Soc. Netw. Anal. Min. (2016) 6:51
DOI 10.1007/s13278-016-0358-5

ORIGINAL ARTICLE

Exploring characteristics of suspended users and network stability on Twitter

Wei Wei[1] · Kenneth Joseph[1] · Huan Liu[2] · Kathleen M. Carley[1]

Received: 18 December 2015 / Revised: 21 June 2016 / Accepted: 24 June 2016 / Published online: 21 July 2016

Abstract Social media is rapidly becoming a medium of choice for understanding the cultural pulse of a region; e.g. 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. Unfortunately, social media is also the home of users who engage 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? Our prior research on this topic showed that Twitter-based network structures and content are unstable and can be highly impacted by the removal of suspended users. Because of this, statistical techniques can be established to differentiate potential types of suspended and non-suspended users. In this extended paper, we develop additional experiments to explore the spatial patterns of suspended users, and we further consider how these users affect structural and content concentrations via the development of new metrics and new analyses. We find significant evidence that suspended users exist on the periphery of social networks on Twitter and consequently that removing them has little impact on network structure. We also improve prior attempts to distinguish among different types of suspended users by using a much larger dataset. Finally, we conduct a temporal sentiment analysis to illustrate differences between suspended users and non-suspended users on this dimension.

✉ Wei Wei
weiwei@cs.cmu.edu

Kenneth Joseph
kjoseph@cs.cmu.edu

Huan Liu
huan.liu@asu.edu

Kathleen M. Carley
kathleen.carley@cs.cmu.edu

[1] School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA

[2] School of Computing, Informatics, and Decision Systems Engineering, Arizona State University, Tempe, AZ, USA

1 Introduction

Undesirable users, those users who deliberately engage in activities that harm either other users (e.g. spammer, network phishing) or larger social systems (e.g. militants, terrorism propagandists), are everywhere on today's social media platforms. For example, 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 (Luxton et al. 2012). Spambots can clog the network of information, providing useless or false information to millions of possibly 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 possibly malicious behaviors on a site.

While important questions exist with respect to if and how such behaviors should be restricted by social media platforms, an indisputable point is that the majority of the behaviors engaged in by these undesirable 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. Undesirable users can harm users in

social media by, for example, sending out phishing links, spamming advertisements, or by sending out false and seditious information that might threaten the stability of the society. Because of that, the undesirable users often distort normal social pulses or even become key players in that process.

The actions of these undesirable users have not gone unnoticed, either in the research community (Thomas et al. 2011, 2013; Miller et al. 2014; Xie et al. 2008) or within the social media industry (Hern 2015). Twitter in particular has been taking positive actions to *suspend* users who are recognized by Twitter to be malicious. These decisions on whether (or not) to suspend such users has the potential to bias conclusions drawn about what the population of Twitter users is saying and the social actions they are engaging in. The impacts of these biases on analyses are important particularly for Twitter data, which is increasingly used to understand the cultural landscape and so to identify who are the key users talking about specific topics, responding to important events, providing early warnings of crises, identifying the major topics about which people are concerned and so on (Carley et al. 2014).

In such analyses, we are generally concerned with understanding the actions of humans (i.e. not bots), both those acting in malicious ways and those who are not adjudged by Twitter. Unfortunately, attempts to suspend users introduce two forms of bias into such analysis. First, some bots are not suspended, leading to the existence of "fake" nodes in the network. Second, the suspension of malicious but human users or "true" users of interest—lead to missing data. Such issues are particularly difficult with respect to network analysis, which is often used to analyze Twitter data. Network metrics are sensitive to changes in the nodes introduced into the network (Frantz et al. 2009). For example, Borgatti et al. (2006) 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.

Although we know such biases exist in both the Twitter data and analyses we use, little is known about the cumulative impact of these undesirable users or how the removal of different types of undesirable users affects the information extracted from social media. Illustrative examples exist showing that one type of undesirable user can have massive consequences; e.g. bots have been used to coordinate hashtag campaigns and so influence trending topics on Twitter (Ratkiewicz et al. 2011). However, there has not been a systematic analysis of the impact of suspensions made. The present work provides a wide set of analyses of suspended users and their effect on analyses performed on both the social network structure and the text of a large set of Twitter data. Our study focuses on four key points.

First, we seek to understand the impact that removing suspended users has on the structure of the social networks that can be extracted from Twitter by considering user mentions. Second, we consider how these suspended users impact our understanding of the topical content of our data. Third, we perform a 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. Finally, we assess the extent to which suspended users affect the overall levels of sentiment in our data. All analyses are conducted with consideration of the spatiotemporal properties of our data.

The efforts in the present work extend prior work on the same topic (Wei et al. 2015a) in four major ways. First, we provide additional evidence supporting our prior conclusions which suggested that most suspended users lie in the periphery of the network, and thus that removing them has little impact on the rest of the network. Second, we include additional analyses that highlight the spatial properties of the data. Third, we develop a more robust approach to analyzing the topical information within our data, which also allows us to significantly increase the number of users we consider in our clustering analysis, resulting in an entirely new set of results. Finally, we include a new section devoted to the analysis of sentiment within our data. The new analysis here serves as further evidence that suspended users show important variability in their actions and can also have significant effects on our understanding of the network, topics, sentiment and important users within a particular Twitter dataset.

2 Related work

2.1 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 sociocultural activities (Anthonisse et al. 1971; Freeman 1979). 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 affect different sociocultural 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 (Yin et al. 2013). 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 (Wei and Carley 2014). When the network changes overtime, dynamic network metrics provide additional insights as to how network change impacts individuals (Wei and Carley 2015). Applications of network analysis include recommendation systems (Xia et al. 2015), community detections (Newman 2006) and network structure predictions (Xia et al. 2014).

From a social media perspective, network analytics have been used to, for example, identify communities (Lim and Datta 2013) and better understand the relationship between social and topical structures (Romero et al. 2013). Network analytics are also increasingly used to support spam detection (Wang 2010) and fraud detection (Bolton and Hand 2002). 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 other undesirable users (Joseph and Carley 2015). Such works suggest 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.

2.2 Topic modeling and content analysis

Topic modeling and content analysis have seen major advances in the last decade. The majority of techniques draw inspiration from one or both of the following two methods: latent Dirichlet allocation (LDA) (Blei et al. 2003) and latent semantic analysis (LSA) (Dumais 2004). Both methods model documents as a "bag of words," in which case only the constitution of words is considered rather than their orders. The goal of topic modeling is to infer latent topics, where a topic can be roughly defined as a set of words that frequently co-occur together within the same document.

LDA and LSA have been widely used in the area of information retrieval and data mining (Wang and Blei 2011; Griffiths and Steyvers 2004; Hong and Davison 2010) with different strengths. LDA is a Bayesian model built based on the probabilistic graphical model (PGM) formalization (Jordan 1998) and can be flexibly integrated into other Bayesian models (Wei et al. 2015b; Hong et al. 2012; Diao et al. 2014). LDA can also be naturally interpreted in a hierarchical Bayesian fashion which enables it to be used in hierarchically structured problems beyond topic modeling (Yuan et al. 2012; Joseph et al. 2012). LSA, on the other hand, relies on an eigenvector technique referred to as singular value decompositions (SVD) (De Lathauwer et al. 1994). While LSA prevails in computationally intensive areas such as recommendation systems (Dumais 2004) because of its efficiency, model results from LSA usually lack a clear interpretation of topic hierarchy.

In addition to LDA-based and LSA-based techniques, content analysts have recently leveraged advances in optimization of neural networks to construct new "deep learning" approaches to extract meaning from text (e.g. Mikolov et al. 2013). While such efforts are promising, the extraction of broad topical focus, as opposed to representation of words themselves, is a relatively nascent field (Le and Mikolov 2014). Consequently, in the present work, we choose to leverage LDA to extract a rough representation of the topical focii of users in our dataset.

2.3 Sentiment analysis

A significant amount of recent work has focused on methodologies for the analysis of sentiment and opinions within text (Lin and He 2009; Liu 2012; Titov and McDonald 2008). In such analyses, models to extract sentiment are developed based on a ground truth sentiment dataset, which contains either human labeled word-level or document-level sentiments. In certain domains, document-level labels can be acquired from meta-data such as the review rating on IMDB (Pang et al. 2002). In most domains, however, humans must annotate tweets and acquire sentiment labels manually (Pak and Paroubek 2010). Unfortunately, Twitter falls into this latter domain, as it does not contain meta-data that can represent sentiment labels.

Within such domains, a popular alternative to hand-labeling data is to utilize word-level lexicons, such as the method proposed in the joint sentiment–topic (JST) model (Lin and He 2009) and the Vader model (Hutto and Gilbert 2014). We use a generalized lexicon that combines several existing lexicons including the one find in JST and Vader model to reflect sentiment from tweets.

