Article

new media & society
2019, Vol. 21(7) 1506–1527

DOI: 10.1177/1461444818825133
journals.sagepub.com/home/nms
SAGE

State-aligned trolling in Iran and the double-edged affordances of Instagram

Simin Kargar
Harvard University, USA

Adrian Rauchfleisch
National Taiwan University, Taipei

Abstract

Online harassment is increasingly applied as a form of information control to curb free speech and exert power in online public spheres. In recent years, states have appeared to be particularly invested in weaponizing information against dissidents in an attempt at dominating social and political discourses. Reports by prominent human rights institutions, as well as anecdotal evidence, indicate that Iran remains among the states with a track record of such actions. The scope of targeted cyber abuse varies by case. This study investigates the size and perpetrators of online violence, harassment, and abuse against critical members of the Iranian diaspora, including journalists, civil society activists, and artists, among many others. This study substantiates findings of qualitative interviews with a quantitative study of Instagram accounts of related individuals and explores the patterns and communities involved in disseminating hate speech in an attempt at manipulating public opinion and suppressing voices of dissidents.

Keywords

Hate speech, Instagram, Iran, online communication, cyber warfare, topic modeling, social network analysis

In September 2016, hackers attacked the Instagram account of a popular Iranian musician, Shahin Najafi. They changed Najafi's bio to display the hackers' contact information and

Corresponding author:
Simin Kargar, Harvard University, 23 Everett St. Cambridge, MA 02138, USA.
Email: skargar@law.harvard.edu

replaced his profile picture with the flag of the Islamic Republic of Iran. Najafi was targeted by state-aligned hackers because of his controversial music. His songs address politically and socially sensitive topics such as censorship, theocracy, homophobia, and sexism. This brought Najafi under attacks of the Iranian clerics, and in 2012, they issued fatwas that declared him guilty of apostasy. Since the incident, Najafi has remained a regular target of hate speech online and cyberattacks. Yet, he is not alone. Najafi's case represents an underlying issue that informs the main concern of this study.

Cyber abuse and online harassment are increasingly applied as a form of information control to curb free speech and exert power in cyberspace (Software Freedom Law Centre, 2016). In recent years, states have appeared to particularly invest in weaponizing information against dissidents and activists in an attempt at dominating social and political discourses (Golkar, 2013; Rahimi, 2003). These practices are often exercised in tandem with other forms of suppression such as intrusion campaigns in the form of state-sponsored hacking of emails, surveillance of communication and devices, and distributed denial-of-service (DDoS) attacks on opposition websites (Guarnieri and Anderson, 2016).

Coordinated harassment of dissidents on social media appears as the most recent form of strategic communication, where particular messages are crafted by state-aligned actors to manipulate public opinion (Chen, 2015). The scope of such targeted abuse varies by case, while evidence is typically scarce and has not been comprehensively substantiated with data. These extrajurisdictional practices turn harassment into a relatively low-cost weapon for targeting the opposition, limiting freedom of speech (Software Freedom Law Centre, 2016) through intimidation and pursuing a "silencing" strategy. These practices are conducted on the same mediums that are designed to give voice to the voiceless.

A growing body of literature recognizes that the affordances of the Internet can indeed strengthen authoritarian regimes both domestically (Morozov, 2011; Pearce, 2015; Roberts, 2018) and internationally (for an overview, see Wooley and Howard, 2018). This study seeks to shed light on the latest practices of the Iranian regime to extend its ideological arms in cyberspace by crafting and disseminating propaganda against its opposition through international platforms. To such ends, we first explore how the Iranian regime targets and suppresses the opposition via online mediums.

To identify the underlying patterns of these practices and develop initial interpretations, we designed a qualitative study to conduct 18 semistructured online interviews with high-profile Iranian journalists, activists, artists, and celebrities between November 2016 and January 2017. These interviews revealed, among other findings, that Instagram is presently the most contested social media platform in Iran. In addition, interviewees suspected that the cyber mobs they encountered on Instagram were infiltrated by accounts affiliated with Iran's intelligence and security forces. These individuals also underscored that the experience of targeted harassment online has led them to take more precautionary measures offline. Some elaborated that as a result of these circumstances, they were more reserved about the topics they chose to publicly speak or write about. These observations align with the repertoires of suppression of a number of other authoritarian regimes.

These considerations ultimately led to a distinct quantitative study. To test the anecdotal evidence and substantiate observations with data, we set to aggregate and analyze data from public profiles that the interviewees own and regularly update on

Instagram.[1] This in particular included identifying any pattern of co-occurrences of comments, politically charged hate speech, and the ideological orientation of mob participants. The second part of the study, therefore, offers a quantitative analysis of Instagram data.

We first discuss the political context in Iran with a particular focus on the regime's online strategy to intimidate, harass, and silence dissidents. We will then refer to the repertoires of suppression that authoritarian regimes utilize to suppress critical voices, followed by the results of our initial qualitative interviews. Finally, we present our quantitative analysis of communications on Instagram based on 7893 posts and over 2.8 million comments aggregated.

The analysis informs our discussion of the repertoire of suppression of the Islamic Republic in online public spheres. We do so by investigating the extent to which pro-regime actors appear to instrumentalize online harassment. By examining past cyber mobs against Iranian dissidents, we explore how social media has evolved the parameters of propaganda and strategic communication.

Status of media and Internet freedom in Iran

Since the Islamic revolution in 1979, the Iranian regime has maintained strict control over the media (Khiabany, 2008). Media personnel are constantly exposed to intimidation, arbitrary arrest, and long jail sentences imposed by revolutionary courts (Committee to Protect Journalists, 2015). The rise of the Internet in Iran in the 1990s led to an extension of state control over digital communications. Content filtering and website blocking remain prevalent and typically follow the political contours (Clark et al., 2017). Most social media platforms and messaging applications are only accessible through proxies (Rahimi, 2008). Bloggers, activists, and journalists are constantly monitored for their online activities (Reporters Without Borders, 2017). Despite a partial improvement in international relations since the moderate president Hassan Rouhani took office, Iran continues to rank as one of the world's biggest prisons for media personnel (Reporters Without Borders, 2017). While the current administration promotes a rhetoric in favor of broadening online freedoms, restrictions on freedom of expression online have continued to grow since the election of president Rouhani in 2013 (ARTICLE 19, 2017).

Repertoires of information control and suppression in post-revolutionary Iran

Any authoritarian regime has a unique history and utilizes context-specific instruments to remain in power. As such, Iran is no stranger to suppressing voices of dissent. Simultaneously, these regimes often learn and "adapt their own repertoires of suppression in response to developments on the ground, regionally and at home" (Heydemann and Leenders, 2011: 649). In addition, authoritarian regimes often learn from the experiences elsewhere and cooperate with one another. For example, Iran's largest telecommunication company purchased surveillance equipment from ZTE, a Chinese company, in 2010 (Vatanka, 2015).

For the purpose of this study, we address current strategies within three main categories: censorship in publishing (Small Media, 2015), Internet filtering (Aryan et al., 2013), and harassment and persecution (Golkar, 2013).

