single user while this type of user dominance occurs only in 3% regular communities. Thus, the ratio of percentages for suspended to regular communities is 12.8.
- The three features with the next highest ratio (i.e. 3 to 4) for suspended versus regular communities are dominant retweet content producer, dominant domain, and dominant hashtag dimensions. For example, Figure 1(c) shows that 8.6% suspended communities have at least 50% of tweets containing URLs from one domain while this type of domain dominance occurs only in 2.3% regular communities.
- The suspended and regular communities are about the same on burstiness dimension. Figure 1(b) shows that 17.2% suspended communities have tweets posted in one hour which is at least 200 times the average tweets per
hour while this type of bursty posting occurs in 13.7%
regular communities. Thus, the ratio of percentages for
suspended to regular communities is only 1.3.

Note that one community can have high values in multiple
dimensions. In total, 57.6% suspended communities have high
devalue in one or more of five investigated dimensions while that
percentage is lower as 17.6% for regular communities. Thus,
the ratio of suspended communities to regular communities
which have high value in one or more of five investigated
dimensions is 3.3. This suggests that high values in multiple
dimensions is a signal for the closer scrutiny on a com-

Takeaway: Overall, our results show that suspended and
regular communities exhibit differences on all five investigated
dimensions. Especially these differences are significant in
dominant hashtag dimension and very significant in dominant
poster dimension. Moreover, compared to regular communities
there are more suspended communities having high values on
all five dimensions. Especially this can be up to one order
magnitude (12.8 times) for the dominant poster dimension.

B. Qualitative Analysis

To illustrate what is going on suspended communities’
activities, we next do qualitative analysis on several repre-
sentative communities which have high values in these five
dimensions. Specifically, we first analyze one representative
community which has a significant high value in only each
dimension. Particularly, for content dimensions such as domain
and hashtag, the analyzed community is one of suspended
communities who posted unique and suspicious dominant
domains and hashtags compared to regular communities. We
then analyze one representative community which has high
values in multiple dimensions. Table III shows statistics of
these six suspended communities.

1) Dominant Poster:

- Trump-IRA Community: Most tweets (147 out of 201)
  are from an account @td21241 (profile name as terri
  in July, mainah4Trump in most of September, and De-
  plorableME4Trump for other times in our collection),
  who retweeted multiple different users (from outside
  the community) with contents supporting Trump and
  attacking Clinton. Aligned with this user’s profile de-
  scription which only contains #MAGA, top hashtags used
  in the community include #MAGA, #Trump2016, and
  #NeverHillary. More interestingly, checking with the list
  of 3.8K RU-IRA accounts on Twitter released by U.S.
  congress [43], we notice that 11 out of 28 members in
  this community are RU-IRA accounts. Overall, we see
  295 RU-IRA accounts present in 165 different suspended
  communities (out of the total 9,554 suspended com-
  munities). This community appears to be the second top
  suspended community in containing the most RU-IRA
  accounts. Looking closely on the community’s network,
  it reveals that the dominant poster happened to play a
  role as filtering messages from external users for some
  RU-IRA accounts to retweet.

2) Dominant Retweet Content Producer:

- GayRights Community: Most tweets (2,716 out of 2,935)
  were posted by an account @DCHomos and retweeted by
  this community’s members. Aligned with this account’s
  profile description which contains “all things LGBT+ in
  DC”, this community discussed mainly about Gay Rights
  policy, with most tweets including keywords as lgbt, gay,
  marriage equality. For example, more than 100 retweets
  with content “hillary clinton mention lgbt rights in her
  opening statement! “do not reverse marriage equality”
  appeared shortly in two hours of October 20. Moreover,
  the community’s messages show that they support Clinton
  and oppose Trump. For example, nearly 100 retweets
  with content “don’t ever forget, republican nominee Don-
  ald Trump mocking a disabled reporter. #demconvention
  #demsinphilly” appeared shortly for one hour in July 26
  - time of the Democratic National Convention.
### TABLE III
STATISTICS OF SIX REPRESENTATIVE COMMUNITIES FOR QUALITATIVE ANALYSIS.

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Community Name</th>
<th>Number of Users</th>
<th>Number of Tweets</th>
<th>Number of Retweets</th>
<th>% Retweets</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dominant Poster</td>
<td>Trump-IRA</td>
<td>28</td>
<td>201</td>
<td>172</td>
<td>85.6</td>
</tr>
<tr>
<td>Dominant Retweet Content Producer</td>
<td>GayRights</td>
<td>1,578</td>
<td>2,935</td>
<td>2,850</td>
<td>97.1</td>
</tr>
<tr>
<td>Burstiness</td>
<td>Vote/BlackLivesMatter</td>
<td>1,544</td>
<td>2,313</td>
<td>2,000</td>
<td>86.5</td>
</tr>
<tr>
<td>Dominant Domain</td>
<td>EbayAds</td>
<td>1,062</td>
<td>2,105</td>
<td>2,105</td>
<td>100.0</td>
</tr>
<tr>
<td>Dominant Hashtag</td>
<td>NoEthicsNoOffice</td>
<td>89</td>
<td>5,532</td>
<td>407</td>
<td>7.4</td>
</tr>
<tr>
<td>All except Dominant Poster</td>
<td>PoliticalPorn</td>
<td>89</td>
<td>5,532</td>
<td>407</td>
<td>7.4</td>
</tr>
</tbody>
</table>

#### 3) Burstiness:
- **Vote!BlackLivesMatter** Community: Most tweets (2K out of 2,313) are retweets from many different users, of which 642 are retweets from a user @BernieSanders. The content of this community’s tweets shows that they first supported Sanders but then gave their support to Clinton after Sanders lost to Clinton for the Democratic nomination. Particularly, there is the highest spike in the late morning of September 22 including 862 retweets from a user @NadelParis with the content as “I#vote! @hillaryclinton: #humanrights w. @berniesanders on #freecollege + #retrainpolice!! #blacklivesmatter!” (Figure 2).

