The Fragility of Twitter Social Networks Against Suspended Users

Wei Wei  
Carnegie Mellon University  
Pittsburgh, PA, 15213  
weiwei@cs.cmu.edu

Kenneth Joseph  
Carnegie Mellon University  
Pittsburgh, PA, 15213  
kjoseph@cs.cmu.edu

Huan Liu  
Arizona State University  
Tempe, AZ 85281  
huan.liu@asu.edu

Kathleen M. Carley  
Carnegie Mellon University  
Pittsburgh, PA, 15213  
kathleen.carley@cs.cmu.edu

Abstract—Social media is rapidly becoming one of the mediums for understanding the cultural pulse of a region; i.e., for identifying what the population is concerned with and what kind of help is needed in a crisis. To assess this cultural pulse it is critical to have an accurate assessment of who is saying what in social media. However, social media is also the home of malicious users engaged in disruptive, disingenuous, and potentially illegal activity. A range of users, both human and non-human, carry out such social cyber-attacks. We ask, to what extent does the presence or absence of such users influence our ability to assess the cultural pulse of a region? We conduct a series of experiments to analyze the fragility of social network assessments based on Twitter data by comparing changes in both the structural and content results when suspended users are left in and taken out. Since an account can be suspended in Twitter for various reasons including spamming or spreading ideas that can lead to extremism or terrorism, we separately assess the impacts of removing apparent spam bots and apparent extremists. Experimental results demonstrate that Twitter based network structures and content are unstable, and can be highly impacted by the removal of suspended users. Further, the results exhibit regional and temporal characteristics that can be triggered by important social events. We also provides guidance on the differential impact of different types of potentially suspend-able users.

I. INTRODUCTION

Disingenuous users are everywhere on today’s social media platforms. The actions of these users, both human and bot, affect genuine users on these social networking sites in a variety of ways. Human “trolls”, individuals who seek out others with the intent of annoying or offending them, can cause irreparable harm to one’s self-confidence and self-concept [1]. Spam bots can clog the network of information, providing useless or false information to millions of possible unsuspecting users. Scam artists can engage in social engineering to extort money from unsuspecting users, and hackers can leverage weaknesses in platform security measures and user passwords to take over user accounts or enact other malicious behaviors on a site.

The actions of these users have not gone unnoticed, either in the research community [2]–[4] or within the social media industry [5]. Twitter has been taking positive actions to suspend users who are recognized to be malicious. While important questions exist with respect to how and if various malicious behaviors should be restricted by social media platforms, an indisputable point is that the majority of the behaviors engaged in by these users are potentially disruptive social behaviors. From hackers’ use of social engineering to intricate manipulations of social relationships by scam artists, these actions affect the social environment of users, potentially resulting in misinformation, disinformation fads, social chaos, and so on. In response to this threat, increasingly research has been directed at identifying such malicious users and removing or suspending such users [6].

The actions of malicious users can harm both the users of social and those who want to use social media to understand the cultural pulse of a region. The suspending of such users, even a fraction of them, has the potential to impact conclusions drawn about what the population of Twitter users is saying, and the way they are using Twitter to effect social action. Little is known about the cumulative impact of such users in general, or suspended users in particular, on the volume of information in social media. Illustrative examples exist showing that one type of malicious user can have massive consequences; e.g., bots have been used to coordinate hashtag campaigns and so influence trending topics on Twitter [7]. While numerous examples point to the negative impact of these malicious users, there has not been a systematic analysis of the impact of these users on the overall social media landscape. Nor is there an understanding of the relative impact of different kinds of malicious users on the behavior observed in social media.