Censorship in publishing

Censorship in Iran is not exclusive to the post-revolutionary era. From monarchs to clerics, politicians have continuously attempted to control the flow of information to consolidate power (Small Media, 2015). Iran's current Press Law and other guidelines (e.g., Supreme Council of Cultural Revolution, 1988) do not tolerate (1) any materials "undermining the political system" or attempts at "infiltrating the pillars of the Islamic Republic" and (2) "unethical artistic products" with "adverse ethical effects" that promote "sexual freedom and indecency" and materials that lead the youth "astray to corruption" (Mollanazar, 2011). Those involved in any form of expression, from journalism to art, are therefore no stranger to state-imposed "redlines."

Internet filtering

As the Internet penetration rate grew in Iran, it became one of the dominant sources of information and a medium for publishers to reach a much broader audience. This led to an extension of the control repertoire over cyberspace. As a result, new regulations on digital communications were introduced (e.g. Cyber Crimes Law passed in 2010; ARTICLE 19, 2017) and several regulatory and monitoring bodies were established (Bowen, 2015; Center for Human Rights in Iran, 2015). With the emergence of social networking platforms, Iran's stance on censorship and Internet access has become further complicated. At politically and socially sensitive momentums, such tension has repeatedly surfaced (e.g. after the disputed presidential elections in 2009: Burns and Eltham, 2009; Morozov, 2011). Utilizing social media to organize street protests was an unprecedented form of resistance against the political establishment and immediately led to Facebook and Twitter being banned in Iran. Despite various restrictions on access to online content and platforms, Instagram curiously remains unblocked (Khodabakhshi, 2017) and continues to grow in popularity in unprecedented ways.[2] Given these circumstances, this study will investigate how Instagram is utilized to disseminate harassment and propagate intimidating messages against the Iranian opposition and other vulnerable communities.

Harassment and intimidation of civil society and other dissidents

In addition to the limitations on access to information, politically charged intrusion campaigns against the opposition and different forms of surveillance remain prevalent (Guarnieri and Anderson, 2016). Political opponents and activists, journalists (Feinstein et al., 2016), members of civil society, religious minorities (Baha'i International Community, 2011), even artists (ARTICLE 19, 2015), writers, and poets have remained primary targets of intimidation, castigation, arrest, and, in some cases, execution (Soleimani, 2016). In recent years, journalists affiliated with international media,

diaspora artists, and activists have endured recurring cyberattacks (Anderson and Sadjadpour, 2018) and some form of harassment. These new forms of abuse are utilized to deter these influential communities from pursuing professional activities that are perceived as "anti-revolutionary" by the Islamic Republic. Like other forms of information control, the extent and duration of harassment typically correlate with the political and social contours (for example, during election periods: Anderson, 2013).

While numerous studies have previously analyzed the role of social media during protests in Iran (e.g. Burns and Eltham, 2009; Morozov, 2011), very few have investigated the repertoires of suppression on an individual level (cf. Feinstein et al., 2016) and their potential impact on online activism in authoritarian regimes. To address this gap, this study investigates strategic communications that often emerge in tandem with, or following, intrusion campaigns or other forms of abusive treatment of specific targets. These in particular include high-profile members of Iranian diaspora pursuing activism in different domains such as sexual minority and women's rights, thereby maintaining remarkable capacity to reach, influence, and even mobilize large audiences inside Iran (see Part 1: Qualitative Interviews and Index 1).

Empirical analysis Part 1: qualitative interviews

This study has leveraged a hybrid methodology comprised of qualitative interviews as well as social media data aggregation and analysis. The former preeminently helped us to develop detailed understanding of the scope of strategic communications by state-affiliated actors and to identify different strategies implemented by these actors in Persian online public spheres. These observations were critical to narrowing down the focus of the quantitative analysis to the platforms where organized cyber mobs were more likely to emerge. The following will first present the results of our qualitative interviews before moving onto the research questions, methodology, and results of the second part of this research—our quantitative analysis.

Between November 2016 and January 2017, we conducted 18 interviews with prominent members of Iranian diaspora. These individuals have repeatedly been targets of cyber mobs that appear to be aligned with, and perhaps directly or indirectly orchestrated by, the Iranian state. In selecting our sample, two factors had to be considered. First, we aggregated available anecdotes from the past decade indicating harassment and cyberattacks as increasingly rampant following the contentious presidential election in 2009 (Fassihi, 2009; Naji, 2018). Most interviewees had been prime targets of such attacks since. Second, selection of final interviewees was a matter of access to these targets of digital harassment and building trust to the extent that they were comfortable with sharing the details of their experiences. Ultimately, we were able to establish contact with 18 individuals who met all these criteria and were willing to take part in our interviews.

The sample of interviewees comprised of seven journalists (two women, five men); three activists of gender equality, all of whom identify as sexual minorities, two musicians; and two entertainment show hosts, one of whom is a political satirist; one member of a religious minority community; one Internet freedom activist; one model; and one media producer who identifies as lesbian. Each of these 18 individuals is known for their

Figure 1. Aram addresses a new wave of harassment by posting a picture of herself and writes about the disappointment she felt after she was aggressively targeted because of her sexual orientation.

professional activities performed outside the jurisdiction of the Islamic Republic. Their activities are typically deemed as "illegal" and "anti-revolutionary" by the state. Their diaspora status adds to the pressure they endure and underscores the extra-territoriality of the harassment that they encounter. Index 2 offers an overview of the individuals who took part in our interviews, mostly under the condition of anonymity, and their affiliations.

The journalists and media producer within our sample have had affiliations with well-known international media such as the BBC, Voice of America (VOA), and popular Persian outlets based outside of Iran, for example, the London-based Manoto TV. The Islamic Republic has repeatedly labeled these media institutions as "adversaries" and has accused them of undermining the religious and revolutionary values (Dehghan, 2017). According to these interviewees, topics that have attracted some of the most aggressive mobs are the very issues that Iran has an established interest in or is in conflict with. These include national security, Iran's nuclear program, Iran's role in the Syrian conflict, gender equality, and sexual minority rights. Figure 1 offers a detailed example.

In September 2016, Aram Bolandpaz[3] publicized her sexual orientation as lesbian by publishing a short video. Within a few hours, she was subjected to over 9000 comments, the majority of which included derogatory language toward her. She responded to these attacks and called out the state-affiliated accounts and their raid on her sexual identity by posting the picture that Figure 1 captures. It attracted over 4000 comments,

