(Dis)Information Wars*

Adrian Casillas† Maryam Farboodi‡ Layla Hashemi§

Maryam Saeedi¶ Steven Wilson‖

August 25, 2023
Preliminary and Incomplete, Please do not distribute

Abstract

With the unprecedented rise of internet access across the globe, social media platforms have emerged as prominent vehicles for displaying dissent. In response, oppressive governments have engaged in spreading fake news on these platforms. We identify a new form of disinformation supply on social media—*disinformation wars*: intentional spread of fake news while pretending to be an ordinary account. We propose a novel approach to curb the spread by assigning a *disinformation score* that represents the propensity of spreading disinformation, intentionally or not, to all the platform users. We demonstrate that this new form of disinformation supply is considerably more effective in spreading fake news on Twitter compared to traditional propaganda, and assess the effectiveness of a score disclosure policy in limiting the spread of disinformation on the platform.

*This paper was prepared for the 2023 AI Authors' Conference at the Center for Regulation and Markets (CRM) for publication by the Brookings Institution. We thank the Brookings CRM for financial support. We would also like to thank Azam Heydari, Maryam Nazari, Ali Shourideh, David Yang, and many others who help us throughout this project. All the remaining errors are ours.

†MIT Sloan

‡MIT Sloan, NBER and CEPR

§George Mason University

¶Carnegie Mellon University

‖Brandeis University

1 Introduction

In countries with authoritarian regimes, traditional means of information dissemination such as newspapers, TV, and radio are heavily controlled by the central government. In the early 21st century, the introduction of social media and decentralized platforms proved to be a groundbreaking development in these countries. The rapid improvement of big data technologies enhanced convenient access at an unprecedented rate and made these platforms prominent vehicles for displaying dissidence during episodes of unrest.

Soon after, authoritarian regimes intervened by limiting internet access and censoring social media. In parallel, they started to spread disinformation on social media through propaganda accounts—accounts who are publicly pro-government and spread false news in an attempt to change the narrative in favor of the government. However, the widespread access of the public to multiple sources of information has reduced the effectiveness of this tactic. As such, governments have turned to a smarter approach to supply fake news. They engage in a *"disinformation war,"* i.e., they create *imposter accounts* who spread fake news on social media platforms while pretending to be unbiased, ordinary accounts.

Disinformation wars have several advantages as they do not require exerting force and they are difficult to trace, yet they disrupt the flow of information. As such, they derange the opposition movement without apparent aggression.[1] Furthermore, it is difficult to identify the pieces of fake news that are originated by imposter accounts, as these accounts imitate the behavior of ordinary accounts in many respects.

Although both strategies spread fake news, unlike classic propaganda, disinformation wars are not intended to control the narrative in favor of the central government. Rather, they are a means to disturb the narrative in order to derail and discredit the protest movements. In order to delineate the supply of fake news on social media platforms, it is crucial to understand both strategies concurrently.

As the prevalence of disinformation wars continues to grow, it becomes crucial to address the challenges they present. Previous studies have highlighted real-time content moderation and ex-post efforts to eliminate audience biases as essential strategies to contain the spread of fake news, but they have limitations. Real-time fact-checking or content moderation is time-consuming, allowing disinformation to go viral before it can be addressed; moreover, ex-post debunking shows limited impact on debiasing the public. In this paper, we propose an alternative approach to restrict the supply of disinformation—ex-ante content moder-

[1]This approach aligns with the recent shift in dictators' tactics, wherein they exert control over the public through the manipulation of truth, rather than relying solely on force Guriev and Treisman (2022, 2019).

ation. We predict accounts likely to engage in spreading disinformation, even before they do so, and explore the effectiveness of a policy that uses this ex-ante information to limit the spread of fake news on social media.

We apply our method to the disinformation war launched on Farsi-Twitter in the wake of recent protests in Iran. On September 16, 2022, a 22-year-old Iranian woman named Mahsa Amini died in a hospital in Tehran, Iran, after being arrested by the religious morality police of Iran's government for not wearing the hijab in accordance with government standards. Eyewitnesses reported that she died as a result of police brutality which was denied by Iranian authorities. Amini's death resulted in a series of widespread protests across Iran. These protests were primarily concentrated among the younger generation and were accompanied by a lot of activity on social media, particularly on Farsi-Twitter, where the hashtag #MahsaAmini was created and widely used. Concurrently, a surge in disinformation supply across Farsi-Twitter occurred during the protests. Multiple sock puppet accounts, which we call "imposter accounts," were created by certain entities to disseminate fabricated, misleading, and false information.

The imposter accounts pretended to be dissidents, posting pro-dissidence content and hashtags to build a network among the protesters. Subsequently, they started posting disinformation. This disinformation was then spread by other imposter accounts as well as some of the normal Twitter accounts. We call the imposter accounts and the normal accounts who actively engaged in the spread of disinformation, collectively, "unsafe accounts." Put differently, the unsafe accounts either intentionally supply the disinformation to the social network or actively facilitate the widespread supply, intentionally or unintentionally.

We first create a unique dataset comprising all Farsi tweets from September 16, 2019, to March 14, 2023. We augment this data with a network of user relationships, including follower-following connections and retweet interactions. We also gather a comprehensive set of user characteristics within this network, such as date created and activity rates. We then assemble a labeled dataset of unsafe, traditional propaganda, and ordinary accounts. The unsafe accounts are all imposters, i.e., accounts known to have previously created verified disinformation while pretending to be dissidents. The traditional propaganda accounts openly support the government, and ordinary accounts are regular trusted accounts.

The paper has two methodological contributions. First, we propose social network-based characteristics that help identify unsafe accounts. We outline and implement an algorithm to construct these "network proximity measures" for our labeled dataset.

Second, in order to shed light on the supply and spread of disinformation, we devise a classifier to determine whether an account on Farsi-Twitter is unsafe, propaganda, or

ordinary. Our core classifier is a multinomial Logit model that uses these network proximity measures along with non-network account characteristics. We train the multinomial Logit model on a portion of our labeled dataset and use the remainder of the labeled data to test the model.

We then use the trained classifier to assign *"disinformation scores"* and *"propaganda scores"* to all Farsi-Twitter accounts. The disinformation (propaganda) score of each Twitter account reflects the probability of that account being an unsafe (or propaganda) account. More specifically, the disinformation score indicates the likelihood of an account acting as an unsafe account and actively engaging in the dissemination of disinformation, intentionaly or not, and even if it has not yet done so. We find that the usafe accounts—those with a high disinformation score, have a disproportionate share in a wide-spread supply of fake news in the social network, compared to the traditional propaganda accounts.

The disinformation score is a useful instrument to guide the activities of ordinary users as well as to design policy. At the same time, it is potentially prone to manipulation by unsafe accounts. In order to address this concern, we propose an alternative classifier that only relies on network proximity measures and find that this limited classifier achieves 90% accuracy of the baseline in detecting unsafe accounts. The significance of this finding is twofold: First, it highlights the importance of the structure of the social network in detecting adversarial behavior on social media and as such, closely ties the literature on social networks with media economics. Second, since the network proximity measures are difficult to manipulate, as they depend on the past and present structure of the social network, it shows that our disinformation scores are robust to manipulation by unsafe accounts.

We propose two policies using our disinformation score and assess their effectiveness in interrupting the flow of disinformation on social media. The first policy involves blocking tweets by unsafe or propaganda accounts, or both, while the second policy involves disclosing either disinformation scores, propaganda scores, or both.

To evaluate the efficacy of our policies, we analyzed verified instances of disinformation campaigns that occurred during the recent period of unrest in Iran. Our findings indicate that blocking unsafe accounts or disclosing disinformation scores significantly reduces the spread of disinformation among ordinary accounts, leading to a faster cessation of disinformation. However, blocking or disclosing information about propaganda accounts does not have a substantial impact. This finding highlights the importance of detecting the unsafe accounts to contain the supply of information on social media.

A critical advantage of our methodology is that it proactively targets accounts with

a high propensity to engage in disinformation campaigns, even before they do so. Early identification of these accounts provides a tangible opportunity to take preventative steps before the disinformation goes viral, making it an effective mechanism to combat the spread of disinformation. Furthermore, due to its high tractability, this methodology can be applied to a wide range of scenarios for which the spread of disinformation is a problem, thereby bolstering our ability to counteract the negative impact of disinformation proliferation.

1.1 Literature Review

Our paper contributes to the extensive body of literature on the economics of media. One strand of this literature considers media capture by governments and its consequences (Besley and Prat, 2006). A second strand studies the political economy of media censorship. Some papers focus on the government obstructing access to valuable information (Schedler, 2010; Shadmehr and Bernhardt, 2015), while others explore the effects of public demand for uncensored and non-ideological information (Gentzkow and Shapiro, 2006; Chen and Yang, 2019; Simonov and Rao, 2022), another strand studies the impact of change in technologies on news production (Cagé et al., 2020b; Angelucci et al., 2020).

We focus on a less explored intervention employed by authoritarian regimes to influence political outcomes: the deliberate spread of disinformation on social media platforms, aimed at distorting waves of unrest. Gottfried and Shearer (2016) emphasize the significance of social media, providing evidence that approximately two-thirds of adults in the United States access their news through these platforms. Cagé et al. (2020a) show that even mainstream media get impacted by social media news. In their research, Allcott and Gentzkow (2017) delve into the theoretical and empirical aspects of fake news dissemination on social media prior to the 2016 election. Moreover, Thomas et al. (2012) and Stukal et al. (2017) present evidence highlighting the extensive use of false information, particularly on Russian Twitter.

Estimating the volume of misinformation circulating on social media between 2015 and 2018, Allcott et al. (2019) found that user interactions with false content increased steadily on Facebook and Twitter until the end of 2016. However, they also discovered a sharp decline in interactions with false content on Facebook since then, while interactions on Twitter continued to rise. Additionally, Bradshaw and Howard (2018) conducted an examination of organized social media manipulation campaigns in 29 countries worldwide, uncovering evidence of governments employing social media as a tactic for manipulation.