Twitter is increasingly used to understand the cultural landscape and so who are the key users talking about specific topics, responding to key events, providing early warnings of crises, indentifying the key topics about which people are concerned, and the implication of such chatter on activities ranging from marketing, to disaster response to state stability [8]. To answer such questions, network analysis is often used. Unfortunately, network metrics are relatively sensitive to erroneous data [9]. For example, Borgatti et al. [10] showed that when “fake” nodes are added or “true” nodes are dropped at random to/from a particular network, the likelihood that one is able to recover the “top” nodes in the “true” network drops precipitously. Estimates suggest that half the Twitter accounts created in 2014 were suspended 1. Since not all bots are suspended, this action results in fake nodes (bots) still existing and ture nodes (malicious humans) being dropped. Accurate estimations of what non-suspended human users are tweeting and to whom they are doing so are needed.

In this article, we study suspended users and their affect on analyses performed on the social structure of a particular

---

1Only 11% of new twitter users in 2012 are still tweeting http://blogs.wsj.com/digits/2014/03/21/new-report-spotlights-twitters-retention-problem
social media platform, Twitter. We note that suspended users are a subset of all malicious users in social media; hence, our results may be underestimating the impact of such users. Our study focuses on three key points. First, we seek to understand the impact that removing these suspended users has on the structure of the social networks that can be extracted via Twitter when two users mention each other. Second, we consider how these suspended users impact our understanding of the topical focus of a collection of users. Finally, we perform an unsupervised clustering analysis on the set of all suspended users in addition to a subset of non-suspended users to better understand the different types of suspended users in our data and the differing roles they might play in the social environment.

II. RELATED WORK

A. Network Analysis

The quantitative study of the patterns of relations among users has been used for the past 70 years to understand and predict human behavior and socio-cultural activities [11], [12]. This area, referred to as social network analysis, dynamic network analysis and network science is concerned with assessing how the patterns of connections among entities constrain and enable behavior, and how different patterns effect different socio-cultural outcomes. While much of the early work focused on interactions among small groups of humans (<50), more recent efforts have focused on the development of scalable methods, often for reasons associated with social media analysis [13]. Many metrics in this area are focused on identifying those nodes that have disproportionate potential influence or power in the overall network; e.g., degree centrality, closeness centrality and betweenness centrality [14]. When the network changes overtime, dynamic network metrics provide additional insights as to how network change impacts individuals [15].

From a social media perspective, network analytics have been used to, for example, identify communities [16] and better understand the relationship between social and topical structures [17]. Network analytics are also increasingly used to support spam detection [18] and fraud detection [19]. Such research has demonstrated that networks in social media, and specifically in Twitter, can be much larger and take on different forms than networks in the real world. Case studies often report having to clean the data significantly to remove bots and malicious users [20]. Such works suggests the possible negative impact of spam on network analyses, but there has been no systematic assessment. We utilize standard metrics and assess how the results vary as suspended users are removed.

B. Spam Detection

Spammers are users in online social networks whose purpose is to distribute advertisements, fraudulent information, or general chaotic information. Various techniques have been developed to detect spammers in an automatic way and to disable the accounts. There are two main categories of these techniques. The first uses network structures and interaction processes to detect fraud and spam [21]. For example, Bolton et al. [19] rely on the fact that most spammers designed computer software to distribute their contents. This software can post information in a speed that is much faster than a human, making the statistical distribution of inter-arrival time of the behaviors look abnormal. For example, sending 10,000 messages in one second definitely does seem to be a human behavior. Another types of spammer detection builds on the fact that spammer contents have much narrow topic sections than normal contents [22]. Using a topic model can effectively pick up users who constantly post spammer topics.

A host of scholars have studied spam on Twitter specifically. Several works have recently considered the problem of determining whether or not particular tweets or users were spam-based [3], [4], [23], [24]. These approaches tend to rely on established patterns of spammers on Twitter, such as the content they utilize, their (lack of) network connections and the prevalence of URLs in their tweets [2]. Thomas and colleagues [4] spent ten months infiltrating the underground marketplace for fraudulent accounts on Twitter and other social media sites, exposing the intricacies of the spam and bot marketplace. Cumulatively this work demonstrates that spammers, and spam bots in particular, have very different profiles than normal users in the way they construct and use tweets.