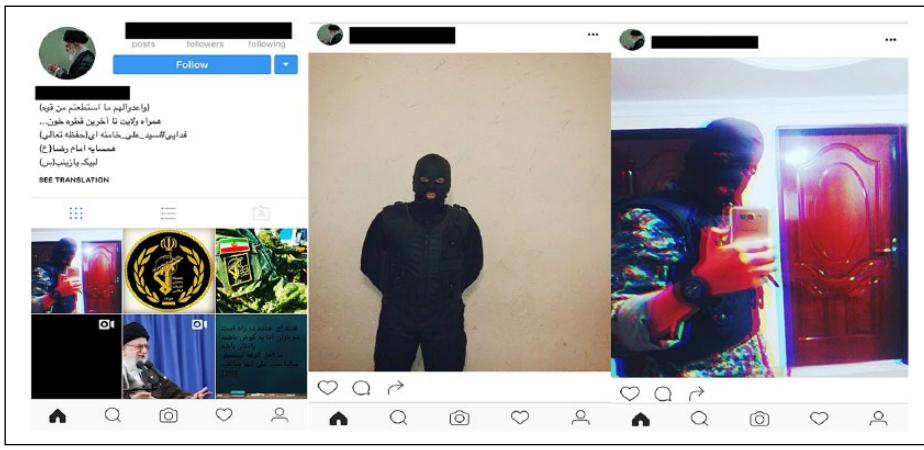

Figure 2. Sample public account of an attacker who directed violent threatening messages to Aram during the second wave of cyber mobs against her. The profile name is a pseudonym in Persian for "I am a soldier of Supreme Leader." It has a picture of the Leader as the account's identifier. The account mostly publishes radical content featuring the Islamic Revolutionary Guard Corps (IRGC), the Supreme Leader, and selfies in a Shrine Guards Corps uniform and a face mask. While the account holder attempts to remain anonymous, he does not seem disturbed by some selfexposure through selfies. The IRGC logo repeatedly appears on the pictures posted to the account—an indication of potential affiliation.

the majority of which questioned her integrity or expressed contempt toward the lesbian, gay, bisexual, transsexual, and queer (LGBTQ) community at large.

During our interview, Aram mentioned that a significant part of these slurs were directed at her as direct messages on Instagram. These messages were repeatedly sent by accounts that appeared to belong to members of Shrine Guards Corps and other entities affiliated with the Islamic Revolutionary Guard Corps (IRGC). Some of these profile pictures displayed men with covered faces who disseminated the most violent threats of murder and slaughter (see Figure 2). Aram suspects that her documentary productions about the intelligence and security apparatus of Iran such as IRGC may have incited such violence and continuous abuse against her.

Civil society activists encounter similar challenges. All interviewees have demonstrated long-standing activism to advocate for sexual minority and women's rights and Internet freedom. These are among the most socially, politically, and religiously contested issues to broach. The Islamic Republic of Iran has frequently accused diaspora activists of cultivating Western norms, subverting social and cultural standards and, ultimately, undermining the religious values of the society. Various attacks on the integrity, work ethics, and credibility of these individuals are therefore inadmissible consequences of their activism and strive for justice.

In one case, a prominent activist of sexual minority rights was targeted by an extensive disinformation campaign. Only several hours after her then new book had been published in May 2015, a website was launched that disseminated fabricated information about the target and her work. The website attacked her work ethics and promoted

an aggressive narrative about her "morally" and "financially" corrupt practices. Shortly after this website was launched, pro-state media began to circulate similar, if not identical, articles in a concerted campaign against the target, amplifying the distorted narrative. In addition to this website, she became a target of multiple fake accounts on Facebook that circulated further rumors and allegations against her. One Facebook account in particular publicized her ID name which is different from the professional name she has been using for over two decades. To the best of her knowledge, her ID name had last been used during her detention in Iran many years ago. Only her former interrogators, affiliated with Iranian intelligence forces, had access to this information—yet another cue that may suggest the involvement of state-aligned actors in this disinformation campaign.

Overall, the interviews elaborated on the patterns of suppressing online expression in different forms. Traditional forms of expression such as writing and public speaking are no longer the primary targets of state-affiliated propaganda campaigns. Instead, bodily expression such as modeling and other forms of manifestation of, e.g., music and media production, provides unprecedented grounds for state suppression. This may provide an explanation for the rise of coordinated cyber mobs against Iranian diaspora who pursue modeling, acting, music, or satire for a living.

Moreover, these hateful attacks demonstrate a gender aspect. Women are far more often targeted by state-aligned media with rumors and smear campaigns about their private lives than their male peers are. Single women journalists, for example, are more susceptible to online harassment than their married peers or male journalists regardless of their marital status. Furthermore, attacks against female journalists include vulgar and misogynistic references, while men are targeted with typical political slurs such as "dirty traitor" and "spy of the Imperialism machinery." This observation was later reinforced by our quantitative analysis. Out of all the pictures and videos in our data set, the top 140 that received the largest number of hateful comments all belonged to women.

The presence of Instagram accounts affiliated with security and intelligence forces was a core theme across the 18 interviews that we conducted. Our anecdotal evidence indicates that state-aligned profiles often feature more than one of the following:

- They rarely use a human image, if any, as their profile picture. We noted multiple accounts with images of the Supreme Leader as profile pictures coupled with religious terms and verses from the Holy Quran in bio captions.
- The language used by these accounts is more violent and vulgar than other users. In particular, these accounts are more likely to disseminate threats of death, rape, and sexual assault against women.
- Account names often carry religious and political connotations.
- Use of female names and pictures on these profiles is very rare.
- Multiple accounts with the same name and sequential, numeric suffixes suggest that they may have been created by the same individual(s). This is particularly the case when older accounts are reported and removed by Instagram due to violating its terms of service.
- Images posted on these accounts often indicate a religious theme. For instance, numerous images are taken at Shi'a holy shrines in religious cities of Iran or Iraq.

- These are often low activity accounts, purportedly created for the sole purpose of participating in online mobs. The users whom these profiles follow typically outnumber their followers. In many cases, they have no followers, follow no one, and have never posted any image.

Empirical analysis Part 2: quantitative analysis of Instagram comments

To examine these underlying patterns of these coordinated attacks further, we set to analyze content that our interviewees had posted on Instagram. Due to the large corpus of comments that these postings had received, a comprehensive quantitative analysis was required. The quantitative analysis seeks to identify groups of users that specifically target diaspora activists, celebrities, and other public figures. Based on the observations from our interviews, we ponder the possibility of "patriotic" perpetrators affiliated with the Iranian state. The analysis investigates four research questions:

RQ1. What topics are discussed in the comments? Analyzing these topics enables the detection of aggressive, insulting, and harassing comments against the account holders.

RQ2. Second, we address the temporal patterns of cyber mobs that the targets have repeatedly experienced on Instagram.

RQ3. Third, we seek a strategic element in these communications—one that leads to different Instagram accounts deploy simultaneous or similar attacks against a target. To examine such coordination, it is necessary to identify any co-occurrence of comments as well as the communities involved in co-commenting.

RQ4. Finally, we turn to analyze the internal structure of these groups through follower networks. We suspect that co-commenting perpetrators of harmful speech are connected to one another and follow the same set of public figures, including our interviewees. The analysis provides insight into the interests, backgrounds, and ideology of the perpetrators.

While attribution of cyber mobs to state actors with certainty remains a challenge, we can still examine the extent of coordination and ideological orientation of mob participants. There might be individual accounts that use similar harmful and/or patriotic language, appear together in the same online mobs, and follow one another, and their profiles suggest a strong ideological orientation. Considered altogether, these features suggest coordination that may have been instructed by authorities or out of radical ideological support for the Leader of the Islamic Republic.

