The Geopolitics of Deplatforming: Which Politically-Interested Iranian Accounts get Suspended on Twitter?

Mehdi Zamani[*] Andreu Casas[†]

Abstract

Citizens increasingly rely on social media to express opinions and engage in politics. In recent years social media companies have played a more active role in regulating freedom of speech by removing content and accounts. Initially companies did so to improve the health of the platforms, battling bots and toxic behavior. However, there are today many discussions around whether mainstream platforms such as Facebook, Instagram and Twitter suspend users for ideological reasons. We tackle this question from a geopolitical angle. The U.S. government pressures US-based platforms to suspend the accounts of people and organizations included in their international sanctioning plans, as well as those who support them. Despite some transparency efforts from the platforms, we know very little about the impact that these suspensions have on the political conversations in the platforms. After identifying about 600,000 Twitter users interested in Iranian politics, for a six-month period we track their activity and whether they get suspended. We then use computational methods to assess differences between suspended and non-suspended users. We find that after accounting for many alternative explanations, conservative (Principlist) users and those supportive of the Iranian government are much more likely to be suspended. From a theoretical perspective, the results contribute to the debate on deplatforming by emphasizing its (geo)political dimension and consequences. From a more practical point of view, the study contributes to keeping accountable the companies regulating our online environment, and to inform public and policy debates on the topic.

[*] University of Amsterdam
[†] Vrije Universiteit Amsterdam

1 Introduction

Today private social media companies play a crucial role in regulating freedom of speech (Balkin, 2017; Rogers, 2020). People around the world increasingly rely on social media to consume news (Shearer and Mitchell, 2021), learn and talk about politics (Barberá et al., 2019), and coordinate political actions (González-Bailón et al., 2011). Despite many initial positive views about the role of social media for enhancing more inclusive, equal and free political conversations (Bennett and Segerberg, 2013; Howard and Hussain, 2012), the platforms are increasingly suspending accounts (a phenomenon commonly known as "deplatforming") to address concerns about incivility, bots, fake news, rumors, and conspiracies (Bastos, 2021; Bay and Fredheim, 2019). Moreover, in recent years many have claimed that widely-used platforms such as Facebook, YouTube and Twitter also suspend accounts for political reasons, allegedly targeting conservatives (Davalos and Brody, 2020) as well as voices supportive of governments involved in a geopolitical rivalry with the West, such as China, Russia, Venezuela or Iran (O'Sullivan and Moshtaghian, 2020). Exploring potential political biases in the suspensions of social media users by these widely-used platforms is crucial for assessing how these companies shape online political environments.

There are many studies on how non-democratic governments (e.g. China) control local social media companies (e.g. Weibo, RenRen) to constrain political speech in their countries (King, Pan, and Roberts, 2013, 2014), yet research on political deplatforming by widely-used social media companies (e.g. Facebook, Twitter) is on its infancy. Existing research finds that accounts in these platforms are most often suspended for engaging in toxic behavior (Bay and Fredheim, 2019), such as using uncivil language, spreading false information or acting an automated way (bots). However, some also find that accounts posting about politics (Chowdhury et al., 2020), and those reflecting particular views (Bastos, 2021), are also suspended at higher rates. These studies however rarely control for potential confounders (e.g. accounts posting about politics may be suspended because they also engage in toxic behavior at a higher rate) and lack clear theoretical explanations for their findings.

Furthermore, existing research has overlooked crucial elements of the geopolitical role of social media and deplatforming. Governments around the world are well aware that information disseminated on social media can have meaningful effects on citizens, such as to motivate them to protest (Larson et al., 2019), turn out to vote (Bond et al., 2012), and spread misinformation on many topics (Guess, Nagler, and Tucker, 2019). In turn, governments engage in strategic social media campaigns to influence attitudes and behaviors at home but also abroad. The actions for example of the Russian Internet Research Agency (IRA) in the last few U.S. elections are well known (Golovchenko et al., 2020).

However we know much less about how Western countries such as the United States influence social media communications for geopolitical reasons; nor their effects on who gets to have a voice in these platforms. As part of its international sanctioning plans, the United States maintains a list of individuals and organizations (SDN: the *Specially Designated Nationals And Blocked Persons List*), whose assets are blocked, and U.S. citizens and organizations are prohibited from dealing with them. In response, US-based social media companies suspend accounts that believe are connected to any person or group in this list. Despite some transparency efforts from the companies (Twitter for example regularly publishes information about accounts they suspend, and the reasons, in different countries – including information on those suspended for supporting "*state-linked information operations*"),[1] it is often unclear how the publicly-released data was selected (we only observe the output of a black box), nor we can fully assess the impact of these suspensions on their respective environments given that we lack information about the non-suspended accounts and the overall political environment they inhabited.

In this paper we study the conditions under which Twitter users interested in Iranian politics are suspended, exploring the ways in which the company shapes the political conversations around Iran on the platform. Several Iranian elites and organizations are in the SDN list, including the *Islamic Revolutionary Guard Corps* (IRGC), a branch of the Iranian Army responsible for protecting the country from foreign interventions. After a U.S. drone strike killed general Qasem Soleimani (January 2020), many reported that the accounts of users who expressed their discontent, as well as the accounts of some journalists simply covering the event, were suspended. The question remained: how do these suspensions affect the political discussions around Iranian politics on the platform? After controlling for engaging in toxic behavior, are users holding particular ideological views still more likely to be suspended than others? If so, do these suspended users differ on the type of political views they discuss in the platform?