While Twitter clearly lays out the reasons that one can be suspended online, the reason why any particular account is suspended is not known to the analyst. Prior research provides guidance on how we might differentiate types of suspended users and the extent to which these individuals might act as extremists. Our work differs from these previous efforts in spam detection in that we are not concerned with detecting spammers, but in assessing the impact of such users on our understanding of who is influential in social media and in what is being said on social media; specifically, we are assessing the impact of the removal of suspended users from the holistic network and topical analytic results commonly performed on data derived from Twitter.

III. DATA AND METHODS

Our dataset contains approximately 87M tweets from April 2010 to November 2013 sent by roughly 3.9M users. Tweets in the dataset are collected via a combination of geo-spatial bounding boxes around the fifteen countries of interest for this analysis (listed in Table V, as well as keyword and user-based searches performed on the Streaming API. We choose this area of the world during this particular time period as social movements and protests occurred frequently during this time period. As we will discuss, this social unrest corresponded to reports of internet censorship and also to the development of extremists and militant actors who were active on Twitter and using it in a way that would later lead their accounts to be suspended. From our dataset, we extract all users whose accounts were suspended by Twitter as of November 2013. Table I provides a summary of our data set.

Our analysis is geared towards better understanding how suspended users affect network and topical analyses. In order to understand how removing suspended users alters network analysis results, we analyze the mention network in our data. In the mention network, a directed link is formed between two users A and B if A includes the username of B in their tweet, prepended with an “@” sign (e.g. “Hey B, what’s going on?”). We create snapshots of the networks for each month, for each country of interest. A tweet is determined to be relevant to a
particular, we would expect that spammers make up a large
portion of suspended users. However, in light of recent events
with ISIS [29] and in the context of reports during the Arab
Spring of accounts being suspended\(^2\), there are also many
accounts that do not explicitly emit spam that are suspended.
We perform a clustering analysis of the 9,603 suspended
accounts that contributed at least one “document” to the LDA
as described above in order to explore the different types of
suspended users.

The features used for the clustering consists of ten types
of information totaling 209 features. We use the Twitter meta-
data and tweet text to construct the first nine features, described
in Table II. These features are based on known properties of
spammers discussed in the previous work mentioned above.
The re-tweet, hashtag, mention, and URL ratios represent the
proportion of a user’s tweets that were re-tweets, or that con-
tained hashtags, mention or URLs, respectively. The followers
ratio and number of followers are indicative of the fact that
most spam users have relatively few followers, and in any event
are likely to have far fewer followers than they themselves
follow. The cosine similarity of a user’s tweets indicate the fact
that average users tend to focus on a very small, particular set
of topics [28], while this may not be true of spammers. Finally,
the number of days active indicates the fact that most spam
users are caught relatively quickly by Twitter, and are thus
active for fairly short periods of time.

In addition to these nine features, we also utilize informa-
tion from the output of the LDA on the topics that users tend
to focus on. The LDA we ran, 200 topics, resulted in a topic
distribution for each document \(\theta\) of size 200. The strength of
\(\theta\) represents how likely the user is to choose the specific topic,
i.e. the correlation between the user and topic. For each user
we use the sum of all the topic strengths over all the tweets
that particular user sent. We use these 200 topics plus the six
features above, as the features for each of our users as the
input to the clustering algorithm.

twitter-says-access-to-service-in-egypt-is-blocked-amid-riots-with-police

\begin{table}[h]
\centering
\begin{tabular}{|c|c|}
\hline
Metric Name & Description \\
\hline
Log Num. Tweets & \(\log(|T|)\) \\
\hline
Cosine Sim. Tweet Text & \(\sum_{t \in T} v_{t,g} \cdot v_{g,t} / (\sqrt{\sum_{t \in T} v_{t,g}^2} \cdot \sqrt{\sum_{g \in G} v_{g,t}^2})\) \\
\hline
RT Ratio & \(\log(\frac{|RTs|}{|T|})\) \\
\hline
Follower Ratio & \(\log(\frac{|Friends|}{|Followers|})\) \\
\hline
Hashtag Ratio & \(\log(\frac{|Hashtags|}{|T|})\) \\
\hline
Mention Ratio & \(\log(\frac{|Mentions|}{|T|})\) \\
\hline
URL Ratio & \(\log(\frac{|URLs|}{|T|})\) \\
\hline
Num. Days Active & Days between first and last tweets in dataset \\
\hline
\end{tabular}
\caption{Custering features captured by Twitter Meta-data}
\end{table}