2.4 Clustering techniques

Clustering algorithms discover latent patterns on data and cluster them into several communities within which members share common patterns. Clustering is a central problem in un-supervised learning and plays an important role in pattern recognition and machine learning. There are two classes of clustering algorithms that are relevant to this paper. First, network-based clustering algorithms such as the Newman algorithm (Newman 2006) can be applied to pair-wise relational data. Examples of pair-wise data are social networks and similarity networks generated by computing cosine similarity between data samples. A more general clustering algorithm treats all inputs as features and

then cluster samples based on their relative similarities. Such algorithms include the well-known K-Means algorithm (Monmarché et al. 1999) and the Gaussian mixture model (GMM) (Reynolds 2009). For a complete review of clustering techniques, we urge the readers to Xu and Wunsch (2005); here, we leverage GMMs to cluster our data.

2.5 Spam detection

Spammers are users in online social networks whose purpose is to distribute advertisements, fraudulent information or to create chaos through misinformation. 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 (Moh and Murmann 2010). For example, Bolton et al. (2002) rely on the fact that most spammers design computer software to distribute their contents. This software can post information at a speed that is much faster than a human, making the statistical distribution of inter-arrival time of the behaviors looks abnormal. For example, sending 10,000 messages in one second definitely does not seem to be a human behavior. Another type of spammer detection builds on the fact that spammer contents have much narrow topic sections than normal contents (Bíró et al. 2008). 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 spammers (Santos et al. 2014; Miller et al. 2014; Amleshwaram et al. 2013; Thomas et al. 2013; Wei et al. 2015a). 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 (Thomas et al. 2011). Thomas and colleagues (Thomas et al. 2013) 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 spambots 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.

3 Data and methods

Our dataset contains approximately 73M tweets from April 2010 to November 2013 sent by roughly 3.8M 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 1), 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 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 October 2014. Table 1 provides a summary of our dataset.

One thing that is worth noting in our dataset is the present of Arabic Tweets. Figure 1 is a visualization of major languages used in the dataset we have colored by suspended and non-suspended users. Here, we see that Arabic (ar) and English (en) constitute of the majority of the tweets in the dataset with English tweets dominating the dataset.

Our analysis is geared toward better understanding how suspended users affect network and topical analyses. The analysis is conducted in four stages as described below.

3.1 Structural impacts

In order to understand how removing suspended users alters network analysis results, we analyze the *mention*

Table 1 Basic statistics for the dataset

Time spread	April 2010–November 2013
Countries studied	Bahrain, Qatar, Libya, Algeria, Tunisia, Oman, Lebanon, Morocco, Jordan, Saudi Arabia, Kuwait, Syria, Iraq, UAE, Egypt, Yemen, Iran
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

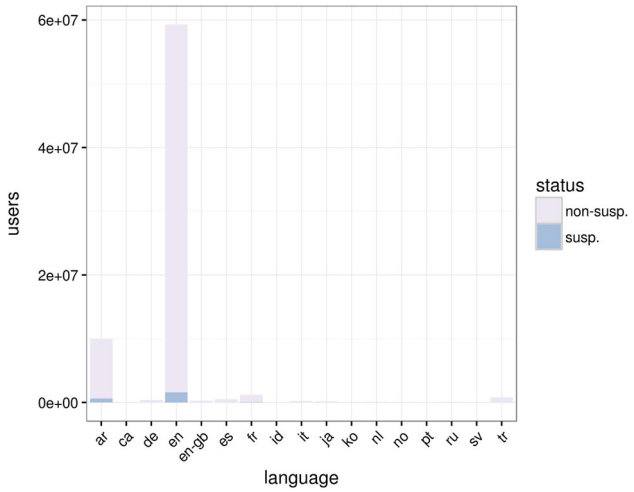

Fig. 1 Languages used by more than 10,000 tweets

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, pretended 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 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 one or more countries.

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 betweenness centrality (which is equivalent to characteristic path length), average closeness centrality, diameter of the network 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 Wei et al. (2011).

3.2 Content-level impacts

In order to understand the effects of suspended users on content-level analyses, we consider both hashtags and topic-model-based conceptualizations of content. With respect to the latter, we use LDA to identify topics. In an LDA model, each tweet has a multinomial distribution over topics, θ. To reduce difficulties poised by the use of common bi-grams and use of common words with similar meanings, we first ran a generic thesauri to clean the data. Finally, to reduce the distraction caused by high levels of nonsense words in Twitter, we removed from each tweet those words that occurred only once in our dataset.

It is well known that LDA provides noisy topic distributions for short texts (Hong and Davison 2010). 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, reasonably consistent topics in nearly all of their tweets (Bosagh Zadeh et al. 2013), this approach helps to alleviate at least this particular problem with applying LDA to Twitter data.

We take this same approach in the present work, with one important difference. To address topical drift over time, we consider the same approach as in Wei et al. (2015a). Basically, we combine user's tweets within each 3-month period and organize these tweets into a single document and use it as training data to LDA. All documents for the training are also restricted to those that contain at least 300 unique words 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. Importantly, different from the approach in Wei et al. (2015a), we back-propagate decisions on the most likely topic for all the tweets that are not restricted to the above limitations. As a result, we can achieve LDA labeling with a robust document set on all the Twitter data we have in the testing stage without the need to lose the quality of LDA analysis in the training stage.

After running LDA on our data, we assess differences in the usage of topics by suspended and non-suspended users, and how these differences can be understood in terms of the change in topical focus as one removes suspended users from the data. To complement this analysis of topics, we also consider how different hashtags are used by suspended and non-suspended users as well.

3.3 Identifying types of suspended users

As noted, different types of suspended users exist. In particular, we would expect that spammers make up a large portion of suspended users. There are also many accounts that do not explicitly emit spam that are suspended. We performed a clustering analysis of the 224,639 users with 83,808 of them are suspended and 140,831 of them are non-suspended users.

The features used for the clustering consist of 11 meta-data-based features along with 200 topical features, which makes a total of 211 features instead of the 209 features used in Wei et al. (2015a). Twitter meta-data-based features are used to construct the first 11 features, described in Table 2. These features are based on known properties of

Table 2 Clustering features captured by Twitter meta-data

Metric name	Description
Num. tweets	$\|T\|$
Cosine sim. tweet text	$\frac{\sum_{t_1,t_2 \in T} \frac{t_1 \dot{t}_2}{\|t_1\|\|t_2\|}}{\|T\|^2 - T}$
RT ratio	$\frac{\|T_{RT}\|}{\|T\|}$
Follower ratio	$\log(\frac{\|\text{Followers}\|+1}{\|\text{Friends}\|+1})$
Number of followers	$\|\text{Followers}\|$
Hashtag ratio	$\frac{\|T_{\#}\|}{\|T\|+\|T_{\#}\|}$
Mention ratio	$\frac{\|T_{@}\|}{\|T\|+\|T_{@}\|}$
URL ratio	$\frac{\|T_{\text{URL}}\|}{\|T\|+\|T_{\text{URL}}\|}$
Num. days active	Days between first and last tweets in dataset
Num. replies	$\|T_{\text{RP}}\|$
Spatial tweet ratio	$\frac{\|T_{\text{Spatial}}\|}{\|T\|}$

spammers discussed in the previous work mentioned above. The hashtag, mention, and URL ratios represent the proportion of a user's tweets that contained 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 indicates the fact that average users tend to focus on a very small, particular set of topics (Bosagh Zadeh et al. 2013), while this may not be true of spammers. 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. Finally, the number of replies and the ratio of spatial tweets are also added to further distinguish human from spammers.

In addition to these 11 features, we also utilize information from the output of the LDA on the topics that users tend to focus on. The LDA we ran has 200 topics, resulting in a topic distribution for each document θ of size 200. The strength of θ represents how likely the user is to choose the specific topic, i.e. the correlation between the user and topic. As is mentioned above, the training set of LDA is restricted to a much smaller set of tweets to achieve a higher quality while the LDA labeling is applied on all the tweet data we have. For each user, we use the sum of all the topic strengths over all the tweets that particular user sent.

3.4 Impact on sentiment

The final section of our analysis addresses how the overall level of sentiment in the data is affected by suspended users. To do so, we apply a sentiment lexicon to the tweets. The lexicon contains binary (+/−) sentiment labels of 25,076 English words, of which 10,182 are labeled as positive and 14,894 are labeled as negative. Many of the English words are also translated into Arabic words using Google translate. If a word does not appear in our lexicon, it is ignored and will not be taken into account in our analysis.

4 Results

4.1 Structural impacts of suspended users

4.1.1 Change in node and edge statistics

Figures 2 and 3 show the percentage change in the number of users and percentage change in the number of edges in the network after suspended users are taken out for each month and for each country. Note that in this particular analysis, we eliminated data before January 2012 or after March 2013, as sampling during these periods was 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. Because of that, suspended users in Algeria also have the most edges. This covers the time frame when Algeria had a protest that last a long period. Other countries that have high proportion of suspended users include Kuwait, Morocco and UAE.