There are many challenges to the study of deplatforming biases (Rogers, 2020). First, some platforms (e.g. Facebook) do not allow independent researchers to collect and analyze user-level data for ordinary users, making it impossible to study deplatforming beyond the suspension of a few salient users. Second, even when looking at platforms that do allow for the study of ordinary accounts (e.g. Twitter), suspensions are likely to be rare, and so a large sample of interest needs to be drawn in order to be able to detect meaningful variations. Finally, accounts may be suspended for many non-political reasons, such as for spreading misinformation and using uncivil language. Researchers need to find ways to control for many potential confounders.

In March 2020 we identified 601,940 Twitter users who followed Iranian politics, and for a six-month period we periodically collected the messages they posted in the platform and checked whether they had been suspended. Most of the accounts (N=594,852) remained *active* after the period of analysis, yet many were (at least temporarily) *suspended* by Twitter (N=3,737), or *deleted* by either the user or Twitter (N=3,351). We use state-of-the-art computational methods to assess potential ideological differences between the *active* and *suspended* users, after controlling for many confounders; and explore the types of conversations that in turn were repressed *vs.* amplified as a result of such suspensions.[2]

[1] See for example information released here: https://blog.twitter.com/en_us/topics/company/2021/disclosing-networks-of-state-linked-information-operations-

[2] Some of the methods used in the analysis are very computationally intensive and unfortunately analyzing the full sample of users and tweets in this first draft of the paper turned out to be unfeasible. Instead, for this draft we randomly sampled 30,000 users (out of the ones that had not been deleted nor suspended) to be

2 Material and Methods

2.1 Sampling

In order to assess the effects of deplatforming on political conversations, we needed to find a set of politically-interested users to study. First, we identified the accounts of a group of Iranian elites on Twitter: the Iranian Supreme Leader, all members of the tenth Iranian Parliament ($N = 136$), a cabinet member of the Rouhani administration ($N = 20$), Iranian news media outlet ($N = 19$), for a total of 175 elite accounts. Then we used the Twitter's REST API to pull the list of followers for each of these elite accounts (a total of 2,410,543 unique followers). To make sure these followers were indeed interested in politics, we sampled for analysis those that followed at least 3 of the 175 elite accounts: 601,940 users in total.

2.2 Data collection

We tracked the activity of these 601,940 users between March 11th and September 10th, 2020. On the starting date of the data collection, we retrieved the last 3,200 tweets posted by the followers prior to this date. Subsequently, every other week[3] we collected every new tweet posted by a follower since the starting date of the data collection; a total of 65,120,890 tweets. In addition, we monitored the activity status of the followers within this time period to record which accounts stopped being active during this time period ($N = 7,088$). However, accounts could have stopped being active either because Twitter suspended or deleted the account, or because the user decided to delete it. For this reason, on October 22nd 2020, we manually checked each of the inactive accounts for whether they were: (a) deleted ($N = 3,351$), (b) suspended ($N = 2,491$), or (c) active again ($N = 1,246$). We dropped the deleted ones from the analysis as we are not certain whether those had been suspended by Twitter (or the user). In the analysis we focus on comparing the ones that had been at some point (temporarily or permanently) suspended ($N = 3,737$), to the remaining still-active accounts ($N = 594,852$).[4]

2.3 Ideology

The key objective of this research was to assess political/ideological biases in the suspension of Twitter accounts. To accomplish this goal we needed to find ways of mea-

compared to the 3,737 suspended accounts. Finally, we dropped from this sample of 33,737 users those that did not send any tweet in 2020, given that part of the analytical relies on the content of the users' tweets (and so on the users being somewhat active). The final sample we use for analysis in this draft is composed of 11,859 users (2,340 suspended accounts and 9,519 non-suspended) and the 5,932,777 tweets they sent in 2020.

[3] We did not automatically set up a script to regularly pull the data, but instead relied on manually executing the script. We did so the following days: 2020-04-16, 2020-04-20, 2020-04-27, 2020-05-06, 2020-05-08, 2020-05-14, 2020-06-08, 2020-06-18, 2020-07-05, 2020-07-10, 2020-07-16, 2020-07-20, 2020-07-31, 2020-08-17, 2020-09-03, and 2020-09-10.

[4] See Footnote 1.

suring the ideology the Twitter user/accounts under study. We measured two key ideological dimensions in Iranian politics: where do users fall in the left-right (Reformist-Principlist) continuum, but also how supportive they are of the Iranian government (which claims to be neutral and to not align with the policy stances of the different reformist-principlist factions in Parliament).

To measure the ideology of the users in the reformits-principlist space, we adapted, to the Iranian context, a validated and widely used method (*Bayesian Spatial Following model*) for measuring the ideology of elite and ordinary Twitter users into a single left-right dimension (Barberá, 2015). The model relies on the homophily assumption that ordinary users follow political elites that better reflect their views, and it has been validated and found to produce accurate ideology estimates for Twitter users in the U.S. context. To estimate the model, we first built a bipartite network graph with information about which of the 167 elite accounts each of the 601,940 users in the full sample followed. Then, in order to be able to compute the model,[5] we randomly sampled 5,000 users, who followed most of the elite accounts in the list (152 out of 167, so 91%). Then, we used the `mediascores` package (Eady et al., 2020) to fit the model, obtaining ideology scores (in the same dimension) for the 5,000 users and the 152 elite accounts; finally using the trained model (and so the scores given to the elite accounts) to estimate the ideology of the remaining ordinary users (based on the elites they followed). We validate the method in Appendix A, where we show how the resulting ideology scores did a good job at distinguishing between known left-leaning (reformist) and right-leaning (principlist) users in our dataset.