Methodology and data

Nine out of 18 interviewees maintain public profiles on Instagram. These individuals agreed to be publicly identified and their Instagram accounts' data be used for this analysis. In addition, we identified three other individuals whose work and "celebrity"

status expose them to similar threats and aggression on social media platforms. Eventually, a total of 12 accounts were selected for our final analysis, all of whom represent civil society and public figures of Iranian diaspora. These individuals' work is generally not welcome by the Islamic Republic, increases the possibility of state-affiliated harassment campaigns against them. To answer the research questions, in March 2017, we downloaded the information of all the publicly available posts (pictures and videos) that the 12 accounts in our sample had uploaded ($n = 7893$).[4] We then downloaded all comments that these posts had received. We trimmed the data set by focusing on users that had commented on at least two different sample accounts. This provided a final data set of 1,243,318 comments by 96,283 unique users.[5]

As part of a multi-step process, we first identified the topics that these commenters had discussed for RQ1. Due to the large corpus of comments, a manual content analysis was not feasible. Therefore, we conducted an automated analysis to identify topics that contained harmful speech. This was possible through topic modeling, a statistical technique that identifies co-occurrence of lexical topics in a large corpus of documents (Blei et al., 2003; Roberts et al., 2014). In any topic model, two general parameters are estimated: the probability of each word belonging to a specific topic (per-topic word distribution) and the probability of the topics for each document (per-document topic distribution). The probabilities are usually normalized for each document and sum up to 1. We used the R package STM (Roberts et al., 2014) to calculate the topic model.

A challenge, however, is that for topic modeling to render reliable results, some information in documents is required. Single comments, in particular, are typically too short and consist only of a few words. For every unique user, we created one document that included all comments written by the same user ($n = 96{,}283$ documents). This is a common approach to addressing said challenge and often results in more comprehensive topic models (Hong and Davison, 2010). We removed punctuation marks and numbers as well as all stop words from the text corpus using different Farsi stop word lists (Kharazi, 2017; Taghva et al., 2003). As the use of emojis in Instagram commenting has become common, we included emojis in our analysis as single words. The number of topics (k) has to be defined before the analysis of topic models. We first identified the statistically optimum number of topics and checked all solutions between 10 and 100. We settled to use a topic model with 60 topics (Murzintcev, 2016).

To identify the temporal dimension of politically charged hate speech for RQ2, we identified pictures from our sample that had received the largest number of comments that contained these topics. Since we used a relatively high number of topics, there are more topics with a lower probability in our data set. Therefore, we decided to use a low threshold of at least 0.1. Users that have at least a probability of .1 for one of the hate speech topics were classified as disseminators of harmful speech.[6] For each user, we then combined the probability of two vulgar harmful topics and three politically charged hate speech topics. If a user had a probability of over .1 for one of the two forms of harmful speech, we classified him or her under the corresponding category. We applied the same threshold to the other topics. While the exact publication date of comments on Instagram is not available, we settled to use the publication date of the posts that had been commented on as a date.

For RQ3, we analyzed the co-commenting structure of the 96,283 unique users through network analysis. For each Instagram post, we confirmed the users who wrote a comment. In this network, every node is a user and every edge indicates that two users have at least commented once on the same post. The edge-weight shows the number of pictures that these users have commented on together. We then identified six distinctive co-commenting groups of users with the Louvain algorithm (Blondel et al., 2008).[7] The co-occurrence edge lists were created with an R script created by the authors. We then created an undirected graph and used the Louvain algorithm for the community detection with the R package igraph. With no additional information, defining these communities was rather convoluted. Therefore, we combined the results of the topic model with the information about co-commenting communities.

Finally, we downloaded for RQ4 all information for a subset of users who appear in a co-commenting community with higher frequency of disseminating politically charged harmful speech to detect any relation in regard to who they followed. This allowed us to further analyze how these users were connected with one another, whether they shared the same interests and ideology, and thus followed the same accounts (e.g. political leaders, singers, or football players).

RQ1: topics in the comments

Most of these 60 topics reflect typical communications on Instagram. Two topics consist almost entirely of emoji. The majority of topics reflect the general themes of what the interviewees publish on Instagram. These are typical Instagram discussions in which users may argue with one another; express reverence for a favorite celebrity, that is, one of the interviewees; or broadly talk about politics, religion, and/or society.

However, five topics could be classified as potential harmful speech to which various users contributed. Two of these used derogatory and explicitly sexist language. Particularly, the names of two women interviewees appeared with high probability in one of these topics. The remaining three topics were far more implicit in their lexicon and used less directly identifiable derogatory terms (see Figure 3 left). These topics mainly included nationalistic and religious vocabulary. One centered on the moral behavior of women (e.g. "promiscuous" and "engaged in wedlock"). The second includes explicit words referring to the sexual behavior of women (e.g. "prostitute," "whore"). The third topic mainly includes nationalistic language (e.g. homeland, nation) and some insults (e.g. "dirty," "treasonous dog"). Separately, we also identified one counter-speech[8] topic and a loosely defined "religion and society" topic, covering politics, religious debates, and social debates (see Figure 3 right).

RQ2: the temporal dimension of harmful speech

We identified two forms of temporal patterns that are illustrated in the following two cases:

> Sadaf Taherian is a former actress who left Iran in 2015 after she published unveiled photos of herself and faced intense backlash from Iranian authorities (*The New York*

Figure 3. One of the harmful speech topics specifically targeting women (left). General religious topic (right). The size of the words indicates the probability of the word in the topic.

Times, 2015). Taherian has been working as a professional model since 2016. She often experiences harsh attacks on Instagram, which contain both politically charged statements as well as explicit sexual harassment. As Figure 4 indicates, on the outset of her controversial move abroad, many users discussed the political implications of her departure under her Instagram posts. These users were often accompanied by others who disseminated political hate speech topics. The pattern shifted over time, and she became a frequent target of vulgar, harmful speech, even surpassing the topic of "religion and society."

In contrast, Aram Bolandpaz attracted very few comments from users using the political hate speech topic. In August 2016, when she made her coming out public on Instagram, many participated in the "religion and society" topic. This post also attracted more comments from those using vulgar, harmful speech (see Figure 4). Yet many commenters with political affiliation implicitly criticized her actions using terms that mostly appeared in the "religion and society" topic. According to Bolandpaz, many also reached out to her through direct messaging with violent, threatening statements.

RQ3: co-commenting patterns

With the help of the Louvain algorithm (Blondel et al., 2008), we identified six distinctive co-commenting communities. With no additional information, defining these communities was rather convoluted. Therefore, we combined the results of the topic model with the information about co-commenting communities. For each community, we calculated the average probability for the relevant harmful speech as well as a few other distinctive topics (see Figure 5). One particularly small community consisted of 1424 users and indicated the highest probability for hate speech and "religion and society" topics. This was a clear distinction from all other groups who had a much higher probability for the topics consisting of positive emojis, for example. Yet, the same community of 1424 users had a notable presence in counter speech topic. This can be explained through the methodological approach: it is highly likely that the content subject to hate speech simultaneously attracted support of the fan base of our interviewees (see more in RQ4: follower networks and backgrounds of users).

