On Twitter Purge: A Retrospective Analysis of Suspended Users

Farhan Asif Chowdhury
University of New Mexico, USA
fasifchowdhury@unm.edu

Lawrence Allen
University of New Mexico, USA
lallen@unm.edu

Mohammad Yousuf
University of New Mexico, USA
myousuf@unm.edu

Abdullah Mueen
University of New Mexico, USA
mueen@unm.edu

ABSTRACT

Abuse and spam in Twitter have long been a pressing issue, and in response, Twitter regularly purges (i.e., suspends in mass) accounts that violate Twitter Rules. However, there is no available information about the characteristics and activities of these regularly purged users. We have developed a novel and comprehensive measurement mechanism to identify millions of purged Twitter users and collect their tweets. We have identified 2.4M purged users and collected 1M tweets made by them over eight months. Using our dataset, we perform a retrospective analysis to characterize their account properties and behavioral activities. We analyze their tweet content to identify their role and abuse strategy over-time.

Our analysis shows that the abuse on Twitter is pervasive globally and not confined in mere spamming. Alarmingly, more than 60% of the purged users survived on Twitter for more than two years. We observe that politics is a major theme among the purged users irrespective of language and location, and these politically motivated users spread controversial content consistently over time. However, the spammers reorient their agenda across time to participate in multiple marketing campaigns. We also discover interaction and associated communities among purged users. Our analysis sheds new light on the evolving nature of abuse in Twitter that can help researchers understanding the characteristics and behavior of emerging malicious users to develop an effective defense system.

CCS CONCEPTS

• **Information systems → Social networking sites**; *Web mining*; • **Security and privacy** → Social engineering attacks.

KEYWORDS

Social Networks, Abuse, Suspension

ACM Reference Format:
Farhan Asif Chowdhury, Lawrence Allen, Mohammad Yousuf, and Abdullah Mueen. 2020. On Twitter Purge: A Retrospective Analysis of Suspended Users. In *Companion Proceedings of the Web Conference 2020 (WWW '20 Companion), April 20–24, 2020, Taipei, Taiwan.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3366424.3383298

WWW '20 Companion, April 20–24, 2020, Taipei, Taiwan

ACM ISBN 978-1-4503-7024-0/20/04.
https://doi.org/10.1145/3366424.3383298

1 INTRODUCTION

Social media platforms were considered to promote the free flow of information and freedom of speech. Still, as social media became increasingly popular, the misuses of these platforms grew too. Previously, spamming was considered as the major abuse in social media. However, during recent political events, these platforms were used in disinformation campaigns and as a medium for manipulation, as indicated in many reports [5, 10, 14]. These abusive and malicious activities raised serious concerns about the vulnerability of these platforms against such campaigns. Since investigations began into the reported misinformation campaigns targeting the 2016 U.S. presidential election, social media companies (i.e., Facebook and Twitter) have claimed to launch several initiatives to counter the spread of misinformation on their platforms [1, 3].

Twitter has announced its improved capabilities to detect and suspend suspicious accounts. In a report [2], Twitter describes, "we are now removing 214% more accounts for violating our spam policies on a year-on-year basis". In 2018, in three months, Twitter reportedly suspended nearly 70M accounts [4]. Although Twitter purge is a significant event, very little has been studied about purged users. Misinformation, spamming, and bot detection are some of the most studied aspects regarding Twitter suspension [15, 18, 22, 30, 33]. An extensive body of research has been carried out to develop algorithms and tools for automated BOT detection and twitter suspension prediction [9, 17, 19, 28, 34]. Previously, Twitter suspension has been thoroughly studied in the context of spamming [11, 32, 36]. After the 2016 U.S. presidential election, emphasis has been put on studying the characteristics of suspended Russian and Iranian troll accounts, and their activities during the election [23, 25, 37]. Recently, [26] studied a more general group of suspended users active during the 2016 U.S. election.

However, there has been no research on users being purged on a regular basis. In general, several questions about purged users remain unanswered. For example, what makes the purged users different from regular Twitter users? What was the role of suspended users before the purge? Were there interactions among the purged users? These questions are important for two reasons. First, neither there is public information available about the characteristics of the purged accounts, nor there is a way to collect them easily. Second, there is a need for transparency about the inner workings of Twitter's suspension policy, which is cascaded to events such as U.S. Congressional hearings. By examining the roles and activities of the purged users, it is possible to identify the topics and events targeted for manipulation.

In this paper, we aim to address the above-mentioned questions by performing retrospective analysis on purged users and their activities. In that regard, we identify 2.4M purged users that were suspended by Twitter in August 2018 by deploying a novel data collection mechanism. We also collect more than 1M tweets previously posted by these purged users. We analyze these two datasets to characterize the purged users and examine their role prior to purge. To characterize these users, we compare their account properties with randomly sampled regular users. We analyze the active duration and follower-friend information to gauge the impact of suspended users during their lifetime. Through language and location analysis, we explore the spread of malice and manipulation across regions.

We have detected a large *Russian Botnet* consisting of 54*K* dormant users with exact same properties across multiple dimensions, which were created sequentially within a very short time period. We examine the shared content of these users to identify their role and the topics targeted for manipulation. We form a hashtag similarity network by training a *word2vec* word-embedding model to cluster hashtags used in a similar context. We identify distinct user groups based on their participation in different hashtag clusters. Later, we analyze their profile information and shared tweet content to characterize their malice. We also analyze their content sharing strategy over time. Using retweet data, we identify interaction and derived communities among purged users. We explore these communities and their activities across multiple dimensions.

Key findings. Our study leads to several key observations.