After constructing the networks for each country, for each
month, we calculate several network metrics on the network
both with and without the suspended users. Specifically, we
consider the number of nodes in the network, average degree
centrality, average closeness centrality and the average
clustering coefficient of the network. As these are traditional
network measures, we do not further describe them. For more
information on these metrics in directed, weighted graphs, we
refer the reader to [25].

In order to understand the effects of suspended users
on the topical focus results, we consider both hashtags and
topic-model based conceptualizations of content. With respect
to the latter, a Latent Dirichlet Allocation (LDA) [26] was
conducted to identify topics. In an LDA model, each tweet has
a multinomial distribution over topics, \(\theta\). To reduce difficulties
posed by the use of common bi-grams and use of common
words with similar meanings we first ran a thesauri to clean the
data. Finally, to reduce the distraction caused by high levels
of non-sense words in Twitter, we removed from each tweet
those words that did not occur more than once.

It is well-known that LDA provides noisy topic distri-
butions for short-texts [27]. One of the primary issues with
applying LDA to short texts, like tweets, is that the assumption
that the text is drawn from a mixture of topics is frequently
violated - short texts often focus on only one concept. To
address this issue, scholars often aggregate all tweets by a
user into a single document. As users tend to focus on a few,
important difference. To address topical drift over time, we
consider as a document all tweets sent by the same user within
a 3 month span. All documents are also restricted to having
to contain at least 300 unique tokens and at least 3 tweets. Thus,
certain users may be responsible for multiple documents in
the LDA, and many users will not be represented at all in the
LDA. Still, we can back-propagate decisions on the most likely
topic for a particular tweet by matching terms in any tweet to
our topic distributions after the LDA has been run.

As noted, different types of suspended users exist. In
particular, we would expect that spammers make up a large

\begin{table}[h]
\centering
\begin{tabular}{|c|c|}
\hline
Country & Time spread \\
\hline
Bahrain, Qatar, Libya, Algeria, Tunisia, Oman, Lebanon, Morocco, Jordan, Saudi Arabia, Kuwait, Syria, Iraq, UAE, Egypt, Yemen, Iran & Apr, 2010 - Nov, 2013 \\
\hline
Num. tweets & 72,722,180 \\
Num. total users & 3,877,141 \\
Num. suspended users & 278,753 \\
Num. hashtags & 1,230,974 \\
Num. LDA topics & 200 \\
\hline
\end{tabular}
\caption{Basic statistics for the data set}
\end{table}

country if the tweet contains a geo-tag pointing to a location
within the country, or if the text of the tweet contains the
country’s name in English or Arabic, or if the text of the tweet
contains any of the five major cities of the country in English or
Arabic. Consequently, a tweet in our dataset may be considered
to be relevant to zero or more countries.
relatively sparse and thus could corrupt a holistic view of the data. We observe that Algeria consistently had the highest level of suspended users - on average, over 25% of the users in any given month in the mention network relevant to Algeria were suspended. This covers the time frame when Algeria has a long standing protests. Other countries that have high proportion of suspended users include Kuwait, Morocco and UAE. Second, there is a very obvious increase in the number of suspended users across all countries around Oct of 2012. This sudden change might well be attributed to the well-publicized attack in Benghazi, Libya in September of 2012, where the U.S. Embassy was attacked by Islamic militants [30]. These attacks led to an increase in American interest in the entire Arab region, both politically as well as socially from the average American citizen.

In addition to the changes in the number of users in the network, one other interesting thing to look at is the robustness of the networks against the suspended users. As noted above, three network metrics are used in this analysis: 1) degree centrality, which is a measurement on the average number of direct neighbors of a user in the network, 2) closeness, a measurement on the average shortest path distance from a particular user to the rest of the network members, and 3) clustering coefficient, which is the proportion of the possible triplets that are formed into a closed triangle in the network.