We used a text-based machine learning method to measure the extent to which the accounts were supportive of the Iranian government. First, from all Farsi, Arabic and English tweets sent in 2020 by the users we tracked, we sampled 2,893 messages, some at random and some following an active-learning procedure. Second, we manually annotated them for whether they were about politics (N = 2,893), and if so (N = 1,228), for whether they were against or in favor of the Iranian government (or neutral). Then, we fine-tuned (twice) a BERT multilingual model, creating two binary classifier capable of accurately predicting whether a message was political, and if so, whether it was supportive of the Iranian government: *Political* (44% true positives) 83% accuracy, 81% precision, and 83% recall; *Pro-Iran* (39% true positives) 72% accuracy, 63% precision, 69% recall. In Appendix C we provide further information about the training and accuracy of these two models. Finally, we used these models to generate probabilistic $(0 - 1)$ political and pro-Iran predictions for all tweets in our dataset that were sent in 2020, and then aggregated the predictions at the user level, measuring the average and maximum predicted support for the Iranian government for each of the users in the dataset.

2.4 Incivility

We wanted to control for the alternative explanation that accounts are suspended for being uncivil. We fine-tuned the same BERT multilingual model to build a binary text

[5] The model is very computationally intensive, so one needs to reduce the size of the input matrix in order to be able to estimate the model in a reasonable time frame.

classifier predicting whether a message was uncivil (10% true positives: 94% accuracy, 66% precision, and 46%recall). We also selected a random set of messages (and some others selected via active-learning) and annotated them (N = 1,151) as uncivil if they contained any of the following markers: (1) made violent threats against an identifiable group, (2) incited fear about a group/community, (3) wished, hoped, or called for serious harm on an individual or group, and (4) made references to violent events. We used the model to classify all tweets sent in 2020 by the users we tracked, and then generated an analytical variable measuring the number of uncivil tweets sent by each user in 2020. In Appendix C we provide further information about the manual annotation, training and accuracy of this model.

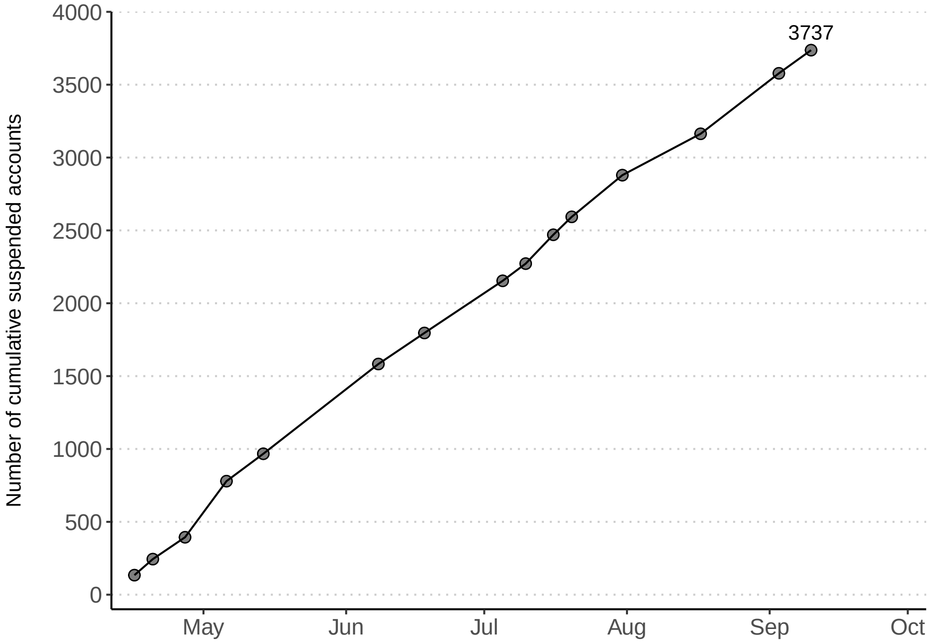
2.5 Coordination

We also wanted to control for the alternative explanation that accounts are suspended for engaging in coordinated (malicious) actions. We used the following empirical strategy to identify coordination among accounts we tracked. Building on the premise that coordinated accounts share very similar (if not the same) content, we developed a four-step protocol to measure messaging similarity between all possible pairs of users. First, we used the same BERT multilingual model to generate (768-size) tweet-level embeddings for all messages sent in 2020 by the users we tracked, by passing the tweets through the pre-trained BERT architecture and pulling the output of the second-to-last (fully connected) layer. Second, for each user we generated a (768-size) user-level embeddings by averaging the indexed embedding values of all the user's tweets. Third, we calculated the cosine similarity between all possible pairs of user embeddings. And finally, we used these cosine similarities to generate two user-level variables measuring the *average* and *maximum* content similarity (and so likely coordination) between a given user and all/any other user/s in the dataset.