- We find that malice in Twitter has spread beyond automated spamming, and politics is a major conversational topic among the suspended users across language and region.
- In contrast with the previously short-lived spammers, more than 60% of the suspended users in our dataset sustained in Twitter for at least two years, meaning they had ample opportunity to abuse the Twitter platform. We have also identified a large cluster of dormant users with suspiciously synchronized profile information.
- The suspended users had a follower base similar to regular users, which implies many of them were able to create a large follower base. These suspended users were well distributed across the world. All these users exploit hashtag and mention as a key tool to disseminate their content, which is a consistent behavior irrespective of language and location.
- In general, we observe two major abuse by these suspended users; (1) politics and (2) viral marketing campaign. Political users were persistent in spreading a common agenda over time (i.e., #QAnon, #FakeNews). However, spammers evolved based on related events.
- Exploiting retweet information, we identified interaction among the purged users who collaborated towards a similar objective. Based on interaction, we detect several user communities of distinct group-level features across multiple dimensions.

The rest of the paper is organized as follows. We discuss related work in Section 2. We describe our novel data collection system and collected dataset in Section 3. We analyze the characteristics of purged users in Section 4. In Section 5, we analyze the contents shared by purged users. We show the interactions among the purged users to detect ideological groups in Section 6. In section 7, we discuss various aspects of our study, and we draw a conclusion.

2 RELATED WORK

Early works on suspension on Twitter [11, 27, 32, 35, 36] focused mostly on spamming and aggressive marketing. Thomas et al. identified 1.1 million Twitter accounts suspended between August 2010 and March 2011. They found 93% of the suspended accounts were spamming, and the remaining 7% were involved in "mimicking news services and aggressive marketing." In [11], analysis has been performed on account properties of spam accounts after categorizing them into two categories. Both these early studies mainly focused on spammers and their behavioral attributes. However, in the recent past, abuse in social media had many dimensions, and our study is inclusive of all types of malice.

Recent works analyzed state-sponsored campaigns and trolls trying to manipulate outcomes of political events such as the U.S. presidential election [13, 23, 25, 37]. These studies mainly focused on characterizing the activities of foreign state-sponsored (Russian and Iranian) accounts during the 2016 U.S. presidential election. Scholars also studied similar suspension actions of other social networking sites (i.e., Reddit), such as removing pages for hate speech [16]. Recently, [26] analyzed nearly one million suspended Twitter accounts that were active during the 2016 U.S. presidential election. They found that suspended users were heterogeneous, meaning they were using Twitter with a variety of objectives. And they were significantly different from regular users "in terms of popular tweeter and hashtags."

The above studies focused on a specific group of suspended users, users who actively participated in the 2016 U.S. presidential election discussion. However, the malice present in Twitter is not limited to political discourse. On the other hand, the previous suspension related studies were focused on spammers. Our dataset consists of a comprehensive set of suspended users irrespective of their participation, which gives us a unique opportunity to provide a holistic overview malice present in Twitter.

Twitter Terms and Rules. Twitter explicitly mentions the reasons behind the suspension of accounts [8]. According to Twitter, suspended accounts are "spammy, or just plain fake, and they introduce security risks for Twitter and all of our users." Twitter also suspends accounts for a reported violation of its rules regarding abusive behavior such as "sending threats to others or impersonating other accounts." Though some accounts are removed permanently, others can be reinstated. Twitter also admits that sometimes it mistakenly suspends accounts belonging to "real" persons. But these accounts can be reactivated later. Twitter may also suspend or terminate accounts because of prolonged inactivity.

3 DATA COLLECTION FRAMEWORK

With the motivation to detect purged users and characterize their activities, we developed two distinct data collection frameworks. One was deployed to detect purged users within a specific time frame, and another was used to collect 1% sample Tweet using *Twitter Streaming API*.

Table 1: Purged and Control Users Statistics

Statistic	Count
# of Purged Users	2, 420, 073 (2.4M)
# of Control Users	1, 973, 700 (2M)

Table 2: Tweet Collection Statistics

Statistic	Count
# of Tweets	1, 078, 727 (1M)
# of Retweets	765, 741 (765K)
# of Tweets/Retweets with a URL	226, 385 (226K)
# of Distinct Purged Users	147, 421 (147K)

Purged Users Detection. To detect purged users, we curated a list of 560M Twitter users by collecting follower information of top 100 most-followed Twitter users. Using the Twitter API, we collect two distinct snapshots of these 560M users starting on 4th August'2018 and 11th September'2018. We compare the earlier user set with the later user set to identify the purged users in between those two dates, which produces a list of 2.4M purged users. We make another round of Twitter API request specifically for these 2.4M users to confirm their suspension as Twitter sends the response code of 63 for the suspended users. This process of contrasting two snapshots to identify purged users is the first of its kind. Note that, purge is not an abrupt process of suspending millions of accounts; it is rather a continuous process of suspending at a higher rate than usual.

Control Dataset. To compare suspended users with regular users and distinguish their characteristics, we randomly sampled 2M users. These users were sampled from the 560M users collected on 4th August 2018, which were not suspended, whom we label as the *control user* set. In Table 1, we report the purged and control users statistics.

Tweet Collection. The major challenge to analyze purged user's activity is the restriction to access the tweets of a user after he/she is suspended. To circumvent this limitation, we deployed a Tweet data collector that continuously collected 1% sample Tweet using *Twitter Streaming API.* In this method, we collect in total 90M tweets made by 19M unique users from 7th December 2017 to 4th August 2018, a total of eight months. After the purge during August 2018, we use this tweet collection to filter out 1M tweets made by 147K unique purged users. In Table 2, we describe key statistics of our purged users tweet collection.