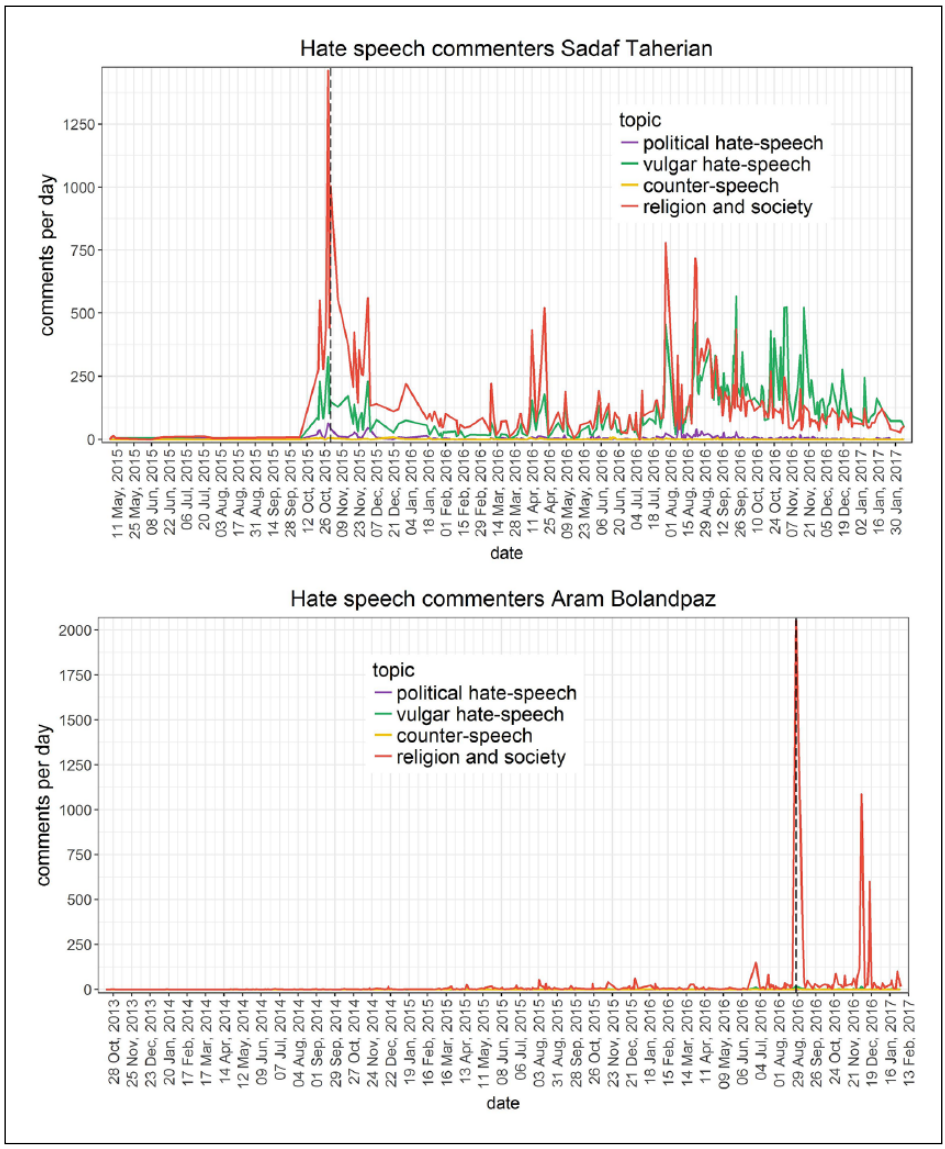

Figure 4. Timeline of different types of harassment on the Instagram account of Sadaf Taherian and Aram Bolandpaz. Dashed line (above) indicates the moment Taherian's unveiled pictures went viral. Dashed line (below) indicates the date Bolandpaz published the video of her coming out on Instagram.

RQ4: follower networks and background of users

To obtain more information about the relatively small community who disseminated political hate speech, we downloaded the follower network of all users in this community.

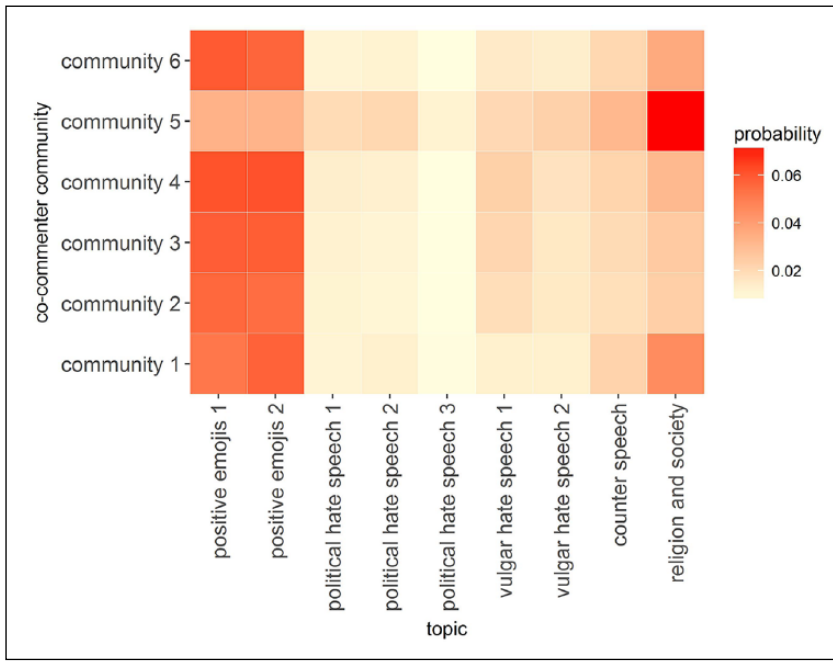

Figure 5. Heatmap with the probabilities for the most relevant topics in the co-commenting communities.

This allowed for distinguishing these users and their interests and backgrounds more clearly. 17% of the accounts in this community had been deleted, and 53% were not accessible due to their privacy settings. Therefore, we could only analyze the follower relations of 429 accounts (30%) of said community.

The internal follower relations affirm a clear divide between two groups of users in this community (see Figure 6). The content posted by the most popular accounts in one group suggests conservative and religious interests of account owners. For example, one of the most prominent accounts in this community mainly posts pictures with negative references to the political opposition in Iran.[9] However, there are also popular accounts with religious usernames in this community that mostly posts pictures about their religious and political ideology. Notably, one of these users commented on three different accounts from our sample, using derogatory language, for example, "the cunt who lies," "you and your goddamn family deserve hell." The second group stood contrary to the first one. They publish content in support of the opposition and post critical content of Iran's political establishment.

