f i ® s t m ¤ ñ d @ ¥
PEER-REVIEWED JOURNAL ON THE INTERNET

Hacktivism and distributed hashtag spoiling on Twitter: Tales of the #IranTalks

by Mahdi M. Najafabadi and Robert J. Domanski

Abstract

The landmark Joint Comprehensive Plan of Action (JCPOA) was the final agreement of a series of tense nuclear negotiations between Iran and EU3+3 countries that started from September 2013, after which the Iranian people had elected a new president, and finalized in 2015. During several rounds of these negotiations, we noticed that some Twitter users were seemingly trying to distract people from the flow of latest news about the most trending negotiation's hashtag, "#IranTalks", by posting irrelevant tweets at a high frequency, that could be categorized as 'spam'. We collected a sampling of all the tweets that contained the #IranTalks hashtag, and marked the distracting tweets based on some criteria. We populated a list of the spammers' accounts and extracted their one-on-one friendship relationship (following/followed by). We applied social and organizational network analysis techniques and found strong evidence for the existence of an organization through which the accounts responsible for such tweeting behavior are connected. We believe the results of this study can stimulate more research about this social and organizational phenomenon and its possible impacts, and can help in better understanding and more accurate analysis of social trends on social media platforms.

Contents

---

1. Introduction

Twitter is one of the most popular social media and social networking platforms that engages users in several ways. Users can share their stories (mainly some information about the event or otherwise users' reactions to it) in up-to-the-second microblogs — *i.e.*, 'tweets'. Tweets can include links, images, and location information, and users can make a tweet viral by marking it 'favorite' and/or 'retweeting' it in a crowdsourced manner. Users can also include one or more relevant keywords — *i.e.*, 'hashtags' (#) — to participate in and take advantage of an automatic indexing mechanism, that makes their tweets more visibility and discoverability by other users. If employed properly, this mechanism helps reporters promote and information consumers seek the latest news, progress, and opinions, per desired topic, around the world, in a matter of seconds.

Twitter users can follow one another to receive each other's tweets on their 'timeline', and can block certain users to disallow them from following their account. Another way to seek tweets by the users is to employ the search mechanism to look for one or more desired hashtags or keywords, that results in a list of most popular and recent tweets that contained the search term. Topics that are viral among a large number of users will have 'trending' hashtags. For every trending hashtag, many users are posting and are viewing the corresponding tweets. If the hashtag is among the most popular hashtags in a certain time, it is displayed to all users who lookup trending hashtags. Thus, tweets that contain trending hashtags receive more attention from a large audience, just because of employing the right hashtag(s) at the right time.

However, users have no restrictions on how to use hashtags, including popular hashtags, for any tweet, even if the tweet is not relevant to the hashtag(s) used. This leaves room for user to exploit trending hashtags for promoting totally irrelevant tweets (about issues that do not have connections to the issues corresponding to the hashtag in the first place). In most of these cases, the users committing this action are considered 'spammers'.

Employing an irrelevant or less relevant hashtag in a tweet can be have variety of motivations such as commercial advertisement of a business, supporting a political rally, or sometimes increasing social awareness around a (forgotten or diminishing) topic in the public sphere, and can be done either as an individual or otherwise as a collaborative effort. Such collaborative efforts sometimes try to 'hijack' a popular hashtag for a different — or maybe less relevant — cause (Campbell, 2013). These efforts in which certain technical capabilities are used without directly violating or breaking any policies or regulations, in an unconventional and unexpected way (compared to the generally accepted ways of using those capabilities), is known as 'hacktivism'. To give their voice leverage, hacktivists sometimes use pieces of software known as 'bots' to post tweets at a higher frequency and from several different 'distributed' accounts [1].

During the negotiations between Iran and the world powers (known as EU3+3) [2], we noticed a situation in which irrelevant tweets were posted by some real users, in a manner different from all categories above.

The pattern we recorded suggests an organized distributed effort mostly to distract users from receiving latest news and reactions. This could not be categorized as a collective hashtag hijacking effort because the spammers were not communicating a single message (although some of the recorded tweets can be categorized as such), nor as a hacktivist bot-based attack because there were indications that the user account are in control of real people. The fundamental difference here was the evidence of a real organized, coordinated effort between users engaging in this type of behavior, which suggested a possible organizational mandate. In fact, several different messages were sent at a very high frequency that would result in 'hashtag spoiling' (Najafabadi, 2017) — a situation in which the real news stories and reactions regarding a viral hashtag are not easy to follow by the mainstream users, because of the high frequency of repeated, irrelevant, outdated, or even false stories posted in tweets that contain the same hashtag and dominates the hashtag search results. This practice of hashtag spoiling, done by real users who are already connected through a traditional organization, with the goal of user distraction on a mainly political issue is, we would argue, a new form of hacktivism.

1.1. Research overview

To examine this initial thought, we recorded sample sets of the tweets containing the '#IranTalks' hashtag during a few weeks at the peak of the negotiations, mostly when the #IranTalks hashtag was trending. These tweets were supposed to be about the nuclear negotiations. We analyzed the contents of a simple random sample of the randomly recorded tweets and marked the ones irrelevant to the ongoing nuclear talks as 'spam' by human judgement. We populated a list of active 'spammers', that

contained all the users responsible for five or more spams. We employed social and organizational network analysis (SNA) techniques to check upon meaningful signs of relationships between those spammers — *i.e.*, a network structure that could signal a corresponding physical organization of the spammers — and made an assertion about the possible motivations for this organized behavior.

Table 1 summarizes the research steps, and the instruments and outcomes per step.

We finish the paper by listing some other possibilities for further analyzing social network actors, and by explaining the possible impact of this phenomenon on analyzing social media — here, Twitter — data.

Table 1: Research steps, and their instruments and outcomes.

Step	Activity	Tool(s)	Outcome
1	We collected sample Twitter data containing the #IranTalks hashtag (in decks of the most recent 100 tweets) at intervals of every few hours during peak times of the #IranTalks. We repeated this procedure 100 times to end up with 100 twitter decks, each containing 100 tweets and its corresponding metadata. (We made simple exclusions in the query to exclude simple and detectable bots' activity.)	Twitter API (Python code)	Initially sampled tweet decks
2	We randomly selected 25 out of 100 decks of tweets (each deck containing 100 actual tweets) for coding and further analysis.	Simple random sampling	Confirmed sampled tweet decks
3	We coded the tweets in each sampled deck based on their relevance to the ongoing nuclear negotiations (0 = relevant; 1 = irrelevant). To minimize human bias, we only included users with more than five spam tweets on our spammers list populated from all our sampled decks (an initial list of 57 users was formed).	Manual coding	List of twitter users with identified spamming activity

4	We extracted one-on-one Twitter relationships (follows and followed by) for every possible pair of the users in our spammers list, and formed a ‘friendship matrix’.	Python code	Friendship matrix
5	Using social and organizational network analysis techniques, we identified, analyzed, and visualized some network characteristics of the spammers network. The findings led to conclusions about possible existence of an organization, and a possible mandate of users’ activity on Twitter regarding the #IranTalks.	Social network analysis	Assertions about the organization of the spammers
6	We took a retrospective look at the un-sampled tweet decks to see what information we can find about our top spammers, to see whether this information is also a match with our findings from previous step or not.	Simple user detection	Added insights about the spammers activity

2. Research question and significance

As explained earlier, during some rounds of nuclear negotiations, we noticed a situation in which some Twitter users were seemingly trying to distract other Twitter users from the flow of latest news about and reactions to the most trending negotiation’s hashtag, #IranTalks. We focused on discovering possible indications of an organized distributed effort to misuse a trending — popular — hashtag, or ‘hashtag spoiling’, targeting the real-time news coverage of the nuclear negotiations. We checked the accounts of a few of these users and found some similarities in their demographics, and their tweeting behavior.

Besides these similarities, the way these users exploited the ‘#IranTalks’ hashtag to post their tweets was abnormal — very frequent tweets, or repetition of tweets by

others (and not retweeting the exact same content from other similar users), less relevant or irrelevant, and outdated or even tweets that contained fake news. We were interested to find out if this was an individual behavior by some random users who just feel that they should engage in such activity to amplify their voice by hijacking the #IranTalks hashtag, or if it was an organized effort launched by an existing institution in the physical world.

Our specific research question in this study was: Are there evidence that there is an organization (or a number of organizations) through which these specific spammers (who are showing this type of unique spamming behavior) are connected? We believed that extracting Twitter relationships between the user accounts who followed this pattern of tweeting could help us in finding an answer to this question.

This research can help us understand the situations in which this behavior is likely to occur, and allows us to recognize it when it takes place, and depending on the situation, make sure that we would account for it when we are conducting social network analysis. This study can generate more insights about this type of behavior's short-term and long-term repercussions, and can possibly help evolve more effective counter mechanisms, enhanced by social media platform algorithms and by participation of social media users individually or collectively. In other words, this study can increase individual users' awareness and inform the user community to keep away from such distraction and keep up with the real flow of an under-attack hashtag, and even further, help with some crowdsourced counter mechanisms.

3. Reviewing the literature and background

In the following subsections, we provide a twofold literature review. The first part covers modern hacktivism in the age of Internet and social media and a background on the various ways in which specific cyberspace can be misappropriated to advance political agendas. The second part explains the history and sociopolitical backgrounds in which the nuclear negotiations took place.

3.1. Hacktivism: Political or ideological motivations for misuse

Social media platforms are now employed by politicians and by officials at different levels of governments, as well as by individuals, businesses, and social activists as an enabler for their social outreach, for promoting, supporting, or opposing a cause or an idea (Sandoval-Almazan and Gil-Garcia, 2014; Jungherr, 2016; Khazraee and Losey,

2016; Rossini, *et al.*, 2017; DePaula, *et al.*, 2018). In some cases, different methods of 'political spamming' is used as an opposing digital action. Political and ideological spams are mostly aimed to promote a perspective or belief, or otherwise, as we observed in this study, distract users from following the normal flow of tweets. In fact, social media is known to have an effect on the evolution of a user's opinions (He, *et al.*, 2017) and thus political causes and figures try to shape or shift public opinion for or against an issue (Dincelli, *et al.*, 2016). Although affecting public opinion is a normal activity in a democratic society, filling the public space with false news is in most cases prohibited by domestic and national laws and regulations and is considered illegal, and in many instance, unethical. Considering the recent findings about how some thousand organized social media pages or bots actively have spread false news or fake opinions to some targeted audiences, for example in the 2016 U.S. elections (Shane and Goel, 2017; Entous, *et al.*, 2017; Bastos and Mercea, 2017), reveals the importance of studying these types of user behavior in social media, and its effects.

As illustrated in Figure 1, there are various ways in which people seek to respond to or protest against their ideological opponents in cyberspace.

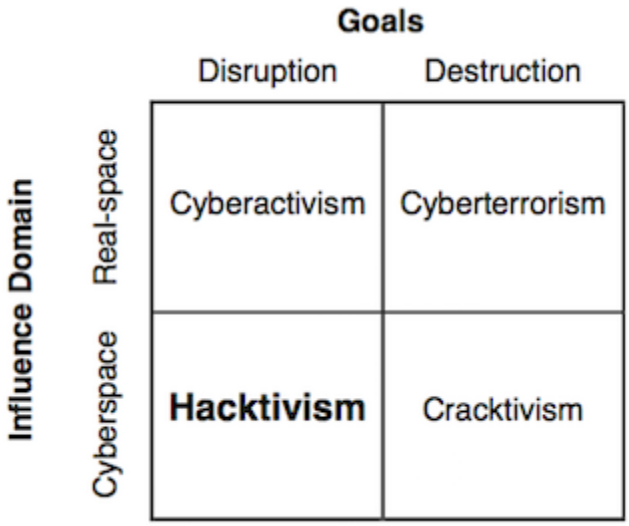

Figure 1: Spectrum of cyber-spatial activities (Domanski, 2015).