2.6 Misinformation

Finally, we wanted to control for the alternative explanation that accounts may be suspended for spreading misinformation. Studies of misinformation or fake news in Western countries usually build upon pre-designed list of known fake-news sites to then check which users share posts linking to those sites. Such curated list does not exist, and is very difficult to put together, for Iran. Instead, we decided to identify some hashtags in our dataset that we could clearly link to misinformation content. This is also a difficult strategy to apply to content on any potential topic. However, given that the platforms were very concerned about the spread of misinformation related to covid, we focus on identifying covid-misinformation hashtags. First, we generated a list of keywords that would help us identify covid-related messages, and the we manually annotated a random sample of 1,000 messages for whether they contained misinformation. Next, we selected the unique hashtags in those coded as containing covid-related misinformation, and went back to the full dataset to pull 10 random messages containing each of those hashtags. After coding those 10 messages per hashtag again for whether they contained misinformation, we treated as clear covid-misinformation

Figure 1: Cumulative number of accounts that we tracked and were suspended during the period of analysis.



hashtags those for which at least 8 of the 10 random messages had been coded as containing misinfo. Finally, we generated the following individual-level variable for all the users in our dataset: the number of messages they sent in 2020 that contained one the covid-misinformation hashtags.

3 Results

In Figure 1 we show the number of cumulative suspensions detected among the 601,940 users that we tracked between May and October 2020, a total of 3,737. Each dot corresponds to a moment in time when we checked whether the accounts were still active (and collected any new message sent since the last check). Hence about 0.7% of the users had been suspended after the five months. Although it represents a non-trivial amount of users, it may seem a small percentage. Yet the clear linear trend in Figure 1 suggests that Twitter assesses and suspends accounts in an incremental fashion, and that we would have uncovered an even larger number of suspensions if we had tracked these accounts for a longer period of time.

Clear differences emerge already when simply comparing the suspended and non-suspended users on many theoretically-relevant descriptives (see Table 1). First, unsurprisingly, the data seems to indicate that at least part of the suspended users were suspended for engaging in malicious behavior. They were much more active than non-suspended ones, indicative of (at least some) bot-like behavior, sending on average 1,147 messages in 2020 (compared to 340 for non-suspended). In fact, about 27% of the suspended accounts were among the 90th percentile of most active users over-

Table 1: Descriptive statistics (with 95% confidence interval) for Suspended and Non-Suspended users.

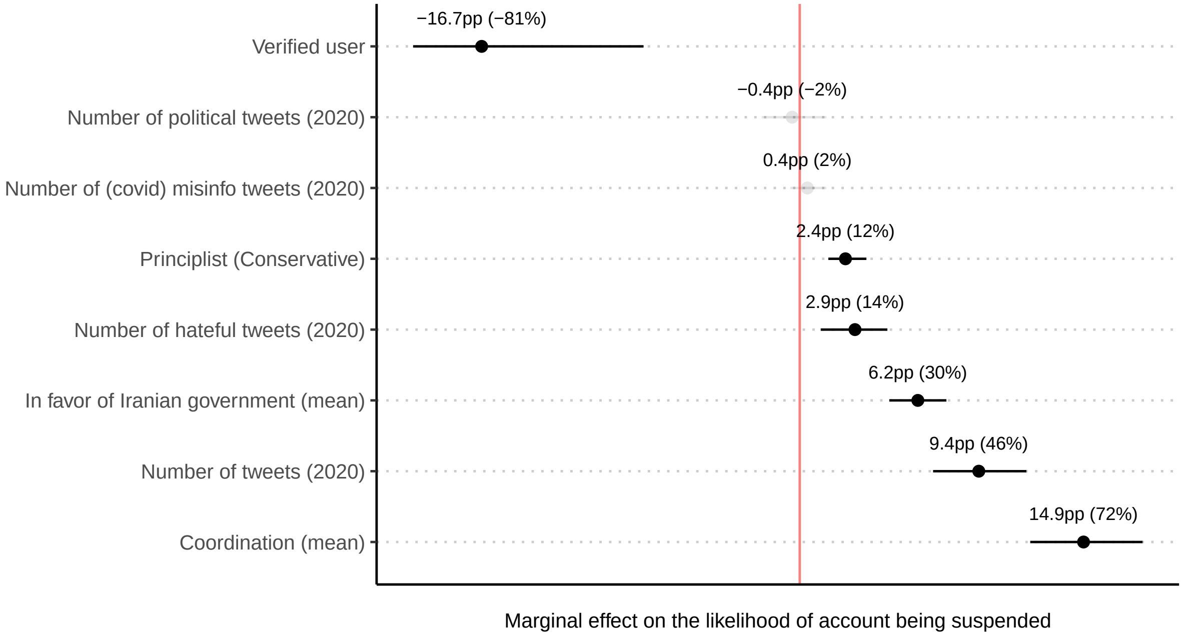
	Non-Suspended	Suspended
Prop. of verified users	0.003	0.001
Avg. Number of tweets (2020)	340 [316-364]	1,147 [1,080-1,213]
Prop. in the 90th most active percentile (2020)	0.06	0.27
Avg. Number of political tweets (2020)	125 [115-135]	412 [384-440]
Avg. Prop. of political tweets (2020)	0.35 [0.34-0.35]	0.37 [0.36-0.38]
Avg. Number of hateful tweets (2020)	1 [1-2]	6 [5-6]
Avg. Prop. of hateful tweets (2020)	0.007 [0.006-0.008]	0.007 [0.005-0.008]
Avg. Number of Covid-Misinfo tweets (2020)	0 [0-0]	1 [0-1]
Avg. Coordination score {0-1}	0.961 [0.959-0.962]	0.978 [0.977-0.98]
Avg. Prop. In favor of Iranian government {0-1}	0.356 [0.351-0.361]	0.446 [0.438-0.454]
Avg. Principlist (Conservative) score {0-1}	0.169 [0.167-0.171]	0.188 [0.183-0.192]