Given the abundance of evidence regarding the use of social media in the manipulation of public opinion, several researchers have explored methods to combat this issue, see

Bak-Coleman et al. (2022). Some researchers have proposed real-time fact-checking and moderation of information. However, Vosoughi et al. (2018) show that false information tends to spread faster than true information. They investigate the differential diffusion of true and false news stories using a comprehensive dataset of fact-checked rumor cascades on Twitter spanning from its inception in 2006 to 2017. Due to the rapid spread of misinformation, real-time moderation appears futile, as disinformation often goes viral before being detected by content moderators. While ex-post rebuttals of disinformation have been extensively studied, their impact has been found to be limited Kunda (1990); Chan et al. (2017); Nyhan and Reifler (2010); Ecker et al. (2022); Kahan et al. (2017).

There is also a strand of theoretical literature on information diffusion, information aggregation, and belief formation. The seminal work of Crawford and Sobel (1982) studies strategic dissemination of information using a cheap talk model. Acemoglu et al. (2010) consider the trade-off between information aggregation and propagation of misinformation on social media. Akbarpour et al. (2020) studies the optimal seeding strategy for information diffusion in networks. Related to our policy experiments, Budak et al. (2011) study theoretical approximation algorithms that can effectively limit the spread of misinformation on social media.

The rest of the paper is organized as follows. Section 2 provides details of the Farsi-Twitter data that we use for estimation. Section 3 describes the algorithm we use to construct the network measures which we then use for estimation, as well as the estimation methodology for scoring. Section 4 presents the results. Section 5 reports the outcome of policy experiments guided by the disinformation scores estimated in the paper. Lastly, Section 6 concludes.

2 Twitter Data

Iran is notorious for its cyber army and their extensive efforts in disseminating disinformation. They employ various tactics, including the use of numerous unsafe accounts on social media platforms to influence the activities of the opposition. In this paper, we focus specifically on their activity on Twitter, as it is often the source of news stories that then spread to other popular social media platforms in Iran such as Instagram or Telegram.

Our analysis focuses primarily on users who were active since before the torrent of protests cascading from the death of Mahsa Amini on September 16, 2022, through the end of Twitter API's v1 streaming service on March 14, 2023.

Our data comes both from Twitter's API and from an archive of all Farsi-language

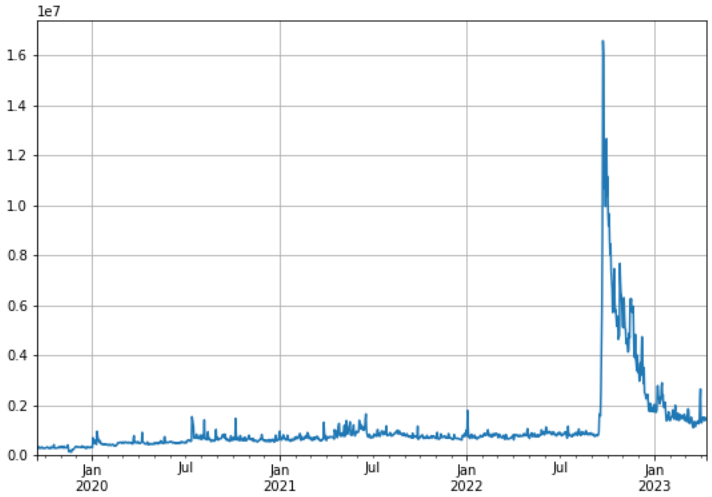


Figure 1: Daily volume of Farsi-language tweets throughout our sample

tweets made after September 16, 2019, which has been stored by Brandeis University. As shown in Figure 1, we observe a substantial jump in the activity on Farsi-language Twitter after the start of the protests in Iran.

Twitter's v1 API tweet stream allowed its users to collect a stream of all incoming tweets filtered on a number of parameters provided that the stream made up less than 1% of Twitter's total tweet volume at any given moment. By including a filter that only lets through tweets tagged as being written in Farsi and given that Farsi language tweets almost never made up more than 1% of Twitter's volume, we were able to archive nearly all Farsi-language tweets since the start of our stream on September 16, 2019.[2]

These tweet objects are rich in detail and vary in kind. All posts contain not just their tweet data but a snapshot of their creator's profile at the time of posting as a user object and, if applicable, a media object and a geo-location object.

All tweet objects contain at least their unique tweet ID, text, timestamp, public metrics (retweet, quote-tweet, reply, and like counts), and creator's information which includes at least their unique user ID, username, screen name, and public metrics (follower, following, and the number of total public tweets) at the time of the tweet's creation. Twitter's API has the courtesy to catalog metadata such as URL links, hashtags, and mentions of other users which can otherwise be scraped from the text. Engagement posts (replies, quote-tweets, and retweets) will also include references to the tweet they are engaging with.

[2]The 1% limit was set by Twitter for streaming tweets using their v1 API.

	Total	Monthly average pre 09/2022	Monthly average post 09/2022
Number of users who have tweeted in Farsi	9,525,800	711,600	1,240,600
Number of active Farsi-Twitter users*	1,767,350	614,000	1,016,300
Number of retweets sent	639,337,800	9,637,000	48,734,300
Number of retweets sent*	589,714,000	8,364,400	44,186,200
Number of quote tweets sent	57,919,800	1,345,300	1,581,500
Number of quote tweets sent*	52,452,100	1,198,945	1,248,300
Number of replies sent	377,199,600	8,264,700	13,278,400
Number of replies sent*	349,949,200	7,840,700	12,080,400

*statistics from users with more than 10% of posts and engagements in Farsi and more then ten Farsi tweets

Table 1: Aggregate and average monthly statistics of Farsi-language tweets pre and post September 2022

Since we collect tweets at the moment they are created, all public metrics are zero. However, given that we have chronicled nearly all Farsi tweets after a given post's creation, we can estimate these values by looking for all future engagements that point to the original tweet. In doing so, we also recreate the entire structure of Farsi-Twitter: every reply thread; every chain of quote-tweets; and every retweet.

Moreover, if we swap each tweet's ID for that of their creator, then we have recounted a history of every engagement between any two Farsi-Twitter users since the start of the stream. Finally, we supplement these engagement networks with the follower-following network of all active Farsi-Twitter users, which is a directed friendship network.

We need to draw a distinction between what we call a Farsi-Twitter user and a user who has posted in Farsi. We limit our sample to users who have at least 10% their posts or engagements in the Farsi language and who have posted at least ten tweets in Farsi and refer to this data as "Farsi-Twitter".[3] Table 1 presents some monthly average statistics of all Farsi tweets before and after the start of the protests in September 2022. The number of Twitter users who have ever posted a tweet in Farsi is very large, with more than nine million users. However, it turns out that a large fraction of these users have posted very few Farsi tweets and do not have any meaningful presence in the Farsi-Twitter network. As such, we limit our sample to users who have at least 10% of their posts or engagements and contributed at least ten posts in Farsi. Table 1 also presents summary statistics limited to this sub-sample of Farsi-Twitter accounts. Comparing the figures makes it clear that

[3]This is the same platform with the same features as the English version of Twitter, except that the text is in Farsi which is the dominant language of Iran.

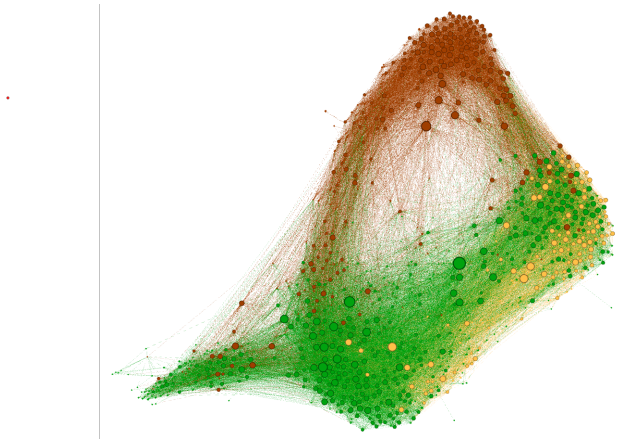


Figure 2: Network structure of followers and followings between the 1,000 most followed Farsi-Twitter users as of December 2022. Yellow nodes represent accounts that are likely to spread disinformation, red nodes represent accounts that are likely to spread propaganda, and green nodes are less likely to engage with either.

although the number of users drops fourfold, the number of interactions barely changes. As such, limiting the data to this sub-sample is without loss of generality.

Figures 2 and 3 sample the structure of the follower-following and retweet engagement networks. Each figure shows the 1,000 largest accounts in each network (i.e., the accounts with the most followers or retweets) and their connections to one another in that network. The size of the node is proportional to the account's centrality in the network and its color is based on the "disinformation score" we compute in the paper. Accounts that have a high propensity to spread disinformation (unsafe accounts) are colored yellow, accounts that spread propaganda are colored red, and ordinary accounts are colored green. There are a few noteworthy observations. First, as shown in Figure 2, most of the accounts of each type are highly interconnected between themselves. However, the exceptions to this are unsafe accounts which follow and are followed by ordinary accounts. This integrated frontier is where disinformation permeates to ordinary users. In this vein, unsafe accounts have been much more successful at blending in with ordinary accounts while most ordinary accounts steer clear of propaganda accounts. This last observation is much more apparent in Figure 3. In this figure, we observe accounts with the highest retweets and engagements, the distinction between this figure and the previous one is staggering. Here, unsafe accounts have a much stronger and integrated presence while the propaganda accounts do not succeed in getting very many engagements from neither ordinary nor unsafe accounts.

Our goal is to use this data to identify accounts that spread disinformation, especially the unsafe accounts that do so with a pointed agenda, as they are more difficult to identify

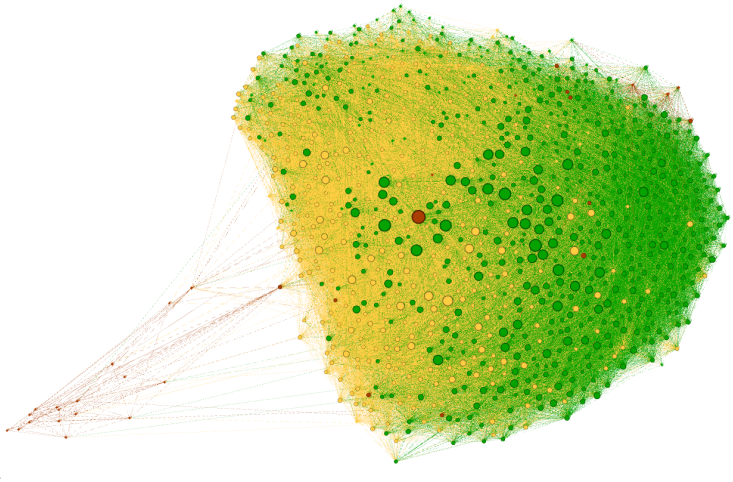


Figure 3: Network structure of retweets between the 1,000 most retweeted Farsi-Twitter accounts since the start of the sample on September 16, 2019. Yellow nodes represent accounts that are likely to spread disinformation, red nodes represent accounts that are likely to spread propaganda, while green nodes are less likely to engage with either.

for the average Twitter user, and, as shown by previous figures, they tend to be very vocal.