#### 4) Dominant Domain:
- **EbayAds** Community: Almost URL-posted tweets in this community (75K out of 77K) used Ebay platform (ebay.com) to promote selling stickers such as “CHICKEN TRUMP” or “HILLARY CLINTON LOVE TRUMP HATE” (Figure 3). This is community supporting the Democratic party, along with Hillary Clinton and Bernie Sanders. Their top used hashtags contain multiple Democrat-supported hashtags such as #dems, #blm (Black Lives Matter), #toporg (Top Progressives), #ctl (Connect The Left) as well as multiple Trump-opposed hashtags such as #trumpdsreceive, #chickentrump, #dumptrump. It is also noticeable that this community posted mostly original tweets, with the retweet rate of 0.1%.

#### 5) Dominant Hashtag:
- **NoEthicsNoOffice** Community: Their dominant hashtags include #hillaryclinton, #hillary, #clinton, #election, #politics. The reason is that near the end of October, from different users in this community about 3.2K tweets (out of 5.5K) contained these dominant hashtags. Moreover, these hashtags were mostly posted together with URLs from domain tumblr.com which conveyed messages to oppose Hillary Clinton such as “No Ethics No Office” or “She sold out Bernie, next she’ll sell out America” (Figure 4).

#### 6) Combination:
The community with the highest values on all five dimensions is the one that supports Trump and Clinton. They use hashtags for different purposes such as promoting votes (e.g. Vote!BlackLivesMatter) or spreading porn (e.g. PoliticalPorn). They engage in the discussion of a diverse set of issues ranging from foreign policy to gay rights (e.g. GayRights). Especially, some of them engage in the discussion of Russian trolls (e.g. Trump-IRA).

#### C. Twitter’s Suspension Algorithm
Accounts that violate Twitter policies are constantly removed from the platform. We waited for more than one year after the 2016 U.S. presidential election (as of February 2018) to get the suspended account collection as described and analyzed in previous sections. However, since May 2018 Twitter announced a number of cleanups, so-called Twitter “purge” [26], which led to suspension of many more Twitter accounts. This happened because Twitter claimed to develop “new measures to fight abuse and trolls, new policies on hateful conduct and violent extremism, and ... new technology and staff to fight spam and abuse” during May-August 2018 [19, 29]. Thus, in January 2019 we rechecked to
see how many more accounts in our election-related tweet collection were suspended. Although these newly suspended accounts may be suspended because of either actions during 2016 or actions taken since the election, we assume these newly suspended accounts might be the result of the new development in Twitter’s suspension algorithm. We want to know whether these newly suspended users had connections to the original suspended ones. We hope our findings will shed light on Twitter’s new development in their account suspension algorithm.

In fact, our 2019 recheck has found an additional newly suspended accounts in our dataset. Specifically, there are 192,415 suspended accounts that had not been suspended earlier. This increases the percentage of suspended accounts in our dataset from 9.5% to 11.6%. We then determine if these newly suspended accounts have connections to the 9,554 communities of the original suspended accounts. To this end, we calculate the maximum retweet/mention connection of each newly suspended account to the original suspended communities. The result shows that more than 90% of the newly suspended accounts had at least one retweet/mention connection to an original suspended community.

Next, since most of new suspended accounts have direct connections to old suspended communities, we want to examine the strength of these connections. Specifically, we ask whether these new suspended accounts connected to major actors in the original suspended communities or is it that there is no pattern in the connections. To this end, since the direct connections from new suspended users to old suspended communities are retweets or mentions, we examine what percentage of these newly suspended users connected to the top-k retweet content producers of the original communities.

Note that when k = 1, the top-k retweet content producers actually is the dominant retweet content producer for that community, one of our five proposed dimensions. Besides, although a newly suspended account may be connected to more than one of the original communities, we consider only the strongest connection (i.e. smallest k) for each newly suspended account. The result shows that 72% of the newly suspended users retweeted or mentioned the dominant retweet content producer of an old suspended community. And this percentage increases to 85% with k = 5, which means that 85% had a connection to one of the top five active users in the original community.

**Takeaway:** Overall, we find that Twitter’s new development on their suspension algorithm helps to detect more malicious users. However, our analysis on Twitter newly suspended users reveals that more than 90% of them have direct (retweet/mention) connections to communities of suspended users that Twitter detected before. More interestingly, a high percentage (>72%) of these new suspended users retweeted or mentioned the top retweet content producers of the old suspended communities that they connect to.

**V. Conclusion**

In this paper, we retrospectively analyze the activities of suspended Twitter accounts that engaged in political discourse during the 2016 U.S. presidential election. By developing community-based method and measures which are new to the field of analyzing suspended accounts, we are able to characterize about a million suspended accounts. In short, we find that (1) suspended communities are different from regular communities in their posting behavior; (2) suspended communities exhibit heterogeneous characteristics; and (3)
newly suspended accounts connect tightly to the old suspended communities.

To the best of our knowledge, we are the first to conduct in-depth postmortem analysis of accounts suspended by Twitter to study their community-level activities as well as assess the effectiveness of Twitter’s new countermeasures. Our community-level analysis of suspended accounts highlights their coordinated behaviors which can be leveraged to develop more effective countermeasures. Although the results in this paper are limited to the 2016 U.S. presidential election, our postmortem analysis approach is broadly applicable to any future election-related events for which social media sites are looking to develop more effective countermeasures.

REFERENCES