Although the number of nodes in Fig. 2 only shows a moderate level change in suspended users in October of 2012, the number of edges decreases as much as 60 % in countries like Morocco and Algeria, showing the strong impact of suspended users on these countries in the

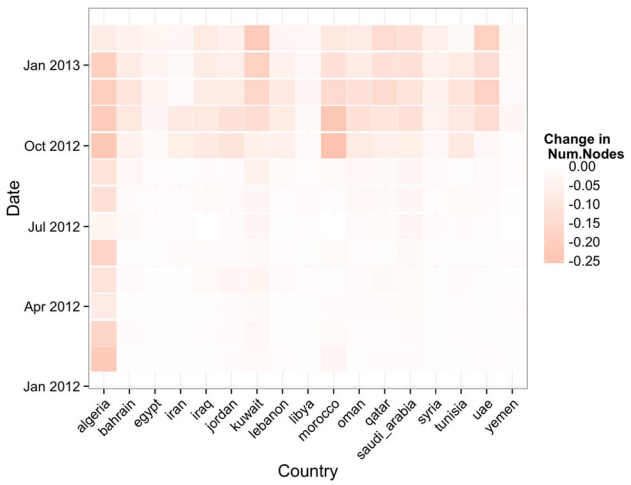

Fig. 2 Change in number of nodes in the networks

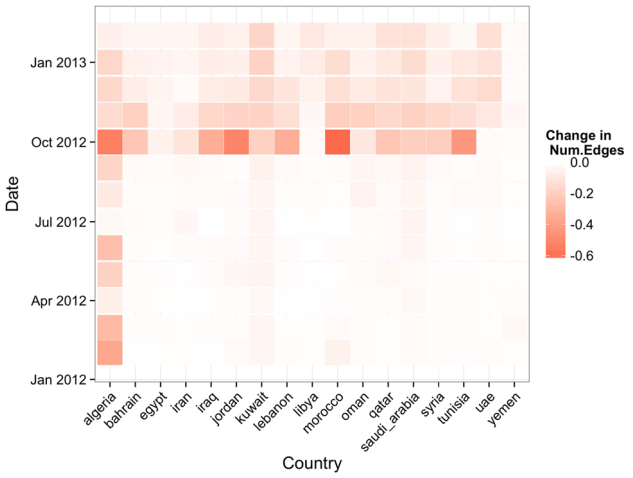

Fig. 3 Change in number of edges in the networks

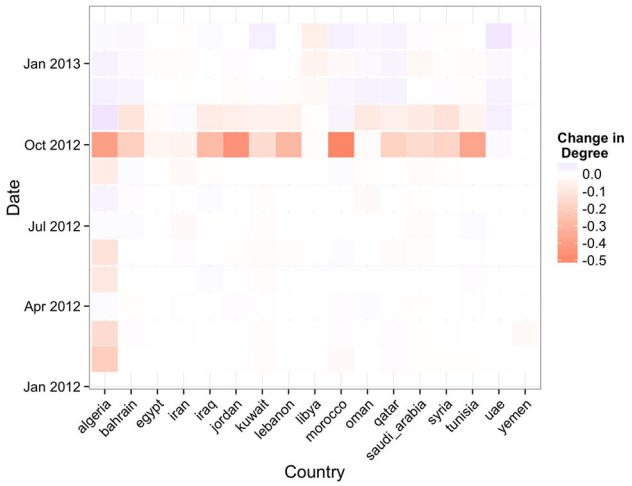

Fig. 4 Change in degree centrality in the networks

mention network. This finding indicates that suspended users are, surprisingly, highly connected to each other in the mention network in these countries. There are several reasons for this, among which are the following three reasons. First, one class of bots—the social bots are designed to mention many other users including each other. Second, a strategy for getting Twitter to suggest that people follow you is to mention many other users, so both bots and human actors often have high levels of mentions so as to attack followers. Third, high levels of mentioning are also used as a strategy by some undesirable users as a way to attract retweets.

In addition to these changes in the number of users and edges in the network, one other interesting thing to look at is the robustness of the networks against the suspended users. As noted above, five network metrics are used in this analysis: (1) degree centrality, which is a measurement of the number of direct neighbors of a user in the network, (2) closeness centrality, a measurement of the average shortest path distance from a particular user to the rest of the network members, (3) the clustering coefficient, which is the proportion of the possible triplets that formed a closed triangle in the user's ego network (4) betweenness centrality, a metric to measure the degree that a particular node is on the shortest paths between other pairs of nodes in the network, and (5) diameter, which measures the longest shortest path in a network. Here, we note that although diameter is not usually considered to be a robustness metric, we here calculate it in order to better understand how robust the pathways from far ends of the network are to the removal of suspended users. Where metrics are at the node level, we take the average over all nodes to determine the value of the metric for the entire network.

4.1.2 Change in network metrics

In Fig. 4, we plot the average change in degree centrality 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, illustrated by change in average degree centrality. As expected, when suspended users 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. This effect begins to fade after October.

When Twitter suspends a user, they can be unsuspended, however, that rarely happens. There is, however, a chance that a suspended user will reappear under a different Twitter account. In our data, we collected the suspended label 1 year after the last of the Tweets were collected; thus, it is not the case that they would have been gone and reappeared under the same Twitter account. Suspending users decreases the average degree centrality as the suspended users generally have more connections in the mention network as previously discussed. When average degree increases, it does so only slightly. One possible reason for this is that a larger fraction of the suspended users in these cases were low in degree centrality in the mentions network. And, as will be seen, many of the cases where there is a slight increase in degree centrality, it is because those users that are suspended are spammers that are high in hashtags but low in mentions.

Figure 5 illustrates the change in closeness centrality in the networks over time and country in a manner similar to

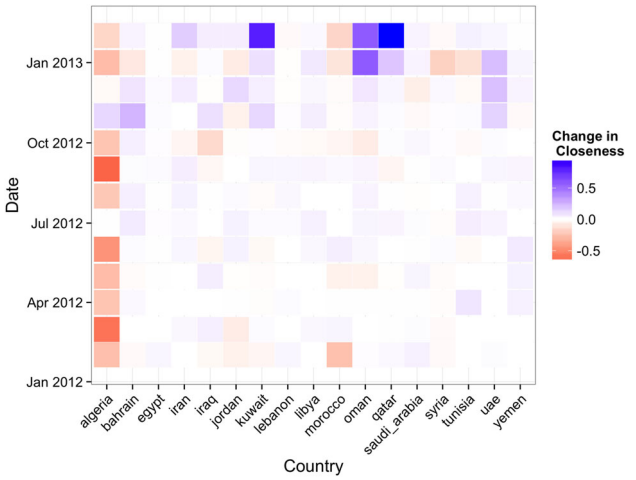

Fig. 5 Change in closeness centrality in the networks

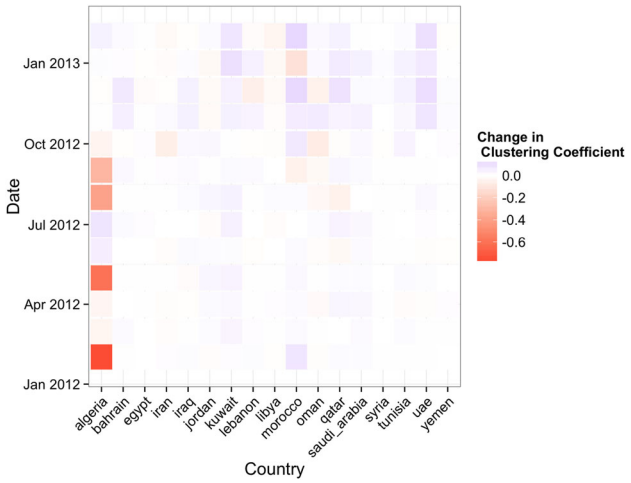

Fig. 6 Change in clustering coefficient in the networks

the previous plot. Since closeness is affected not only by the direct neighbors but also by long distance network structures, its value is more sensitive to the removal of suspended users. This can be validated in the mixed change in patterns that appears in most time steps across all countries, which contains both positive and negative changes.

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 consists of mostly 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 *decreases* by a factor of 2 after the removal of suspended users.

This finding provides further evidence that suspended users are in the peripheral of the social networks. Thus, removing these 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 words, the suspended users are being excluded from the main mention network component, composed of individuals who are tightly connected. 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 Fig. 6. Overall, suspended users have little impact on the local clustering structures of the network, further suggesting their existence on the periphery. The only exception is in Algeria, where we see a nearly 80 % decrease in the clustering coefficient when suspended users are removed. Most other changes are positive, meaning that the elimination of suspended users makes the local structures of the

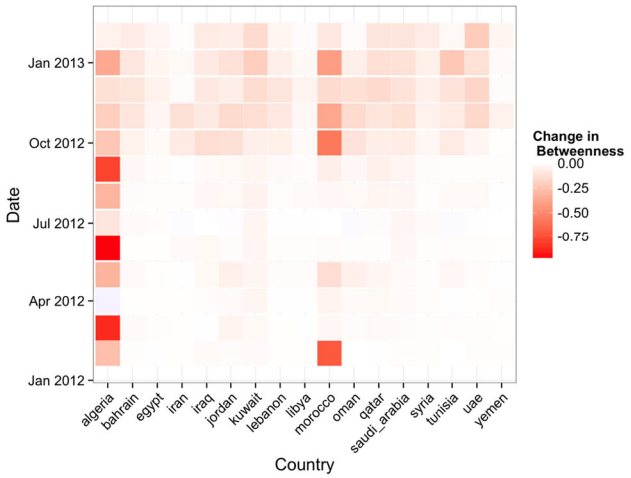

Fig. 7 Change in betweenness centrality in the networks

mention network more cohesive. After October 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 normalized clustering coefficient, making the overall metric increase.