all (*vs.* 6% for non-suspended). In addition, we find that suspended users posted a much larger number of uncivil tweets (6 on average, compared to 1 for non-suspended), posted content more similar to what other accounts also posted (potentially indicative of malicious coordination: 0.978 *vs.* 0.961 coordination score), and posted more tweets promoting covid-related misinformation (1 *vs.* 0, on average).

More importantly, these bivariate comparisons also reveal substantive ideological differences. In the last two rows of Table 1 we observe suspended users to be substantially more supportive of the Iranian government as well as more ideologically conservative. On average, we observe for example that 45% of the political tweets sent by suspended users were supportive of the Iranian government, compared to 36% for non-suspended users. The Principlists (conservatives), as well as those supportive of the Iranian government, particularly stand for a stronger and more independent Iran at the international arena, specially *vis-a-vis* the United States.

We provide more stringent evidence for these differences in Figure 2, where we show the results of a multivariate logistic regression predicting suspensions. In particular, we show the marginal effect (expressed both as percentage-point and as percentage difference) of a one standard deviation change for numeric variables, and of being a verified (*vs.* non-verifed) user. In line with Tab. 1, we find that verified users are less likely to be suspended. We also observe that engaging in malicious behavior (posting more tweets, content similar to other accounts (coordination), and uncivil messages) is highly predictive of an account being suspended. More importantly, in line with our theoretical expectations, and with what we also observed in Tab. 1, we find that the two ideological measures of interest (conservatism and support for the Iranian government) are also strong predictors of suspension; findings that are robust to many model specifications (see Appendix B). A one standard deviation increase in conservatism (Principlism) is correlated with a 12% increase in the likelihood of suspension (2.4 percentage points). The same increase in support for the Iranian government is also predictive of a 30% increase in the chances of being suspended (6.2 percentage points). First, the findings show that accounts are in part suspended to reduce toxic and malicious behavior and improve the health of the platform. However,

Figure 2: Logistic regression predicting whether an account was suspended. Marginal effects expressed in percentage points (pp) and percentual change (%).