Cyberactivism refers to the use of the Internet and social media tools to communicate with others in support of or in opposition to a cause (Jordan, 1999). Examples include online petitions, organized information campaigns in social media forums like Facebook groups, Web sites that seek to disseminate information about a cause or an upcoming event, general recruitment to an activist organization, etc. These are all legally and constitutionally accepted forms of political activism whose implementations have simply migrated to cyberspace. They might seek to disrupt the sociopolitical status quo, but do so legally and with an eye towards organizing and mobilizing people to collective action.

Cracktivism, on the other hand, is where the boundary is crossed from legal acts of political expression over to illicit and criminal actions. It refers to unauthorized intrusions of private software systems, defacement, the destruction and/or theft of information, and more. While admittedly often still motivated by political ideology, cracktivism does value destruction as its goal. If sponsored by governmental actors, it then falls into the classification of cyberwarfare (Rosenzweig, 1998; Curran and Gibson, 2013).

Hacktivism, by contrast, refers to a similar type of largely benign, but disruptive, not destructive, form of political expression. Acts of hacktivism, by definition, do not seek to destroy their targets. It is the act of computer hacking for political purposes and stems from the original meaning of the term ‘hacking’ — utilized since the 1970s — meaning to make use of something for a purpose other than for which it was originally intended (Jordan and Taylor, 2004).

There is an important distinction between hacking and hacktivism. Hacking is a clever use of technology for virtually any purpose, often self-interested (Bell, 2001). Hacktivism, on the other hand, is a clever use of technology specifically to express oneself politically. It can be thought of as a form of political speech (Coleman, 2013), often administered through code and automation, designed as an unexpected exploitation of technological capabilities.

Hacktivism has a history reaching back at least to the early 2000s (Morozov, 2011) and its current incarnations include relatively infamous collectives such as Anonymous and 4Chan (Coleman, 2015). An early prominent example of hacktivism was the introduction of the ‘Googlebomb’ [3]. and some unconventional exploitation of the Google’s PageRank algorithm to effectively manipulate the search results relating to specific political figures. For example, beginning in 2003, a Googlebomb made it so that whenever a user searched for the phrase “miserable failure”, a biography of George W. Bush was displayed as the first result (Cohen, 2007). Similarly, in January 2009, before even taking the oath of office, opponents of President-Elect Barack Obama made it so that he sat atop the search results for

"failure" while, simultaneously, his supporters produced the same effect for "cheerful achievement" (Taylor, 2009). There are several more examples of such activity (Havenstein, 2008; Amira, 2011).

Hacktivism also includes misusing some mechanisms that are in place to avoid this type of behavior. For instance, by using a social media platform's crowdsourced 'report' mechanism which is originally to help stopping users from violating laws and acceptable use policies, some accounts might be attacked with false reports, and thus might be suspended, at least temporarily. There was an instance of this in the Democratic Party's 2016 presidential primaries, where Hillary Clinton's campaign was accused of buying social media trolls (May, 2016) and made efforts to shut down some of Bernie Sanders' accounts on social media using this technique (Taylor, 2016; Zimmerman, 2016).

The legality of hacktivism sometimes finds itself in a gray zone, as such acts are not overtly criminal by nature (Domanski, 2015). Rather, like cyberactivism, they seek to disrupt the political status quo through expression, often with the ultimate goal of gaining mainstream attention to their advocated position on an issue. Nevertheless, such activities raise numerous problematic challenges for regulators (Spinello, 2002). The difference between cyberactivism and hacktivism is that, whereas the targets of a cyberactivist lie in real space, for example in an online petition to change a local school board ordinance, environmental regulation, or some other type of public policy, the targets of hacktivism exist in cyberspace. For example, in redirecting visitors trying to reach the White House Web site to a different site expressing opposition to the President's position on an issue, no actual damage is being done to the White House site, no information is being stolen, and their administrators will surely 'fix' the redirect in a brief amount of time. However, the story would (hopefully, in the hacktivist's mind) garner media and public attention to the cause he or she was trying to advocate for, and might ultimately have repercussions in real space politics.

Individuals and different groups of people (*e.g.*, people being affected by a policy in a certain part of a country or in a certain industry, or likewise) can have political and ideological motivations for hacktivism. They sometimes find themselves collaborating in an act of hacktivism due to popularity of a cause or purpose. In some instances of the group action however, acts of hacktivism is performed by a group of people are already connected and organized through an existing institutional mechanism in the real space. In this case, the users can simply organize their online activities — *e.g.*, spamming — for greater effectiveness, but most often their activities will contain enough footprints that would indicate the existence of such organization.

3.2. Related work

Political spamming has been a hot research topic recently (see several related references in the previous section and in the current section).

Previous studies have reported instances of hacktivism in which social media capacities are used in large scale in an unexpected way (Taylor, 2005; Hampson, 2012; Deseriis, 2017; Solomon, 2017). There are also many studies that include cases in which spamming is performed by programmed software bots (Sedhai and Sun, 2015; Deseriis, 2017; Bessi and Ferrara, 2016; Ferrara, *et al.*, 2016; Woolley, 2016; Suárez-Serrato, *et al.*, 2016; Stukal, *et al.*, 2017), or by real people (Irani, *et al.*, 2010; Song, *et al.*, 2011; Thomas, *et al.*, 2011; Lee, *et al.*, 2012; Hadgu, *et al.*, 2013; Yang, *et al.*, 2014; Almaatouq, *et al.*, 2014; Wu, *et al.*, 2017), often organized by an entity such as a government, or a social or political group (Qtiesh, 2011; Emspak, 2011; Finley, 2015; King, *et al.*, 2017; Stukal, *et al.*, 2017; Leonnig, *et al.*, 2017; Finley, 2015).

One popular way of propagating spam in Twitter, is by misusing a 'hashtag' that is already designated to a viral conversation. This is mostly done by including the hashtag, in a high volume of tweets that are irrelevant to the corresponding conversation. Although some spammers try to 'hijack' the 'trending' hashtags to communicate their own message to the users who are engaged with the original conversation, some other spammers (mostly with political aims) do not necessarily communicate a specific message in their spams, and they only want to distract users from following that conversation. This type of spamming would result in disrupting the formation of social conversations (around that issue), interfering with the normal shaping of public opinion, and deflecting public opinion by using propaganda to amplify a point of view that is not prolifically supported, and at the bottom line, affecting people's choices based on the wrong perceptions they would receive from spammers' activities.

In different studies, those instances are referred to as 'hashtag spamming' (Hyun and Kim, 2016; Sedhai and Sun, 2015), 'hashtag hijacking' (or 'hashjacking' for short) (Hadgu, *et al.*, 2013; Jackson and Foucault Welles, 2015; Virmani, *et al.*, 2017), or 'trend stuffing' (Irani, *et al.*, 2010), based on the details of the case study. In our case though, the characteristics are slightly different: It is done in a distributed and organized manner, it is performed by real people who (as we will see) are connected through a traditional organization which is not a government, and their aim seems not to be promoting a unique message or Web site, but to distract the public and interfere with the normal shaping of public opinion. We call this specific type of spamming as 'hashtag spoiling' (Najafabadi, 2017) [4].

3.3. The sociopolitical grounds of the phenomenon under study

After the Iranian people elected a new president in 2013, within a matter of months, the long-halted negotiations restarted between Iran and the EU3+3 countries and escalated to the 'ministry' level. These negotiations eventually created "the longest and most complex international arms-control agreement ever, at 159 pages" (Porter, 2015). It took almost two years to achieve a final agreement known as the Joint Comprehensive Plan of Action (JCPOA), which was backed by a United Nations Security Council resolution [5]. It imposed major restrictions on the Iranian nuclear program and put Iranian nuclear facilities under strict and thorough surveillance. In return, the deal mandates that the international community would accept a safe level of uranium enrichment on Iranian soil for civilian and research purposes, and would lift (and ultimately revoke) all nuclear-related embargoes, sanctions, and corresponding restrictive regulations [6].

This was the result of several rounds of tense negotiations closely followed by the media, journalists, advocacy groups and lobbyists, investors, researchers, and also individual people. Twitter was a popular tool for these stakeholders to keep up with the most recent developments in the talks. In each round of the negotiations, formal or leaked news about a possible failure or success of the negotiations was highly impactful on the popularity of the #IranTalks (and some other relevant) hashtag(s) [7].

During the two years of negotiations, some external entities tried to disrupt the progress, because any kind of agreement that would end the isolation of Iran from the global market — regardless of the terms of the agreement — would have been counted a game-changing loss for the entities benefiting from the status quo. These entities form a wide spectrum ranging from some countries in the region and in the world, to some powerful lobbies and influential groups both in Iran and the U.S. Newspaper/Internet/TV ads, significant investments in lobbying efforts, several rounds of meetings between high officials of the countries, public speeches, and the propagation of falsified information, all were among the techniques used. However, the negotiations finally produced the JCPOA as an agreed upon roadmap for all negotiating parties. Given this atmosphere, it was not unexpected to find that there were some politically motivated organized efforts to misuse Twitter capabilities to hijack negotiation's hashtag(s).

4. Research method

Politically motivated activities that fit into hacktivism and similar categories are more likely to be committed by groups of people and thus can be studied for 'organized'

behavior. Thus, social and organizational networks analysis (SNA) tools and techniques can help us reveal some information about the users (hereafter, actors) based on their position in the identified network. In Twitter — as well as many other social media platforms in which users can post almost anything at their discretion — the complexities and conflicts of information rights, user rights, and acceptable use policies makes it hard to distinguish between normal activity and spam. This is because some users might see a tweet as a spam (*i.e.*, irrelevant or less relevant) to the employed hashtag(s), and some may not, and even more, some might even believe that the distraction goal and its repercussions on public opinion and awareness are always relevant. Thus, categorizing tweets that contain a specific hashtag as spam would be more valid if more distinctive criteria were employed, although any criteria would still be debatable. Thus, we tried to employ techniques to make sure that we identified spammers in a way that was acceptable by a larger community of Twitter users (see subsection 4.5. Inclusion criteria).

As illustrated earlier in Table 1, we wrote a Python script and used it to collect a sample of #IranTalks tweets on occasions when the hashtag was highly used in tweets. We coded a sample of the tweets and identified some of them as spam in cases where we found a disconnect between the content of the tweet and the #IranTalks hashtag. Next, we populated a list of the user accounts responsible for five or more spams in at least one dataset (tweet deck). For the second round of data collection, we evaluated the Twitter relationship (following and followed by) for all pairs of user accounts in our spammers list from the previous step, and formed a 'friendship matrix'. Finally, we used the UCINET software (v. 6.611) to analyze this network. We have highlighted our major findings of this network assessment below and also in the Findings section.

4.1. Preparation and logistics

To avoid biases from Twitter algorithms that favor some tweets based on users' historical activity patterns, we created a new Twitter account and made some anonymity precautions to avoid any of such heuristics tailoring and filtering algorithms applied to our search results. We also used a new workstation and Internet connection for account creation while no other Twitter login session was active, to make sure Twitter would not bind this new account to any existing account behavior.