We also look at the change in betweenness centrality illustrated in Fig. 7. Although both betweenness and closeness measures are based on shortest path calculations, they reveal different information on how the shortest paths changed in the network. Recall in Fig. 5, average closeness centrality decreased in some of the countries at particular time, indicating that the suspended users had been making the network appear larger and less connected. In Fig. 7, however, the average betweenness shows a decrease in all

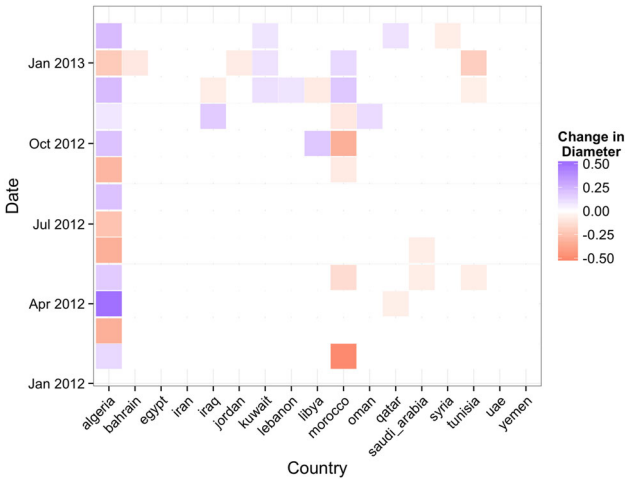

Fig. 8 Change in diameter in the networks

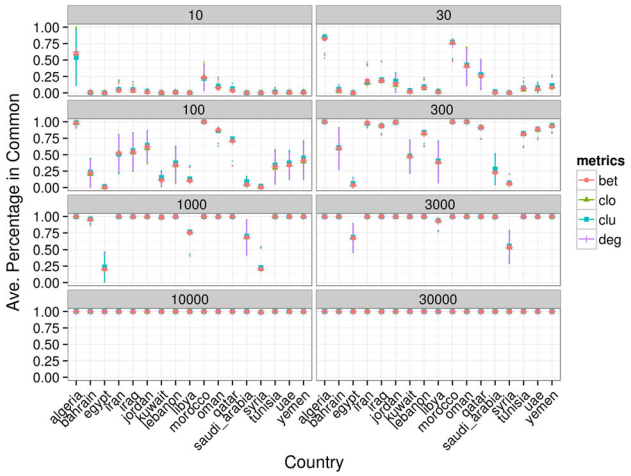

Fig. 9 Percentage in common of the top nodes ranked list

of the countries at all the time steps. This means that the suspended users were possibly connecting disconnected groups or were on the fringe of the social media society and made the overall network appear larger with longer paths between users. Again, Algeria and Morocco are the most affected countries in this analysis due in part to the larger number of suspended accounts.

Furthermore, we look at network diameter, which is a network-level metric in Fig. 8. In the plot, we see that for most countries, network diameter does not change. In a few cases, e.g. Algeria and Morocco, the diameter changed as much as 50 %. Prior to October 2012, there is little change except in these two countries. However, after October of 2012, the removal of suspended actors tends to increase the network diameter. The increase in diameter indicates that suspended users lie in critical positions along the longest shortest path and were connecting groups that are now possibly disconnected. In contrast, in those cases where diameter decreased this indicates that the suspended users were on the fringe and actually increasing the size of the network without increasing connectivity. All of which suggests that there are multiple types of users being suspended with multiple types of network profiles.

In summary, these findings suggest three broad conclusions about the impact of suspended users on the mentions networks. First, the removal of suspended users can have dramatic consequences on the mentions network. Hence, it may be difficult to recover the network properties of the network prior to user suspension if you only use data after the suspensions have occurred. Second, there appear to be at least two different network profiles for suspended users—those who mention many others and serve to connect disconnected groups and those that mention few others and served to increase the size of the network while reducing its connectivity. Knowing which type of actors is likely to have been suspended is critical to understanding the network structure and may be key to determining the chance that users of interest were suspended. Thirdly, October 2012 signaled a marked change in the impact of suspensions; this may be partially due to changes what types of users were suspended and/or changes in types of bots that were appearing.

4.1.3 Change in metrics ranking of network members

In addition to changes on average at the network level, we also evaluate the effects of removing suspended users on the distribution of "top" nodes in our dataset, as measured by the metrics noted above. Figure 9 provides results from an analysis of the change in the top-k non-suspended nodes rank based on their individual network metrics. For each country and each month, network metrics are generated for each node 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 compute 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 k. 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 9 shows that when k is small (e.g. 10), the chance that we will see common nodes in both of the ranked lists is low. This suggests that analysis of the top nodes in a network are highly vulnerable to the addition or removal of suspended users, a finding consistent with popular press on the extent to which followers of prominent political actors were bots[1]. Thus, while aggregate network-level metrics

[1] http://www.politico.com/story/2014/06/twitter-politicians-107672.

may show little change, the perceived importance of nodes can be affected significantly by the removal of suspended users, and so by 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 nodes 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 nodes begins to increase and eventually becomes close to 1. A balance must thus be struck between larger k, where such patterns emerge, and smaller k, where only the important nodes are considered, when attempting to select important nodes that are also robust against suspension.

4.2 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. θ). 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 θ_i of this particular topic i across all documents. For hashtags, we use $\mathrm{RH} = \{h_1, h_2, \ldots, h_{L-1}, h_L\}$ to denote the rank list of hashtag on all the users, while using RH^- to denote the rank list of only non-suspended users on hashtag. Here, h_i has a higher or equal importance factor than h_j if $j > i$. Similarly, we can define $\mathrm{RT} = \{t_1, t_2, \ldots, t_{K-1}, t_K\}$ to be the rank list of topics on all the users and RT^- to be the topic rank list on only non-suspended users.

4.2.1 Top hashtag/topics for suspended and non-suspended accounts

Tables 3 and 4 show 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 of 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 3 shows that suspended users refer to the hashtags of a host of nations, as well as to the CIA and CNN. Table 3 shows that the focus of our data additionally centered on topics such as "pain" and "killing."

Table 3 Top hashtags used in the dataset

Hashtag	S_{H}	Hashtag	S_{H}
Syria	162,271	Saudi Arabia	21,922
CIA	80,390	Kuwait*	18,580
Syria*	65,307	Country*	18,013
Egypt	39,229	CNN	15,840
Egypt*	30,931	Benghazi	13,297

Table 4 Top LDA topics found in the dataset

S_T	Terms		
44,4898	community*	kuwait	wall*
	kuwait	egypt	
145,323	wall*	syria	rehab*
	amayadeentv1	flag	
145,323	almayadeentv1	bahrain	collection*
	enemy*	killing*	
75,518	Egypt*	ordered*	race*
	Islam*	pain*	
48,649	Kuwait	egypt	UAE
	qatar	bahrain	

4.2.2 Ranking of hashtag/topic for suspended and non-suspended accounts

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 $\mathrm{RH}_{q*L} = \{h_1, h_2, \ldots, h_{q*L}\}$ and only the active users $\mathrm{RH}^-_{q*L} = \{h^-_1, h^-_2, \ldots, h^-_{q*L}\}$. The matching score $m_{\mathrm{H}}(q)$ 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 Eq. (1). If the elements in RH_{q*L} are exactly the same as the ones found in RH^-_{q*L}, then $m_{\mathrm{H}}(q) = \frac{qL}{L} = q$. Otherwise, $m_{\mathrm{H}}(q) < q$. Similarly, one can define the corresponding matching score for topics, which we refer to as $m_{\mathrm{T}}(q)$.

$$m_{\mathrm{H}}(q) = \frac{\left|\mathrm{RH}_{q*L} \cap \mathrm{RH}^-_{q*L}\right|}{L} \tag{1}$$

We vary q from 0 to 100 % to see how the matching score changes. Figure 10 shows both the results of hashtags and the LDA topics. The horizontal axis is q, while the vertical axis is the matching score [either $m_{\mathrm{H}}(q)$ or $m_{\mathrm{T}}(q)$]. A reference line with a slope of 1 is also plotted. If the suspended users had no affect on the ranking list, the ranking list before and after suspended users are removed and would align with the reference line. The more the matching

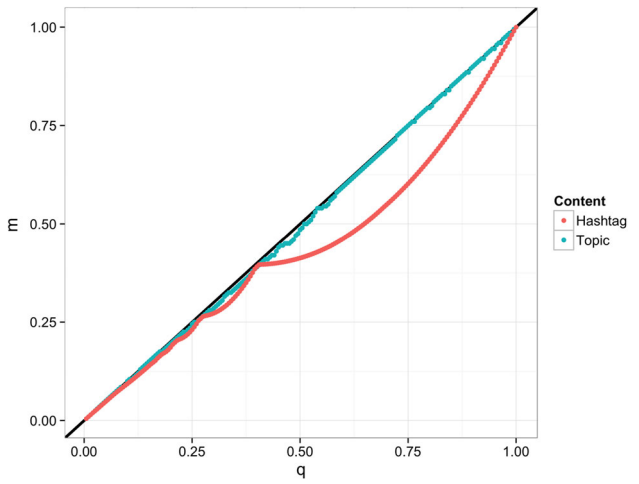

Fig. 10 Change of LDA topics and hashtags made by suspended users

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 only small deviations. Those deviations appear when q is between 25 and 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, are 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 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 suggests 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 begins to show up. However, if q reaches the two ending points of the subgroups, the difference returns to nil and the matching score returns to the reference line.