Ethical Concerns. We use the Twitter API keys only for passive data collection, and we do not engage in any posting activity. We do not redistribute Twitter data, and we maintain ethical research standards [31]. Hence, the project was exempted from formal *IRB* review.

4 ACCOUNT CHARACTERISTICS

We analyze the account properties of the purged users in comparison with the control user set that includes account creation time, follower-following relationship, activity rate, account location, and language.

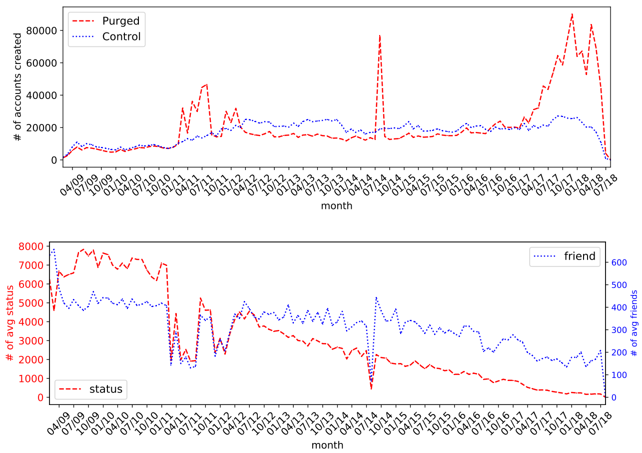

Figure 1: (top) Number of accounts created per month for purged and control user set. (bottom) Number of average status and friend of purged users grouped by account creation month.

4.1 Account Creation

In Figure 1(top), we plot the number of accounts created per month for both purged and control user set. A few important observations can be made from this figure. Firstly, although the account creation date is well distributed across time for the control users, nearly 40% purged users were created in the last two years (2017 and 2018), which resonates with the fact that twitter is becoming more proactive in response to terms and rules violation. However, 60% of the purged users were active for at-least two-years, which is significantly different from previous research on malice in Twitter [11, 32]. In [32], for spamming related suspended accounts, it has been reported that spammer accounts have a very short life span, however, as evident from our findings, the face of malice in Twitter has taken new turns.

Russian Botnet. A prominent anomaly in Figure 1(top) is the stark presence of few clustered high account-creation months in the mid and early years. To uncover the reason behind this, we examine the average status count and friend count of purged users grouped by their creation week. We specifically choose these two properties as these are controllable only by the account owner. In Figure 1(bottom), we observe that for the accounts created in clusters with high volume, the average status and friend count is significantly lower than the nearby months. As a case study, we examine accounts that were created in July'2014 (the month with the highest account creation), and filtered out 54, 266 accounts that had exactly seven friends, and zero status, follower, favorite count. Although these accounts had English as their default language, all the account names were detected as Russian. Moreover, most of these accounts were created sequentially in seconds apart and did not mention anything regarding automated (bot) account in their description. Although these accounts did not share any content, their malice

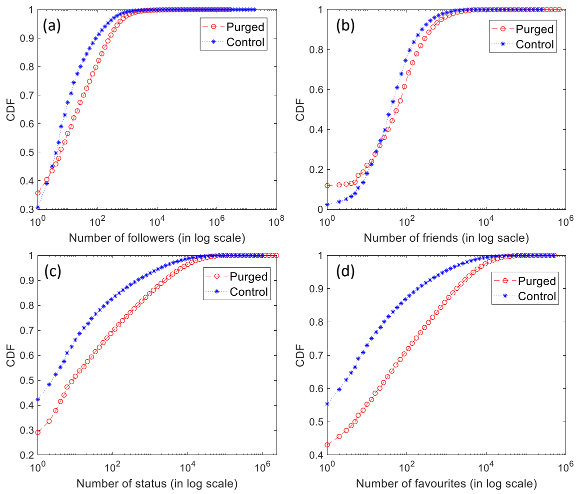

Figure 2: Cumulative Distribution Function of no of (a) Followers, (b) Friends, (c) Status and (d) Favourites for purged and control set users (note the log-scale x-axis).

can be attributed to their account creation mechanism. Previously, [21] have detected a similar *Star Wars* Botnet with suspiciously similar account properties, which only tweeted randomly chosen quotes from the *Star Wars* novel.

4.2 Distributional Properties

In Figure 2, we show the cumulative distributions of the number of *followers, friends, statuses, and favorites* for purged and control users. Alarmingly, the purged user's follower distribution looks similar to control users, which implies that many of the purged users were able to build a large follower base, although straightforward spammer accounts reportedly had fewer follower count [32]. Also, 15% purged users do not follow any user, compared to only 1% of control users, which indicates that many purged accounts remained inactive right after account creation.

In general, the purged users were notably more active than the control users based on status and favorite count as expected and reported in [32]. For example, as shown in Figure 2(c), 20% of the purged users had 1,000 or more status, while for the control set, it is only 10%. The purged users also more active to favorite other user's tweets, which is often employed by automated users to exploit *homophily* [20].

4.3 Language and Location

We use the language and location information associated with each account to obtain a generalized overview of their target audience. It is to be noted that these two are self-reported properties, and not-mandatory.