In addition, we compared the average probability for the hate speech topics between the two groups. The community that hinted a strong religious ideology (yellow: $M_1 = .019$, $SD_1 = .014$; $M_2 = .025$, $SD_2 = .014$) had a significantly higher probability for both vulgar and politically charged harmful speech topics than the group of users supporting the opposition, orange: $M_1 = .013$, $SD_1 = .008$; $M_2 = .017$, $SD_2 = .01$; Welch's t test: $t(150.92) = 3.53$, $p < .001$; $t(116.84) = 3.97$, $p < .001$.[10]

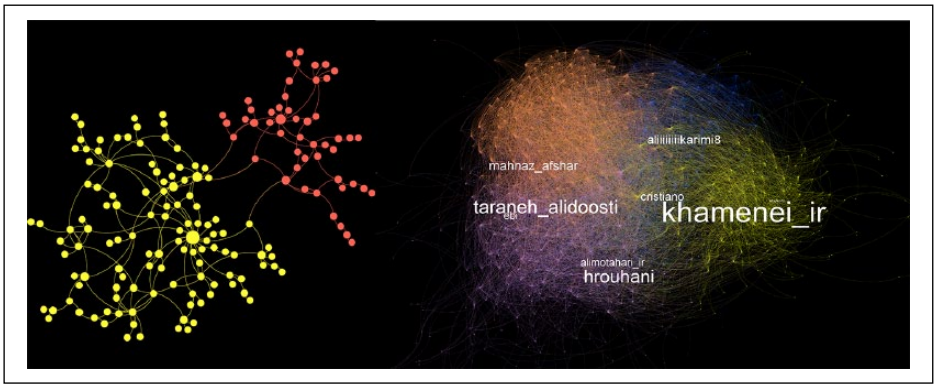

Figure 6. Left: Internal follower network of the co-commenting community. Religious accounts are on the left side (yellow), liberal accounts on the right side (orange). Right: Follower network including the additional 200 most followed accounts. The yellow community shows strong religious orientation; the blue mainly consists of men interested in football; the orange is a more progressive community interested in art and culture; and the violet consists mostly of conservative politicians. Visualizations were created in Gephi.

To further identify the political orientation of these accounts, we identified who they followed and added 200 of the most followed accounts to the analysis. With this additional information, the political divide became even more evident. The results revealed that the majority of accounts that engage in politically charged hate speech follow the highest authorities of Iran. This in particular includes the Supreme Leader of Iran (*khamenei_ir*) whose ideological fanatics follow his lead on many topics (see Figure 6, yellow community). Further inspection of these accounts disclosed that these are mostly religious or were related to the security and intelligence forces of Iran.

As captured in Figure 6, next to the religious community (yellow), we identified a community with mostly men interested in football (blue) following the football players such as *cristiano*—the Portuguese football star, Cristiano Ronaldo—and *aliiiiikarimi8*—one of the most famous Iranian football players—and a conservative political community (violet) following *iran_alimotahar*—a conservative member of the Parliament—and *hrouhani*—the current president. The orange community consists of users interested in pop culture, including popular singers and actors. Users from the non-religious community in Figure 6 were identified commenting on the same pictures as the religious users did. But, they did not participate in disseminating hate speech. The presence of these ostensibly progressive users explains why the counter speech topic also appeared in this community.

Discussion

Our study examined the extent to which harassment can be weaponized against vulnerable communities, particularly, in an extrajurisdictional manner. The following discusses the potential alignment of perpetrators analyzed herein with state actors in Iran and the impact of these repertoires of suppression.

Weaponized harassment and disinformation

State-imposed Internet filtering in Iran remains a key factor in choices that the public makes to adopt and use any social medium. Circumvention tools are widely accessible and their application is commonplace (ASL19, 2013). However, adoption of these tools is subject to a number of factors: (1) censorship apparatus shifts gears rather quickly to restrict access to circumvention tools that bypass censorship; (2) While using circumvention tools is not illegal, it is perceived as a means to accessing outlawed content. The availability of Instagram in Iran can therefore explain the appeal for state-aligned actors to coordinate harassment campaigns. Perpetrators prefer non-filtered platforms to demonstrate their adherence to the rule of law and the leadership of Iran.

Evidence of this form of abuse includes the terms and tone used by various accounts, timing of comments, and choice of profile pictures and captions, some of which signal a vivid religious or political affiliation. In this study, we identified a group of perpetrators that were highly likely to appear together in hateful attacks against the individuals in our sample. As many interviewees indicated, several attackers had signaled in their threatening messages that they had learnt about their targets in private forums, for example, Telegram groups. According to a few interviewees, several accounts had attempted to convince them that to avoid further harassment, they better remain silent about sensitive topics. Such evidence suggests that the abusive behavior documented herein may be an organized effort to collectively silence these targets.

Nevertheless, attribution of attacks remains a challenge. In some cases, it is highly complex to differentiate between patriotic fanatics and those officially organized by the state. In many cases, state-backed media such as the Islamic Republic of Iran Broadcasting (IRIB) or websites affiliated with IRGC incited violence against our sample by covering fabricated or manipulated news about these individuals. This may have triggered extreme reactions from the patriotic and religious audiences of these media.

While the financial motivation of perpetrators cannot be confirmed, there is evidence that the Islamic Republic has an abundance of volunteers to bolster its ideology on social media. In February 2017, Iran announced the recruitment of 18,000 "volunteers" to monitor and report "unlawful content" to the Iranian authorities (BBC Persian, 2017). This may signal other potential recruitments for participation in mobs and other abusive behavior in cyberspace. Our findings indicate that propaganda and ideology are potent drives for targeting of influential diaspora figures. And social media offers a more affordable and pernicious way of suppressing these voices in a manner that was not possible prior to the proliferation of networked communications.

The evolving dynamics of repertoires of suppression

Practices outlined herein demonstrate a shift in the way that the Islamic Republic engages in strategic communications. As critical voices extend their protest repertoires to nontraditional platforms like Instagram, the state also learns to adapt its suppression repertoire to these new mediums (Heydemann and Leenders, 2011). For example, faux reports of abuse or violation of terms of service by anonymous users

are increasingly common practices that aim to take dissidents content down or have their accounts suspended. Multiple journalists and activists whom we interviewed confirmed that they had been subjected to similar silencing tactics. In response, targets often try to verify their accounts with social media platforms—one of the only available remedies against impersonation and false flags (Kargar, 2017). However, not many succeed due to the cumbersome, and sometimes contradictory, practices of these platforms to grant verification badges to applicants.

Polarized debates about the potential of the Internet in the context of authoritarian regimes are fairly common (Rauchfleisch and Schäfer, 2015; Sullivan, 2014). China, for example, has its own unique repertoires of suppression (for an overview, see Roberts, 2018). Optimists like Shirky (2011) consider the Internet's potential and its affordances for protesters and dissidents living under authoritarianism. Skeptics like Morozov (2011), however, anticipate that authoritarian regimes eventually turn the Internet into the "Spinternet," a "Web with little censorship but lots of spin and propaganda" (p. 117). The two positions are less contradictory if the dynamic nature of protest repertoires is considered: dissidents in authoritarian regimes continuously adapt their strategies—a process that can be best described as a "never-ending cat and mouse game" (Endeshaw, 2004: 41). In the absence of an official ban, Instagram has become an open battleground between the arms of propaganda, that is, state-aligned users and ideological fanatics, on one hand, and loyal fans and ordinary witnesses on the other.

In this context, weaponized harassment and disinformation seek two objectives: first, to strategically control the information flow to manipulate public opinion, garner support, and discredit opponents (Nyst and Monaco, 2018) and second, to curb "dissenting" speech and behavior. In these cases, political violence is instrumentalized to undermine trust in the work, ethics, and personalities of targets (Rid, 2013). Unlike other repertoires of suppression, for example, the Chinese model, propaganda, and state-approved content are unlikely to survive the diversity of communities and points of views on the Persian-language online spheres. Therefore, manipulation of public opinion may not have a potent chance to succeed.