4.2. Round 1 of data collection

Round 1 data collection for this study was a two-step process.

We programmed and used a Python script to collect data directly from Twitter using a Twitter application programming interface (API), facilitated by a public Python

library. We executed this Python script in certain intervals (*i.e.*, every two to four hours) at times when the negotiations were active and the #IranTalks hashtag was at the peak of its use by many Twitter users worldwide. We recorded one hundred tweets per run — the maximum that the Twitter API allows us to collect in every attempt free of charge. The data collected contained the body (*i.e.*, text) of the tweets plus some metadata such as timestamps, number of followers of the sender of the tweets, date of account creation, and in some cases even more information like geographical location (based on user preferences and privacy settings of the corresponding user account). Upon fetching a tweet deck, the Python code also parsed the tweets, and exported selected elements we deemed important for our analysis into MS Excel worksheets (one Excel worksheet per tweet deck or dataset).

4.3. Sampling

As explained earlier, a part of the sampling was done by Twitter, due to the Twitter API's 'rate limit' on free Twitter data. This API returns 100 tweets per call, which seemed sufficicent for this study. After it reached its predefined maximum capacity (in most cases, 15 API function calls per user), it stopped for a 15-minute window before it functioned again. Each of the collected tweet decks was stored in a separate Excel file. We used simple random sampling and selected 25 out of 100 for this study.

4.4. Coding

We used human judgment to identify spam tweets. We marked a tweet as spam if it did not contain something directly relevant to the nuclear talks between EU3+3 and Iran. These criteria were judged based on familiarity with the issue. We did not mark a tweet as spam unless there was strong evidence for it based on our criteria both in content and tweeting/retweeting behavior, *e.g.*, false information, outdated posts, tweets containing links to non-credible Web sites, and very high frequency of tweets.

4.5. Inclusion criteria

To make sure we were focusing on certain spamming behavior, we defined another inclusion criterion for coded datasets. We decided to only include datasets that contained at least one user with five or more tweets coded as spam. Based on this new criterion, seven of the datasets did not meet inclusion criteria at this stage and the remaining 18 other datasets met the criteria.

4.6. Data cleansing

In the middle of the data collection phase, we noticed some efforts to hijack the #IranTalks hashtag. These efforts were conducted by different interest groups and

were also facilitated by some Twitter bots. Among the interest groups we detected were some pro-Iran bot spams, pro-Saudi bot spams, and some bot spams and real-person spams from within U.S. political coalitions.

In the middle of the data collection stage, for example, some pro-Saudi spammers bots started to use the #IranTalks hashtag to propagate supportive messages about the war in Yemen conducted by the coalition of Kingdom of Saudi Arabia (KSA) and its allies. From the frequency of the tweets and the similarity of the contents (although they were not retweets of each other) and some other footprints, it was clear that they were generated and posted by bots, not by real people.

The focus of this study was to detect organizational characteristics of real users with prearranged mandates on misusing Twitter capabilities. Therefore, we intentionally excluded the bots' activities. We could exclude most of the bots' activities with just a little twist to the original search query by which we fetched tweet samples. All the bots we detected were promoting specific hashtags, which we employed as a filtering mechanism to exclude bots from our initial data collection query. Despite these precautions, some of these bot-propagated tweets passed through our filters and ended up in the sampled datasets, but not at a large scale.

In addition, we realized that in some of our sampled datasets, there were human spammers with different interests and patterns of activity — *e.g.*, business interests — than the ones who were the subject of this study. Human spammers with different interests were almost impossible to exclude from recorded data, but they were usually posting spam tweets at a much lower rate, not affecting the results of our analysis. Here, in most instances, these spammers were detected with only one spam in a 100-tweet dataset, which increases the chances that it was an individual user. Our final inclusion criterion excluded such individuals — *i.e.*, they did not appear in our final list of spammers. We also noticed that two of the sampled tweet decks had a time overlap with each other, but since the identification of the spammers was done based on each tweet deck, it did not violate the integrity of our method.

Before the analysis, we removed four isolated actors which did not seem a part of this spammer network. However, there could be other possibilities: 1) A false positive in our spam criteria; or, 2) Sampling error which implies that our sampling technique did not capture spammer actors who could bridge a relationship between the main network and the isolates (*i.e.*, corresponding dyadic connections were absent only in our sample).

4.7. Identifying most active spammers

To decrease the likelihood of false positive spam detection, we defined a new criterion for counting an account as a spammer: each spammer had to have five or more tweets marked as spam in any of our sample datasets of 100 tweets. Thus, any user with four or fewer recorded spam tweets per dataset were not counted as a spammer. These criteria captured those actors that were doing a recognizable spamming job, which made our analysis more robust. This also made it more suitable for our SNA at this stage, since we were trying to analyze the network of the spammers at the dyadic level. Finally, we created one list of all the accounts responsible for five or more spam tweets in each of our 25 datasets.

4.8. Round 2 of data collection: Extracting the friendship matrix

We used another Python script to extract Twitter friendship relationships (following/followed by) between every pair of actors in our spammer lists. This formed a bidirectional social network which enabled us to employ social network analysis (SNA) tools and techniques to acquire more information about the network. The friendship data was extracted from Twitter in spring 2016, based on the list of spammers that was populated in winter and spring 2015 rounds of data collection.

4.9. Social network analysis

We employed social network analysis (SNA) concepts to find meaningful insights about our hypothesis, *i.e.*, existence of the spammers' organization based on our sample data. In our analysis, we exclusively relied on Twitter data. Knowing nothing about the true relationships, while trying to come up with assertions solely based on the data, we used a black box analysis. We started by using simple network characteristics, then we identified the most influential network actors (based on the Twitter data we extracted), and then we tried to identify some cohesive subgroups in the network, which basically showed subgroups of the network who were very well connected to each other.

4.10. Final round of analysis: A retrospective look at the unsampled datasets

Although the original Twitter data was collected in the spring of 2015 and the friendship matrices were extracted in the spring of 2016, we decided to conduct a retrospective analysis in the datasets that were not selected in our initial sample for the study. Thus, having populated the spammer list from our sample, we checked the rest of the datasets (the unselected tweet decks) to find if the identified spammers appeared in some tweets. There was no coding involved in this step, and we simply counted the number of appearances of each of the spamming user accounts, to find if there was a meaningful pattern of activity related to our research.

5. Findings

To have an assertion about the organization of user accounts, we considered several aspects of user account characteristics and patterns of behavior. We documented patterns of account activities and demographics, and employed SNA techniques to find meaningful clues about an organization network related to spamming activities.

5.1.Data description and statistics

As illustrated in Figure 2, some of the sampled tweet decks have significantly shorter durations (vertical timescale equaled one hour). This was due to the fact that the Twitter API collects the last 100 tweets at each call. Thus, in peak times of the #IranTalks hashtag to collect the most recent 100 tweets required going back to a shorter period of time, because tweets were posted more frequently.

The data description and statistics — even before employing any SNA techniques — often can generate insights about the phenomenon. We divided these insights into two types: insights related to characteristics of the data, and insights generated after coding took place.

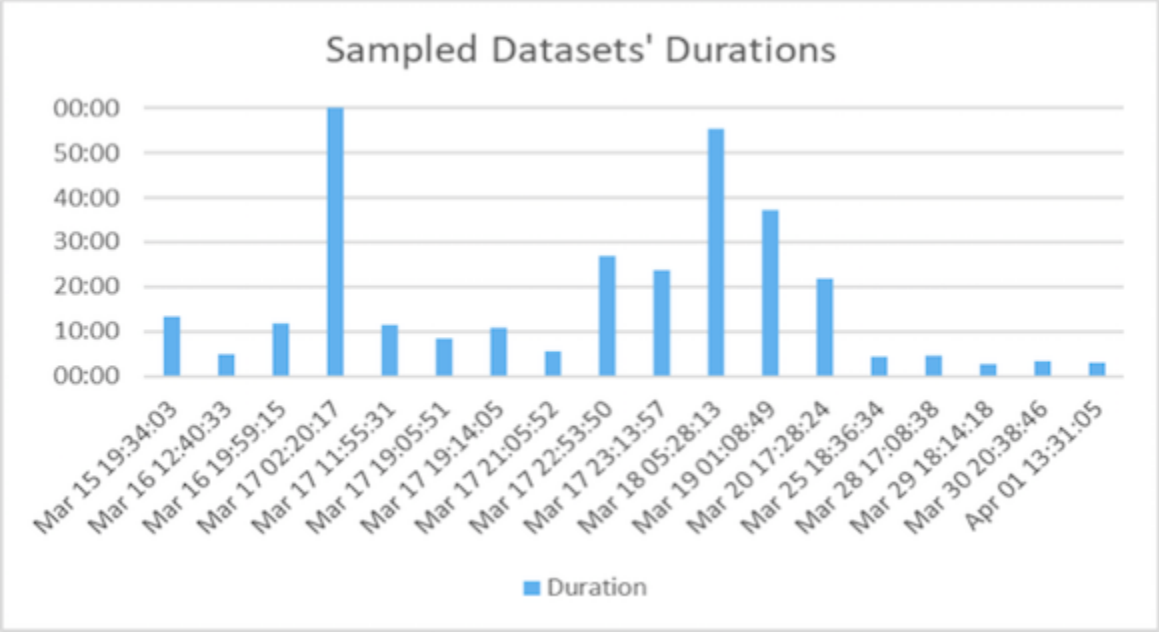

Figure 2: Sampled tweet decks timestamp and duration.

Note: Larger version of figure available here.

Even before we began coding, we found some signs of abnormal activity. For instance, regardless of the content of tweets, the volume of tweets and the high frequency of posting tweets using a specific hashtag (around one tweet every six seconds) could hardly be considered a normal user activity. As we found patterns of activities that did not fit into Twitter bot activities, we believe it was an effort to help with distraction of Twitter users from following the #IranTalks hashtag, as discussed earlier. For instance, in our sampled data, we found that certain Twitter accounts were bombarding the #IranTalks hashtag. Here are a few instances to support this assertion.

- A total of 43 tweets in 4 minutes and 53 seconds by one user (out of 100, dataset 02), of which 12 had met our inclusion criteria as spam;
- A total of 37 tweets in 4 minutes and 18 seconds by another user (out of 100, dataset 15), of which 16 had met our inclusion criteria as spam; and,
- A total of 38 tweets in 4 minutes and 38 seconds by another (out of 100, dataset 18), of which 35 had met our inclusion criteria as spam.

To come up with these numbers, we initially created a list of all identified spammers. This list contained 286 accounts that 1) had posted at least one spam tweet according to our inclusion criteria; 2) were responsible for 814 spam tweets; and, 3) comprised a total of 1,041 tweets (in the datasets they had at least one tweet labeled as spam). However, we discovered that 19 accounts out of 286 (less than seven percent) were responsible for roughly 50 percent of the tweets, and that five of these accounts appeared more than once (one account appeared three times) in our sampled datasets. On the demographics, the accounts of over 85 prcent of identified ‘active’ spammers were created either late 2013 or early 2014, the first months that the new Iranian government took office.