Figure 3 shows circles with a radius proportional to the frequency of user location. Self-reported locations are often not machine-readable, and they provide variable spatial resolution. We show the country-based frequency at the center of each country and state or city-based frequency at their respective locations. The results show that the highest number of purged users are from the United

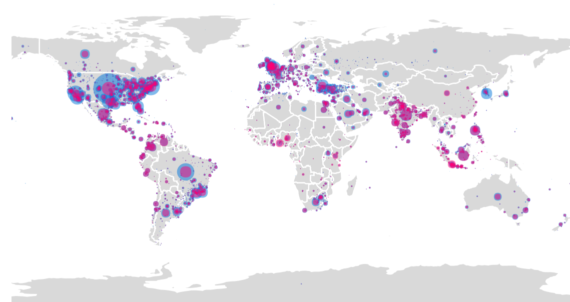

Figure 3: Locations of the purged and control users. Radius is proportional to count. Blue represents the purged users while red represents the control users.

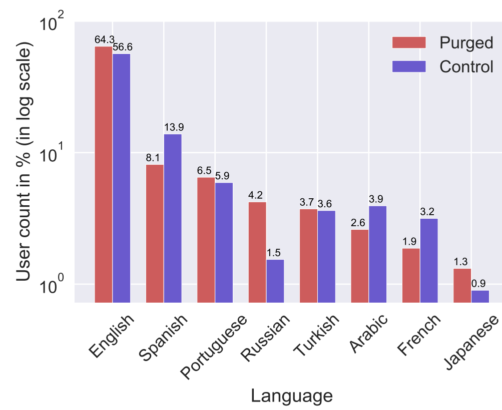

Figure 4: Distribution of the eight most-used languages on Twitter by purged and control users. Note the log scale on the y-axis.

States. Brazil and Turkey got the second and third highest numbers of purged users. Other countries with a good number of purged users include the U.K., Mexico, and Japan. In the United States, the highest numbers of purged users were from California, Florida, and Texas.

In Figure 4, we show the user distribution across top eight languages that constitutes more than 90% of the purged users. Although the top three languages have similar occurrence in purge and control, Russian is fourth among the purged accounting for 4.2% users, where else it is only 1.5% for the control users, which makes the control to the purged ratio 1 to 2.8, much higher than any other languages.

5 CONTENT ANALYSIS

In this section, we perform analysis on the content shared by the purged users to observe their role in the last eight months prior to the purge. By examining the role of these purged users, we can identify the topics and discourse that are being targeted by malicious users. Our purged user set contains a diverse group of users who use various languages and are spread across the world, which helps us in obtaining a holistic view of malice on Twitter.

Table 3: Top hashtags across five most used languages (Arabic and Turkish are Translated)

English	(%)	Arabic	(%)	Portuguese	(%)	Spanish	(%)	Turkish	(%)
iHeartAwards	1.18	Friday	1.35	BBB18	3.42	DebateINE	1.35	Election2018	3.32
BestFanArmy	0.93	Saudi ArabiaEgypt	0.71	TheVoiceKids	1.86	iHeartAwards	1.12	PresidentErdoğan	1.49
BTSARMY	0.89	SaudiArabia	0.67	GleiciDoRetorno	1.76	KCAMexico	1.20	NewEraWithErdogan	1.43
WorldCup	0.81	Eid'sPrizes	0.66	MasterChefBR	1.58	MTVHottest	0.96	Maltepe	1.40
MondayMotivation	0.78	WorldCup	0.65	BrasilComBolsonaro	1.30	Rusia2018	0.79	WeWillNotForget	1.29

Table 4: Key statistics of tweets in five most used languages

language	Users	Tweets	Retweet	Hashtags	URLs	Mentions
(1) English	$85K$	$480K$	$365K$	$325K$	$114K$	$501K$
(2) Arabic	$13K$	$144K$	$81K$	$170K$	$25K$	$107K$
(3) Portuguese	$18K$	$131K$	$100K$	$70K$	$15K$	$115K$
(4) Spanish	$20K$	$124K$	$97K$	$88K$	$18K$	$128K$
(5) Turkish	$10K$	$59K$	$41K$	$36K$	$10K$	$56K$

To obtain a comprehensive overview of the purged user's activity, first, we group the tweets based on the language (as detected by Twitter), and we perform separate analyses on each tweet group. Later, we analyze the tweets in English with in details to identify suspended user groups with distinct motivation and manipulation strategies.

5.1 Across Language

We select the tweets written in the five most used languages in our tweet collection for analysis. The five most used languages in our tweet collection are; English, Arabic, Portuguese, Spanish, and Turkish. We list the top hashtags used in the top five languages in Table 3. From the data, we can observe that a few hashtags related to various awards and reality-shows are present in all English, Spanish, and Portuguese. In English tweets, #iHeartAwards, #BestFanArmy, #BTSARMY were used to show support for Korean boy-band *BTS* in the *iHeartAwards* award. The Portuguese hashtags #BBB18, #TheVoiceKids, #MasterChefBR were related to three different television-shows in Brazil. In Spanish, both #MTVHottest, #KCAMexico were associated with two different musical award show. Political hashtags were also present across multiple languages. Four of the top hashtags in Turkish are political, i.e; #Election2018, #PresidentErdogan, #NewEraWithErdogan, #WeWillNotForget (translated from Turkish). The top hashtag in Spanish #DebateINE is related to the Mexican presidential election of 2018. In Portuguese, #BrasilComBolsonaro is related to the 2018 Brazil president election candidate (later elected) *Jair M. Bolsonaro*.

In Table 4, we describe the key statistics of the tweets for each group. Although Arabic is in the sixth position based on user profile language, in the tweet collection, it jumps to the second position, which indicates a high activity rate of Arabic users prior to purge. Also, Arabic has the lowest amount of retweet percentage, meaning they shared more original content. However, one common property present across all five languages is the high usage of hashtags, URLs, and mentions, although English and Spanish tweets contained a considerably higher amount of mentions.