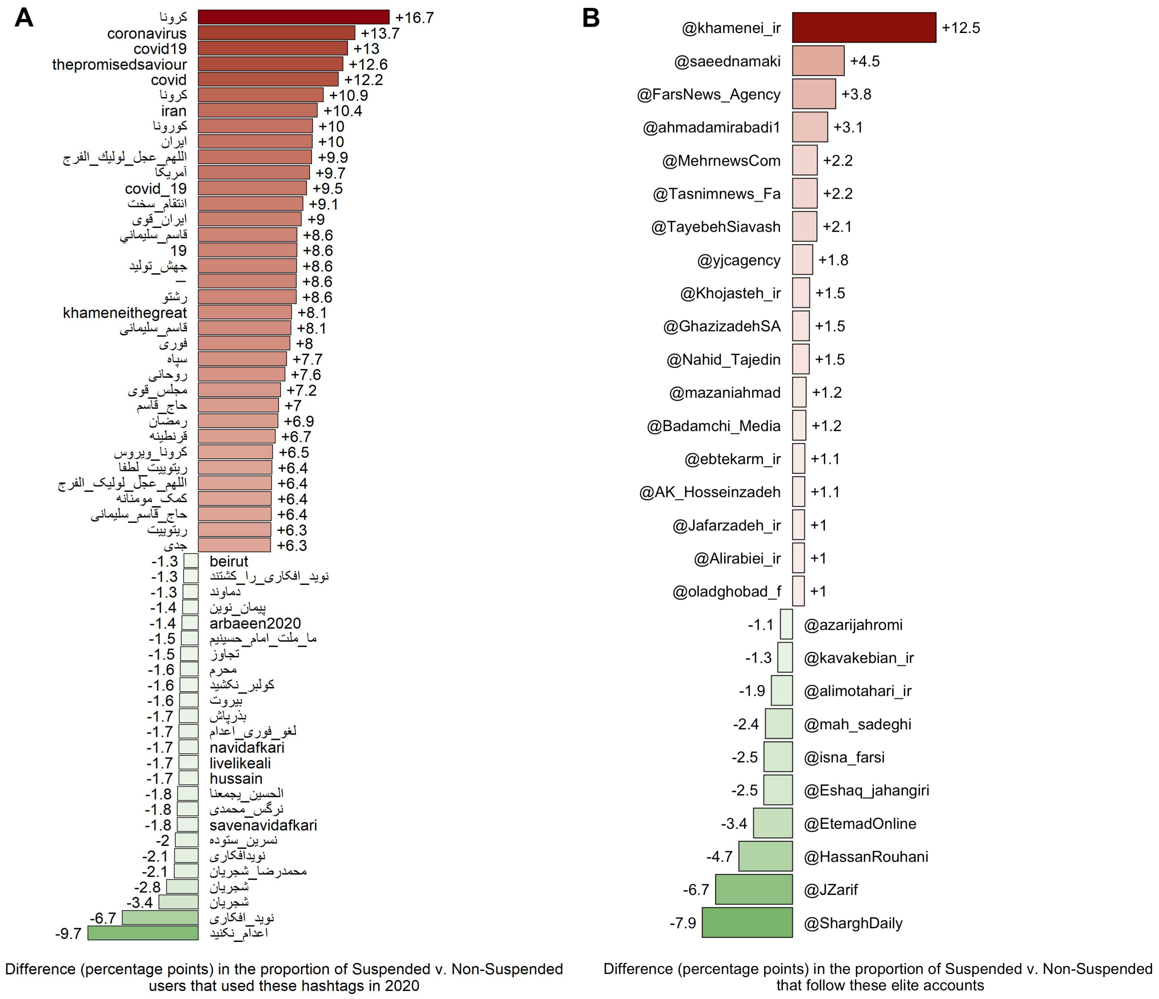


the findings also show some clear political biases in the suspension of users. These political biases are likely to be the result of Twitter suspending accounts that mention, engage with, praise, etc., people and organizations included in the SDN list (e.g. IRGC or Qasem Soleimani). However, in Tab. 1 and Fig. 2 we observe that by doing so, the company is clearly influencing which ideological views get to have a voice in the platform; advancing the geopolitical interests of the United States.

To shed more light into these ideological biases, in Figure 3.A we analyze the content of the users' posts and explore the hashtags most often used by suspended *vs.* non-suspended users. For each hashtag used by any of the users under analysis, we first calculated the proportion of suspended and non-suspended users who used the hashtag in any of their tweets in 2020, and then calculated the difference between the suspended and non-suspended proportions. In Fig. 3 we also analyze their networks and use the same procedure (comparing the proportion that follows each elite) to explore which elite accounts are most often followed by suspended *vs.* non-suspended users. The positive (and red) bars are hashtags and elite accounts most often used/followed-by suspended, and the green ones are most often used/followed-by non-suspended.

Fig. 3.A illustrates the type of content that was repressed *vs.* emphasized as a result of the suspensions. First, we observe that (at least some) suspended users posted about covid at a much higher rate than non-suspended users. Many of the hashtags at the

Figure 3: Differences in hashtag usage (A), and elite following (B), between suspended and non-suspended users.



top of Fig. 3.A are related to coronavirus, such as کرونا, coronavirus, and covid19. In line with Tab. 1 and Fig. 2, and given that we know that social media platforms strongly battled covid-related misinformation around that period, we believe that this is an indication that some of the accounts were suspended for engaging in malicious activities.

However we also observe many relevant political and ideological differences. Among the hashtags most often used by the suspended users, we can also find some that allude to Qasem Soleimani (the general of the IRGC killed by a U.S. drone: e.g. قاسم سلمانی) and that ask for *hard revenge* (انتقام سخت). We also observe other hashtags representing some of the common Principlist narratives, such as such as ایران قوی (*strong Iran*) and جهش تولد *(production growth)*. On the contrary, we can see that many hashtags critical with the Iranian government were dis-proportionally used by non-suspended users, and so prompt to be amplified by the suspensions. For example, hashtags against the execution of Navid Akfari, a wrestler who was executed in 2020 for murdering a security guard in 2018, such as اعدام■نکنید (*do not execute*), نوید■افکاری (*Navid Afkari*), savenavidafkari, livelikeali, and navidafkari.

We find similar ideological biases in Fig. 3.B. Among the most followed elites by suspended users, we find the Supreme Leader of Iran (Khamenei), some of his close cabinet members, as well as some state-own media corporation (@FarsNews_Agency). On the contrary, among the most followed elites by the non-suspended users, we find the former President Hassan Rouhani and members of his cabinet, who negotiated the nuclear deal with the United States and who defend a softer foreign policy agenda, compared to the current administration; as well as some independent and privately-owned media companies (e.g. @SharghDaily).

4 Discussion

The political relevance of social media platforms is in the rise: an increasing number of citizens around the world use them to consume news and to learn and engage in politics. To combat malicious behavior, the platforms suspend accounts that e.g. use hateful language and spread misinformation. Many in the last years however also accuse Western social media platforms, such as Facebook and Twitter, to suspend accounts for political reasons. We looked into this question from a geopolitical perspective. Although there is been a lot of research on how non-Western countries use social media for (geo)political reasons (e.g. Russia, China), little is known about how a Western country such as the United States can leverage its international sanctioning plans to force US-based social media companies to suspend accounts (and influence political conversations) to advance its geopolitical interests.

For a five month period in 2020, we tracked about 600,000 Twitter accounts interested in Iranian politics. About 4,000 of them had been suspended after the period of analysis. We find two overarching patterns when comparing suspended and non-suspended accounts, and when using multivariate regressions to model suspension. First, accounts that engaged in different kinds of toxic/malicious behavior (e.g. posted messages at an abnormal/bot-like rate, used uncivil and hateful language) were more likely to be suspended. But then, even after accounting for these malicious activities, we find clear ideological suspension biases: Principlists (conservative) accounts and those supportive of the Iranian government were more likely to be suspended.