Twitter's administrators already work to identify and disrupt bot networks. Traditionally, these are large networks that are run automatically by a small number of machines or manually by a few users in order to boost their engagements. Early on, bot accounts were very easy to spot. They typically had few, if any, followers; their only engagements were retweets; and their activity was often implausible for human users (e.g., retweeting a post the moment after it was created). As they were discovered and banned, they changed their behavior to become less conspicuous, more organic, and harder to discover.

As a baseline, our models consider basic account features and activity statistics such as the number of followers and followings, follower-to-following ratio, account age, rate of posting by post type, and post composition by type. However, we should expect these measures to become less predictive as the arms race between administrators and disinformants forces these accounts to strategically change their behavior to impede detection.

Given that we want to take advantage of the explosion of Farsi-language activity that took place on Twitter following the death of Mahsa Amini in order to identify disinformants who were caught off guard by the reaction, we look at account features and activity before and after the event. We indicate whether an account was created after the start of the protests. We look at the changes in the activity and account features throughout the protests, and how these changes depend on the account having been created after the start of the protests.

We want to top this more traditional model with the keystone of our analysis: measures of the proximity of an account to known unsafe accounts. These measures are constructed from the retweets and following networks illustrated above as we explain in detail in Section 3.2. As mentioned before, knowing not only the direct connections to known unsafe accounts but also the structure of engagements with them can expose both the practice and strategy behind purposefully spreading disinformation. No matter how well unsafe accounts disguise their proxy accounts, their inorganic goal to enhance their engagements will result in inorganic structures. If an account is conspicuously situated to engage with content created by known unsafe accounts, then its intentions should be called into question.

3 Methodology

In order to understand the supply of disinformation on Farsi-Twitter, we classify Farsi-Twitter accounts into three disjoint sets. First, unsafe accounts that consists of two groups. One group consists of imposter accounts who pretend to be part of the opposition, but are in fact operated by government or other agencies. The other group consists of normal Farsi-Twitter accounts who actively participate in spread of disinformation. We do not argue that these latter accounts intentionally spread the fake news, rather, they might be fooled by the imposter accounts impersonating protesters. Second, propaganda accounts, who employ more traditional forms of propaganda and openly support the Islamic Republic of Iran (IRI).[4] The third group consists of users who do not belong to the former two groups, which we call ordinary accounts.

In order to classify accounts, we start with a training dataset and then classify every account based on their similarity to our training data. We use various characteristics of the accounts and their activities for this classification. One crucial set of variables we consider is the position of an account within the network based on four different measures. We provide a detailed explanation of the construction of these variables in Section 3.2.

Besides these network variables, we use a wide range of variables that experienced Twitter users employ to identify unsafe accounts anecdotally. Table 2 lists all of the characteristics of labeled Farsi-Twitter accounts and their mean and standard deviation, across all labeled accounts and for each type separately. The first set of rows are variables related to the position of the accounts in the network, for which one can easily see a difference between unsafe and propaganda accounts relative to ordinary accounts. This is true for

[4]Given that Twitter is banned in Iran, we suspect that most of these accounts will be controlled by the same government agencies.

some other variables as well. For instance, unsafe accounts tend to be newer and more active on Twitter.

The variations among different groups along with the commonalities within a single group allow us to detect unsafe and propaganda accounts even before they start disseminating disinformation.

3.1 Labeled Dataset

In order to test the algorithm and run our policy experiments, we need two pieces of data. First, we need a set of labeled unsafe, propaganda, and ordinary accounts on which to train and test the algorithm. Second, we need a set of confirmed disinformation campaigns for our policy experiments.

We identify eight pieces of news that widely spread on Farsi-Tweeter at some point during the protests but were later publicly refuted. We use our extensive Farsi-Twitter data to construct the diffusion network of each piece of fake news. The diffusion network consists of all the original tweets, retweets, replies, and quoted-tweets related to that fake news. We keep track of who replies to whom, the length of conversation chains, the followers of all rumor participants, and the disinformation and propaganda scores of all participants and their followers. We call each of these diffusion networks a "disinformation campaign."

The labeled set of "unsafe" accounts is comprised of labeled imposter accounts only. The labeled set of imposter accounts consists of the early initiators of each verified disinformation campaign. In other words, labeled imposter accounts are those who have supplied at least one piece of disinformation on Farsi-Twitter in this time period; thus, we know that they are unsafe. To construct the set, we collect the first 5% of accounts who have contributed original tweets to each disinformation campaign. To ensure that we only include accounts who have intentionally initiated the spread of fake news rather than those who have simply shared it in error, we review each account in this set holistically. This involves going through the account timeline, replies and likes for evidence of suspicious behavior and common features of unsafe accounts, which include copying tweets related to the same fake new verbatim; spamming the same content; spamming pro-resistance hashtags in content unrelated to the ongoing protests; liking spammed content; or engaging with disinformation impossibly quickly. All labeled unsafe accounts exhibits some of these suspicious behaviors.[5] This procedure gives us a list of about 350 unsafe accounts that have initiated some verified

[5]Our analysis indicates that these rumors typically begin with a few accounts located in different parts of the network, all sharing a piece of fake news with almost identical phrasing, with slight variations in word choice.



Figure 4: Diffusion of information surrounding the death of Hana Duzduzani

piece of disinformation in the past.

We next label an initial set of "propaganda" accounts. In contrast with the unsafe accounts, propaganda accounts do not hide their true allegiance, and, therefore, it is easy to distinguish them and create a training set around them. Lastly, we gather a set of "ordinary" Twitter accounts from a diverse group of individuals, ranging from well-known figures to our own friends and family members and label them as ordinary accounts in our labeled dataset.

Figure 4 depicts one of our disinformation campaign. In this figure, red, yellow, and green nodes are accounts that have been classified by our model as propaganda, unsafe, and ordinary, respectively. An edge in this graph represents a user retweeting, quote-tweeting, or replying to a tweet created or shared by another user (i.e., engaging with their content). It is directed from the user who posts the content (source) to the user engaging with the content. Edges are colored based on the classification of their source. A node's size is based on its out-degree centrality, i.e., the total number of accounts that engage with its content. The node's position in the graph is central to all other accounts it engages with or is engaged by.

3.2 Construction of Network Measures: Algorithm

We next describe the algorithm used to define four network "proximity measures." We construct two sets of these measures, one set for proximity to known unsafe accounts and

	All		Ordinary		Unsafe		Propaganda	
	mean	stdev	mean	stdev	mean	stdev	mean	stdev
log(Unsafe Followers Measure)	3.25	2.70	1.53	1.55	5.77	2.25	2.09	1.74
log(Unsafe Following Measure)	3.40	2.60	3.45	2.33	5.13	2.20	1.15	1.42
log(Unsafe Retweets Measure)	3.58	3.25	3.78	3.06	5.66	2.77	0.70	1.36
log(Unsafe Retweeted Measure)	3.23	3.22	2.29	2.18	6.39	2.43	0.32	0.68
log(Propaganda Regime Followers Measure)	2.52	3.40	0.38	0.86	0.71	1.43	7.37	1.86
log(Propaganda Regime Following Measure)	2.53	3.00	1.37	1.38	0.37	1.02	6.68	1.65
log(Propaganda Regime Retweets Measure)	1.95	2.93	0.64	1.07	0.19	0.56	5.76	2.75
log(Propaganda Regime Retweeted Measure)	2.46	2.97	0.40	0.79	1.36	1.69	6.33	2.15
Degree Followers Centrality	1.35E-04	4.28E-04	1.65E-04	5.59E-04	6.47E-05	2.14E-04	1.90E-04	4.42E-04
Eigen Followers Centrality	1.41E-03	4.02E-03	2.44E-04	5.70E-04	2.21E-04	6.32E-04	4.31E-03	6.59E-03
Betweenness Followers Centrality	8.40E-05	1.62E-03	2.04E-04	2.77E-03	1.06E-05	6.90E-05	3.41E-05	1.37E-04
Followers to Following Ratio	761.44	9,964.05	96.02	703.59	251.36	3,334.70	2,204.43	18,092.18
Account Age	943.27	482.93	1,190.02	401.85	709.60	460.62	947.74	452.31
Is New Account	0.25	0.43	0.07	0.25	0.46	0.50	0.20	0.40
% Change Followers	10.31	73.30	10.62	98.49	10.72	50.87	9.41	61.75
% Change Following	-0.36	0.48	-0.12	0.32	-0.61	0.49	-0.34	0.48
New Followers Rate	11.93	42.97	15.29	52.46	5.90	26.13	15.62	46.83
New Following Rate	0.55	1.66	0.43	1.11	0.29	1.20	1.01	2.45
New Retweets Sent	2,350.82	5,500.13	984.53	1,822.08	3,760.69	7,104.38	2,177.96	5,649.98
New Retweets Received	41,869.67	266,603.60	56,726.90	326,954.56	52,774.94	302,666.62	10,268.69	36,164.38
New Quote Tweets Sent	422.39	4,735.54	319.60	490.23	753.68	7,782.55	121.97	240.60
New Quote Tweets Received	2,481.56	16,389.50	4,393.53	26,494.51	1,667.71	7,345.73	1,244.30	4,680.45
New Replies Tweets Sent	1,332.72	2,335.05	1,827.87	2,664.59	1,028.85	2,006.24	1,131.10	2,214.05
New Replies Tweets Received	5,606.32	18,666.75	6,024.32	13,600.28	3,588.03	14,865.52	7,684.34	26,434.71
New Tweets Sent	4,506.81	8,627.65	3,490.20	4,112.87	6,160.72	11,950.01	3,606.30	7,115.84
Tweet Rate	8.66	15.13	6.60	7.77	12.25	20.99	6.53	11.60
Retweets Sent Rate	4.57	11.68	1.47	3.39	8.15	16.84	3.68	8.35
Retweets Received Rate	49.21	337.62	51.73	314.31	68.16	460.61	22.02	79.61
Quote Tweets Sent Rate	0.61	3.42	0.57	0.83	0.93	5.55	0.25	0.50
Quote Tweets Received Rate	3.30	18.89	4.90	27.99	2.57	13.66	2.33	8.02
Replies Sent Rate	2.63	4.64	3.67	5.21	2.07	4.16	2.10	4.31
Replies Received Rate	8.01	23.29	8.79	15.57	5.14	20.41	10.75	32.35
Proportion Retweets	0.40	0.32	0.25	0.24	0.52	0.33	0.41	0.31
Proportion Quote Tweets	0.06	0.07	0.09	0.06	0.05	0.08	0.04	0.06
Proportion Replies	0.35	0.27	0.46	0.24	0.22	0.21	0.40	0.29
Followers to Following Near 1	0.03	0.18	0.03	0.17	0.01	0.11	0.07	0.25
New Account x Followers Rate	0.69	8.76	0.22	2.59	1.47	14.17	0.25	1.17
New Account x Retweets Sent Rate	2.11	10.17	0.28	2.80	5.07	15.62	0.51	4.57
New Account x Retweets Received Rate	4.16	80.75	0.49	3.78	10.56	132.83	0.37	5.42
New Account x Quote Tweets Sent Rate	0.14	0.68	0.06	0.45	0.31	0.98	0.03	0.29
New Account x Quote Tweets Received Rate	0.31	6.37	0.05	0.45	0.77	10.47	0.03	0.41
New Account x Replies Sent Rate	0.39	2.16	0.15	0.96	0.75	3.03	0.20	1.77
New Account x Replies Received Rate	0.42	3.74	0.21	1.30	0.79	5.60	0.19	2.52
New Account x Proportion Retweets	0.13	0.29	0.04	0.16	0.27	0.37	0.07	0.22
New Account x Proportion Quote Tweets	0.01	0.04	0.01	0.03	0.02	0.04	0.01	0.05
New Account x Proportion Replies	0.06	0.18	0.02	0.10	0.08	0.15	0.10	0.25
Propaganda Hashtag Measure	0.10	0.08	0.10	0.07	0.09	0.06	1.13	0.09