For each dataset (each file containing 100 tweets), once the coding of the data was completed, we created an overview table to summarize demographics and information about patterns of activity in that dataset: specifically, the identified spammer accounts’ names; total number of their tweets in the that dataset; and, number of tweets marked as spam. Table 2 is a sample of the overview tables that summarized a deck of 100 tweets that were sent in 4 minutes and 52 seconds. In this deck, the user in the first row posted 43 total tweets (average of one tweet per 6.79 seconds), 31 of which we coded as spam. This user and the user in the second row, accounted for 67

percent of the total #IranTalks tweets recorded in that short time period, of which 54 percent were coded as spam. The accounts were created on two consecutive days in November 2013. The overview tables revealed that some spammers posted tweets (either detected as spam or not) more often than others, at least in the recorded samples. Further assessment showed that most of the tweets by the spammers' accounts — even some that did not meet our inclusion criteria and thus were not marked as spam — were in fact negative posts opposing the Iranian government and/or against making progress in the talks.

Table 2: Overview table of a sample tweet deck.

User ID	# Tweets	# Spams	Followers	Friends	Account created on
karemym11	43	31	345	266	20 Dec 2013
amir43872139	26	23	231	570	19 Dec 2013
habibsaaid52	4	4	463	371	12 Nov 2013
			Total	1,039	1,207

Table 3 shows simple data description for the same deck of tweets as in Table 2. Looking at some of the numbers in the table — such as number of total tweets or the number of total spams by identified spammers, out of the total 100 tweets — provides insight suspicious or at least abnormal activities.

Table 3: Simple statistics of the sample deck of tweets from Table 2.

	# Tweets	# Spams	Followers	Friends
Total tweets accounted for by identified spammers	73 (out of 100)	58 (out of 100)	N/A	N/A
Standard deviation	15.97	11.32	94.72	126.07
Mean	24.33	19.33	346.33	402.33

The time of account creation for 100 percent of the accounts that were identified as spammers were between 2012 and 2014 and the number of their followers were in the order of tens and hundreds (only a few outliers exceeded one thousand followers).

Figure 3 illustrates the number of users recorded per sampled tweet deck vs. the number of active spammers identified. This figure suggests that as the number of total users in a dataset decreases, the portion of tweets by spammer accounts increases. In other words, it seems there was a base of spamming activity in each dataset, regardless of whether other users were tweeting about the hashtag or not.

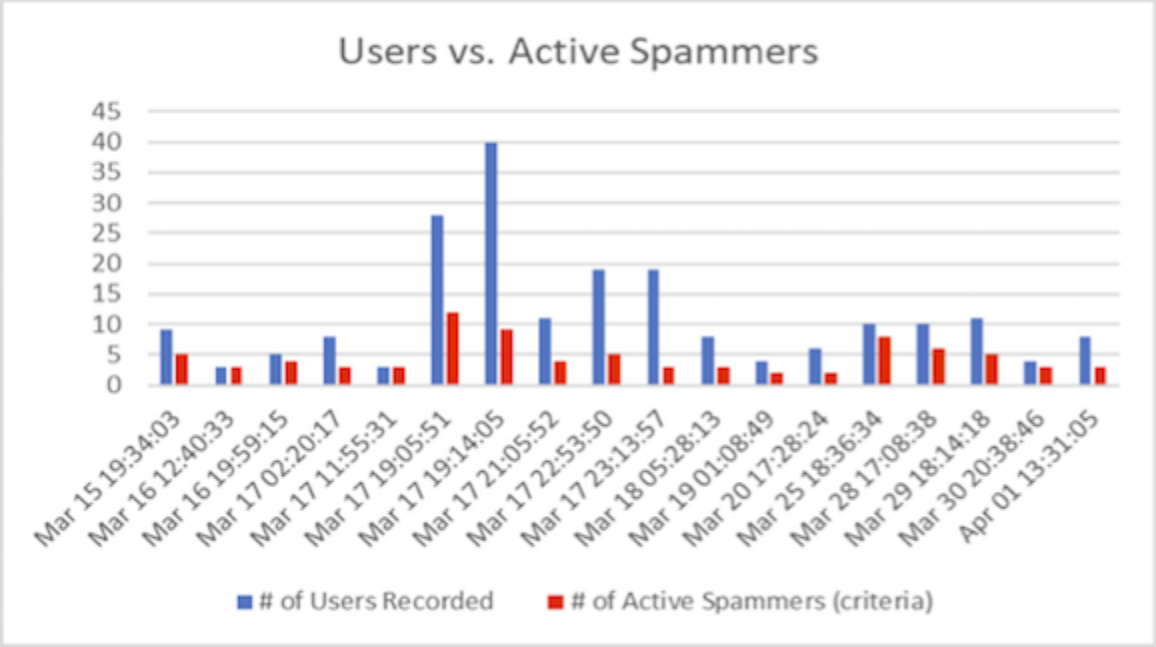

Figure 3: Number of users recorded per sampled tweet deck vs. the number of active spammers identified.

Note: Larger version of figure available here.

We also examined the time factor. Figure 3 shows that as the #IranTalks hashtag was used by more users in a shorter time period (*i.e.*, the hashtag was increasingly trending), the number of spammers was a smaller percentage of total users. In other words, the number of spammers did not increase proportionally to the number of the users in a dataset. This suggests a limited number of spamming accounts, and a

pattern that was different from general Twitter users reacting to the #IranTalks hashtag.

To better illustrate this, we plotted the percentage of spammers by the duration of each of our sampled datasets (Figure 4). It clearly suggests that as the duration of the dataset increases identified spammers were a smaller portion of the total users.

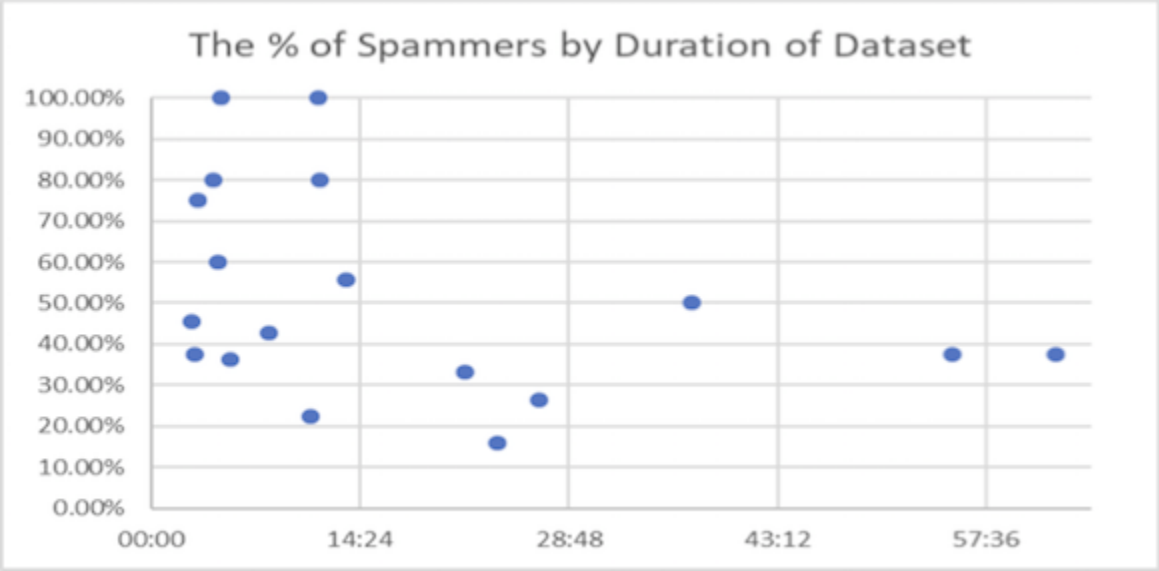

Figure 4: Percentage of spammers by duration of dataset.

Note: Larger version of figure available here.

Our analysis of the identified active spammers (the spammers who met our selection criteria), as illustrated in Figure 5, shows that given the number of spammers that account for a lot of tweeting activity, a high frequency tweeting pattern suggests spamming behavior. Comparing Figure 5 and Figure 2 gives a better understanding of the average time between each of the tweets, as explained in the sample data at the beginning of this section. To better illustrate this, we calculated the average number of seconds between each tweet of the most active spammers (Figure 5).

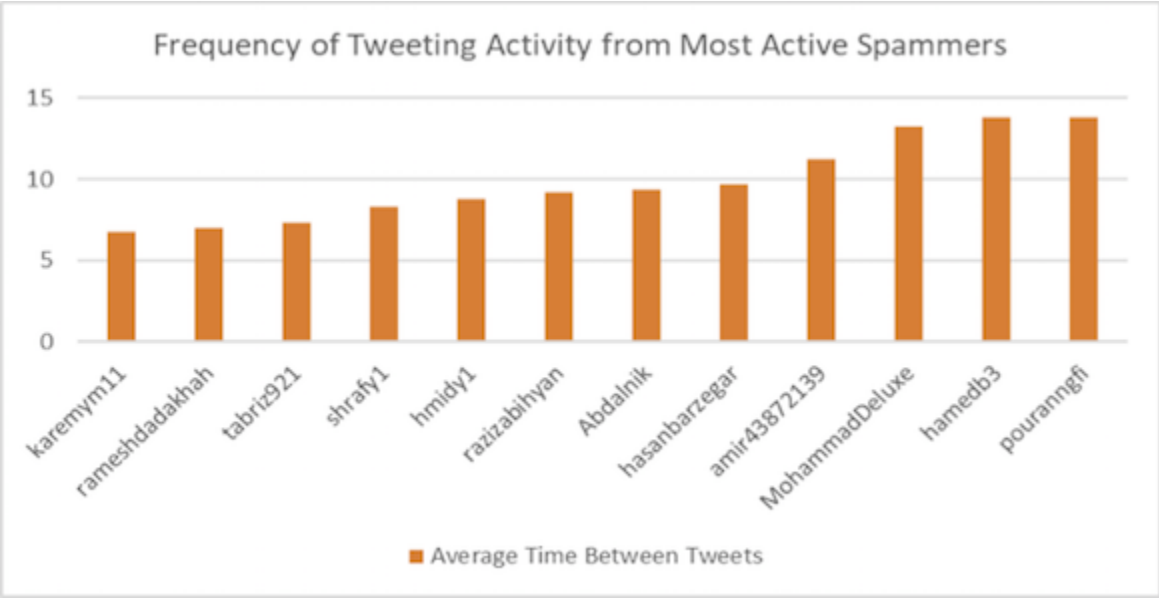

Figure 5: Frequency of tweeting activity from most active spammers.

Note: Larger version of figure available here.

As illustrated in Figure 5, the most active spammers were tweeting on average one tweet every seven to 14 seconds. That was only their recorded activity within a sampled dataset of 100 tweets, at a specific time in negotiations. Figure 6 shows the aggregate number of tweets and the number of tweets that were marked as spam, per identified active spammer in the sampled dataset. These numbers are clearly abnormal for a typical Twitter user. The number of identified spams, illustrated in Figure 7, also confirm this finding.