4.2.3 *Change in spatial concentrations on hashtag/topic*

In addition to the aggregate analysis above, we also analyze how the concentrations of hashtags and topics change between the tweets sent by suspended users and non-suspended users in different countries. Different from the analyses conducted in Sect. 4.1 which compares the full dataset and the dataset after suspended users are removed; here, we analyze the percentage change inthe hashtag–topic concentrations between non-suspended users and suspended users.

To aggregate the results by geo-regions, we extract the latitude and longitude coordinates from Twitter JSON, representing the geospatial coordinates of the mobile device when a tweet is being sent out. Tweets that do not have a geo-location are ignored in this analysis. After tweets that belong to each country are collected, they are normalized across hashtag/topics so that the aggregated value for each country represents the probability of a tweets coming from this country utilizing a specific hashtag/topic.

Figure 11 shows the change in top hashtags between non-suspended users and suspended users in different geo-regions. Here, we apply the same three-color scheme in the previous analysis where white denotes neutral or no change, red denotes negative change and blue denotes positive change. Colors are scaled to their maximum–minimal limits for each subplot. The hashtags are selected to be the top 5 hashtags in the dataset, and each sub-plot represents their geo-spatial spreads of the strength that belongs to each specific hashtag. The first thing we see is that most countries such as Iran, Sudan, and Saudi Arabia have constantly negative changes, which means moving from non-suspended users to suspended users, and these countries have a much lower likelihood of utilizing these specific hashtags. The reduced probability masses must be relocated to hashtags that are less popular, meaning that

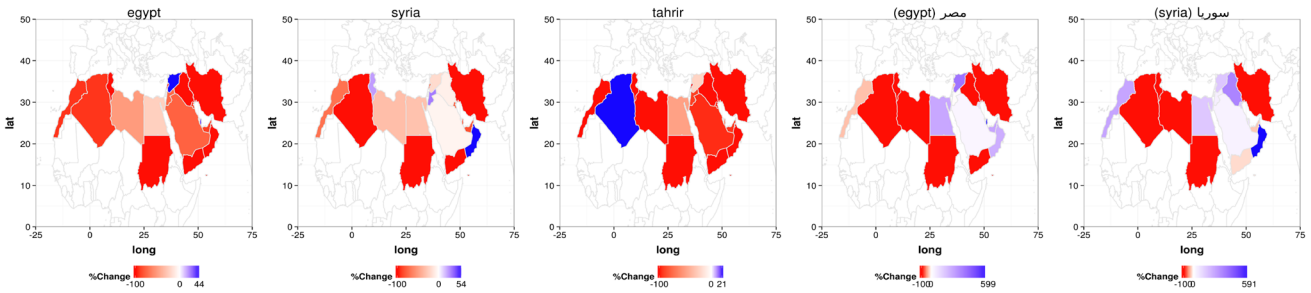

Fig. 11 Spatial visualization of change in top hashtags

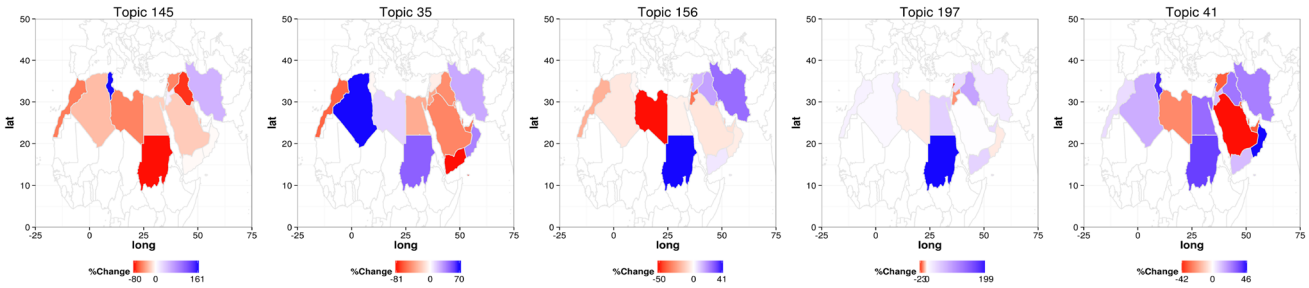

Fig. 12 Spatial visualization of change in topics

Table 5 Top LDA topics found in spatial dataset

Topic	Terms		
145	terrorist	syrian arab republic	disagree
	syrian	washington dc	houthis
35	communicate	Washington DC	uncertain
	disagreement	about	
156	los angeles	australia	illinois
	language	acknowledge	
197	hamad*	machine*	throne*
	pray*	unite*	
41	love	:)	humor
	lol	good news	

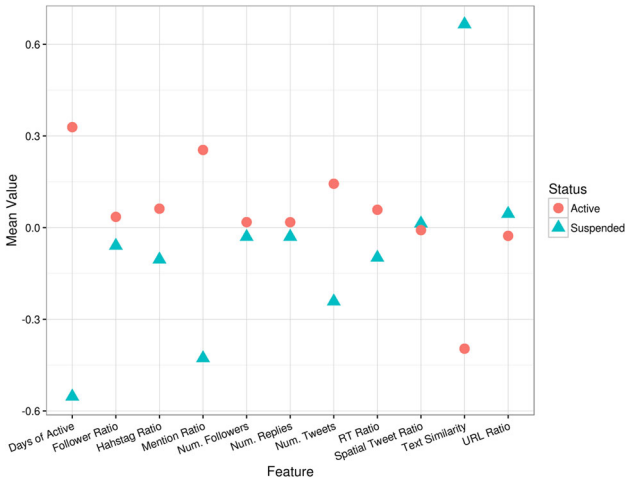

Fig. 13 Mean value of the 11 text/meta-data metrics for suspended and non-suspended users. 95 % bootstrap confidence intervals are plotted, but are often smaller than the point presented

suspended users from those countries will usually not target the most popular hashtags. For countries such as Algeria, Oman, Qatar and Syria, however, their change in hashtag probability varies and depends on specific hashtags. For example, the right-most subplot of Fig. 11 showed that suspended users in Oman are 591 % more likely to send out a tweet about the hashtag Syria than non-suspended users.

Figure 12 shows the change in usage of the top LDA topics between non-suspended users and suspended users in different geo-regions. The most related words for each topic in the visualization are illustrated in Table 5. Figure 12 shows that topic distributions are much more diverse than patterns in the hashtag analysis. As we move from normal tweets to tweets sent by suspended users, we get significant increase in the probability of sending tweets in topic 35 in Algeria, which is a topic showing the concern of existing political situations. Similarly, there is also significant increase in the interests of suspended users on topic 156 and topic 197 in Sudan.

While patterns between the topical and hashtag results differ in sign, one interesting aggregate result is that the magnitude of change is generally smallest in locations relevant to the hashtag/topic and the larger in external places. This observation suggests two things. First, unsurprisingly, suspended users are generally more focused on globally relevant topics/hashtags rather than hashtags relevant to any particular locale. Second, and more interesting, is that our data suggests that the removal of suspended users will have a smaller impact on content analysis if the analysis is focused on a locally relevant conversation, as their "noise" may be drowned out by the "signal" coming from non-suspended users.

4.3 Identifying types of suspended users

In this section, we perform a clustering analysis on 224,639 suspended users, of which 83,808 are suspended users that tweeted more than once in our dataset and the other 140,831 are randomly sampled non-suspended users in our dataset that tweeted at least once.

4.3.1 Distributions of features on labels

Figure 13 shows the mean value of the 11 text-based and meta-data-based metrics for suspended and non-suspended users. All metrics have been centered and scaled by two standard deviations, which facilitates comparison of the

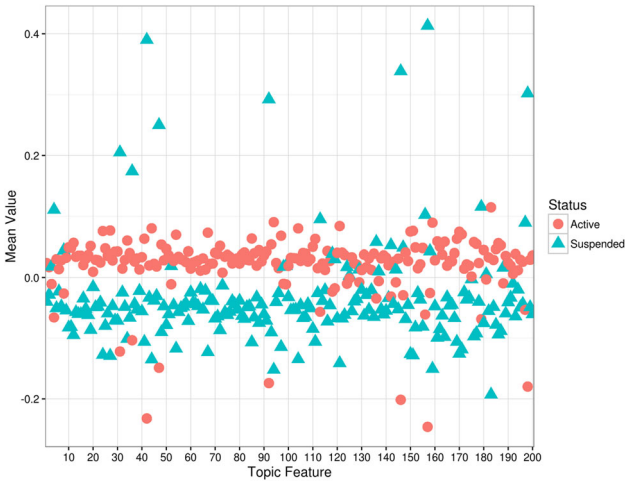

Fig. 14 Mean value of the 200 topical strength suspended and non-suspended users. 95 % bootstrap confidence intervals are plotted, but are often smaller than the point presented

more extreme ends of the distribution (Gelman 2008). Our data generally fits with expected differences between suspended and non-suspended users. Namely, 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.