Table 2: List of account characteristics of Farsi-Twitter accounts with mean and standard deviation across labeled accounts.

Statistics are reported for all the labeled accounts and for each type of labeled accounts separately.

one for proximity to known propaganda accounts.

Consider the unsafe accounts. We posit that unsafe accounts are close to one another in networks constructed using a number of relationships. For instance, when an unsafe account starts its activity on Twitter, it starts by following other accounts in the hope that they follow it back. It makes sense to start this process with the help of other unsafe accounts, which will result in a very tightly-knit set of accounts. Furthermore, unsafe accounts try to help each other to disseminate disinformation by retweeting the tweets of other unsafe accounts or by reposting their content verbatim.[6] The same goes for propagandists. We now explain the construction of network proximity measures to unsafe accounts in detail; the procedure for proximity measures to propaganda accounts follows the same steps.

We create four graphs in which each node represents a Farsi-Twitter user, and the edges represent one of four binary relationships between any two users. Two users denoted as u and v, are connected in each graph if the corresponding condition is met:

1. User u currently follows user v;
2. User u is currently being followed by user v;
3. User u has retweeted user v; or
4. User u has been retweeted by user v.

For the sake of simplicity, we will refer to these relationships as follows: "following," "followers," "retweets," and "retweeted," respectively.

We define our disinformation proximity measurement algorithm as follows. Given a random set of known disinformants (unsafe) and regular accounts, at $t = 0$, we initiate our set of disinformation proximity measures on relationship-R (either following, followers, retweets, or retweeted) by assigning a score of one to known disinformants and zero to all other accounts. This scoring, L_t^R, maps users to their disinformation proximity measure based on relationship-R at iteration-t of the algorithm. Define $U_{t=0}$ as an empty set of exhausted users and select the natural number constant-k as the exit threshold. At each iteration t, we loop through the following steps:

1. Randomly choose a user, u, from the set $\{v|L_t^R(v) = \max_w L_t^R(w)\}$ (the set of users with the highest proximity measure). If $u \in U_t$, go to $t + 1$.

[6]For the purposes of measuring propensity for spreading disinformation, the only engagements we consider are retweets. This is because retweets are the only form of engagement whose unambiguous purpose is the amplification of the original post and are not context-dependent like quote-tweets or replies.

2. Get the set of users who share the relationship-R with u, R_u. These are all the users who share an edge with user-u in the graph defined on relationship-R.

3. Define L^R_{t+1} by starting with L^R_t and increment the measure of each user in R_u by one if they are already in the domain of L^R_t, or otherwise add them to the domain of L^R_t with a measure of one.

4. $U_{t+1} = U_t + u$.

The algorithm terminates after T-iterations when $\|L^R_t\| = \|L^R_{t+k}\|$. The domain of the disinformation proximity measure map has not changed after k-iterations. User-u's final disinformation proximity measure defined by relationship-R is $L^R_T(u)$.

Let's try to provide some intuition about this simulation. We start with a set of disinformants. At time $t = 1$, we choose one of them at random and assume that they post a piece of disinformation. We also assume that this piece might be seen by those who follow them or have retweeted them in the past. These users have a higher proximity measure for disinformation due to their close proximity to the original poster.

At $t = 2$, a high-proximity user is chosen again. This user could be proximate to the user who created the original piece of disinformation, or they could be another disinformant who is already inclined to share the piece as they come across it. It could also be the original poster trying to amplify its message to the same set of users.

It's important to note that these relationships are reciprocal. Disinformants not only create disinformation but also receive and amplify it, so we need to consider users who are being followed and retweeted by known disinformants as well.

At each iteration, the random selection of a high-proximity user represents a user spreading disinformation. Their proximity measure can be thought of as a frequency of observing or likelihood of sharing the disinformation.

We posit that a user is more likely to share disinformation if they are more exposed to it, which is a reasonable assumption since an unsafe can't spread disinformation if they never come across it. Likewise, a well-meaning or ordinary user will inevitably share disinformation if they receive it very often. On the other hand, users on the periphery are less likely to be reached by unsafe accounts, and the more paths they have to the source of the disinformation, the more likely it is to reach them.

3.3 Estimation

This section outlines our estimation methodology, wherein we calculate a triplet propensity score for each account on Farsi-Twitter as the probability of belonging to the ordinary, unsafe, or propaganda class of users. For each account with attributes $x \in X$ we estimate $(p_s(x), p_i(x), p_p(x))$ where p_s is the probability of this account being an ordinary account, p_i is the probability of being an unsafe account, and p_p is the probability of being a propaganda account, and $p_s + p_i + p_p = 1$. In the rest of the paper, we will refer to p_i and p_p as *disinformation score* and *propaganda score*, respectively.

The estimation is analogous to propensity score matching, a widely adopted method for estimating treatment effects in the literature Heckman et al. (1997); Smith and Todd (2001); Becker and Ichino (2002); Dehejia and Wahba (2002); Hirano et al. (2003). We use a multinomial Logit to estimate the propensity scores. This model assumes that the error terms in the classification come from an extreme value distribution. Given a set of explanatory variables $x \in X$, the propensity score of being an ordinary account, i.e. the probability that an account is an ordinary account is given by

$$p_s(x) = \frac{exp(\beta_s x)}{exp(\beta_s x) + exp(\beta_i x) + exp(\beta_p x)}.$$

The probability of being an unsafe or propaganda account follows similar expressions.

Estimating the propensity scores requires training and testing the Logistic regression, which in turn needs a set of labeled accounts and a set of explanatory variables. We use two-thirds of the labeled accounts as explained in Section 3.1 to train the model. For the explanatory variables, we use various non-network characteristics, as explained in Section 2, and network variables, as explained in Section 3.2. In order to avoid over-fitting, we apply a regularized fit. This fit is an elastic net with equal weights on L1 and L2 penalties. The resulting set of non-zero regressors is then passed into the logistic regression. We then test the estimated model on the remaining one-third of our labeled data that we excluded to assess the effectiveness of our classification.

The estimated Logistic regression provides us with three propensity scores (p_s, p_i, p_p) for each account. We classify each account as a member of the group with the highest score. For example, if an account has $p = (0.5, 0.2, 0.3)$ we classify it as an ordinary account. Along with the propensity scores, we report the "confusion matrix" constructed from applying the estimated classification algorithm on the labeled testing data. The confusion matrix reports how often we (correctly) classify ordinary accounts as ordinary, (incorrectly) classify them as unsafe, or (incorrectly) classify them as propaganda and the parallel statistics for unsafe

and propaganda accounts.

This methodology provides a standard approach with which to evaluate the performance of alternative classification models and allows us to adjust our models as needed to improve their accuracy.

4 Results

In this section, we detail the results of our estimation procedure. Section 4.1 reports our baseline classification results; Section 4.2 describes the trade-off between type I and type II errors and how they evolve as we increase the size of our training set; and Section 4.3 discusses the possibility of manipulative behavior by the users and the robustness of the estimation results to such manipulations.

4.1 Account Classification

Table 3 shows our baseline estimation results using two-thirds of the labeled account as the training data. A notable finding is that network proximity measures are consistently chosen by the elastic net to have some explanatory power in the classification of all categories. The coefficients associated with these variables align with expectations, indicating that unsafe accounts both closely follow one another and boost each other's content. Likewise, propaganda accounts demonstrate these same two connections. These results highlight the significance of network proximity measures in understanding the relationships between unsafe accounts and propaganda accounts.

A couple of other observations warrant additional emphasis. First, being a new account tends to be a strong predictor of being an unsafe account. Yet, this variable is not significant in predicting propaganda accounts, suggesting distinct strategies for creating and administrating these two classes of accounts. Second, all the non-network variables chosen by the elastic net for both ordinary and unsafe categories have opposite signs. Although it is in the interest of unsafe accounts to mimic ordinary accounts to avoid suspicion, either their agenda or technical shortcomings leads them to behave in a distinct and identifiable way according to our model.

To assess the performance of our model, we utilize the remaining one-third of the labeled accounts, not used in training, as the test data to calculate the probability of account misclassification. Figure 5 represents the general structure of the confusion matrix as it relates to type I and type II error rates.