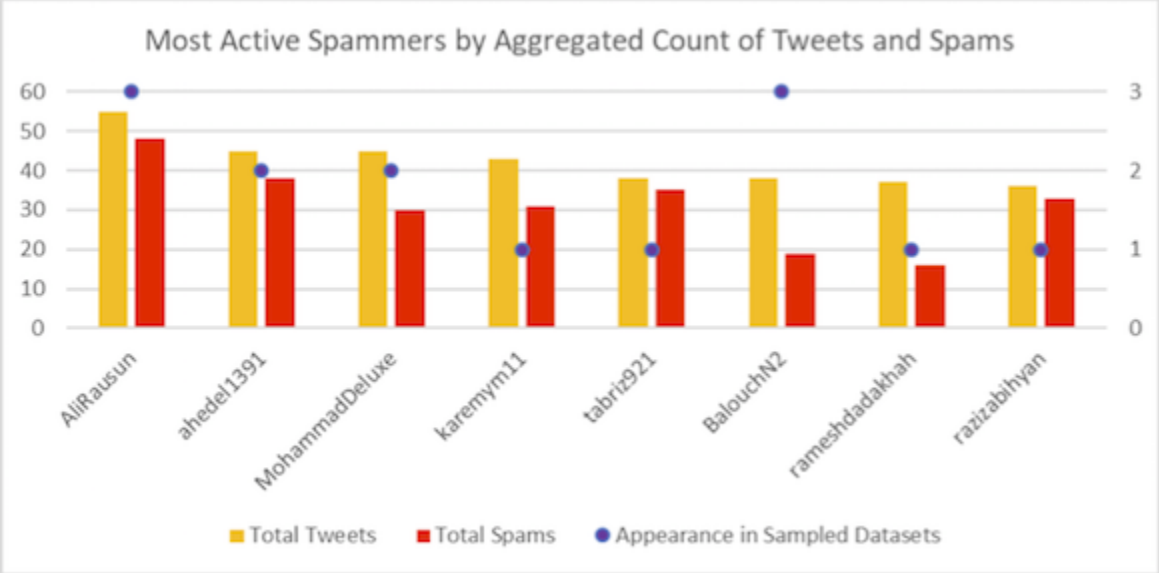

Figure 6: Most active spammers by aggregated count of tweets and spams.

Note: Larger version of figure available here.

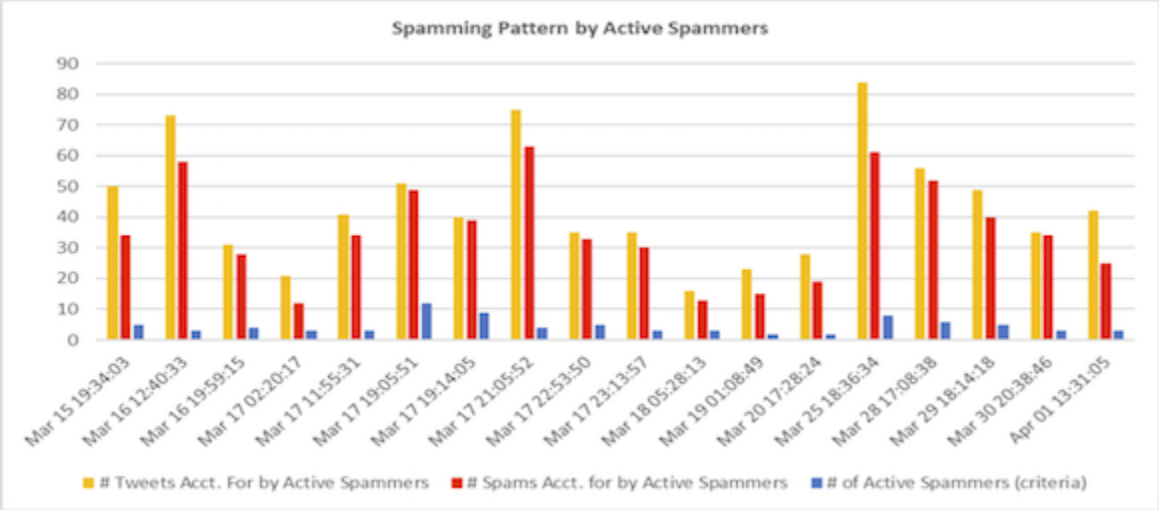

Figure 7: Spamming pattern by active spammers per sampled dataset.

Note: Larger version of figure available here.

We extracted tweeting patterns of identified spammers from all datasets. Although we had collected and selected a very sparse sample of datasets, three spammers appeared in three of the datasets and another 20 had appeared twice.

A deeper look at the contents of the tweets revealed that many of the spam tweets mentioned the Mojahedin-e-Khalgh organization (MKO or MEK) leader's Twitter handle in many of their tweets in an irrelevant manner to the negotiations (probably to make her more visible and give her credit for being mentioned several times, or maybe to report the progress they made toward an assigned duty).

These clues indicated that this tweeting behavior was probably part of an organizationally mandated agenda, most likely attributed to the members of the MEK, which also introduced themselves as the People's Mojahedin of Iran (PMOI), National Liberation Army of Iran (NLA), National Council of Resistance for Iran (NCRI), Muslim Iranian Student's Society, and even some more. This is the largest anti-Iranian terrorist organization, with a history that goes back to the time of the Shah in the 1960s (Mejdini, 2016; Pike, 2017) [8]. The U.S. had listed the MEK as a foreign terrorist organization on 8 October 1997 and removed it from the list on 28 September 2012 [9].

We found that they repeatedly retweeted or even copied their fellow tweets at a high frequency (copying the exact content of someone else's tweet on a large scale, instead of 'retweeting' it, is not considered a normal activity). They also included in their tweets, Web URLs [10] to one of their organizational Web sites, thus, bombarding the #IranTalks hashtag with tweets containing fake news and opinions unrelated to the subject.

5.2. Basic network characteristics: Network connectedness and geodesic distances

Our active spammers list of 57 users, contained only a few 'isolates', *i.e.*, actors who were not connected to the rest of the network. Best practice in social network analysis is to remove isolates if they are not a major part of the network. After elimination of the isolates, we had a directed (non-symmetric) friendship matrix corresponding to our list of active spammers, extracted from the 'following' (out-degree) and 'being followed by' (in-degree) relationships on Twitter.

The overall average degree for the actors was 17.49 (out of possible 52 connections per actor), which indicated a highly connected network with a normalized network density of 33.64 percent. This shows the percentage of all possible links in the network that actually exist; in other words, on average, the portion of all possible one-on-one friendship relationships that are in place. For interpreting this number in a directed graph, we should remember that each pair of actors can have two

relationships: actor A following actor B, and actor B following actor A. Each of these relationships either exists or is missing. In our data, 21 percent of the actors were fully connected (they followed each other) and 46.15 percent of actors had at least one form of direct connection (following, being followed by, or both), which we considered a significant finding for our hypothesis. This density indicated that almost one-third of all possible dyads were present in this network (927 ties out of the total 2,756 total possible dyads in our directed graph). This number clearly suggests a well-connected network. Another factor illustrating the connectedness of this network was the average distance of 1.68 which shows that every actor was within reach by 1.68 dyads (*i.e.*, edges, or ties between actors) on average from any point in the diagraph. The diameter of the network — *i.e.*, the longest path between any two actors in the network (after removing the isolates) — was only four edges. By looking at the geodesic distances matrix, it appears that except for one pair of actors, the diameter of the rest of the network is three. These findings also support the presence of a well connected network.

5.3. Centrality and prestige measures

There are several methods to calculate the centrality and prestige measures of a social network. Most often, the results of one method is congruent with other methods. For our application of centrality here on Twitter friendship relationships (in which network actors can follow each other asymmetrically), the degree centrality is a good fit. The basic idea is to find the most influential actors of the network, in terms of connections to other actors. Other forms of centrality calculations include betweenness centrality and closeness centrality. Since we are dealing with a directed graph corresponding to the Twitter friendship (following and followed by) of the spammers' network, we may also refer to the incoming centrality measures (here, the in-degree measure) as the 'prestige' of the actor, that is an indication of how popular an actor is among all other network members.

In our case, the degree centrality analysis depicted a few outstanding accounts, of which some had high out-degree 'dyads' to the network (following), some had high in-degree dyads from the network (being followed by), and some had both. For instance, the users 'BalouchN2' and 'IranArabSpring' had the highest in-degrees of 45 (86.5 percent) and 42 (80.8 percent) out of 52, which could be interpreted as accounts associated with more popularity, or those with a higher rank in an organizational order. The users 'no2censorship' and 'peymaneh123' each had 43 out-degrees (82.7 percent) out of 52 which could be interpreted as their extensive activity in following other accounts — which in some cases could persuade the followed account to 'follow back', helping to spread tweets of the original account. The accounts 'IranArabSpring' and 'peymaneh123' had the highest numbers for both in-

degree and out-degree centrality. Besides the degree centrality analysis, other popular centrality measures also illustrated very similar results.

Although the degree centralization is mostly calculated for individual actors within a network graph, the whole network can also be assigned a network centrality measure. This number is maximized in a star network (Wasserman and Faust, 1994), in which only one actor is connected directly to all other actors. For our network of identified spammers, the normalized degree centralization of the network was 51 percent which is significant and implies a connected set of actors. This suggests that no one actor completely dominated other actors in having direct connections to them.

5.4. Cohesive subgroups

Analyzing the cohesive subgroups helped us understand the internal structure of a network. The most cohesive subgroups — by using the 'cliques' procedure — returns no results in a directed graph. However, if we symmetrize the data (using the "Max Sym." procedure), there will be a huge number (218) of cliques of size 3 or more. We decided to increase the 'n' to find the largest cliques in the symmetrized network. We ended up with 18 as the size, which returned six cliques. In other words, there are at least six different sets of actors in which they were fully connected to each other, and each of these sets contained 18 actors (either by following, by being followed, or both, since we symmetrized the friendship matrix for this procedure). The 'clique participation scores' revealed that 13 actors were present in all these cliques. Not surprisingly, the actors already identified in the previous section in the centrality and prestige calculations also appeared in this list. Some of these 13 very well-connected actors also had a very high frequency of posting tweets in the recorded datasets, and six of them appeared more than once.

In our cohesive subgroup analysis over the directed graph, we also observed that there were three 2-cliques in the network (subgroups which only lack two dyads to become maximally connected). These numbers were a strong indication of some type of organizational relationships among the actors we identified as spammers. Another interpretation was that there is a set of very well connected — central — actors in the network, and that other actors in the network were connected to the rest of the network through them. The following core/periphery analysis also supported this interpretation.

5.5. Core/periphery analysis

The core/periphery analysis is a helpful tool that allows splitting network actors into two groups, of which, the 'core' actors have strong network ties and are considered central to the network, while the 'periphery' actors have fewer ties with the core, and

much fewer ties among themselves. In our spammers network, the core actors formed a subnetwork of 23 actors in which the density of the network was 77 percent (*i.e.*, 77 percent of all possible 'following' and 'followed by' ties in fact exist). Sixty-one percent of actors were fully connected (they both follow each other) and there was at least one form of direct connection between 93 percent of the actors. The density of the periphery network containing the rest of spammers (30 actors) was only seven percent. This could be an indication of a specific role that the core actors had undertaken in managing social media activities or some other form of affiliation between core actors. It also might indicate that some of the actors in the periphery group might not be a member of a specific organization, and might have been acting like one of them, for some reason.

5.6. Network diagram

A network diagram (Figure 8) can clarify how intensely these spammers were connected to each other. A network diagram visualizes friendship (following/being followed by) relationships among actors. As Figure 8 illustrates, many actors in the network had direct connections to each other. This network contains only the spammers identified from a simple random sample of collected tweets, so this diagram could be an indication of the organization network we were testing for. This network diagram as well as the high measure of network density, both support our initial hypothesis of the existence of an organized network of spammers.