This research makes many relevant contributions to the emergent literature on political deplatforming. First, by emphasizing its geopolitical role, it provides (and illustrates) a clear theoretical framework and expectations about the conditions under which accounts may be suspended for ideologicla reasons. Second, it puts forward a research designs that not only allows for clear comparisons between suspended and non-suspended accounts, but that it also does not rely on data publicly-released by the platforms, which we never fully know how it was curated. Third, it contributes to inform the normative debate about the role that private companies in general, and social media platforms in particulars, should play in the regulation of (political) speech. Finally, the paper also contributes to reveal crucial public information and to keeping accountable the companies regulating our online environment.

References

Balkin, Jack M. 2017. “Free speech in the algorithmic society: big data, private governance, and new school speech regulation.” UCDL Rev. 51: 1149.

Barberá, Pablo. 2015. “Birds of the same feather tweet together: Bayesian ideal point estimation using Twitter data.” Political analysis 23(1): 76–91.

Barberá, Pablo, Andreu Casas, Jonathan Nagler, Patrick J Egan, Richard Bonneau, John T Jost, and Joshua A Tucker. 2019. “Who leads? Who follows? Measuring issue attention and agenda setting by legislators and the mass public using social media data.” American Political Science Review 113(4): 883–901.

Bastos, Marco. 2021. “This Account Doesn’t Exist: Tweet Decay and the Politics of Deletion in the Brexit Debate.” American Behavioral Scientist 65(5): 757–773.

Bay, Sebastian, and Rolf Fredheim. 2019. Falling Behind: How Social Media Companies Are Failing to Combat Inauthentic Behaviour Online. NATO StratCom COE.

Bennett, W Lance, and Alexandra Segerberg. 2013. The logic of connective action: Digital media and the personalization of contentious politics. Cambridge University Press.

Bond, Robert M, Christopher J Fariss, Jason J Jones, Adam DI Kramer, Cameron Marlow, Jaime E Settle, and James H Fowler. 2012. “A 61-million-person experiment in social influence and political mobilization.” Nature 489(7415): 295–298.

Chowdhury, Farhan Asif, Lawrence Allen, Mohammad Yousuf, and Abdullah Mueen. 2020. On Twitter Purge: A Retrospective Analysis of Suspended Users. In Companion Proceedings of the Web Conference 2020. pp. 371–378.

Davalos, J., and B. Brody. 2020. “Facebook, Twitter CEOs Sought by Senate Over N.Y. Post Story.” Bloomberg: https://www.bloomberg.com/news/articles/2020-10-15/facebook-twitter-chided-anew-by-repu
.

Eady, Gregory, Richard Bonneau, Joshua A Tucker, and Jonathan Nagler. 2020. “News Sharing on Social Media: Mapping the Ideology of News Media Content, Citizens, and Politicians.” OSF Preprint: osf.io/ch8gj (Nov).

Golovchenko, Yevgeniy, Cody Buntain, Gregory Eady, Megan A Brown, and Joshua A Tucker. 2020. “Cross-platform state propaganda: Russian trolls on Twitter and YouTube during the 2016 US presidential election.” The International Journal of Press/Politics 25(3): 357–389.

González-Bailón, Sandra, Javier Borge-Holthoefer, Alejandro Rivero, and Yamir Moreno. 2011. “The dynamics of protest recruitment through an online network.” Scientific reports 1(1): 1–7.

Guess, Andrew, Jonathan Nagler, and Joshua Tucker. 2019. "Less than you think: Prevalence and predictors of fake news dissemination on Facebook." Science advances 5(1): eaau4586.

Howard, Philip, and Muzammil Hussain. 2012. Democracy's Fourth Wave?: Digital Media and the Arab Spring. Oxford University Press.

King, Gary, Jennifer Pan, and Margaret E Roberts. 2013. "How censorship in China allows government criticism but silences collective expression." American political science Review 107(2): 326–343.

King, Gary, Jennifer Pan, and Margaret E Roberts. 2014. "Reverse-engineering censorship in China: Randomized experimentation and participant observation." Science 345(6199).

Larson, Jennifer M, Jonathan Nagler, Jonathan Ronen, and Joshua A Tucker. 2019. "Social networks and protest participation: Evidence from 130 million Twitter users." American Journal of Political Science 63(3): 690–705.

O'Sullivan, D., and A. Moshtaghian. 2020. "Instagram says it's removing posts supporting Soleimani to comply with US sanctions." CNN: https://edition.cnn.com/2020/01/10/tech/instagram-iran-soleimani-posts/index.html .