5.2 Topics and User Groups

In this section, we perform an extensive analysis of tweets in the English language to undercover the motivation of purged users tweeting in English. We group users based on their participation in co-related topics and events. We exploit the usage of different hashtags by the purged users in a similar context to group hashtags. Hashtags are used as a conversational tool to indicate participation in a specific event or context. In the recent past, hashtags were instrumental in creating political and social movement [12, 24]. Once we cluster hashtags and identify the related themes, we group users based on their participation in these specific events or topics.

Hashtag Network. In order to identify hashtags used in a similar context, first, we create a hashtag similarity network based on word2vec word-embedding as described in [37]. We train a *word2vec* model using the English language tweets as labeled by Twitter. Before training the model, we pre-process every tweet text, which includes removing non-alphanumeric characters, tokenization, stop-word, and low-frequency word removal (at least 100 occurrences was selected as the threshold). To create the hashtag similarity network, we select the hashtags that appear at least 800 times, and afterward, we use the trained word2vec model to calculate cosine distances between each pair of hashtags. An edge is formed between two hashtags if the distance is less than a selected threshold. We perform community detection on this network using a community detection algorithm [29]. In Figure 5, we show the produced hashtag similarity network where each hashtag community is labeled in a distinct color, and the node size is proportional to occurrence frequency.

Hashtag Clusters. In figure 5, we observe two major hashtag communities, one includes hashtags related to U.S. politics (labeled in green and termed as political), and another consists of hashtags broadly representing few musical award show and musical fan base (labeled in purple and termed as musical). In the *political* hashtag cluster, there are 14 hashtags that include #QAnon, #GreatAwakening, #FakeNews, #ReleaseTheMemo, and #ObamaGate which in general refer to various conspiracy theory ideologies related to U.S. politics. #QAnon is well-positioned at the center of the political hashtag community, which is a far-right conspiracy theory that talks about a supposed secret plot against U.S. President Donald Trump and his supporters by the so-called deep state [7]. #GreatAwakening is another such conspiracy theory based hashtag that often co-occurs with #QAnon. Most of the key hashtags indicate a far-right dominance in this community. In the *musical* hashtag community, there are 16 hashtags, which include #iHeartAwards, #BestFanArmy, #BTSArmy, #MTVHottest, and #BestBoyBand. Most of these hashtags

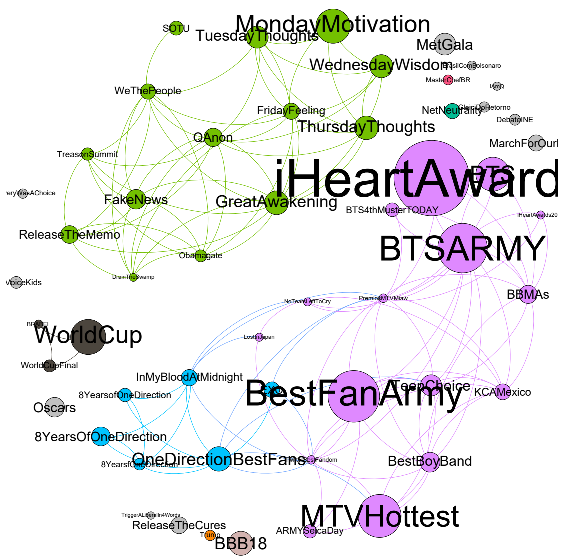

Figure 5: Hashtag similarity network.

in this community are related to several music events. Tweets containing #iHeartAwards talk about the iHeartRadio Music Awards that celebrates the consumption of music through iHeartMedia radio stations [6].

User Groups. Based on the participation in the two aforementioned thematically distinct tweet topics, we identify two user groups. 4,777 and 5,837 numbers of users participated in political and musical topics, consecutively. Among them, only 194 user participated in both. In Figure 6, we show the 25 most popular words used in these two user group's profile descriptions. We use word bi-grams instead of uni-grams to filter out filler words. From the profile description, we can observe that *political*-users heavily expressed their political agenda, which is mostly pro-conservative. On the other hand, *musical*-users explicitly asked for followers, expressed their support for various musical artists, and talked about their personality. We examine the tweet content shared by the users from the two user groups. In Figure 7, we show the top-25 word bi-grams from their tweets, where we observe that both user groups maintained their distinct conversational topics. The musical users mostly talked about various music festivals and music artists. The political users mainly engaged in topics related to U.S. President Donald Trump and few political conspiracy theory, i.e, obamagate, greateawakening, qanon, etc.

Content Dynamics. We inspect how users from the aforementioned groups shared tweet content over-time relating to different events and agendas. In figure 8, we plot the per week usage of the top ten hashtags from the political and musical hashtag cluster. The hashtags from politics are, in general, steady over time with a few exceptions. The exception (i.e., #releasethememo) is occurring in accordance with a particular political event. However, in the musical hashtag cluster, most hashtag usages are time frame limited. For instance, #iHeartAwards, #BestFanArmy, and #BTSArmy had a high presence only from December 2017 to March 2018, till

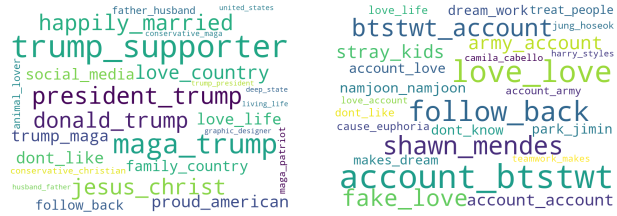

Figure 6: (left) Top-25 word bi-grams from *political* user's description. (right) Top-25 word bi-grams from *musical* user's description.