Figure 14 shows the mean values of 200 LDA topics for suspended and non-suspended users. We see that most of the topics have equal mean values between suspended and non-suspended users. For suspended users, however, particular topics are more likely to present than others, while non-suspended users do not tend to exhibit such differences. These topics can be used to distinguish between suspended and non-suspended users.

4.3.2 Clustering results

Having observed differences between suspended and non-suspended users, we now turn to a cluster analysis of this same set of users. To perform the clustering, we utilize `scikit-learn` (Pedregosa et al. 2011) to perform Gaussian mixture modeling (Reynolds 2009). We select the best number of clusters via comparison of model Bayesian information criterion (BIC). 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 clustering results and their distribution on suspended and active users are illustrated in Table 6. Clusters

Table 6 Number of (non-)suspended users in each cluster found by the mixture model

Clust. num.	0	1	2	3	4
Active	26,624	14,572	20,642	841	11,197
Susp.	13,439	3701	5968	18,213	1672
Clust. num.	5	6	7	8	
Active	10,778	24,060	20,146	11,971	
Susp.	2296	4167	4978	29,374	

3 and 8 contain mostly suspended users and are likely to be obvious spammers. Clusters 4 and 5 contain mostly active users and are likely to be standard Twitter users. Clusters 0, 1, 2, 6, 7, on the other hand, contain a mixture of suspended and active users and are likely to contain both smarter spammers and militants mixed in among regular Twitter users.

The semantic meanings of those clusters are more obvious when we look at the text of the tweet in Table 7. Based on Twitter's terms, we can not release the actual texts of the tweets. However, we made some synthetic tweets based on the patterns we see in tweets fall into similar user groups. Firstly, clusters 3 and 8 contain tweets with repeated patterns of gibberish text, random hashtags and a link. The majority of the tweets in these clusters are in almost the same pattern over and over again. We conclude that these are obvious spammers. Clusters 4 and 5 contain mostly normal tweets, and their texts belong to a diverse range of topics. Clusters 0, 1, 2, 6 and 7 contain tweets that are sent by both normal users and suspended users. Tweets such as the one sent by the suspended user in Table 7 are most likely an individual suspended for promoting or encouraging violence.

We plot the proportions of users that fall into different clusters across countries and over time in Figs. 15 and 16, respectively. Here, we see that clusters 3 and 8 have relatively stable proportions of users over country and over-time. The lack of variance over space and time is yet another evidence that it should be regarded as obvious spammers. Clusters 0, 1 and 2 which contains both suspended and active users have increased proportions over time, while the clusters that contain most normal users (i.e. 4 and 5) has a much lower proportion in the recent years.

Another way to look at the clustering result is through Figs. 17 and 18, which is the same feature distribution as detailed above but plotted with different clustering classes rather than suspended–active labels. Here, we observed that clusters 3 and 8 have an abnormally high text similarity, low mention rates, low follower ratios and low days of active. All of these again indicated that they are obvious spammers. They also constituted to the spikes in several

Table 7 *Synthetic*, prototypical messages by clusters for suspended or non-suspended users

Cluster(s)	Susp?	Representative synthetic text
0, 1, 2, 6, 7	Yes	RT @Barackobama: you murderer.#HT [[link]]
0, 1, 2, 6, 7	No	Merry Christmas #HT1 #HT2
4, 5	No	War reporter kidnapped a second time [[link]]
3, 8	Yes	[[link]] gibberish text #HT1 #HT2 #HT3 #HT4 #HT5 [[link]]

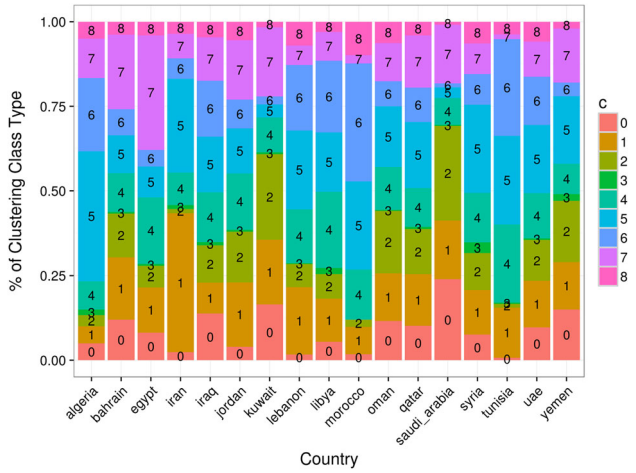

Fig. 15 Percentage of cluster distribution of each country

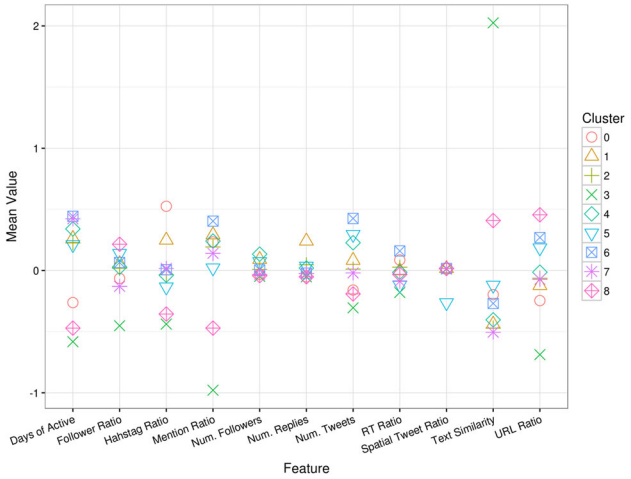

Fig. 17 User feature distribution of different clustering classes

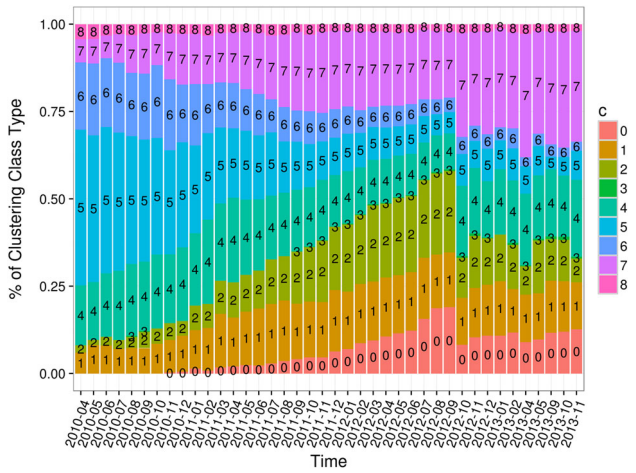

Fig. 16 Percentage of cluster distribution over time

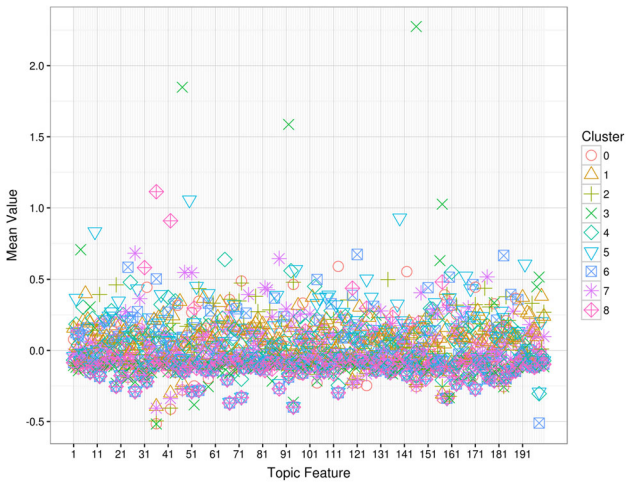

Fig. 18 LDA topics of different clustering classes

topics in the topic features as well. Cluster 4 and 5 have mean values that are close to each other on most of the metrics, while cluster 0, 1, 2, 6 and 7 have values in between the normal clusters and the obvious spammers.

These analyses suggest that although nine clusters were found to fit the data best, an obvious pattern arises in which three "kinds" of clusters can be observed. The first kind of cluster contains mostly stock Twitter users, the second, mostly "dumb" spammers, and the third a mix of "smarter spammers", regular Twitter users and, we believe, individuals who were using the service as a human but that were suspended for other reasons. Future work will more carefully consider this third kind of cluster and how best to differentiate between these various sorts of users contained within them.