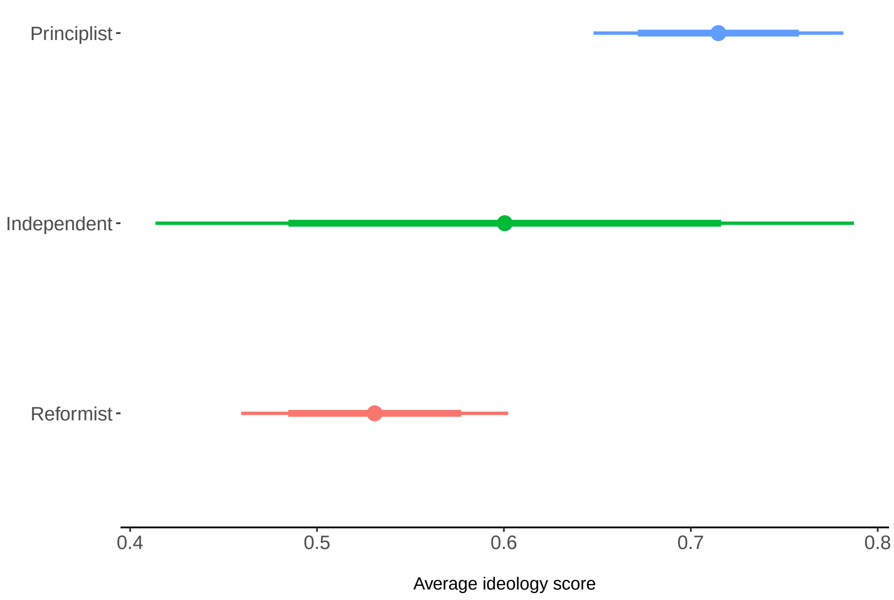
Rogers, Richard. 2020. "Deplatforming: Following extreme Internet celebrities to Telegram and alternative social media." European Journal of Communication 35(3): 213–229.

Shearer, Elisa, and Amy Mitchell. 2021. "News use across social media platforms in 2020.".

Appendix A Validation of the ideology score

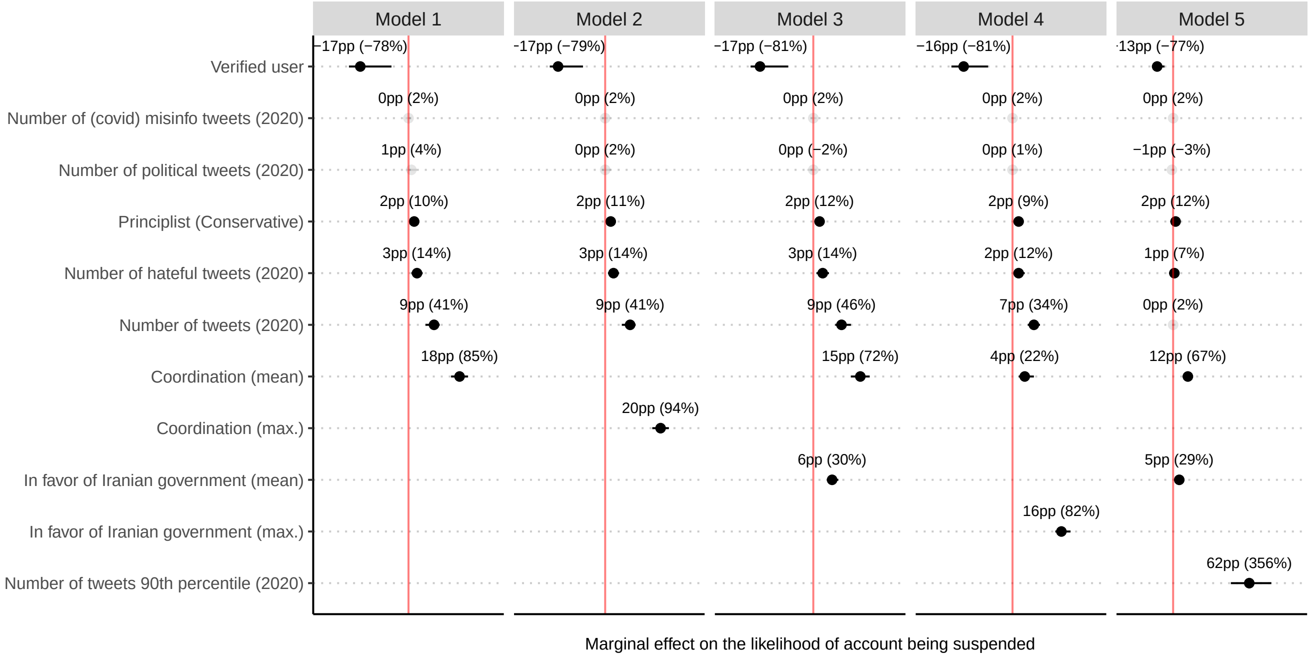
We conducted the following validation exercise to make sure that the *Bayesian Spatial Following model* adapted well to the Iranian context. We checked the average ideology score given by the model to members of Parliament that are known to be affiliated to either the Reformist, Independent, and Principlist factions in the chamber. We report the average ideology scores for these three groups in Figure A1, where we observe the method to perform as expected and to generate ideological scores that do a good job at distinguishing between reformists and principlists. Also as expected, the model estimated independents to have an average ideology between the averages estimated for reformists and principlist. The confidence interval is rather large for the Independents due to the low number of members in this group.

Figure A1: Average ideology score for elite accounts known to be Principlists (conservative), Reformists (liberal) and Independent.



Appendix B Alternative model specifications

Figure B1: Logistic regressions predicting whether the account was suspended.



Appendix C Performance of the text classifiers

Table C1: Performance of 3 BERT-multilingual models predicting political, hateful, and pro-Iran tweets.

	Labeled	**Negative**	**Positive**	**Epochs**	**Accuracy**	**Precision**	**Recall**	**F-Score**
Political	2,893	64%	44%	6	83%	81%	83%	82%
Hateful	1,228	90%	10%	10	93%	66%	46%	53%
Pro-Iran	607	61%	39%	8	72%	63%	69%	66%