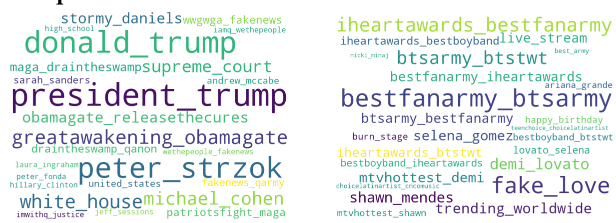

Figure 7: (left) Top-25 word bi-grams from *political* user's tweets. (right) Top-25 word bi-grams from *musical* user's tweets.

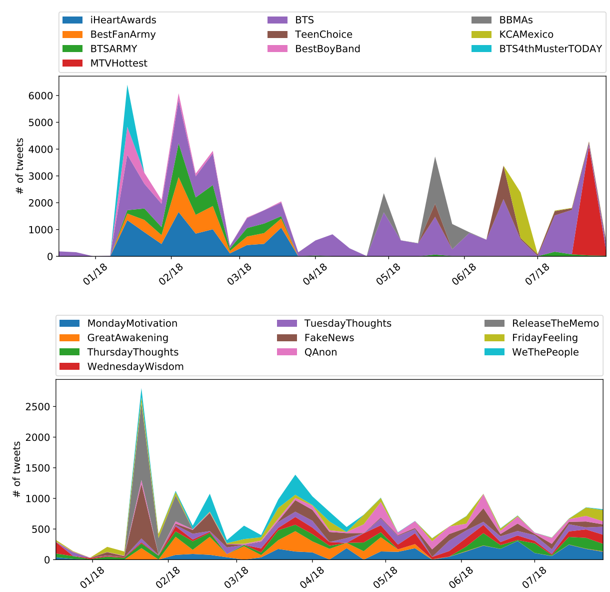

Figure 8: (top) Musical hashtags over time. (bottom) Political hashtags over time.

the iHeartAwards took place, and afterward, they did not reoccur. However, we see such a high presence of musical hashtag in June and July relating to other music award show i.e., #BBMA, #teenchoice. We found that 23% of the accounts that had tweeted about the iHeartAwards later used hashtags related to other music awards. In general, political hashtags are *thematic* in nature, meaning they are related to specific ideologies or themes and has been used persistently to disseminate related content. On the contrary, musical hashtags are largely *episodic*, meaning they were generated ahead of specific events to be used in a confined period.

Table 5: Key Statistics of five Largest Communities Identified from the Retweet Network

Community Name	# of Users	# of Tweets	Language (%)	Hashtags (% of total hashtags)	Retweeted
(1) US Politics	940	37718	English (86.2)	MAGA, GreatAwakening, ReleaseTheCures, FakeNews, WeThePeople (19.8)	ScottPresler
(2) Arabic	347	18623	Arabic (95.2)	Friday, SaudiArabiaEgypt, SaudiaArabia, Eid'sPrizes, WorldCup (5.42)	CiC678
(3) Music-1	334	11107	Portugese (79.34)	InMyBloodAtMidnight, BBB18, MTVHottest, WangoTango, NoTearsLeftToCry (13.7)	arianagrandebr
(4) Music-2	323	16378	Spanish (72.4)	iHeartAwards, BTSARMY, BestFanArmy, BTS, TM88xBTS (18.1)	JIMINILOVE95
(5) Music-3	284	11701	English (76.4)	iHeartAwards, BestFanArmy, BTSARMY, BTS, BTS4thMusterTODAY (29.2)	jhopesgalaxy

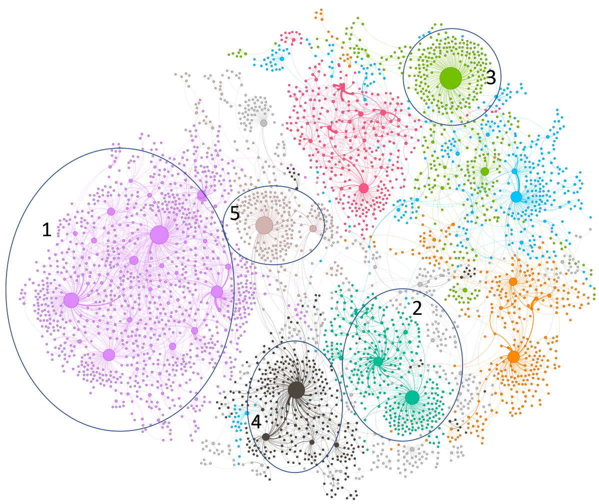

Figure 9: Purged User's Retweet Graph (Ten largest clusters are shown). Each node represents a purged user and an edge implies one purged user retweeted another purged user. Here node size is proportional to in-degree (no of times retweeted), and similar color implies same cluster.

6 INTERACTION AND COMMUNITIES

In this section, we examine the interaction among the purged users and detect communities based on the interaction to identify group-level characteristics. We use retweet activity among the purged users as an indication of interaction as it portrays explicit engagement. Based on the retweet activities that occurred between purged users, we form a retweet graph. In this graph, each node represents a purged user, and an edge implies one purged user retweeted another purged user. We use the modularity based community detection algorithm [29] to identify communities in this retweet graph. In Figure 9, we show the 10 largest communities identified from the retweet graph. We use these interaction-based communities to explore distinct group-level characteristics.