4.4 Sentiment impacts of suspended accounts

In this section, we will illustrate how sentiment change over time and country in the dataset. In order to determine

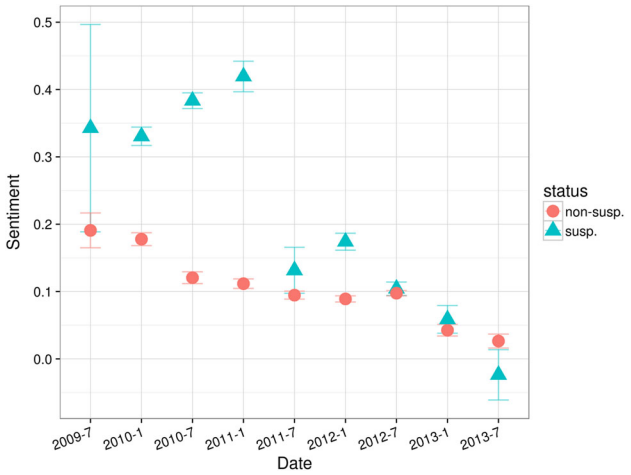

Fig. 19 Change of LDA topics and hashtags made by suspended users

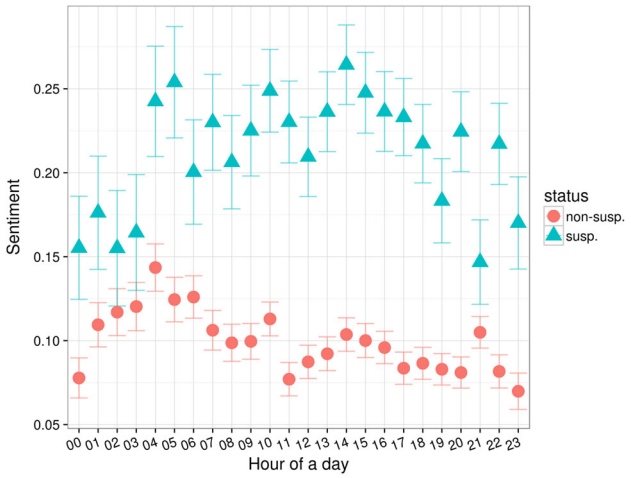

Fig. 20 Mean values of the 11 text/meta-features for suspended users and non-suspended users colored by latent clusters

the sentiment, we use lexicon developed by taking a list of all terms in three major sentiment tools—Sentiwordnet (Esuli and Sebastiani 2006), Lyke (Pennebaker et al. 2007), Vader (Hutto and Gilbert 2014), and ACT (Heise 1987). The valence of the term was set as positive or negative if all 4 of the dictionaries agreed. For all disagreements, the terms were checked by a set of coders, and valence was added. All terms were then translated to multiple other languages using Google translate, and a sample of the translations checked by local language speakers. Development was done jointly by Carnegie Mellon University and Netanomics.

Figure 19 is a overtime visualization of the average sentiment for tweets within the given month for non-suspended and suspended users. In general, we see that the mean value of sentiment in the tweets sent from suspended users is actually much higher than the sentiment of tweets sending from non-suspended users. This suggests that suspended users usually use positive words to persuade people to go into specific link or accept a specific idea. For example, the following is a synthetic tweet that is similar to those observed in our data from suspended users "Good News Everyone! #Egypt develops sequel to #Jan25..." Over time, however, our results suggest that suspended users may have become more intelligent with their use of sentiment. This might be an effort for those spammers to improve their skills in order to avoid being caught by spamming detection software enforced by Twitter.

In Fig. 20, we see another temporal visualization with horizontal axis being the time of a day. The time reported here is the local time for each country. The differences between the suspended user and non-suspended users in terms of sentiments are obvious during the normal business hours. When it is close to the mid-night, the sentiment of suspended and non-suspended users tends to look similar. Thus, an interesting feature to be explored for bot detection is the pattern of sentiment usage as it relates to generally observable human usage of sentiment on Twitter (Golder and Macy 2011).

To see whether the sentiment of tweets can exhibit spatial patterns, we also compared the average sentiment before and after suspended users are taken out as shown in Fig. 21. Here, we use the same color scheme as we use in the previous sections with white being neutral change and blue/red being the positive and negative change. As we can see, differences in mean sentiment usage between suspended and non-suspended users are restricted to a particular set of countries For example, Iran has much positive change in sentiment in the 2010 data after suspended users are taken out. This means suspended users are the main reason to drag down the average sentiment of this country. As we move to 2011, this change becomes less obvious and in 2012, we see a significant negative change in the sentiment. This means as time moves, the sentiment of suspended users are becoming more positive in Iran.

5 Limitation

There are several limitations to the present work. First, the dataset used is collected using country-specific keywords and bounding boxes from Twitter. These countries have different levels of Internet usage, different levels of accessibility to and use of Twitter, and different government regulations, sanctions and oversight of Twitter usage. Such differences create sampling biases that may impact our analyses. For example, in our analysis, we observed that certain countries had a higher percentage of suspended

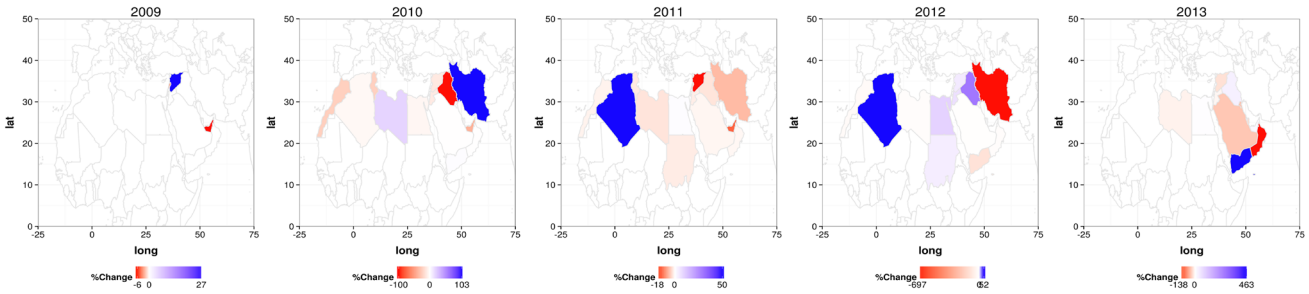

Fig. 21 Mean values of the 200 topical features for suspended users and non-suspended users colored by latent clusters

users than others while having the least number of Twitter users. Such statistical patterns might be difficult to discern due to the small amount of data. It is also possible that these usage differences between countries impact the structure of the mentions and hashtag networks. Indeed, in our prior work, we find that countries with high volatility are more likely to broadcast, and those with low volatility are more likely to engage in discussions where small groups mention each other. That being said, it is unlikely that these country differences will impact the types of bots present but they may impact the way in which humans who end up being suspended present themselves on Twitter. Whether or not these country differences are sufficient to account for differences in the relative number of suspended users that are highly connected and connecting or are more peripheral and expanding the mention network is a point for further study. Such a study might also look at other social media for additional verification.

A second limitation is the way in which users were identified as suspended. As noted, whether an account was suspended was determined at a single point in time after the tweets were collected. Hence, issues such as the impact of rolling suspensions and recovery from suspension could not be addressed. In addition, although the list of suspended users was collected by verifying the status of users using Twitter's service, it is possible that even more of these accounts have now been suspended. Thus, the full impact of suspension may not be known. Future work should examine suspensions from a more temporal perspective.

A third limitation is that the data used drew from two different Twitter selection strategies—the 10 % feed and a directed bounding-box/hashtag data pull from the Twitter API. Each collection strategy is limited. The key though is that while the topics are likely to be representative of the regions, the mentions networks are possibly too sparse. The most likely users to not be included, however, are those that rarely tweet and rarely mention others. Thus, we may not be capturing data on low activity suspended users. These factors are unlikely to impact the results of these analyses.

6 Conclusion

In this paper, we conducted several analyses to understand the impact of removing suspended 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 the identification of what is being talked (tweeted) about. In general, these impacts are strongest in countries experiencing more civil unrest. We find evidence that different classes of suspended users, e.g. bots and extremists (or militants or activists), may be differentiable via meta-data and topical analysis. The removal of these different classes of undesirable users has 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.

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 suffer a bias such that the results are dominated by the activities of undesirable users. Second, improved bot detection techniques that could be used in real time will substantially alter results. Third, different types of undesirable users appear to cluster together and be creating different biases in the data. For example, the removal of non-bot undesirable users may be more likely to impact results observed in areas of high social and political conflict, and in association with extremist events. We note that while bots may truly be noise and may be exerting little influence, undesirable 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 HDTRA11010102, 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

Amleshwaram AA, Reddy N, Yadav S, Gu G, Yang C (2013) CATS: characterizing automation of twitter spammers. In: Communication systems and networks (COMSNETS), 2013 fifth international conference on, IEEE, pp 1–10

Anthonisse JM (1971) The rush in a directed graph. Stichting Mathematisch Centrum Mathematische Besliskunde (BN 9/71):1–10

Bíró I, Szabó J, Benczúr AA (2008) Latent dirichlet allocation in web spam filtering. In: Proceedings of the 4th international workshop on adversarial information retrieval on the web, ACM, pp 29–32

Blei DM, Ng AY, Jordan MI (2003) Latent dirichlet allocation. J Mach Learn Res 3:993–1022

Bolton RJ, Hand DJ (2002) Statistical fraud detection: a review. Stat Sci 17:235–249

Borgatti SP, Carley KM, Krackhardt D (2006) On the robustness of centrality measures under conditions of imperfect data. Soc Netw 28(2):124–136