In Figure 2, we plot the change in average degree of users in a network over time on each country. Here we use a two color scheme to define a positive (blue) or negative (red) change in the value. A white cell indicates little or no change. Algeria has the most suspended users and the degree centrality results are the most fragile. Further, as expected, when suspended actors were removed the average degree centrality decreased. This impact is most pronounced in September and October of 2012, which covers the time when the Benghazi consulate was attacked. Countries that were most vulnerable to the suspended users at this time were Algeria, Jordan, Morocco and Tunisia. The effect of this event begin to fade after October. The fact that this change becomes positive illustrates that the suspended users have fewer interactions with the rest of the network during this period. The elimination of these suspended users increases the connectivity, on average, of the network.

Figure 3 illustrates the change of closeness centrality in the networks over time and country in a manner similar to the previous plot. Since closeness is affected by not only the direct neighbors but also long distance network structures, its value is more sensitive to the removal of suspended users. This can be validated in the diverse change in patterns that appears in most time steps across all countries.

The majority of these changes are moderate. However, there are two regions of the plot that have a significant level of change. One region is concentrated on the country of Algeria again, which is consists of mostly the negative changes. Perhaps more interestingly, huge positive increases in closeness centrality occur in 2013 in Kuwait, Oman and Qatar. The largest change goes up to 100%, meaning that average shortest path in the networks are decreased by a factor of 2 after the removal of suspended users. This means that suspended users are in the peripheral of the social networks. Removing the users will not impact the shortest paths of the normal users but will save the additional path length that extended to the peripheral area where the suspended users are located in. In other word, the suspended users are being excluded from the main mention network component and are strongly contributing to the level of connectivity. It is important to note, though, that even moderate changes in average metrics may have significant impacts in the relative ranking of nodes.

We look at the change of clustering coefficient in Figure 4. Overall, the suspended users made little impacts to the local clustering structures of the network, further suggesting their existence on the periphery. The only exception is Algeria, where we see a nearly 80% decreased in the clustering coefficient when suspended users are removed. The rest of the changes remain mostly positive, which means the elimination of suspended users make the local structures of the mention network more cohesive. After Oct 2012, changes occur in a much more positive way than those in the previous time steps, e.g., see Morocco, Kuwait and UAE. This suggests that the suspended users are clustered together. Deleting suspended users has little impact on the local structure of active, “normal” users but decreases the denominator of the clustering coefficient, making the overall metric increase.

Finally, we evaluate the effects of suspended users on the distribution of “top” actors in our dataset, as measured by the metrics utilized here. Figure 5 reports an analysis of the change in the top-k non-suspended agents rank based on their individual network metrics. For each country and each month, the network metrics are generated for each agent in each network before and after suspended users are removed. Non-suspended users are ranked in both networks and for a given k. We then determine a ratio r which defines the number of common users in the top k list in the networks before and after the suspended users are removed and divided this value by the total number of non-suspended users. We generate such a ratio r on the network generated by each country/time pair and aggregate them over time, providing 95% Confidence Intervals across all time points.

Figure 5 shows that when k is small (e.g. 10), the chance that we will see common agents in both of the ranked list is low. This suggests that analysis of the top agents in a network are highly vulnerable to the addition or removal of suspended users. Thus, the perceived importance of actors can be affected significantly by the analysts’ choice of whether or not to keep...
suspended actors, and by transition the point at which the analyst collects the data (i.e. in real-time or post-hoc via the REST API). There is also significant variance across countries, with the ranking of the top agents being more robust in those countries that are undergoing less civil unrest e.g., Morocco. We do note, however, that as $k$ increases, the percentage of common agents begin to increase and eventually becomes close to 1. Thus, one recommendation may be to consider larger $k$ when attempting to discern between important and non-important actors in Twitter networks.