We analyze these communities across multiple feature dimensions, such as the most used language, hashtags, and retweeted users. We focus on the top five largest communities for an in-depth analysis, which consists of in total 2228 users. In Table 5, we describe the summary statistics of the five largest communities. From the data presented in the table, it is evident that each of these five communities represents distinct communities. None of the two communities have similar values across all features. For example, there are two communities with English as the most used language, but their hashtag usage is different. We manually name these five communities based on their content and language usage. In the largest community named US-Politics, we observe a high volume of tweets made in English, and the hashtags are related to political propaganda. In the Arabic community, we see a dense group of Arabic users who talked about regional issues. The other three communities have shared content in three different languages, but in general, focused on music issues.

It is evident from our analysis that the purged users had interaction and formed communities with heterogeneous characteristics. It can be inferred that there exist other similar interaction communities in each language group. However, as we could collect only 1% sampled tweet data, our communities are mostly sparse. Our observed communities can be perceived as a sampled representation of a broader community. A retweet network formed on a particular topic of discourse could be utilized to quantify the level of interaction among purged users better.

7 DISCUSSION & CONCLUSION

Limitations. Our scope of the study was mainly limited due to the unavailability of large scale datasets about purged users and their previous tweets. In our study, we analyzed 2.4M purged users, where else Twitter reportedly suspended 70M users in the three months of 2018 [4]. A more extensive purged user set would provide a precise characterization of suspended users. Moreover, Twitter restricts access to past tweets of a user after the account is suspended. A comprehensive tweet collection of purged users would lead to a better understanding of their role and agenda over-time. Due to the sparsity in the collected tweets, the exact reason for suspension for a particular account could not be identified. However, our tweet collection was adequate to obtain a comprehensive characterization of malicious activity on Twitter. Again, our performed study related to suspension is limited to Twitter. Similar malice and the responsive suspension is taking place in other social platforms (i.e., Reddit). Due to data collection limitations, we could not gather such suspension related data in a corresponding time frame.

Future work. Our research opens up many future directions to detect, examine, and prevent abuse of social media. In this study, we focused on content shared on Twitter. However, many external contents are shared using URL embedding. Further investigation on the shared content is mandatory, as Twitter is often used as a *fishing* medium. A similar analysis can be carried out on other social media as well. Also, cross-platform analysis of malice can shed new light on the inner workings of such campaigns. Our demonstrated approach has the potential to be routinely used to catalog similar abuse of social media platforms that could be useful for political scientists, social scientists, and financial analysts, among many others. As abusive and malicious activities in social media are continuously evolving, the regular analysis would lead to a better detection and prevention methodology. In future, our goal is to develop a suspension prediction system based on our observed characteristics and behavior of the purged users.

Conclusion. Twitter purge is a significant event on which very little is known. This paper shows a systematic approach to identify a set of purged users and perform retrospective analysis to uncover detected abuse attempt. Our performed study has several major implications. Firstly, there are a significant number of purged users who survived on Twitter for a long time, which can be interpreted in several ways. Either these users suddenly turned malicious, or they were malicious all along, but Twitter was not able to detect them. We also identified several dormant user groups who were created in clusters in an automated way, which implies there might be other such inactive user groups reserved as a future abuse tool. Secondly and alarmingly, many a purged user established a high number of social relationships, which provided them the opportunity to propagate content towards a large portion of users on Twitter. Also, politically motivated users were successful in spreading political conspiracies for a long time. However, it is evident from our study that Twitter has been preemptively performing routine clean-up to remove such malicious and abusive accounts (in cases potential) to create a safer online environment. Further regular research on similar suspension would shed new light on the evolution of malice to evade detection.