However, the second objective, that is, suppressing critical voices is more complex to evaluate. While there is limited evidence on the emotional and psychological effects of online harassment (Feinstein et al., 2016), our interviews suggest that targets have internalized a long-term sense of precaution about their professional activities and the ways in which they engage in public debates. Several journalists in our sample emphasized that they have become more conservative in addressing socially and politically challenging topics such as lesbian, gay, bisexual, and transgender (LGBTQ) rights. They believed that exerting more caution would protect their families from suffering because of their sensitive work (Human Rights Watch, 2018). Curiously, none attributed such reservation to self-censorship.

As Rid (2013) argues, violence induced through computer codes in cyberspace is "physically, emotionally, and symbolically limited" (pp. 20–21). While this was immediately evident in the case of male interviewees, women still endured higher emotional and physical pressure as a result of weaponized, and in many instances sexualized, harassment. For women, cyber violence was not only virtual. It could have real implications

for their physical safety. This gendered feature of cyber abuse observed herein corresponds to the experience of women in other authoritarian context and their encounter with weaponized harassment (Fincher, 2018; Sperling, 2014).

Final note

We explored the affordances of social media not only as they relate to dissidents but also in regard to the repertoires of suppression and how repressive regimes may utilize new media to their own interests. Further research on the insidious effects of weaponized harassment is required to shed light on the long-term, tacit consequences that targets, and in particular women, bear.

Our study is limited in scope to a specific subset of Iranian diaspora as targets of state-aligned harassment and to perpetrators who collectively engage in these campaigns. Future research can be illuminating for the formation of harmful speech, counter speech, and the communications that occur in between. While automated classification of users through topic models worked well in this study, it has some limitations. For example, more granular and implicit hate speech topics cannot be captured with our method. In multiple instances, users could not be appropriately identified as they used potential hate speech in only one of their many comments. As a result of using normalized probabilities, other non-hate speech topics stand out with a higher probability. In multiple other cases, distinguishing counter speech from hate speech could only be done through manual analysis. An example is when the same words are used by different commenters, some attacking the target, while others defending the same individuals. In such cases, only an extensive manual analysis can help to distinguish one use case from any other.

Finally, social media platforms have established policies, community standards, and algorithms to weed out harassment and abuse. However, it has become increasingly complicated to detect state-sponsored, state-executed, state-affiliated, or state-encouraged harmful speech that primarily targets the opposition groups abroad. Our study only partially explored the challenges that platforms encounter, in particular, when their technical solutions can be gamed by adversaries and their policies may not sufficiently cover the emerging threat models. Extensive multi-stakeholder collaboration can pave the way toward addressing these challenges.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD


Adrian Rauchfleisch https://orcid.org/0000-0003-1232-083X

Supplemental material

Supplemental material for this article is available online.

Notes

1. All interviewees explicitly consented to the analysis of their Instagram communication.
2. During the presidential election in 2017, then incumbent president Rouhani and his rivals utilized Instagram features such as Instagram Live to maximize spreading of their campaign messages (Tabrizy, 2017). Amid protests that broke out in December 2017, Instagram and Telegram were temporarily blocked between 31 December 2017 and 13 January 2018 (Deutsche Welle, 2018).
3. We interviewed Aram Bolandpaz in December 2016. She expressly authorized authors to use the interview transcript for any research purposes to become available to the public, academics, and media. She was previously a documentary producer with the popular Persian television, Manoto TV, which is based in London, UK.
4. See Appendix 2 of Supplemental material for an overview of the data sampling procedure.
5. All data were publicly available and downloaded with a script written in R. See Appendix 1 of Supplemental material for more information.
6. We manually checked 50 randomly selected users who were classified as hate speech users as well as 50 randomly selected users who were classified as users without hate speech. In both cases, we identified five misclassified users. This is a built-in limitation of this method that can be considered as its margin of error.
7. The modularity score is .3.
8. Counter-speech is a common, crowd-sourced response to extremism or hateful content directed at undermining such speech. Extreme posts are often met with disagreement, derision, and counter-campaigns (Bartlett et al., 2015).
9. The username is a reference to Iran's controversial presidential election in 2009. The username refers to the "Alzheimer" disease and the discourse of oblivion. The account frequently publishes radical comments such as "I sacrifice my life for Khamenei."
10. The political hate speech topics also had a significantly higher probability in the religious group.

References

Anderson C (2013) Dimming the Internet: detecting throttling as a mechanism of censorship in Iran. *Arxiv*. Available at: https://arxiv.org/abs/1306.4361

Anderson C and Sadjadpour K (2018) Iran's cyber threat: espionage, sabotage, and revenge. Available at: http://carnegieendowment.org/2018/01/04/iran-s-cyber-threat-espionage-sabotage-and-revenge-pub-75134

ARTICLE 19 (2017) *Iran: Backsliding on UN Free Expression Commitments*. Available at: https://www.article19.org/resources/iran-backsliding-on-un-free-expression-commitments/

Aryan S, Aryan H and Halderman JA (2013) Internet censorship in Iran: a first look. Available at: https://www.usenix.org/conference/foci13/workshop-program/presentation/aryan

ASL19 (2013) Information controls: Iran's presidential elections. Available at: https://asl19.org/cctr/iran-2013election-report/

Baha'i International Community (2011) Inciting Hatred; Iran's media campaign to demonize Bahá'ís. Available at: https://www.bic.org/sites/default/files/pdf/inciting-hatred-book_0.pdf

Bartlett J and Krasodomski-Jones A (2015) *Counter-Speech Examining Content That Challenges Extremism Online*. London: Demos.

BBC Persian (2017) Virtual Basij; 18,000 volunteers to report online violations. Available at: http://www.bbc.com/persian/iran-38909342

Blei DM, Ng AY and Jordan MI (2003) Latent Dirichlet allocation. *Journal of Machine Learning Research* 3: 993–1022.

Blondel VD, Guillaume J, Lambiotte R, et al. (2008) Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment* 2008(10): P10008.

Bowen K (2015) Rewiring Iran's supreme council of cyberspace. Available at: https://smallmedia.org.uk/news/rewiring-irans-supreme-council-of-cyberspace

Burns A and Eltham B (2009) Twitter free Iran: an evaluation of Twitter's role in public diplomacy and information operations in Iran's 2009 Election Crisis. *Communications Policy & Research Forum* 2009: 322–334.

Center for Human Rights in Iran (2015) *Khamenei Consolidates Power Over Internet Policy in Hard Line Council He Controls*. Tehran, Iran: Center for Human Rights in Iran.

Chen A (2015) The agency. Available at: https://www.nytimes.com/2015/06/07/magazine/the-agency.html

Clark JD, Faris RM, Morrison-Westphal RJ, et al. (2017) *The shifting landscape of global internet censorship*. Berkman Klein Center Research Publication 2017–4. Berkman Klein Center. Available at: https://dash.harvard.edu/bitstream/handle/1/33084425/The%20Shifting%20Landscape%20of%20Global%20Internet%20Censorship-%20Internet%20Monitor%202017.pdf

Committee to Protect Journalists (2015) 10 most censored countries: Iran. Available at: https://cpj.org/2015/04/10-most-censored-countries.php#7

Dehghan SK (2017) BBC appeals to UN over Iran's crackdown on journalists. Available at: https://www.theguardian.com/media/2017/oct/25/bbc-pleads-with-un-over-iran-crackdown-on-journalists

Deutsche Welle (2018) Iran unblocks Telegram messenger service shut down during country-wide protests. Available at: http://www.dw.com/en/iran-unblocks-telegram-messenger-service-shut-down-during-country-wide-protests/a-42141829

Endeshaw A (2004) Internet regulation in China: the never-ending cat and mouse game. *Information & Communications Technology Law* 13(1): 41–57.