### B. Content Impacts of Suspended Users

Apart from structural impacts, suspended users also may impact the observed content. We use two indicators to detect content changes: the use of hashtags and the LDA topic concentration of tweets (i.e. $\theta$). To evaluate the impact of suspended users, we ranked the hashtags and LDA topics by an importance factor $S$. For hashtags, the importance factor $S_H$ is simply defined to be the number of times that this particular hashtag appears in the tweets. For LDA topics, the importance factor $S_T$ is defined to be the accumulation of document topic concentration $\theta_i$ of this particular topic $i$ across all documents. For hashtags, we use $RH = \{h_1, h_2, ..., h_{L-1}, h_L\}$ to denote the rank list of hashtag on all the users while using $RH^-$ to denote the rank list of only the active users on hashtag. Here $h_i$ has a higher or equal importance factor than $h_j$ if $j > i$. Similarly, we can define $RT = \{t_1, t_2, ..., t_{K-1}, t_K\}$ to be the topic rank list on all the users and $RT^-$ to be the topic rank list on only the active users.

Table III and Table IV shows some of the top hashtags and topics used by suspended users along with their importance factors. Since data collection is focused on the Arab world and the MENA region, a large collection of tweets are in Arabic, which results in the existence Arabic terms in both hashtags and LDA topic terms. We translated Arabic terms into English using Google translate and annotate these terms with a star ("*"). Table III show that suspended users refer to the hashtags of a host of nations, as well as to the CIA and CNN. Table III shows that the focus of our data additionally centered on topics such as “pain” and “killing”.

To measure the impact of the suspended users on mid and low ranked hashtags and topics, we conducted a numerical analysis on the hashtags and topic terms found in the top $q\%$ of rank list and see how much they overlap. Taking hashtag
for example, for a given $q$ we obtain a subset of the rank list on both all the users $RH_{q+L} = \{h_1, h_2, \ldots, h_{q+L} \}$ and only the active users $RH^-_{q+L} = \{h_1^-, h_2^-, \ldots, h_{q+L}^- \}$. The matching score $mH(q)$ is calculated to be the number of elements in the intersection of two subsets divided by the total length of the set $L$, which is defined in Equation 1. If the elements in $RH_{q+L}^-$ are exactly the same as the ones found in $RH^-_{q+L}$, $mH_q(q) = q \ast L/L = q$. Otherwise, $mH_q(q) < q$. Similarly, one can define the corresponding matching score for topic $mT(q)$.

$$mH(q) = \frac{|RH_{q+L} \cap RH^-_{q+L}|}{L}$$

(1)

We vary $q$ from 0% to 100% to see how the matching score changes. Figure 6 shows both the results of hashtags and the LDA topics. The horizontal axis is $q$ while vertical axis is the matching score. A reference line with a slope of 1 is plotted as well. If the suspended users had no affect on the ranking list, the ranking list before and after suspended users would align with the reference line. The more the matching score diverges from the reference line, the greater the impact of suspended users on the top $q\%$ of the content.

We see that suspended users have little impact on topic concentrations. The data generally remain close to the reference line with albeit small deviations. Those deviations appear when $q$ is between 25% to 60%, meaning that suspended users impact the relative standing of moderately popular topics (rather than high or low popular topics). The changes in hashtags, however, is more significant than those found in the topics. Similar to the changes in topics, the divergence does not appear until $q$ reaches around 25%. The difference between the reference line and the data point begin to widen after $q$ reaches around 45%. We also observe a unique pattern of hashtag usage divergence. The gaps between the reference line and the matching score are separated into several different major gaps across the range of $q$.

The existence of these gaps indicates that there may exist subgroups of hashtags that are frequently used mainly by suspended users. The suspended users lead the use of these hashtags in the subgroups but never impact hashtags outside the subgroup. When $q$ reaches in the middle of the subgroup, the difference begin to show up. However, if $q$ reaches the two ending points of the subgroups, the difference will resume to 0 and matching score will adhere to the reference line.