REFERENCES

[1] 2017. Facebook Election 2106 Update. https://about.fb.com/news/2017/09/information-operations-update/
[2] 2018. How Twitter is fighting spam and malicious automation. https://blog.twitter.com/official/en_us/topics/company/2018/how-twitter-is-fighting-spam-and-malicious-automation.html
[3] 2018. Twitter Election 2016 Update. https://blog.twitter.com/en_us/topics/company/2018/2016-election-update.html
[4] 2018. Twitter is sweeping out accounts like never before. https://www.washingtonpost.com/technology/2018/07/06/twitter-is-sweeping-out-fake-accounts-like-never-before-putting-user-growth-risk/
[5] 2019. Mueller report. https://www.justice.gov/storage/report.pdf
[6] 2020. iHeartRadio Music Awards. https://en.wikipedia.org/wiki/IHeartRadio_Music_Awards
[7] 2020. #Qanon. https://en.wikipedia.org/wiki/QAnon
[8] 2020. Twitter terms of services. https://twitter.com/en/tos
[9] Noor Abu-El-Rub and Abdullah Mueen. 2019. BotCamp: Bot-driven Interactions in Social Campaigns. In *The World Wide Web Conference*. 2529–2535.
[10] Hunt Allcott and Matthew Gentzkow. 2017. Social media and fake news in the 2016 election. *Journal of economic perspectives* 31, 2 (2017), 211–36.
[11] Abdullah Almaatouq, Ahmad Alabdulkareem, Mariam Nouh, Erez Shmueli, Mansour Alsaleh, Vivek K Singh, Abdulrahman Alarifi, Anas Alfaris, and Alex Sandy Pentland. 2014. Twitter: who gets caught? observed trends in social micro-blogging spam. In *Proceedings of the 2014 ACM conference on Web science*. ACM, 33–41.
[12] Ahmer Arif, Leo Graiden Stewart, and Kate Starbird. 2018. Acting the part: Examining information operations within# BlackLivesMatter discourse. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 20.
[13] Adam Badawy, Emilio Ferrara, and Kristina Lerman. 2018. Analyzing the digital traces of political manipulation: The 2016 russian interference twitter campaign. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE, 258–265.
[14] Alessandro Bessi and Emilio Ferrara. 2016. Social bots distort the 2016 US Presidential election online discussion. *First Monday* 21, 11-7 (2016).
[15] Qiang Cao, Xiaowei Yang, Jieqi Yu, and Christopher Palow. 2014. Uncovering large groups of active malicious accounts in online social networks. In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*. 477–488.
[16] Eshwar Chandrasekharan, Mattia Samory, Shagun Jhaver, Hunter Charvat, Amy Bruckman, Cliff Lampe, Jacob Eisenstein, and Eric Gilbert. 2018. The Internet's Hidden Rules: An Empirical Study of Reddit Norm Violations at Micro, Meso, and Macro Scales. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 32.
[17] Nikan Chavoshi, Hossein Hamooni, and Abdullah Mueen. 2016. DeBot: Twitter Bot Detection via Warped Correlation.. In *ICDM*. 817–822.
[18] Michael D Conover, Jacob Ratkiewicz, Matthew Francisco, Bruno Gonçalves, Filippo Menczer, and Alessandro Flammini. 2011. Political polarization on twitter. In *Fifth international AAAI conference on weblogs and social media*.
[19] Clayton Allen Davis, Onur Varol, Emilio Ferrara, Alessandro Flammini, and Filippo Menczer. 2016. Botornot: A system to evaluate social bots. In *Proceedings of the 25th international conference companion on world wide web*. 273–274.
[20] Munmun De Choudhury. 2011. Tie formation on twitter: Homophily and structure of egocentric networks. In *2011 IEEE third international conference on privacy, security, risk and trust and 2011 IEEE third international conference on social computing*. IEEE, 465–470.
[21] Juan Echeverria and Shi Zhou. 2017. Discovery, Retrieval, and Analysis of the'Star Wars' Botnet in Twitter. In *Proceedings of the 2017 IEEE/ACM international conference on advances in social networks analysis and mining 2017*. 1–8.
[22] Emilio Ferrara, Onur Varol, Clayton Davis, Filippo Menczer, and Alessandro Flammini. 2016. The rise of social bots. *Commun. ACM* 59, 7 (2016), 96–104.
[23] Jane Im, Eshwar Chandrasekharan, Jackson Sargent, Paige Lighthammer, Taylor Denby, Ankit Bhargava, Libby Hemphill, David Jurgens, and Eric Gilbert. 2019. Still out there: Modeling and Identifying Russian Troll Accounts on Twitter. *arXiv preprint arXiv:1901.11162* (2019).
[24] Sarah J Jackson, Moya Bailey, and Brooke Foucault Welles. 2018. # GirlsLikeUs: Trans advocacy and community building online. *New Media & Society* 20, 5 (2018), 1868–1888.
[25] Zhiwei Jin, Juan Cao, Han Guo, Yongdong Zhang, Yu Wang, and Jiebo Luo. 2017. Detection and analysis of 2016 us presidential election related rumors on twitter. In *International conference on social computing, behavioral-cultural modeling and prediction and behavior representation in modeling and simulation*. Springer, 14–24.
[26] Huyen Le, GR Boynton, Zubair Shafiq, and Padmini Srinivasan. 2019. A Post-mortem of Suspended Twitter Accounts in the 2016 US Presidential Election. (2019).
[27] Po-Ching Lin and Po-Min Huang. 2013. A study of effective features for detecting long-surviving Twitter spam accounts. In *2013 15th International Conference on Advanced Communications Technology (ICACT)*. IEEE, 841–846.
[28] Amanda Minnich, Nikan Chavoshi, Danai Koutra, and Abdullah Mueen. 2017. BotWalk: Efficient adaptive exploration of Twitter bot networks. In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*. 467–474.
[29] Mark EJ Newman. 2006. Modularity and community structure in networks. *Proceedings of the national academy of sciences* 103, 23 (2006), 8577–8582.
[30] Jacob Ratkiewicz, Michael D Conover, Mark Meiss, Bruno Gonçalves, Alessandro Flammini, and Filippo Menczer Menczer. 2011. Detecting and tracking political abuse in social media. In *Fifth international AAAI conference on weblogs and social media*.
[31] Caitlin M Rivers and Bryan L Lewis. 2014. Ethical research standards in a world of big data. *F1000Research* 3 (2014).
[32] Kurt Thomas, Chris Grier, Dawn Song, and Vern Paxson. 2011. Suspended accounts in retrospect: an analysis of twitter spam. In *Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference*. ACM, 243–258.
[33] Onur Varol, Emilio Ferrara, Clayton A Davis, Filippo Menczer, and Alessandro Flammini. 2017. Online human-bot interactions: Detection, estimation, and characterization. In *Eleventh international AAAI conference on web and social media*.
[34] Svitlana Volkova and Eric Bell. 2017. Identifying effective signals to predict deleted and suspended accounts on twitter across languages. In *Eleventh International AAAI Conference on Web and Social Media*.
[35] Wei Wei, Kenneth Joseph, Huan Liu, and Kathleen M Carley. 2015. The fragility of Twitter social networks against suspended users. In *2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE, 9–16.
[36] Chao Yang, Robert Harkreader, Jialong Zhang, Seungwon Shin, and Guofei Gu. 2012. Analyzing spammers' social networks for fun and profit: a case study of cyber criminal ecosystem on twitter. In *Proceedings of the 21st international conference on World Wide Web*. 71–80.
[37] Savvas Zannettou, Tristan Caulfield, William Setzer, Michael Sirivianos, Gianluca Stringhini, and Jeremy Blackburn. 2019. Who Let The Trolls Out?: Towards Understanding State-Sponsored Trolls. In *Proceedings of the 10th ACM Conference on Web Science*. ACM, 353–362.