Fassihi F (2009) Iranian crackdown goes global. Available at: https://www.wsj.com/articles/SB125978649644673331

Feinstein A, Feinstein S, Behari M, et al. (2016) The psychological wellbeing of Iranian journalists: a descriptive study. *JRSM Open* 7(12): 2054270416675560.

Fincher LH (2018) *Betraying Big Brother: The Feminist Awakening in China*. London: Verso Books.

Golkar S (2013) *The Islamic republic's art of survival: neutralizing domestic and foreign threats*. Policy Focus no. 125, June. Washington, DC: The Washington Institute for Near East Policy.

Guarnieri C and Anderson C (2016) Iran and the soft war for internet dominance. Available at: https://iranthreats.github.io/us-16-Guarnieri-Anderson-Iran-And-The-Soft-War-For-Internet-Dominance-paper.pdf

Heydemann S and Leenders R (2011) Authoritarian learning and authoritarian resilience: regime responses to the "Arab awakening." *Globalizations* 8(5): 647–653.

Hong L and Davison BD (2010) Empirical study of topic modeling in Twitter. In: Melville P, Leskovec J and Provost F (eds) *Proceedings of the First Workshop on Social Media Analytics.* New York: ACM, pp. 1–9.

Human Rights Watch (2018) Iran: activists' families facing harassment: authorities using state TV to discredit journalists, dissenter. Available at: https://www.hrw.org/news/2018/08/09/iran-activists-families-facing-harassment

Kargar S (2017) Verification as a remedy for harmful speech online. In: Jansen Reventlow N, et al. (eds) *Perspectives on Harmful Speech Online*. Harvard Berkman Klein Center for Internet & Society Research Publication, pp. 46–48.

Kharazi V (2017) Persian stop words list. Available at: https://github.com/kharazi/persian-stopwords (accessed 28 October 2017).

Khiabany G (2008) The Iranian press, state, and civil. In: Semati M (ed.) *Media, Culture and Society in Iran: Living with Globalization and the Islamic State*. London: Routledge, pp. 17–36.

Khodabakhshi L (2017) Iran's Instagram election sees rivals battle on social media. Available at: http://www.bbc.com/news/world-middle-east-39947080

Mollanazar H (2011) Text screening (Censorship) in Iran: a historical perspective. *Iranian Journal of Applied Language Studies* 3(2): 159–186.

Morozov E (2011) *The Net Delusion: How Not to Liberate the World*. London: Allen Lane.

Murzintcev N (2016) ldatuning: tuning of the Latent Dirichlet allocation models parameters. Available at: https://rdrr.io/cran/ldatuning/man/ldatuning.html

Naji K (2018) BBC UN appeal: stop Iran harassing Persian service staff. Available at: https://www.bbc.com/news/world-middle-east-43334401

Nyst C and Monaco N (2018) State-sponsored trolling: how governments are deploying disinformation as part of broader digital harassment campaigns. Available at: http://www.iftf.org/fileadmin/user_upload/images/DigIntel/IFTF_State_sponsored_trolling_report.pdf

Pearce KE (2015) Democratizing kompromat: the affordances of social media for state-sponsored harassment. *Information, Communication & Society* 18(10): 1158–1174.

Rahimi B (2003) Cyber dissent: the internet in revolutionary Iran. *Middle East Review of International Affairs* 7(3): 101–115.

Rahimi B (2008) The politics of the Internet in Iran. In: Semati M (ed.) *Media, Culture and Society in Iran: Living with Globalization and the Islamic State*. London: Routledge, pp. 37–56.

Rauchfleisch A and Schäfer MS (2015) Multiple public spheres of Weibo: a typology of forms and potentials of online public spheres in China. *Information, Communication & Society* 18(2): 139–155.

Reporters without Borders (2017) One of the world's biggest prisons for journalists. Available at: https://rsf.org/en/iran

Rid T (2013) *Cyber War Will Not Take Place*. New York: Oxford University Press.

Roberts ME (2018) *Censored: Distraction and Diversion inside Chinas Great Firewall*. Princeton, NJ: Princeton University Press.

Roberts ME, Stewart BM, Tingley D, et al. (2014) Structural topic models for open-ended survey responses. *American Journal of Political Science* 58(4): 1064–1082.

Shirky C (2011) The political power of social media: technology, the public sphere, and political change. *Foreign Affairs* 90: 28–41.

Small Media (2015) Writer's block: the story of censorship in Iran. Available at: https://smallmedia.org.uk/work/writers-block-the-story-of-censorship-in-iran

Software Freedom Law Centre (2016) *Online Harassment: A Form of Censorship*. New Delhi, India: Software Freedom Law Centre.

Soleimani S (2016) A history of censorship: Iran's religious dictatorship and the ruling thought police. *Infinite Mile*. Available at: https://infinitemiledetroit.com/A_History_of_Censorship,_Irans_Religious_Dictatorship_and_the_Ruling_Thought_Police.html

Sperling V (2014) *Sex, Politics, and Putin: Political Legitimacy in Russia*. Oxford: Oxford University Press.

Sullivan J (2014) China's Weibo: is faster different? *New Media & Society* 16(1): 24–37.

Supreme Council of Cultural Revolution (1988) Goals, policies and guidelines for book publishing. Available at: https://ketab.farhang.gov.ir/fa/principles/bookprinciples67

Tabrizy N (2017) Live on Instagram at 3 a.m.: Iranian presidential candidates. Available at: https://www.nytimes.com/2017/05/19/insider/live-on-instagram-at-3-am-iranian-presidential-candidates.html

Taghva K, Beckley R and Sadeh M (2003) *A list of farsi stopwords*. ISRI Technical Report 2003-01, 2003. Las Vegas, NV: Information Science Research Institute, University of Nevada.

The New York Times (2015) 5 uncovered Iranian actress who posted photos online not wearing a Hijab forced to flee country. Available at: http://nytlive.nytimes.com/womenintheworld/2015/11/04/iranian-actress-who-posted-photos-online-not-wearing-a-hijab-forced-to-flee-country/

Vatanka A (2015) Iran abroad. *Journal of Democracy* 26(2): 61–70.

Woolley SC and Howard PN (2018) *Computational Propaganda: Political Parties, Politicians, and Political Manipulation on Social Media*. Oxford: Oxford University Press.

Author biographies

Simin Kargar is a researcher on human rights and technology and a research affiliate of Berkman Klein Center for Internet and Society at Harvard University. She focuses on hate speech and gender-based violence online, and the interplays of social media, power and propaganda. Her current research addresses information operations by state and non-state actors designed to manipulate public opinion and silence dissidence. Email: skargar@law.harvard.edu

Adrian Rauchfleisch is an assistant professor at the Graduate Institute of Journalism at the National Taiwan University. He is currently writing a book with Jonas Kaiser for Oxford University Press. It deals with the far-right in Germany and the U.S. and how they (ab)use social media platforms. Email: adrian.rauchfleisch@gmail.com