### C. Identifying Types of Suspended Users

In this section, we perform a cluster analysis on the 9,603 suspended users that contributed a “document” to the LDA described above. To facilitate our understanding of the data, we also include in the clustering a set of approximately 6000 randomly sampled users who were not suspended. Figure 7 shows the distribution of the nine text-based and metadata-based metrics for suspended and non-suspended users. All metrics have been centered and scaled by two standard deviations, which allows for an easier comparison of the more extreme ends of the distribution [31]. Our data generally fits with expected differences between suspended and non-suspended users. Namely, non-suspended users are active for far fewer days, have fewer followers relative to the number of users they follow, use more tweets and more hashtags, use fewer mentions and fewer retweets, and have far less cohesiveness in the text of their tweets.

Having observed differences between suspended and non-suspended users, we now turn to a cluster analysis of this same set of approximately 15K users. To perform the clustering, we utilize the mclust package in R [32] to perform Gaussian

![Image](image-url)
Mixture Modeling. We select the best number of clusters via comparison of model BICs. Because of the volume of data studied, we only consider the possibility of up to 9 clusters in our data. The model selection process suggested that indeed, 9 clusters was the most appropriate number of clusters for the data studied. While this may raise concerns that even greater numbers of clusters are necessary, we leave this to future work and concern ourselves here with exploratory results.

The numbers of suspended and non-suspended users in each cluster are displayed in Table V. As is clear, clusters 1, 3, 6, 8, and 9 all contained virtually no active users. Not coincidentally, these clusters qualitatively appeared to contain almost entirely spammers. We refer to the suspended users in these five clusters as potential bots; whereas, we refer to suspended actors in clusters 2, 4, 5, and 7 as potential militants or smartbots. Table VI, shows a representative synthetic tweet from the different clusters. Clusters 1, 3, 6, 8, and 9 all show highly similar, spam like tweets, which contain one or more links, several unrelated hashtags and a block of nonsense text. Interestingly, users in cluster 9 all said virtually the same exact thing in each tweet, suggesting that it may have been a single, or a select few, individuals or machines running a large number of accounts [4]. The same thing, but with a different textual focus, happens for the data in cluster 8.

Clusters 4 and 7 appear to be “slightly smarter” spam-bots as these suspended accounts heavily utilized retweeting of popular content thus appearing more like human users, mixing these tweets in with more traditional spam tweets. Because of this, these spam users were grouped together with many non-suspended accounts. These users qualitatively appear to follow the general trends of the topical focii in the data, and thus were meshed with spam-bots who did the same.

In contrast to clusters 4 and 7, clusters 5 and 2 appear to exhibit features in line with extremist behavior. While future work will need to explore these clusters in more detail, we observe qualitatively that suspended and non-suspended users in these two clusters seem to be heavily focused on activism, both peaceful and violent. Consider the synthetic tweets from the suspended and active users in Cluster 2 in Table VI. The example tweet from a non-suspended user provides fairly straightforward information about an event occurring at a particular location. In contrast, the suspended user is advocating for a means to bypass government-sponsored Internet blackouts, mentioning the hashtag “#Opsyria”, which was a calling card for the hacktivist group Anonymous in Syria. Similarly, the tweet from the suspended user in Cluster 5 claims to have knowledge of violence actions in the region.

Figure 8 illustrates proportion of tweets sent by the 9,603 suspended users we study here that were attributed to each of the different countries. We observe a high variance in distribution of the clustering classes across different country. In particular, we see three classes dominate the distribution: clusters 2, 4, and 5. Cluster 2 and 4 seem to compliment each other. Those countries that have high concentrations of tweets from users in Cluster 2 have few tweets from users in Cluster 4 and vice versa. On the other hand, Clusters 1, 3, 6, 8, and 9, which we appear to be virtually all spam accounts, do not exhibit geographical characteristics and have a fairly even concentrations across all the countries.

While not conclusive, this initial clustering analysis presents evidence that suspended users may be separated into spam and possibly human accounts via the methods described here. This is critical as each class of suspended user brings a different type of bias to the results. Analysts may be able to go beyond simply removing suspended users and instead retain interesting human users that were suspended for “interesting” reasons (e.g. violent activism) versus spam data, which is generally uninteresting for most analyses.