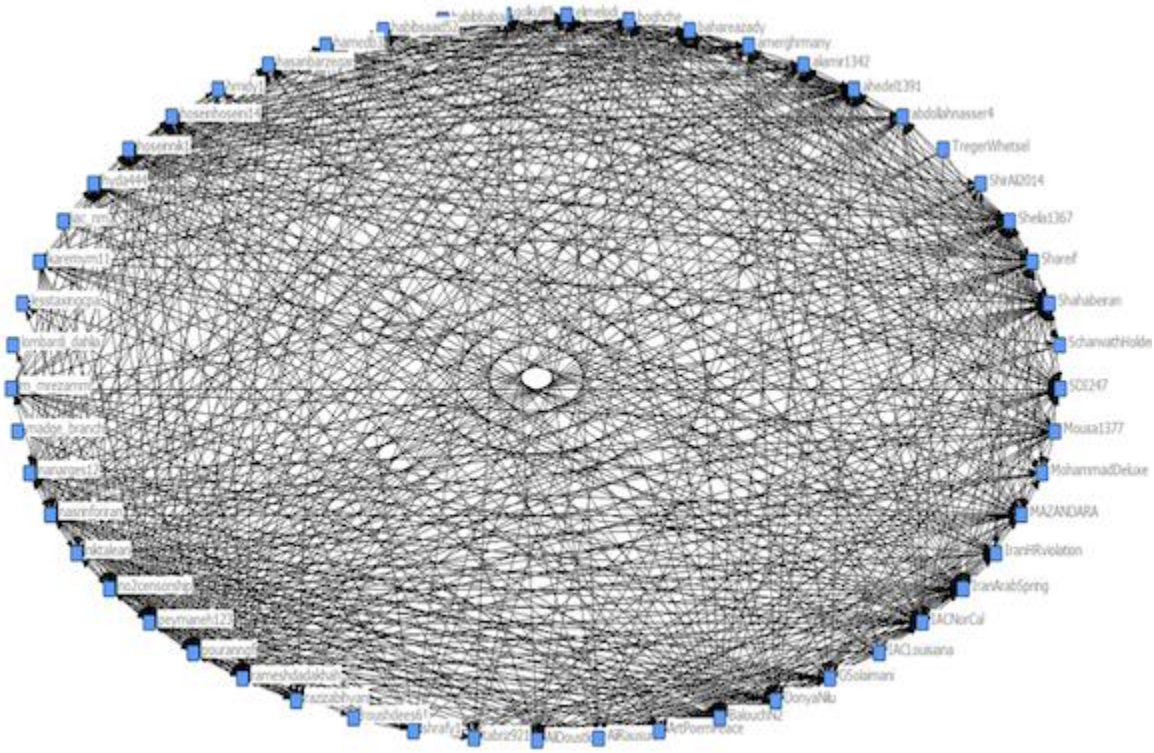

Figure 8: Network diagram of the dyadic relationships.

Note: Larger version of figure available here.

5.7. A retrospective look at the unsampled datasets

We coded a sample subset of our collected tweet decks due to the difficulty of manually coding all the collected tweets, and because only a sample seemed a sufficient proof of concept to determine whether if our hypothesized pattern of behavior was taking place or not. Once we had generated a list of active spammers from the sampled data, we could take a retrospective look at the rest of the data and find confirmatory or non-confirmatory evidence, with no need to code more tweets (75 remaining tweet decks, each containing the 100 most recent tweets with the #IranTalks hashtag, at the time we ran our data collection Python script). Thus, we analyzed the rest of the datasets against our list of active spammers to discover: 1) how many times each spammer appeared in another dataset; 2) how many tweets the spammer posted that were recorded in the rest of the data; and, 3) what was the approximate average interval of time between the spammer's tweets. Here are the highlights:

- The top numbers for appearance in other datasets were: 13, 9, 9, 8, 7, 7, several 6 and several 5 appearances and so on;
- The maximum number of tweets posted by one of the spammers in a single dataset: 52, 40, 39, 28, 26, 25, 22, 21, 19, and so on;

However, there is a time factor for each dataset which is variable. That means the numbers reported here might be the tweets that were posted over a short period of time, or they might have been posted over a longer period of time. Thus, we also looked into the duration of each dataset (the difference between the timestamp of the first and the last tweet recorded in each dataset) and calculated the approximate average intervals between a spammer's tweet postings. We say approximate because duration is a factor of the dataset, while the duration between a spammer's first and last tweets in each dataset is probably a smaller number, and thus, if the frequency is calculated based on that number, we would get a smaller number as the intervals between tweets, which is even more significant.

- The lowest average time lapse (in seconds) between tweets of spammers in a single dataset is: 5, 8, 9, 11, 11, 13, and so on. We are just reporting the lowest number for each spammer that was seen among all 75 datasets. In fact, there were spammers who had a record of having more than just one of these very low time intervals between their tweets. For instance, the spammer who was accounted for the five-second interval between tweets, also had a record of 12 and another record of 13 in two other datasets, or similarly, the spammer with nine seconds between tweets was also seen in another dataset with a time interval of 11 seconds between tweets.

We found these findings confirmatory with our findings that the aim of this pattern of activity on Twitter was to distract users from following updates on the #IranTalks hashtag. Since a similar pattern of activity was recorded from users who happened to have very similar account demographics and characteristics, we can assert that there had been some intentional and organized effort of disruption.

6. Limitations and further research

For conducting this study, we had to rely on the free Twitter API for our data collection, as Twitter does not allow access to a complete set of tweets resulting from a search query, if there are more than a certain number within a given interval of time

(this is called the 'rate limit' of the Twitter API). Therefore, a researcher cannot use his own method of sampling all tweets without invoking some charges and fees. Thus, in most cases when a researcher hits the rate limit cap in the free version of the Twitter API, the API makes a first-level sampling of the data. In this study, we relied on this mechanism every time we executed our Python script to collect the latest 100 tweets on the #IranTalks hashtag. Moreover, the Twitter account used for fetching tweets was created specifically for this project to avoid any effects of previous activities on Twitter filtering mechanisms. But even for brand new accounts, Twitter uses some location and demographic information from the hosting computer and network to tailor tweets returned for a hashtag search, and that also can have an impact on fetched tweets.

Another limitation resulting from this mechanism is that since the Twitter API returns 100 tweets every time it is called, the timespan for each of the fetched datasets containing 100 tweets varies. For instance, some of our sample 100 tweets are posted in a timespan as short as less than three minutes, while some others are posted in over 40 minutes. This disparity can result in inconsistencies that need to be addressed in analysis (as we did), but makes some time-specific analyses more difficult and less reliable.

Another limitation in this study was that since we had a two-phase data collection with almost a year interval, some of the spammers from our original list of spammers were already identified and removed by Twitter, and thus we had to exclude them from our network analysis because no information about their friendships was available or could be fetched at the time of analysis. In at least one instance, the same account handle was used by a new user after the original account had been deleted by Twitter; we discovered this from the date of creation of the account. Friendship relationships might also have been changed during the data collection time interval but we considered this insignificant in our network analysis.

In this study, we focused on what users were doing and how they acted or reacted to each other on social media. However, there are some directions for future research that could enhance our understanding of social networks:

- *Including affiliations in the analysis:* Looking into the accounts that operate together and in the same time intervals could be useful in some instances. For instance, the network might be composed of some subnetworks and this affiliation analysis could open a window to new understandings of network structure. In our sampled data, we might look into datasets and try to figure out which accounts usually appeared together on the sampled tweet decks. Unfortunately, the sampling by Twitter places a barrier on the extent to which this information can be

elicited. Therefore, although a great way of finding out more about an unknown network of actors, an analysis of affiliation networks cannot result in any significant findings in our datasets in this study. However, if the full datasets were representatively sampled and coded for spam, it added value.

- *Including reactions in the analysis:* All activities other than tweeting could also be considered for the analysis. For instance, marking tweets as favorite, retweeting, or mentioning were not part of this study, due to a lack of data, again resulted from limitations and restrictions of Twitter's free API. Also, one thing that was less common was that some of these users copied the exact content of another user's tweet, instead of simply retweeting it. These reactions, if studied, could add to our understanding of how these accounts were related, and how they interacted with each other.
- *More rigorous spammer detection algorithms:* While we had limited our detection of spammer accounts to the coding of tweets and flagged a user as a spammer only if they had five or more tweets that were identified as spam, there could have been some alternatives for identifying spammers, even if they did not meet this study's criteria. For instance, as we described in the findings, some users were as active in posting a tweet every five seconds, and some were responsible for a great majority of the total tweets that were recorded in some datasets. Although most of these users were already recorded as spammers in our datasets, more criteria could have given us a richer list of spammers.
- *Expanding spammer detection criteria to friendships and reactions:* Another spammer detection practice could be that once a tweet was marked as a spam, users who reacted to that tweet were placed on a list of 'potential spammers' and watched for spammer-like behavior. This can be a 'greedier' algorithm for spammer detection. The same logic could also apply to friendship relationships for an account already been identified as a spammer. In these cases, once a potential spammer is associated with sufficient red flags, that actor is then labeled as a spammer. Once this decision has been reached, the same analysis can be done for that user's tweets and/or friends, to identify new potential spammers in the pool.
- *Other network analysis techniques:* There are also other social network analysis techniques that a researcher might find useful, such as block model and positional analysis to find if a network has positional equivalences to the network positions each of the actors have, transitivity analysis and clustering coefficient to find more about indirect friendships

among users, network partitioning and clustering to find the most influential actors in terms of keeping the network connected, and so on.

An interesting issue for future research relates to the perception of disengaging tweets — as we call 'spam' here — by Twitter users. We noticed that while some users are more inclusive in their definition of what they would consider as spam, other users have a much narrower definition of such, *i.e.*, they might find a tweet disengaging and perceive it as spam while another use might accept a loose connection between the tweet and the hashtag it was posted with. In this study, we used a clear two-step criterion which we believe helped us keep false-positive detections to a minimum. A survey of different social media users could help in understanding inclusiveness choices, and find out how they become distractions while users are following trending news and updates.

There also seems to be a growing interest in terrorist groups and violent organizations that employ social media capabilities to support their activities (Bodine-Baron, *et al.* 2016). This calls for further study on the subject to connect hashtag spoiling behavior to known ways terrorist organizations and violent groups use social media, to better understand different possible aspects of the problem, and help social media platforms and users mitigate the impact of such activities. With more information on hand about the situations in which this spamming behavior is likely to take place, and with sufficient samples to give a good sense of how these spamming behaviors occur and spam tweets appear, unsupervised machine learning algorithms could be employed to detect them. This could be beneficial for social media applications that want to identify and block spammers, in order to allow users legitimately take advantage of social platform capabilities.

7. Concluding remarks

In this paper, we presented our findings from a research on organized distributed hashtag spoiling activity in Twitter. We used Twitter data to show relationships between accounts responsible for high frequency and disengaging tweets, in a sample of recorded tweets, coded based on a two-step mechanism and stringent criteria. We employed social media tools and techniques to show the cohesive subgroups and seniority in the network of relationships. We analyzed the micro social network composed of dyadic connections between the accounts responsible for tweets marked as spam in our sample. Although we analyzed a small — and assumingly non-representative — sample of all such account spoiling using this hashtag, we believe

this analysis can at least be a proof of concept in case it actually revealed evidence of some meaningful relationships between these accounts.

The results illustrate that organized hashtag spoiling is actually taking place in some instances, and due to the context, we believe political motivations could be its main cause.