	Ordinary		Unsafe		Propaganda	
	coef	std err	coef	std err	coef	std err
log(Unsafe Followers Measure)	-2.11439	0.628486	3.541529	0.839115	-0.44028	0.963798
log(Unsafe Following Measure)	0.186315	0.550844	0.330225	0.880747	-1.50338	0.89241
log(Unsafe Retweets Measure)	0.746272	0.42047	-	0.57521	-1.6217	1.004109
log(Unsafe Retweeted Measure)	-0.06512	0.489888	2.560686	0.704379	-1.50871	1.175905
log(Propaganda Followers Measure)	-1.59015	1.016807	-	0.82322	2.266125	1.238197
log(Propaganda Following Measure)	-	0.902377	-1.57503	0.988994	2.493949	1.187729
log(Propaganda Retweets Measure)	-0.38259	0.992275	-0.65124	1.521964	2.020674	0.551717
log(Propaganda Retweeted Measure)	-1.69351	1.023339	-	0.525415	1.146343	0.69714
Account Age	0.421371	0.527623	-	0.658766	-	0.731921
Is New Account	-0.92772	1.039615	1.288596	0.673299	-	0.87866
% Change Following	0.470648	0.386961	-	0.35788	-0.05966	0.427943
New Replies Tweets Sent	0.541096	5.345739	-0.13032	4.600996	-	8.285697
Replies Sent Rate	0.38886	7.06761	-0.01817	5.744808	-	12.46526
Proportion Retweets	0.096475	0.620144	-0.29029	0.605048	-	0.720971
Proportion Quote Tweets	0.756749	1.538613	-0.20113	1.621937	-	3.497823
Proportion Replies	1.086651	0.543335	-1.95833	0.924509	-	0.737067
Followers to Following Near 1	0.157597	0.572026	-	0.556499	-	0.354685
New Account x Quote Tweets Sent Rate	0.519923	5.067271	-0.4483	3.207097	-	14.48649
New Account x Proportion Retweets	0.834277	1.274338	-0.26393	0.734334	-	0.989379
Propaganda Hashtag Measure	-0.68003	1.153996	-0.270686	1.269335	-	0.972238
No. Observations:	556					
Log-Likelihood:	-0.15697					
Pseudo R-squ.:	0.92269					

Table 3: Baseline classifier. Multinomial Logistic regressions using all variables.
To choose the explanatory variables, we apply an elastic net with equal L1 and L2 penalties.

Table 4 presents the confusion matrix of the baseline model in terms of the number of accounts. Each row corresponds to the actual (labeled) account type: ordinary, unsafe, and propaganda from top to bottom. Meanwhile, each column represents the outcome of the model's classification: ordinary, unsafe, and propaganda from left to right. As such, the diagonal counts correctly classified accounts, and the off-diagonal counts misclassified accounts.

As expected, our classification performance exhibits a high degree of accuracy in identifying propaganda accounts, with minimal errors: in the third row, out of 108 test propaganda accounts only one is misclassified, as an unsafe account. However, differentiating between the unsafe and ordinary accounts accurately is more challenging as unsafe accounts actively attempt to avoid detection. Table 4 shows that we misclassify 3% of ordinary accounts as unsafe accounts and 10% unsafe accounts as ordinary accounts. Nonetheless, our overall accuracy exceeds 94%, which is encouraging.

We next use the multinomial Logit to classify all the Farsi-Twitter accounts into ordi-

	classified ordinary	classified unsafe	classified propaganda
ordinary accounts	correct	type I error (I) type II error (S)	type I error (P) type II error (S)
unsafe accounts	type I error (S) type II error (I)	correct	type I error (P) type II error (I)
propaganda accounts	type I error (S) type II error (P)	type I error (I) type II error (P)	correct

Figure 5: General structure of confusion matrix

$$\begin{bmatrix} 124 & 4 & 0 \\ 14 & 123 & 0 \\ 2 & 0 & 106 \end{bmatrix}$$

Table 4: Confusion matrix of the baseline classifier
Total accuracy: 94.64%

nary, unsafe, and propaganda accounts and report the results in Table 5. We classify about one fifth of accounts active on Farsi-Twitter as unsafe accounts, i.e. accounts that play a role in supplying disinformation to the social media users.

In the next section, we delve into the trade-offs between the two types of classification errors and explore the influence of the size of the training set on the accuracy of our classification.

Account Type	Count
Ordinary	1,193,671
Unsafe	348,086
Propaganda	225,593

Table 5: Account classification

4.2 Type I and Type II Errors

Using different explanatory variables to categorize the Farsi-Tweeter accounts affects type I and type II errors differently. The classification process always has a trade-off between minimizing type I and II errors. For instance, if our objective is to accurately identify

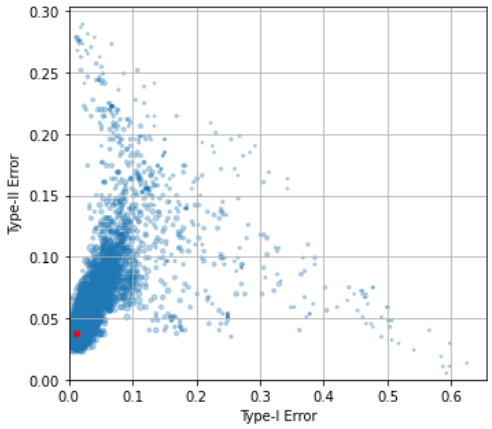


Figure 6: Trade-off between Type I and Type II errors in classifying unsafe accounts using different combinations of account characteristics.
The red point represents the baseline classifier in Table 3.

all unsafe accounts, it is inevitable to classify some ordinary accounts mistakenly as unsafe. Conversely, we might prioritize avoiding misclassifying any ordinary accounts, due to concerns associated with freedom of speech, which would, in turn, increase the chance of misclassifying unsafe accounts as ordinary.

To explore this trade-off, we estimate the classifier, i.e. the Logistic regression, repeatedly. For each estimation, we use a different subset of account characteristics listed in Table 2 for the explanatory variables. We then compute the performance of the estimated classifier on the labeled test data and report the total type I and type II errors in classifying the unsafe accounts.

Figure 10 illustrates the result of this exercise visually. Each point indicates the pair (type I, type II) of errors for each of the estimated classifiers. The points closer to the origin represent better-performing classifiers and the red point corresponds to the values corresponding to the baseline estimation results in Table 3. Figure 10 shows that attempting to reduce one type of error without considering its impact on the other type can lead to drastic and undesirable outcomes. For instance, if we solely focus on minimizing type II errors without any other constraints, we would end up with a very high level of type I error, approximately 60%. Such an outcome is highly undesirable and carries significant implications.

One approach to simultaneously reduce both error types is to expand the size of our training set. To assess the benefits of increasing the training set size and its impact on different error types, we hold the testing set constant and take increasingly large subsamples of users in our training set.



Figure 7: Impact of increasing the size of the training set on classification accuracy

For each training subset, we repeat the entire procedure, including the calculation of network proximity measures. The left panel in Figure 7 illustrates that as we increase the size of the training set, the test accuracy also increases.[7]

Another observation from the left panel is that when the sample size of the training set is low, an increase in the sample size leads to a significant and rapid improvement in test accuracy. However, even with a training set of 300 samples, we observe a plateau in accuracy. This finding has two implications. On one hand, it suggests that the benefits of further increasing the sample size become minimal, although still positive. On the other hand, it indicates that we can achieve a good level of accuracy without relying on an excessively large training dataset. This observation is particularly valuable when applying this method to similar applications, as it demonstrates that substantial improvements can be obtained without requiring a prohibitive amount of training data.

Test accuracy provides an overall perspective on the effect of increasing the sample size on classification performance. However, we also delve into the specific impact of sample size on type I and type II errors when classifying unsafe and propaganda accounts. The right panel in Figure 7 demonstrates a notable decrease across both type I and type II errors. The most significant drop is observed in type II errors for classifying propaganda accounts. This indicates that, early on, adding more data points shifts the marginal improvement from propaganda to unsafe accounts. Even with a small training sample, misclassifying other accounts as propaganda accounts is rare.

In the case of identifying unsafe accounts, we eventually observe a decline in both

[7]Note that the observed drop in accuracy is due to the randomness inherent in the sub-sampling of data for smaller sets. Since calculating the proximity measures is time-consuming, we have not conducted any simulations on the subsets selected for each sub-sample size. However, we anticipate that introducing a simulation would eliminate any non-monotonicities.

	Ordinary		Unsafe		Propaganda	
	coef	std err	coef	std err	coef	std err
log(Unsafe Followers Measure)	-2.6423	0.542045	4.060845	0.678709	-0.42707	0.734537
log(Unsafe Following Measure)	1.848231	0.503302	-	0.687516	-1.23555	0.71667
log(Unsafe Retweets Measure)	1.026578	0.382612	-	0.465351	-1.61311	0.695229
log(Unsafe Retweeted Measure)	-0.00202	0.406024	2.524237	0.530179	-1.53075	0.964246
log(Propaganda Followers Measure)	-0.95294	0.954552	-0.35138	0.710379	2.295785	0.953228
log(Propaganda Following Measure)	-	0.765551	-2.81514	0.834769	2.463936	0.90742
log(Propaganda Retweets Measure)	-0.7099	0.813306	-0.41985	1.329567	2.121217	0.474619
log(Propaganda Retweeted Measure)	-1.48446	0.966687	-	0.430487	1.250272	0.577244
Degree Followers Centrality	0.766002	4.518648	-	6.238704	-0.17735	4.435063
Eigen Followers Centrality	-3.0262	2.179573	-1.90009	2.302874	5.912917	0.917656
No. Observations:	556					
Log-Likelihood:	-0.23041					
Pseudo R-squ.:	0.85684					

Table 6: Non-manipulable classifier. Multinomial Logistic regressions using network-based characteristics only

To choose the explanatory variables, we apply an elastic net with equal L1 and L2 penalties.

types of errors as we increase the size of the training sample. We also observe that our classification errors tend to be higher for unsafe accounts compared to propaganda accounts, which as noted before, is expected as the unsafe accounts try to hide their allegiance while propaganda accounts do not.

4.3 Manipulating Account Characteristics

What makes the detection of unsafe accounts difficult is that they imitate ordinary accounts in certain respects. As such, one concern around revealing these scores to Farsi-Twitter users or using the scoring mechanism to take punitive action is factor manipulation. Unsafe accounts would have the incentive to change their strategies in order to manipulate their disinformation score and reduce the accuracy of our classification algorithm.