V. Conclusion

In this paper, we conducted several analyses to understand the potential impact of malicious users on the results derived from assessing social media data. Using a large dataset containing data from multiple countries over multiple months, we find that the removal of suspended users can have profound impacts on what users are defined as influential, the overall topology of the mentions and co-topic network, and less impact on what users are defined as influential, the overall topology of the mentions and co-topic network, and less impact on the identification of what is being talked (tweeted) about. In general, these impacts are strongest in countries experiencing more civil unrest. We find that different classes of malicious users, e.g., bots and extremists (or militants or activists) can be differentiated using meta-data and topical analysis. The removal of these different classes of malicious users have differential impacts on the results. This analysis sheds light on a new procedure by which analysts can understand the impact of suspended users on their data and an approach for how to differentiate those users they may want to retain from those that simply corrupt understandings of true social processes existent in their data.

<table>
<thead>
<tr>
<th>Cluster(s)</th>
<th>Susp?</th>
<th>Representative Synthetic Text</th>
</tr>
</thead>
<tbody>
<tr>
<td>1,3,6,8,9</td>
<td>Yes</td>
<td>giberish text #HT1 #HT2 #HT3 #HT4 #HT5 [link]</td>
</tr>
<tr>
<td>2</td>
<td>No</td>
<td>Something bad is happening on the street near my home cairo</td>
</tr>
<tr>
<td>2</td>
<td>Yes</td>
<td>The internet has been blocked in #Syria, here is a link to information on how to get access again. Please send to your friends #Opsyria [link]</td>
</tr>
<tr>
<td>4,7</td>
<td>Yes</td>
<td>RT [screen_name]: The American news can’t simply ignore what happened in #Benghazi @USNewsCompany1 @USNewsCompany2</td>
</tr>
<tr>
<td>5</td>
<td>Yes</td>
<td>There will be an attack at location XYZ at time ABC tomorrow!</td>
</tr>
</tbody>
</table>

TABLE VI: Synthetic, prototypical messages by clusters for suspended or non-suspended users.

3Real tweets are not shown in order to conform to the Twitter TOS

<table>
<thead>
<tr>
<th>Clust. Num.</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Active</td>
<td>2</td>
<td>533</td>
<td>1</td>
<td>2297</td>
<td>1466</td>
</tr>
<tr>
<td>Susp.</td>
<td>2081</td>
<td>183</td>
<td>1505</td>
<td>968</td>
<td>894</td>
</tr>
<tr>
<td>Clust. Num.</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>Active</td>
<td>8</td>
<td>1929</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Susp.</td>
<td>1226</td>
<td>1758</td>
<td>491</td>
<td>483</td>
<td></td>
</tr>
</tbody>
</table>

TABLE V: Number of (Non-)suspended users in each cluster found by the mixture model.
There are multiple implications of these results. First, analyses done prior to accounts being suspended are likely to either need higher levels of cleaning or may be dominated by the activities of malicious users. Second, improved bot detection techniques that could be used in real time will substantially alter results. Third, different types of malicious users appear to cluster together and be creating different biases in the data. For example, the removal of non-bot malicious users may be more likely to impact results observed in areas of high social and political conflict, and in association with extremist events. We end by noting that while bots may truly be noise and may be exerting little influence; malicious users that are not bots may actually be exerting true social influence. It is an open question whether removal of such users is impeding their influence, or just impeding the ability to understand the breadth and nature of their influence.

ACKNOWLEDGMENTS

This work was supported in part by the Office of Naval Research (ONR) through a MURI N000140811186 on adversarial reasoning, DTRA HDTA11010102, by the Department of Defense under the MINERVA initiative through the ONR N000141310835 on Multi-Source Assessment of State Stability, and by Center for Computational Analysis of Social and Organization Systems (CASOS). The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Office of Naval Research, the Department of Defense, or the United States government.

REFERENCES