Knowing about networked relationships among spammer accounts makes their tweeting patterns and similarities in demographics more meaningful. We can now assert that these accounts belong to individuals in organizations who had a planned mandate of spoiling the "#IranTalks" hashtag.

In the context of our #IranTalks case study, a few insights and observations stand out.

First, insofar as we have demonstrated that hashtag spoiling around #IranTalks has proven to be an organized campaign with the goal of disengaging Twitter users from following news related to the nuclear negotiations, this case clearly falls under the domain of hacktivism. The patterns of behavior that were undertaken by the hashtag spoilers were disruptive though non-destructive, and while they may have violated Twitter's terms of service agreement, were not overtly criminal.

Second, the case of the #IranTalks hashtag may signal a new evolutionary step in hacktivism. In the past, hacktivists deployed their tactics and disruptive acumen in order to draw attention to their political causes, or to amplify a particular voice to influence public opinion in some way. What makes this case different is that a similar methodology was applied, but in a way that resulted in a different effect — namely, user distraction. The goal of user distraction is a marked shift away from the goal of political expression. This is not only significant in its own right, but also could potentially serve to shift public opinion and, additionally, reshape how researchers contextualize the role of spam, botnets, collective blocklists, online collective actions, and a host of other objectives of study on this front.

Third, in the hopes of being prescriptive, #IranTalks also highlights the ever increasing importance of algorithms as a meta-regulator. As hashtag spoiling and wider campaigns that target user distraction become more prevalent on social media platforms, their effects can only be mitigated through better detection algorithms. Consequently, those algorithms, their technical design, and their potential politicization, become central to this discussion. FM

About the authors

Mahdi M. Najafabadi is currently pursuing his Ph.D. studies by focusing on open data ecosystems. He is interested in applying system thinking to address complex sociotechnical problems in public and private sectors. Mahdi is also interested in studying how information technologies and social media is exploited by people, organizations, and governments for a better outcome, and the implications of this attitude towards information technology.
Web: http://www.albany.edu/faculty/mahdi
E-mail: mnajafabadi [at] albany [dot] edu

Robert J. Domanski holds his Ph.D. and is an instructor of both political science and computer science at the City University of New York. He is the author of the book, *Who governs the Internet? A political architecture*, and his ongoing research focuses on the topics of hacktivism and the politics of algorithms.
E-mail: Robert [dot] Domanski [at] csi [dot] cuny [dot] edu

Notes

1. Twitter runs automated algorithms to detect such activities especially if they are enhanced by bots and bans them, as it believes this is against Twitter Rules. Twitter rules are accessible here at http://support.twitter.com/articles/18311.

2. These countries were U.K., Germany, France (EU3), and U.S., China, and Russia (+3).

3. Google now has a policy of neutralizing Google bombs once the company is made aware of them. See "Detecting new 'Googlebombs'," *Google Public Policy Blog*, by Matt Cutts (2009).

4. For several examples (news reports) of cases in which governments commit a similar activity to distract (disengage) users from following updates on a particular issue, see Najafabadi (2017).

5. Nuclear negotiating parties had good reasons not to trust each other, and thus each party wanted to put enough mechanisms and measures in the final agreement to limit future interpretations or deterioration of the terms by the other party. Thus, unconventionally, instead of a treaty, the negotiating parties were required to implement this deal by an unanimous United Nations Security Council (UNSC) resolution that required the participating countries to implement the deal (BBC News, 2015). However, the Trump administration has since taken action which leaves the agreement's future in doubt.

The UNSC has designated a special subsection on their Web site to the resolution 2231 (2015), that includes many related resources, accessible here: http://www.un.org/en/sc/2231. The complete resolution is accessible via: the United Nation's Web portal on http://www.un.org/press/en/2015/sc11974.doc.htm and also through the International Atomic Energy Agency (IAEA) on http://www.iaea.org/sites/default/files/unsc_resolution2231-2015.pdf

6. The Trump administration though has recently taken some actions which leaves the agreement's future in doubt.

7. Based on the city in which the negotiations were taking place, some alternative hashtags were also used. For a large majority of tweets though, other hashtags were overlapping with the #IranTalks hashtag as most users used more than one hashtag in their tweets to communicate about the topic.

8. Founded in 1965 as an opposition to the Shah's regime, it took part in the 1979 Islamic revolution that replaced the Shah, and then it turned against the Islamic Republic following the 1979 revolution. The group supported Iraq in its eight-year war against Iran, and "targeted Iranian government officials and government facilities in Iran and abroad; during the 1970s, it attacked Americans in Iran" (https://www.globalsecurity.org/military/world/para/mek.htm). According to Iran, "out of the nearly 17,000 Iranians killed in terrorist assaults" since Iran's 1979 Revolution, "about 12,000 had fallen victim to MKO's terrorist attacks".

9. U.S. Department of State, Bureau of Counter-Terrorism, at http://www.state.gov/j/ct/rls/other/des/123085.htm, accessed 1 October 2017.

10. Uniform resource locator, or simply a Web link.

References

Abdullah Almaatouq, Ahmad Alabdulkareem, Mariam Nouh, Erez Shmueli, Mansour Alsaleh, Vivek K. Singh, Abdulrahman Alarifi, Anas Alfaris, and Alex (Sandy) Pentland, 2014. "Twitter: Who gets caught? Observed trends in social micro-blogging spam," *WebSci '14: Proceedings of the 2014 ACM Conference on Web Science*, pp. 33–41.
doi: https://doi.org/10.1145/2615569.2615688, accessed 16 March 2018.

Dan Amira, 2011. "Rick Santorum has come to terms with his Google problem," *New York Magazine* (16 February),

at http://nymag.com/daily/intelligencer/2011/02/rick_santorum_has_come_to_term.html, accessed 9 February 2016.

Marco T. Bastos and Dan Mercea, 2017. “The Brexit botnet and user-generated hyperpartisan news,” *Social Science Computer Review* (10 October). doi: https://doi.org/10.1177/0894439317734157, accessed 16 March 2018.

BBC News, 2015. “UN Security Council endorses Iran nuclear deal” (20 July), at http://www.bbc.com/news/world-middle-east-33594937, accessed 1 October 2017.

David Bell, 2001. *An introduction to cybercultures*. New York: Routledge.

Alessandro Bessi and Emilio Ferrara, 2016. “Social bots distort the 2016 U.S. Presidential election online discussion,” *First Monday*, volume 21, number 11, at http://firstmonday.org/article/view/7090/5653, accessed 16 March 2018. doi: http://dx.doi.org/10.5210/fm.v21i11.7090, accessed 16 March 2018.

Elizabeth A. Bodine-Baron, Todd C. Helmus, Madeline Magnuson, and Zev Winkelman, 2016. *Examining ISIS support and opposition networks on Twitter*. Santa Monica, Calif.: RAND Corporation, at https://www.rand.org/pubs/research_reports/RR1328.html, accessed 16 March 2018.

Anita Campbell, 2013. “What is hashtag hijacking?” *Small Business Trends* (19 August), at http://smallbiztrends.com/2013/08/what-is-hashtag-hijacking-2.html, accessed 15 December 2016.

Noam Cohen, 2007. “Google halts ‘miserable failure’ link to President Bush,” *New York Times* (29 January), at http://www.nytimes.com/2007/01/29/technology/29google.html, accessed 9 February 2016.

E. Gabriella Coleman, 2015. *Hacker, hoaxer, whistleblower, spy: The many faces of Anonymous*. London: Bloomsbury.

E. Gabriella Coleman, 2013. *Coding freedom: The ethics and aesthetics of hacking*. Princeton, N.J.: Princeton University Press.

Giorel Curran and Morgan Gibson, 2013. “WikiLeaks, anarchism and technologies of dissent,” *Antipode*, volume 45, number 2, pp. 294–314. doi: http://dx.doi.org/10.1111/j.1467-8330.2012.01009.x, accessed 16 March 2018.

Matt Cutts, 2009. “Detecting new ‘Googlebombs’,” *Google Public Policy Blog* (24 January), at http://publicpolicy.googleblog.com/2009/01/detecting-new-googlebombs.html, accessed 14 October 2017.

Nic DePaula, Ersin Dincelli, and Teresa M. Harrison, 2018. “Toward a typology of government social media communication: Democratic goals, symbolic acts and self-presentation,” *Government Information Quarterly*, volume 35, number 1, pp. 98–108. doi: https://doi.org/10.1016/j.giq.2017.10.003, accessed 16 March 2018.

Marco Deseriis, 2017. “Hacktivism: On the use of botnets in cyberattacks,” *Theory, Culture & Society*, volume 34, number 1, pp. 131–152. doi: https://doi.org/10.1177/0263276416667198, accessed 16 March 2018.

Ersin Dincelli, Yuan Hong, and Nic DePaula, 2016. “Information diffusion and opinion change during the *Gezi Park protests*: Homophily or social influence?” *Proceedings of the Association for Information Science and Technology*, volume 53, number 1, pp. 1–5. doi: https://doi.org/10.1002/pra2.2016.14505301109, accessed 16 March 2018.

Robert J. Domanski, 2015. *Who governs the Internet? A political architecture* Lanham, Md.: Lexington Books.

Jesse Emspak, 2011. “Russian protesters get Twitter-bombed,” *Discovery News* (10 December), at http://news.discovery.com/tech/russian-protesters-get-twitter-bombed-111209.htm, accessed 1 May 2016.

Adam Entous, Craig Timberg, and Elizabeth Dwoskin, 2017. “Russian operatives used Facebook ads to exploit America’s racial and religious divisions,” *Washington Post* (25 September), at https://www.washingtonpost.com/business/technology/russian-operatives-used-facebook-ads-to-exploit-divisions-over-black-political-activism-and-muslims/2017/09/25/4a011242-a21b-11e7-ade1-76d061d56efa_story.html, accessed 1 October 2017.

Emilio Ferrara, Onur Varol, Clayton Davis, Filippo Menczer, and Alessandro Flammini, 2016. “The rise of social bots,” *Communications of the ACM*, volume 59, number 7, pp. 96–104. doi: https://doi.org/10.1145/2818717, accessed 16 March 2018.

Klint Finley, 2015. “Pro-government Twitter bots try to hush Mexican activists,” *Wired* (23 August), at http://www.wired.com/2015/08/pro-government-twitter-bots-try-hush-mexican-activists, accessed 1 May 2016.

Asmelash Teka Hadgu, Kiran Garimella, and Ingmar Weber, 2013. “Political hashtag hijacking in the U.S.,” *WWW ’13: Companion Proceedings of the 22nd International Conference on World Wide Web*, pp. 55–56.
doi: https://doi.org/10.1145/2487788.2487809, accessed 16 March 2018.

Noah C.N. Hampson, 2012. “Hacktivism: A new breed of protest in a networked world,” *Boston College International and Comparative Law Review*, volume 35, number 2, at http://lawdigitalcommons.bc.edu/iclr/vol35/iss2/6, accessed 16 March 2018.

Heather Havenstein, 2008. “Blogger launches ‘Google bomb’ at McCain,” *Computerworld* (19 June), at http://www.computerworld.com/article/2534384/web-apps/blogger-launches--google-bomb--at-mccain.html, accessed 9 February 2016.