To address this concern, in this section, we re-estimate the model exclusively using the network-based characteristics, which are more complicated to manipulate. The rationale behind this approach is that these characteristics depend on the behavior of all accounts. At the account level, even if unsafe accounts are aware that they will be exposed when they follow or are followed by other unsafe accounts, their options for evasion are limited. Furthermore, in order to establish or promote a legitimate account they need to locate their accounts centrally in the wider network of ordinary users, which is captured by our network centrality measures. Alternatively, manipulating these network measures at an

aggregate level requires large-scale coordination among all unsafe accounts which we do not find plausible. Therefore, network-based characteristics allow us to effectively identify unsafe accounts while being difficult to manipulate.

Table 6 reports the results of estimating our baseline model with network-based account characteristics only.[8] The coefficients of these variables tend to be similar to the ones in the baseline model which included a wider set of variables.

$$\begin{bmatrix} 114 & 14 & 0 \\ 25 & 112 & 0 \\ 1 & 0 & 107 \end{bmatrix}$$

Table 7: Confusion matrix for classifier non-manipulable classifier
Total accuracy: 89.28%

To assess the impact of restricting the set of explanatory variables on the classification performance, Table 7 reports the confusion matrix for the estimation with the restricted set of network-based characteristics. Importantly, we find that excluding explanatory variables that are susceptible to manipulation does not result in a significant loss of accuracy. The total accuracy of the classifier is around 89%, below the baseline accuracy of 94.64%. This decline is primarily attributable to the lower performance in distinguishing the unsafe and ordinary accounts in the absence of the non-network variables. Since differentiating unsafe and ordinary accounts is at the core of this classification, we believe the non-network-based variables play an important role in the classification exercise.[9]

We additionally conduct the same analysis as that of Section 4.2 to analyze the trade-off between type I and type II errors in the classification among the network-based characteristics only. Figure 8 illustrates the results. The red point corresponds to the classifier estimated in Table 6. It minimizes both the I and type II among all the classifiers that only use network-based characteristics and brings us very close to the baseline results.[10]

An alternative approach is to follow the theoretical insight of the studies on optimal signals in the presence of manipulation and use an opaque scoring scheme that is difficult

[8]Some of the other variables that were included in the baseline model may not be very easy to manipulate either. We chose to only use the network-based characteristics to be conservative.

[9]Table 16 reports the confusion matrix for a parallel exercise that only uses non-network-based characteristics to classify the accounts. We find that excluding the network-based characteristics leads to a significant decrease in test accuracy, dropping to approximately 67.76%. The substantial decline in test accuracy when excluding these characteristics emphasizes their crucial role in classifying accounts.

[10]Figure 12 in the appendix illustrates the trade-off between type I and type II errors in classifying unsafe accounts using combinations of non-network-based account characteristics only. It shows that using any subset of non-network-based characteristics alone results in a significant loss of accuracy for the classification of both unsafe accounts and propaganda accounts.

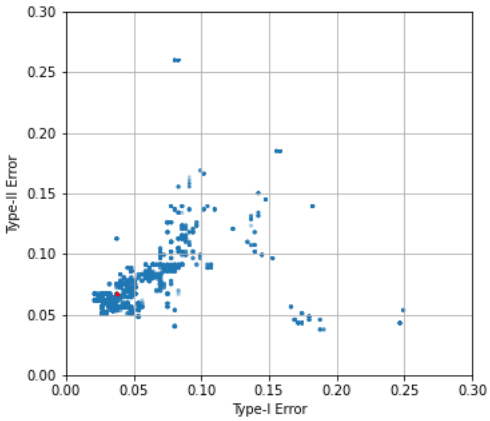


Figure 8: Trade-off between Type I and Type II errors in classifying unsafe accounts using different combinations of network-based account characteristics only
The red point represents the network-based classifier in Table 6.

to manipulate, as introduced in Saeedi and Shourideh (2023). That would enable us to maintain a higher level of accuracy in our classification algorithm while mitigating the risk of unsafe accounts adapting their strategies.

5 Policy Experiments

We propose two policies to restrict the supply of disinformation in the social network using our disinformation and propaganda scores. The first policy involves blocking tweets of unsafe and or propaganda accounts, while the second policy involves disclosing the disinformation and propaganda scores.

We simulate three different variations of each policy on our collected set of eight confirmed disinformation campaigns and assess their effectiveness in restricting the flow of disinformation. We are particularly interested in how each of these policies affects the spread of disinformation among ordinary users on Farsi-Twitter.

The first variation of each policy applies only to unsafe accounts, the second only to propaganda accounts, and the third to both types of accounts. These experiments enable us to separately measure the role of each type of account in spreading disinformation. In particular, we will show that unsafe accounts have a disproportionately large role in disseminating disinformation to ordinary accounts, while propaganda accounts are quite ineffective in engaging with them.

For each confirmed disinformation campaign, we take the following precautionary steps to avoid introducing biases into our estimation from our training set. We remove the unsafe

accounts in our training dataset that are related to that particular disinformation campaign and recalibrate the baseline classifier with the remaining training data. We then run the three variations of each policy on that disinformation campaign and measure the impact of each on the spread of the rumor using the different measures reported in Table 8.

The left and right blocks of Table 8 report the different effects of each policy, via blocking and information disclosure, respectively. The impact is reported as the percentage decline of each indicator relative to the realized information campaign. Each variation of a policy is reported in the column with the corresponding header. For instance, the first column in each block reports the effect of the policy when it is applied to only unsafe accounts.

The first two rows report different measures of the percentage change in the length of tweet threads as a result of the two policies. Rows $3-5$ report the percentage decline in participants based on their classification by our model. A participant refers to an account that tweets or engages with (replies, retweets, or quote-tweets) any content related to that disinformation campaign. Rows $6-8$ report the percentage decline in the number of expected total views that a rumor receives, grouped by the type of viewer as classified by the model. We do not observe the number of views that a post receives in our data. Thus, we use a proxy to capture its percentage change. In particular, we proxy the percentage change in views that a rumor receives as a result of policy implementation as the percentage change in the total number of followers of accounts who participated in the rumor.[11] We further partition the followers of each account by their classification and report the percentage change separately for each class.

Finally, the last 3 rows focus on the same statistics but restricted to tweets that are posted by ordinary accounts. Each of these rows reports the decline in views grouped by "follower type views via source type". For instance, the first column of the last row reports the views of "Propaganda views via Ordinary" declined by 38%. That means that the number of propaganda accounts that viewed a post because they were following an ordinary participant, declined by 38%

It is worth mentioning that we focus on this limited set of disinformation campaigns because they are instances of what we believe are the most disruptive type of disinformation. These are rumors that aim to discredit the protest movement by either attributing murders that have not happened to the IRI or by falsely attributing violence to non-violent protesters.

[11] Twitter started reporting the number of views each tweet has received as a part of the tweet object recently. To confirm the validity of our proxy, we calculated the correlation between the number of views and the number of followers of the account who posts the tweet for a small random subsample of recent posts on Farsi-Twitter, and we found a 92% correlation.

	block			Information Disclose		
	Unsafes	Propaganda	Both	Unsafes	Propaganda	Both
Average Thread Size	-67.0%	-2.8%	-69.4%	-1.9%	-0.2%	-2.2%
Max Thread Size	-48.5%	0.0%	-54.2%	-4.2%	0.0%	-4.2%
Ordinary Participants	-40.6%	-3.3%	-43.7%	-36.4%	-1.9%	-38.3%
Unsafe Participants	-100.0%	-0.9%	-100.0%	-0.3%	0.0%	-0.3%
Propaganda Participants	-22.5%	-100.0%	-100.0%	-0.2%	-0.4%	-0.5%
Ordinary Views	-84.1%	-1.8%	-85.5%	-82.2%	-1.5%	-83.6%
Unsafe Views	-59.5%	-0.7%	-60.1%	-2.6%	-0.2%	-2.8%
Propaganda Views	-49.3%	-18.7%	-66.9%	-5.0%	-0.2%	-5.2%
Ordinary Views via Ordinary	-42.0%	-1.5%	-43.5%	-36.9%	-1.4%	-38.3%
Unsafe Views via Ordinary	-44.1%	-1.6%	-45.7%	-37.6%	-1.4%	-39.0%
Propaganda Views via Ordinary	-38.0%	-1.3%	-39.3%	-33.2%	-1.2%	-34.4%

Table 8: Average impact of policy treatments on key diffusion statistics

Blocking Unsafe and Propaganda Accounts One straightforward yet strict policy intervention is to block the posts (tweet, retweet, or any other form of engagement) of accounts identified as unsafe and/or propaganda. A blocked account is able to view posts made by accounts that are not blocked The first three columns of Table 8 report the impact of the three variations of this policy intervention on the spread of disinformation.

An important observation from the first three columns of Table 8 is that blocking propaganda accounts has minimal impact on the diffusion of disinformation. This is largely due to the fact that ordinary accounts do not frequently engage with content posted by propaganda accounts as they are easy to identify. Unsafe accounts do not typically engage with this content either as they are trying to mimic ordinary accounts. Alternatively, blocking unsafe accounts has a sizable effect on the spread of disinformation among ordinary users, as ordinary accounts can easily mistake unsafe accounts as one of themselves.

Notably, while the average thread length does not change significantly, the engagement of ordinary accounts in the diffusion network declines drastically, with a 40.6% drop when blocking only unsafe accounts and a 43.7% drop when blocking both unsafe and propaganda accounts. Additionally, the number of ordinary accounts that are exposed to disinformation experiences a substantial decrease. For instance, exposure to disinformation among ordinary accounts via any participant drops 84.1% when only unsafe accounts are blocked.