Bosagh Zadeh R, Goel A, Munagala K, Sharma A (2013) On the precision of social and information networks. In: Proceedings of the first ACM conference on Online social networks, pp 63–74

Carley KM, Pfeffer J, Morstatter F, Liu H (2014) Embassies burning: toward a near-real-time assessment of social media using geo-temporal dynamic network analytics. Soci Netw Anal Min 4(1):1–23

De Lathauwer L, De Moor B, Vandewalle J, by Higher-Order BSS (1994) Singular value decomposition. In: Proceedings of the EUSIPCO-94, Edinburgh, Scotland, UK, vol 1, pp 175–178

Diao Q, Qiu M, Wu CY, Smola AJ, Jiang J, Wang C (2014) Jointly modeling aspects, ratings and sentiments for movie recommendation (jmars). In: Proceedings of the 20th ACM SIGKDD international conference on knowledge discovery and data mining, ACM, pp 193–202

Dumais ST (2004) Latent semantic analysis. Ann Rev Inf Sci Technol 38(1):188–230

Esuli A, Sebastiani F (2006) Sentiwordnet: a publicly available lexical resource for opinion mining. In: Proceedings of LREC, Citeseer, vol 6, pp 417–422

Frantz TL, Cataldo M, Carley KM (2009) Robustness of centrality measures under uncertainty: examining the role of network topology. Comput Math Organ Theory 15(4):303–328

Freeman LC (1979) Centrality in social networks conceptual clarification. Soc Netw 1(3):215–239

Gelman A (2008) Scaling regression inputs by dividing by two standard deviations. Stat Med 27(15):2865–2873

Golder SA, Macy MW (2011) Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures. Science 333(6051):1878–1881. doi:10.1126/science.1202775, http://www.sciencemag.org/content/333/6051/1878

Griffiths TL, Steyvers M (2004) Finding scientific topics. Proc Natl Acad Sci 101(suppl 1):5228–5235

Heise DR (1987) Affect control theory: concepts and model. J Math Sociol 13(1–2):1–33

Hern A (2015) Twitter CEO: we suck at dealing with trolls and abuse. http://www.theguardian.com/technology/2015/feb/05/twitter-ceo-we-suck-dealing-with-trolls-abuse

Hong L, Ahmed A, Gurumurthy S, Smola AJ, Tsioutsiouliklis K (2012) Discovering geographical topics in the twitter stream. In: Proceedings of the 21st international conference on world wide web, ACM, pp 769–778

Hong L, Davison BD (2010) Empirical study of topic modeling in twitter. In: Proceedings of the First Workshop on Social Media Analytics, ACM, pp 80–88

Hutto C, Gilbert E (2014) Vader: a parsimonious rule-based model for sentiment analysis of social media text. In: Eighth international AAAI conference on weblogs and social media

Jordan MI (1998) Learning in Graphical Models: [proceedings of the NATO Advanced Study Institute...: Ettore Mairona Center, Erice, Italy, September 27-October 7, 1996], vol 89. Springer Science & Business Media

Joseph K, Carley KM (2015) Culture, networks, twitter and foursquare: testing a model of cultural conversion with social media data. In: Proceedings of the 7th international AAAI conference on weblogs and social media (ICWSM)

Joseph K, Tan CH, Carley KM (2012) Beyond local, categories and friends: clustering foursquare users with latent topics. In: Proceedings of the 2012 ACM conference on ubiquitous computing, ACM, pp 919–926

Le QV, Mikolov T (2014) Distributed representations of sentences and documents. arXiv preprint arXiv:1405.4053

Lim KH, Datta A (2013) A topological approach for detecting twitter communities with common interests. In: Atzmueller M, Chin A, Helic D, Hotho A (eds) Ubiquitous social media analysis. Springer, Berlin Heidelberg, pp 23–43

Lin C, He Y (2009) Joint sentiment/topic model for sentiment analysis. In: Proceedings of the 18th ACM conference on information and knowledge management, ACM, pp 375–384

Liu B (2012) Sentiment analysis and opinion mining. Synth Lect Hum Lang Technol 5(1):1–167

Luxton DD, June JD, Fairall JM (2012) Social media and suicide: a public health perspective. Am J Public Health 102(S2):S195–S200. doi:10.2105/AJPH.2011.300608

Mikolov T, Chen K, Corrado G, Dean J (2013) Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781

Miller Z, Dickinson B, Deitrick W, Hu W, Wang AH (2014) Twitter spammer detection using data stream clustering. Inf Sci 260:64–73

Moh TS, Murmann AJ (2010) Can you judge a man by his friends?-enhancing spammer detection on the twitter microblogging platform using friends and followers. In: Information systems, technology and management. Springer, pp 210–220

Monmarché N, Slimane M, Venturini G (1999) Antclass: discovery of clusters in numeric data by an hybridization of an ant colony with the kmeans algorithm

Newman ME (2006) Modularity and community structure in networks. Proc Natl Acad Sci 103(23):8577–8582

Pak A, Paroubek P (2010) Twitter as a corpus for sentiment analysis and opinion mining. In: LREC, vol 10, pp 1320–1326

Pang B, Lee L, Vaithyanathan S (2002) Thumbs up?: Sentiment classification using machine learning techniques. In: Proceedings of the ACL-02 conference on empirical methods in natural language processing-volume 10, association for computational linguistics, pp 79–86

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V et al (2011)

Scikit-learn: machine learning in python. J Mach Learn Res 12:2825–2830

Pennebaker JW, Booth RJ, Francis ME (2007) Linguistic inquiry and word count: Liwc. Liwc net, Austin

Ratkiewicz J, Conover M, Meiss M, Gonçalves B, Flammini A, Menczer F (2011) Detecting and tracking political abuse in social media. In: ICWSM

Reynolds D (2009) Gaussian mixture models. In: Encyclopedia of biometrics. Springer, pp 659–663

Romero DM, Tan C, Kleinberg J (2013) On the interplay between social and topical structure. In: Proceedings of the 7th International AAAI Conference on weblogs and social media (ICWSM)

Santos I, Miambres-Marcos I, Laorden C, Galn-Garca P, Santamara-Ibirika A, Bringas PG (2014) Twitter content-based spam filtering. In: International joint conference SOCO13-CISIS13-ICEUTE13. Springer, pp 449–458

Thomas K, Grier C, Song D, Paxson V (2011) Suspended accounts in retrospect: an analysis of twitter spam. In: Proceedings of the 2011 ACM SIGCOMM conference on internet measurement conference, ACM, pp 243–258

Thomas K, McCoy D, Grier C, Kolcz A, Paxson V (2013) Trafficking fraudulent accounts: the role of the underground market in twitter spam and abuse. Presented as part of the 22nd USENIX security symposium (USENIX Security 13). USENIX, Washington, D.C., pp 195–210

Titov I, McDonald RT (2008) A joint model of text and aspect ratings for sentiment summarization. In: ACL, Citeseer, vol. 8, pp 308–316

Wang AH (2010) Don't follow me: spam detection in twitter. In: Security and cryptography (SECRYPT), proceedings of the 2010 international conference on, IEEE, pp 1–10

Wang C, Blei DM (2011) Collaborative topic modeling for recommending scientific articles. In: Proceedings of the 17th ACM SIGKDD international conference on knowledge discovery and data mining, ACM, pp 448–456

Wei W, Carley K (2014) Real time closeness and betweenness centrality calculations on streaming network data.

Wei W, Carley KM (2015) Measuring temporal patterns in dynamic social networks. ACM Trans Knowl Discov Data (TKDD) 10(1):1–27. doi:10.1145/2749465

Wei W, Joseph K, Liu H, Carley KM (2015a) The fragility of twitter social networks against suspended users. In: Proceedings of the 2015 IEEE/ACM international conference on advances in social networks analysis and mining 2015, ACM, pp 9–16

Wei W, Joseph K, Lo W, Carley KM (2015b) A bayesian graphical model to discover latent events from twitter. In: Ninth international AAAI conference on web and social media

Wei W, Pfeffer J, Reminga J, Carley KM (2011) Handling weighted, asymmetric, self-looped, and disconnected networks in ora. Tech. rep., DTIC Document

Xia P, Jiang H, Wang X, Chen C, Liu B (2014) Predicting user replying behavior on a large online dating site. In: Proceedings of 8th international AAAI conference on weblogs and social media

Xia P, Liu B, Sun Y, Chen C (2015) Reciprocal recommendation system for online dating. arXiv preprint arXiv:150106247

Xie Y, Yu F, Achan K, Panigrahy R, Hulten G, Osipkov I (2008) Spamming botnets: signatures and characteristics. In: ACM SIGCOMM computer communication review, ACM 38:171–182

Xu R, Wunsch D et al (2005) Survey of clustering algorithms. Neural Netw IEEE Trans 16(3):645–678

Yin J, Ho Q, Xing EP (2013) A scalable approach to probabilistic latent space inference of large-scale networks. In: Advances in neural information processing systems, pp 422–430

Yuan J, Zheng Y, Xie X (2012) Discovering regions of different functions in a city using human mobility and pois. In: Proceedings of the 18th ACM SIGKDD international conference on kowledge discovery and data mining, ACM, pp 186–194