Xiaoyun He, Amir Karami, and Chaoqun Deng, 2017. “Examining the effects of online social relations on product ratings and adoption: Evidence from an online social networking and rating site,” *International Journal of Web Based Communities*, volume 13, number 3, pp. 344–363.
doi: https://doi.org/10.1504/IJWBC.2017.086591, accessed 16 March 2018.

Yoonjin Hyun and Namgyu Kim, 2016. “Detecting blog spam hashtags using topic modeling,” *ICEC ’16: Proceedings of the 18th Annual International Conference on Electronic Commerce,* article number 43.
doi: https://doi.org/10.1145/2971603.2971646, accessed 16 March 2018.

Danesh Irani, Steve Webb, Calton Pu, and Kang Li, 2010. “Study of trend-stuffing on Twitter through text classification,” *CEAS 2010: Seventh annual Collaboration, Electronic messaging, Anti-Abuse and Spam Conference*; version at https://www.cc.gatech.edu/projects/doi/Papers/DIrani_CEAS_2010.pdf, accessed 16 March 2018.

Sarah J. Jackson and Brooke Foucault Welles, 2015. “Hijacking #myNYPD: Social media dissent and networked counterpublics,” *Journal of Communication*, volume 65, number 6, pp. 932–952.
doi: https://doi.org/10.1111/jcom.12185, accessed 16 March 2018.

Tim Jordan, 1999. *Cyberpower: The culture and politics of cyberspace and the Internet*. London: Routledge.

Tim Jordan and Paul A. Taylor, 2004. *Hacktivism and cyberwars: Rebels with a cause?* London: Routledge.

Andreas Jungherr, 2016. “Twitter use in election campaigns: A systematic literature review,” *Journal of Information Technology & Politics*, volume 13, number 1, pp. 72–91.
doi: https://doi.org/10.1080/19331681.2015.1132401, accessed 16 March 2018.

Emad Khazraee and James Losey, 2016. “Evolving repertoires: Digital media use in contentious politics,” *Communication and the Public*, volume 1, number 1, pp. 39–55.
doi: https://doi.org/10.1177/2057047315625076, accessed 16 March 2018.

Gary King, Jennifer Pan, and Margaret E. Roberts, 2017. “How the Chinese government fabricates social media posts for strategic distraction, not engaged argument,” *American Political Science Review*, volume 111, number 3, pp. 484–501.
doi: https://doi.org/10.1017/S0003055417000144, accessed 16 March 2018.

Kyumin Lee, James Caverlee, Krishna Y. Kamath, and Zhiyuan Cheng, 2012. “Detecting collective attention spam,” *WebQuality ’12: Proceedings of the Second Joint WICOW/AIRWeb Workshop on Web Quality*, pp. 48–55.
doi: https://doi.org/10.1145/2184305.2184316, accessed 16 March 2018.

Carol D. Leonnig, Tom Hamburger, and Rosalind S. Helderman, 2017. “Russian firm tied to pro-Kremlin propaganda advertised on Facebook during election,” *Washington Post* (6 September), at http://www.washingtonpost.com/politics/facebook-says-it-sold-political-ads-to-russian-company-during-2016-election/2017/09/06/32f01fd2-931e-11e7-89fa-bb822a46da5b_story.html, accessed 1 October 2017.

Ken May, 2016. “Pro-Clinton super PAC caught spending $1 million on social media trolls” (22 April), at http://kenmay.net/pro-clinton-super-pac-caught-spending-1-million-on-social-media-trolls/, accessed 16 March 2018.

Fatjona Mejdini, 2016. “Iranian opposition ex-fighters ‘transferred to Albania’,” *Balkan Insight* (26 August), at http://www.balkaninsight.com/en/article/albania-continues-to-accept-iranian-mujahedins-08-26-2016, accessed 1 October 2017.

Evgeny Morozov, 2011. *The net delusion: The dark side of Internet freedom*. New York: Public Affairs.

Mahdi M. Najafabadi, 2017. “A research agenda for distributed hashtag spoiling: Tales of a survived trending hashtag,” *dg.o ’17: Proceedings of the Eighteenth Annual International Conference on Digital Government Research*, pp. 21–29.
doi: https://doi.org/10.1145/3085228.3085273, accessed 16 March 2018.

John Pike, 2017. “Mujahedin-e Khalq organization,” *GlobalSecurity.org*, at http://www.globalsecurity.org/military/world/para/mek.htm, accessed 1 October 2017.

Gareth Porter, 2015. “Behind the scenes: How the US and Iran reached their landmark deal,” *Nation* (5 September), at http://www.thenation.com/article/behind-the-scenes-how-the-us-and-iran-reached-their-landmark-deal, accessed 15 December 2016.

Anas Qtiesh, 2011. “Spam bots flooding Twitter to drown info about #Syria protests” (18 April), at http://www.anasqtiesh.com/2011/04/spam-bots-flooding-twitter-to-drown-info-about-syria-protests, accessed 1 May 2016.

Roy Rosenzweig, 1998. “Wizards, bureaucrats, warriors, and hackers: Writing the history of the Internet,” *American Historical Review*, volume 103, number 5, pp. 1,530–1,552.
doi: https://doi.org/10.2307/2649970, accessed 16 March 2018.

Patrícia G.C. Rossini, Jeff Hemsley, Sikana Tanupabrungsun, Feifei Zhang, Jerry Robinson, and Jennifer Stromer-Galley, 2017. “Social media, U.S. presidential campaigns, and public opinion polls: Disentangling effects,” *#SMSociety17: Proceedings of the Eighth International Conference on Social Media & Society*, article number 56.
doi: https://doi.org/10.1145/3097286.3097342, accessed 16 March 2018.

Rodrigo Sandoval-Almazan and J. Ramon Gil-Garcia, 2014. “Towards cyberactivism 2.0? Understanding the use of social media and other information technologies for political activism and social movements,” *Government Information Quarterly*, volume 31, number 3, pp. 365–378.
doi: http://doi.org/10.1016/j.giq.2013.10.016, accessed 16 March 2018.

Surendra Sedhai and Aixin Sun, 2015. “HSpam14: A collection of 14 million tweets for hashtag-oriented spam research,” *SIGIR ’15: Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 223–232.
doi: http://doi.org/10.1145/2766462.2767701, accessed 16 March 2018.

Scott Shane and Vindu Goel, 2017. “Fake Russian Facebook accounts bought $100,000 in political ads,” *New York Times* (6 September), at http://www.nytimes.com/2017/09/06/technology/facebook-russian-political-ads.html, accessed 1 October 2017.

Rukundo Solomon, 2017. “Electronic protests: Hacktivism as a form of protest in Uganda,” *Computer Law & Security Review*, volume 33, number 5, pp. 718–728. doi: https://doi.org/10.1016/j.clsr.2017.03.024, accessed 16 March 2018.

Jonghyuk Song, Sangho Lee, and Jong Kim, 2011. “Spam filtering in Twitter using sender-receiver relationship,” In: Robin Sommer, Davide Balzarotti, and Gregor Maier (editors). *Recent advances in intrusion detection*. *Lecture Notes in Computer Science (LNCS)*, volume 6961. Berlin: Springer-Verlag, pp. 301–317. doi: https://doi.org/10.1007/978-3-642-23644-0_16, accessed 16 March 2018.

Richard A Spinello, 2002. *Regulating cyberspace: The policies and technologies of control*. Westport, Conn.: Quorum Books.

Denis Stukal, Sergey Sanovich, Richard Bonneau, and Joshua A. Tucker, 2017. “Detecting bots on Russian political Twitter,” *Big Data*, volume 5, number 4, pp. 310–324. doi: https://doi.org/10.1089/big.2017.0038, accessed 16 March 2018.

Pablo Suárez-Serrato, Margaret E. Roberts, Clayton Davis, and Filippo Menczer, 2016. “On the influence of social bots in online protests: Preliminary findings of a Mexican case study,” In: Emma Spiro and Yong-Yeol Ahn (editors). *Social informatics*. *Lecture Notes in Computer Science*, volume 10047, Cham, Switzerland: Springer International, pp. 269–278. doi: https://doi.org/10.1007/978-3-319-47874-6_19, accessed 16 March 2018.

Colin Taylor, 2016. “Hillary trolls just got Facebook to shut down Bernie groups by reporting them as pornography,” *Occupy Democrats* (26 April), at http://occupydemocrats.com/2016/04/26/hillary-trolls-just-got-facebook-shut-bernie-groups-reporting-pornography, accessed 10 May 2016.

Marisa Taylor, 2009. “Google-bombing moves from Bush to Obama,” *Wall Street Journal* (23 January), at http://blogs.wsj.com/digits/2009/01/23/google-bombing-moves-from-bush-to-obama, accessed 9 February 2016.

Paul A. Taylor, 2005. “From hackers to hacktivists: Speed bumps on the global superhighway?” *New Media & Society*, volume 7, number 5, pp. 625–646. doi: https://doi.org/10.1177/1461444805056009, accessed 16 March 2018.

Kurt Thomas, Chris Grier, Dawn Song, and Vern Paxson, 2011. “Suspended accounts in retrospect: An analysis of Twitter spam,” *IMC ’11: Proceedings of the 2011 ACM SIGCOMM Conference on Internet Measurement Conference*, pp. 243–258. doi: https://doi.org/10.1145/2068816.2068840, accessed 16 March 2018.

Deepali Virmani, Nikita Jain, Ketan Parikh, and Abhishek Srivastava, 2017. "HashMiner: Feature characterisation and analysis of #Hashtag hijacking using real-time neural network," *Procedia Computer Science*, volume 115, pp. 786–793. doi: https://doi.org/10.1016/j.procs.2017.09.174, accessed 16 March 2018.

Stanley Wasserman and Katherine Faust, 1994. *Social network analysis: Methods and applications*. New York: Cambridge University Press.

Samuel C. Woolley, 2016. "Automating power: Social bot interference in global politics," *First Monday*, volume 21, number 4, at http://firstmonday.org/article/view/6161/5300, accessed 15 May 2016. doi: http://dx.doi.org/10.5210/fm.v21i4.6161, accessed 16 March 2018.

Tingmin Wu, Sheng Wen, Shigang Liu, Jun Zhang, Yang Xiang, Majed Alrubaian, and Mohammad Mehedi Hassan, 2017. "Detecting spamming activities in Twitter based on deep-learning technique," *Concurrency and Computation: Practice and Experience*, volume 29, number 19, e4209. doi: http://dx.doi.org/10.1002/cpe.4209, accessed 16 March 2018.

Chao Yang, Jialong Zhang, and Guofei Gu, 2014. "A taste of tweets: Reverse engineering Twitter spammers," *ACSAC '14: Proceedings of the 30th Annual Computer Security Applications Conference*, pp. 86–95. doi: http://dx.doi.org/10.1145/2664243.2664258, accessed 16 March 2018.

Neetzan Zimmerman, 2016. "Hillary supporters accused of taking down Bernie FB pages in porn attack," *The Hill* (26 April), at http://thehill.com/blogs/blog-briefing-room/news/277657-hillary-supporters-take-down-bernie-fb-pages-in-coordinated, accessed 10 May 2016.

Editorial history

Received 2 March 2018; accepted 14 March 2018.

Hacktivism and distributed hashtag spoiling on Twitter: Tales of the #IranTalks
by Mahdi M. Najafabadi and Robert J. Domanski.
First Monday, Volume 23, Number 4 - 2 April 2018
https://journals.uic.edu/ojs/index.php/fm/article/download/8378/6663
doi: http://dx.doi.org/10.5210/fm.v23i4.8378