Disinformation and Propaganda Score Disclosure An alternative policy is to diminish the probability that an ordinary account shares each tweet by using the disinformation score, the propaganda score, or both scores of the account that posts it. In order to measure the impact of this policy on the spread of disinformation, we need to make a behavioral

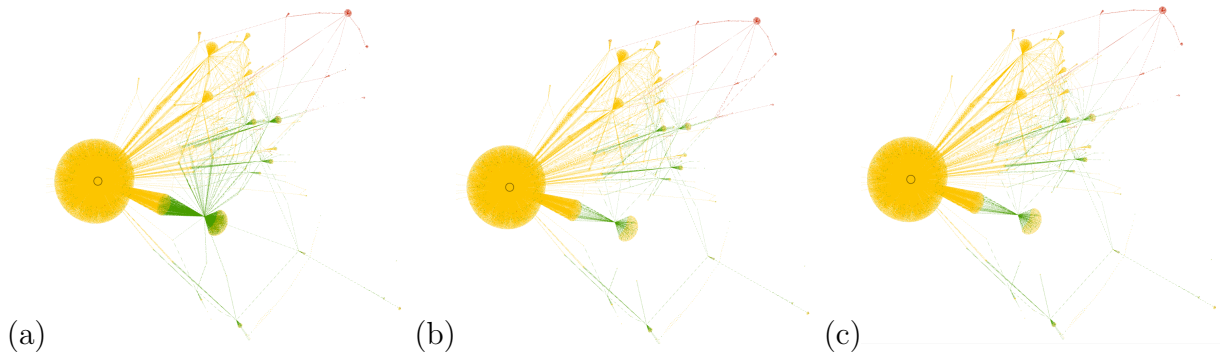


Figure 9: Hypothetical diffusion of information surrounding the death of Hana Duzduzani with Disclosure. (a) corresponds to disclosing only propaganda accounts, (b) corresponds to the case of disclosing only unsafe accounts, and (c) corresponds to disclosing both unsafe and propaganda accounts.

assumption about Farsi-Twitter users. In particular, unlike in the case of blocking, the effectiveness of information disclosure relies heavily on the behavior of the (ordinary) accounts on Twitter.

Intuitively, an ordinary user would not want to spread disinformation. When ordinary users are better able to deem tweets from accounts with high disinformation or propaganda scores as unreliable, they engage with these accounts less frequently and avoid spreading disinformation or propaganda. Solving for the optimal behavioral response of users in this network model theoretically is an interesting question of its own, which is beyond the scope of this exercise. We use a much more limited approach and simulate the effect of the policy of disclosing the scores by positing a specific individual behavioral response.

For each of our confirmed disinformation campaigns, we simulate the spread of disinformation under three variations of the disclosure policy that correspond to disclosing different information. Either, each tweet is accompanied by the disinformation score p_i, the propaganda score p_p, or both disinformation and propaganda scores (p_i, p_p) of the account that posts it in the first, second, and third disclosure policy, respectively.

In the simulation, we use the following behavioral response to model that ordinary accounts are less likely to retweet posts of unsafe, propaganda, or both accounts because they are concerned that it might be a piece of disinformation. Consider account k who has retweeted a tweet of account j, with disinformation and propaganda scores (p_i^j, p_p^j), on the realized diffusion network of the rumor. We assume that if account k is an unsafe or propaganda account, it will still retweet account j's tweet in the policy experiment. Alternatively, if k is an ordinary account, it will retweet the original tweet with probability

$1 - p_i^j$, $1 - p_p^j$, and $1 - (p_i^j + p_p^j)$ corresponding to the relevant version of the information disclosure policy.

The second block of columns in Table 8 reports the average impact of each information disclosure policy across the identified disinformation campaigns. Interestingly, the impact of the information disclosure policy closely follows those of the blocking policy, albeit 25% smaller on average. Although disclosing propaganda scores has little impact on the diffusion of disinformation, disclosing disinformation scores is a very effective way of stopping disinformation from spreading among ordinary accounts, we observe a 36.4% drop in engagements among ordinary accounts.

Another interesting observation is the large decline in the last two rows of the table. Note that the account whose views we count is either a propaganda or unsafe account, thus they do not change their behavior at all when they get information about the source of the tweet. However, these are propaganda and unsafe accounts that follow ordinary accounts, and ordinary accounts do reduce their retweeting activity when disinformation, propaganda, or both sets of scores are disclosed. As such, the number of views that unsafe and propaganda accounts who follow ordinary accounts contribute to the entire rumor declines in tandem with ordinary account participation in the disinformation campaign, reported in the 3$^{\text{rd}}$ row of the table.

Figure 9 illustrates the differential impact of disclosing disinformation and propaganda scores on the spread of disinformation for the specific disinformation campaign depicted in Figure 4. Consistent with the average statistics reported in Table 8, the disclosure of propaganda scores, panel (a), does not have a considerable influence on the diffusion of disinformation among ordinary accounts (green). Alternatively, disclosure of disinformation scores decreases the engagement by ordinary accounts by more than 35%, panel (b), and is only marginally improved by also disclosing propaganda scores, panel (c).

Lastly, recall that for each disinformation campaign, we omitted the corresponding labeled unsafe accounts from our training data. We then use the retrained classifier to classify these omitted accounts. On average, we identify 94.2% of them as unsafe. This high out-of-sample prediction accuracy, along with the above statistics point to information disclosure being a potentially useful tool in preventing the spread of disinformation on social media, which is an interesting insight as some recent arguments hint at efforts to fight misinformation through fact-checking as being ineffective (Levin, 2017).

6 Conclusion

In this paper we propose the concept of "(dis)information wars," the intentional spread of disinformation on social media platforms, often by oppressive governments, in order to counter the growing use of these platforms as vehicles of dissidence across the globe. We take advantage of data from Farsi-Twitter during the recent wave of social unrest in Iran, starting on September 16, 2022, to identify accounts that engage in spreading disinformation and use the characteristics that predict such engagement to assign a disinformation score to all Farsi-Twitter accounts.

We propose two policy interventions to limit the spread of disinformation on social networks. One of the policies involves disclosing accounts' propensity to spread misinformation along with each of their posts, and we measure its effectiveness by positing a behavioral response by network participants and ignoring any possible reactions from disinformation accounts. However, implementing such a policy can lead to gaming by disinformant accounts, which in turn reduces the effectiveness of the policy. Designing the optimal disclosure policy is not straightforward and requires additional theoretical research.[12]

We believe that the classification algorithm can be improved in a few dimensions. First, Natural Language Processing (NLP) techniques can be employed to better classify unsafe and ordinary accounts. Second, adding a clustering step to the classification algorithm can help group the unsafe accounts by the over-arching entity that controls each group. Not all unsafe accounts act the same, as they might be run by different entities within governments, foreign agencies, or even some opposition groups. This can help us expand our algorithm further.

Finally, an alternative algorithm to our multinomial Logistic classifier is a deep neural network. This generalization could identify other characteristics that have predictive power over an account being unsafe. It can also inform us more directly about the non-linearities in the prediction algorithm.

[12]For instance, Hopenhayn and Saeedi (2023) show that the information loss by moving to coarse ratings is limited.

References

Acemoglu, D., Ozdaglar, A., and ParandehGheibi, A. (2010). Spread of (mis) information in social networks. *Games and Economic Behavior*, 70(2):194–227.

Akbarpour, M., Malladi, S., and Saberi, A. (2020). Just a few seeds more: value of network information for diffusion. *Available at SSRN 3062830*.

Allcott, H. and Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of economic perspectives*, 31(2):211–236.

Allcott, H., Gentzkow, M., and Yu, C. (2019). Trends in the diffusion of misinformation on social media. *Research & Politics*, 6(2):2053168019848554.

Angelucci, C., Cagé, J., and Sinkinson, M. (2020). Media competition and news diets. Technical report, National Bureau of Economic Research.

Bak-Coleman, J. B., Kennedy, I., Wack, M., Beers, A., Schafer, J. S., Spiro, E. S., Starbird, K., and West, J. D. (2022). Combining interventions to reduce the spread of viral misinformation. *Nature Human Behaviour*, 6(10):1372–1380.

Becker, S. O. and Ichino, A. (2002). Estimation of average treatment effects based on propensity scores. *The stata journal*, 2(4):358–377.

Besley, T. and Prat, A. (2006). Handcuffs for the grabbing hand? media capture and government accountability. *American economic review*, 96(3):720–736.

Bradshaw, S. and Howard, P. N. (2018). The global organization of social media disinformation campaigns. *Journal of International Affairs*, 71(1.5):23–32.

Budak, C., Agrawal, D., and El Abbadi, A. (2011). Limiting the spread of misinformation in social networks. In *Proceedings of the 20th international conference on World wide web*, pages 665–674.

Cagé, J., Hervé, N., and Mazoyer, B. (2020a). Social media influence mainstream media: Evidence from two billion tweets. *Available at SSRN 3663899*.

Cagé, J., Hervé, N., and Viaud, M.-L. (2020b). The production of information in an online world. *The Review of economic studies*, 87(5):2126–2164.

Chan, M.-p. S., Jones, C. R., Hall Jamieson, K., and Albarracín, D. (2017). Debunking: A meta-analysis of the psychological efficacy of messages countering misinformation. *Psychological science*, 28(11):1531–1546.

Chen, Y. and Yang, D. Y. (2019). The impact of media censorship: 1984 or brave new world? *American Economic Review*, 109(6):2294–2332.

Crawford, V. P. and Sobel, J. (1982). Strategic information transmission. *Econometrica: Journal of the Econometric Society*, pages 1431–1451.

Dehejia, R. H. and Wahba, S. (2002). Propensity score-matching methods for nonexperimental causal studies. *Review of Economics and statistics*, 84(1):151–161.

Ecker, U. K., Lewandowsky, S., Cook, J., Schmid, P., Fazio, L. K., Brashier, N., Kendeou, P., Vraga, E. K., and Amazeen, M. A. (2022). The psychological drivers of misinformation belief and its resistance to correction. *Nature Reviews Psychology*, 1(1):13–29.

Gentzkow, M. and Shapiro, J. M. (2006). Media bias and reputation. *Journal of political Economy*, 114(2):280–316.

Gottfried, J. and Shearer, E. (2016). News use across social media platforms 2016.

Guriev, S. and Treisman, D. (2019). Informational autocrats. *Journal of economic perspectives*, 33(4):100–127.

Guriev, S. and Treisman, D. (2022). *Spin dictators: The changing face of tyranny in the 21st century*. Princeton University Press.

Heckman, J. J., Ichimura, H., and Todd, P. E. (1997). Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. *The review of economic studies*, 64(4):605–654.

Hirano, K., Imbens, G. W., and Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71(4):1161–1189.

Hopenhayn, H. and Saeedi, M. (2023). Optimal simple ratings.

Kahan, D. M., Peters, E., Dawson, E. C., and Slovic, P. (2017). Motivated numeracy and enlightened self-government. *Behavioural public policy*, 1(1):54–86.

Kunda, Z. (1990). The case for motivated reasoning. *Psychological bulletin*, 108(3):480.

Levin, S. (2017). Facebook promised to tackle fake news. but the evidence shows it's not working. *the Guardian*, 16.

Nyhan, B. and Reifler, J. (2010). When corrections fail: The persistence of political misperceptions. *Political Behavior*, 32(2):303–330.

Saeedi, M. and Shourideh, A. (2023). Optimal rating design under moral hazard. *arXiv preprint arXiv:2008.09529*.

Schedler, A. (2010). Democracy's past and future: Authoritarianism's last line of defense. *Journal of democracy*, 21(1):69–80.

Shadmehr, M. and Bernhardt, D. (2015). State censorship. *American Economic Journal: Microeconomics*, 7(2):280–307.

Simonov, A. and Rao, J. (2022). Demand for online news under government control: Evidence from russia. *Journal of Political Economy*, 130(2):259–309.

Smith, J. A. and Todd, P. E. (2001). Reconciling conflicting evidence on the performance of propensity-score matching methods. *American Economic Review*, 91(2):112–118.

Stukal, D., Sanovich, S., Bonneau, R., and Tucker, J. A. (2017). Detecting bots on russian political twitter. *Big data*, 5(4):310–324.

Thomas, K., Grier, C., and Paxson, V. (2012). Adapting social spam infrastructure for political censorship. In *LEET*.

Vosoughi, S., Roy, D., and Aral, S. (2018). The spread of true and false news online. *science*, 359(6380):1146–1151.

Appendix

A Additional Regressions

Table 9 reports the results of a non-regularized classifier (multinomial Logistic regression) with all the account characteristics from Table 2

	Ordinary		Unsafe		Propaganda	
	coef	std err	coef	std err	coef	std err
constant	2.05	0.47	-0.60	0.70	-1.46	0.76
log(Unsafe Followers Measure)	-8.73	0.63	3.46	0.84	5.27	0.96
log(Unsafe Following Measure)	7.61	0.55	10.51	0.88	-18.12	0.89
log(Unsafe Retweets Measure)	16.24	0.42	14.81	0.58	-31.04	1.00
log(Unsafe Retweeted Measure)	9.08	0.49	17.70	0.70	-26.78	1.18
log(Propaganda Regime Followers Measure)	-10.45	1.02	-9.86	0.82	20.30	1.24
log(Propaganda Regime Following Measure)	-9.33	0.90	-6.97	0.99	16.30	1.19
log(Propaganda Regime Retweets Measure)	-4.58	0.99	-21.13	1.52	25.71	0.55
log(Propaganda Regime Retweeted Measure)	-5.35	1.02	4.26	0.53	1.09	0.70
Degree Followers Centrality	5.07	5.02	-4.93	8.21	-0.14	9.87
Eigen Followers Centrality	-18.42	4.34	12.73	6.40	5.70	1.68
Betweenness Followers Centrality	7.43	29.15	-6.79	70.57	-0.64	47.72
Followers to Following Ratio	-2.62	2.91	0.05	10.81	2.57	1.81
Account Age	7.27	0.53	5.45	0.66	-12.72	0.74
Is New Account	-0.23	1.04	4.64	0.67	-4.41	0.88
% Change Followers	-2.21	5.97	-4.21	2.74	6.43	1.86
% Change Following	0.16	0.39	2.48	0.36	-2.65	0.43
New Followers Rate	39.12	8.23	-37.21	13.94	-1.91	6.49
New Following Rate	19.00	3.51	-15.54	1.87	-3.47	1.02
New Retweets Sent	-10.39	16.88	10.74	8.74	-0.35	20.93
New Retweets Received	-4.83	8.37	5.12	11.08	-0.29	30.35
New Quote Tweets Sent	0.32	83.72	-0.22	104.17	-0.10	155.38
New Quote Tweets Received	17.50	25.54	-17.12	36.12	-0.38	48.82
New Replies Tweets Sent	6.81	5.40	-6.35	4.62	-0.46	8.34
New Replies Tweets Received	-1.96	7.14	1.17	12.31	0.79	7.85
New Tweets Sent	-5.07	36.42	5.47	20.06	-0.39	59.85
Tweet Rate	-7.48	21.93	8.88	10.52	-1.39	40.76
Retweets Sent Rate	9.83	23.68	-9.45	11.71	-0.38	29.11
Retweets Received Rate	-13.56	12.88	13.64	14.97	-0.08	42.22
Quote Tweets Sent Rate	-4.92	38.97	5.19	51.82	-0.28	70.28
Quote Tweets Received Rate	3.54	28.67	-3.36	39.38	-0.19	44.17
Replies Sent Rate	1.88	7.07	0.69	5.72	-2.57	12.41
Replies Received Rate	-17.29	7.90	16.62	13.90	0.67	9.32
Proportion Retweets	-3.58	0.62	-10.67	0.61	14.24	0.72
Proportion Quote Tweets	5.07	1.54	-3.92	1.62	-1.15	3.50
Proportion Replies	1.77	0.54	-11.49	0.92	9.72	0.74
Followers to Following Near 1	2.76	0.57	-1.60	0.56	-1.15	0.35
New Account x Followers Rate	3.92	11.62	-3.45	8.53	-0.47	15.39
New Account x Retweets Sent Rate	-3.01	11.82	3.30	6.07	-0.28	10.28
New Account x Retweets Received Rate	-24.95	97.13	25.05	46.47	-0.10	534.07
New Account x Quote Tweets Sent Rate	7.39	5.03	-6.98	3.15	-0.41	14.46
New Account x Quote Tweets Received Rate	-24.63	96.93	24.73	43.57	-0.10	526.28
New Account x Replies Sent Rate	7.77	9.11	-7.57	3.38	-0.20	11.00
New Account x Replies Received Rate	30.65	17.91	-30.23	6.62	-0.42	33.19
New Account x Proportion Retweets	8.11	1.27	6.50	0.73	-14.61	0.99
New Account x Proportion Quote Tweets	-2.63	3.26	3.06	1.86	-0.43	3.60
New Account x Proportion Replies	-1.29	1.51	5.00	1.20	-3.71	1.01
Propaganda Hashtag Measure	-19.07	1.15	10.34	1.27	8.72	0.97
No. Observations:	556					
Log-Likelihood:	-0.06980					
Pseudo R-squ.:	0.96278					

Table 9: Non-regularized multinomial Logistic regressions with all account characteristics

Table 10 reports the confusion matrix of the non-regularized classifier with all account characteristics.

$$\begin{bmatrix} 121 & 6 & 1 \\ 14 & 123 & 0 \\ 1 & 0 & 107 \end{bmatrix}$$

Table 10: Confusion matrix of non-regularized classifier with all account characteristics
Test accuracy: 94.10%

Table 11 reports the results of a non-regularized classifier (multinomial Logistic regression) with only network-based account characteristics from Table 2

	Ordinary		Unsafe		Propaganda	
	coef	std err	coef	std err	coef	std err
constant	3.40	0.15	2.15	0.22	-5.56	0.33
log(Unsafe Followers Measure)	-35.79	0.54	-25.78	0.68	61.57	0.73
log(Unsafe Following Measure)	44.01	0.50	41.15	0.69	-85.16	0.72
log(Unsafe Retweets Measure)	189.78	0.38	188.75	0.47	-378.53	0.70
log(Unsafe Retweeted Measure)	152.14	0.41	154.29	0.53	-306.43	0.96
log(Propaganda Regime Followers Measure)	-76.44	0.95	-77.03	0.71	153.46	0.95
log(Propaganda Regime Following Measure)	25.14	0.77	19.49	0.83	-44.63	0.91
log(Propaganda Regime Retweets Measure)	-122.27	0.81	-122.45	1.33	244.73	0.47
log(Propaganda Regime Retweeted Measure)	-3.48	0.97	-1.71	0.43	5.20	0.58
Degree Followers Centrality	14.35	4.52	2.16	6.25	-16.51	4.44
Eigen Followers Centrality	-53.30	2.18	-27.60	2.30	80.90	0.92
Betweenness Followers Centrality	1.54	25.42	-0.37	56.76	-1.17	34.01
No. Observations:	556					
Log-Likelihood:	-0.18620					
Pseudo R-squ.:	0.87116					

Table 11: Non-regularized multinomial Logistic regressions with network-based account characteristics

Table 12 reports the confusion matrix of the non-regularized classifier with network-based account characteristics.

$$\begin{bmatrix} 112 & 12 & 4 \\ 22 & 114 & 1 \\ 0 & 0 & 108 \end{bmatrix}$$

Table 12: Confusion matrix of non-regularized classifier with network-based account characteristics
Test accuracy: 89.54%

Table 13 reports the results of a non-regularized classifier (multinomial Logistic regression) with only non-network-based account characteristics from Table 2

	Ordinary		Unsafe		Propaganda	
	coef	std err	coef	std err	coef	std err
constant	-0.34	0.56	0.66	0.44	-0.33	0.62
Followers to Following Ratio	-2.89	3.72	-2.93	4.38	5.81	1.38
Account Age	0.29	0.55	-0.13	0.55	-0.16	0.62
Is New Account	-0.71	1.03	0.78	0.52	-0.07	0.92
% Change Followers	-16.90	2.22	13.71	2.94	3.19	2.29
% Change Following	0.76	0.41	-0.69	0.34	-0.06	0.37
New Followers Rate	22.32	7.02	-10.20	9.39	-12.12	4.82
New Following Rate	-4.14	1.98	9.67	1.93	-5.53	0.95
New Retweets Sent	-1.09	11.99	5.04	9.17	-3.95	14.07
New Retweets Received	-1.15	6.21	12.26	9.47	-11.11	7.06
New Quote Tweets Sent	2.02	80.64	-1.56	92.00	-0.46	126.83
New Quote Tweets Received	10.64	25.11	-6.03	34.87	-4.61	24.76
New Replies Tweets Sent	3.25	3.94	1.48	4.57	-4.73	5.76
New Replies Tweets Received	-1.91	5.79	2.92	10.64	-1.01	6.24
New Tweets Sent	-5.21	26.31	9.13	22.12	-3.93	39.37
Tweet Rate	-6.95	21.22	5.31	11.03	1.64	25.85
Retweets Sent Rate	-5.70	19.56	-3.36	11.42	9.06	19.02
Retweets Received Rate	-3.04	9.16	10.24	12.33	-7.20	8.35
Quote Tweets Sent Rate	3.51	38.17	-3.53	46.33	0.02	61.19
Quote Tweets Received Rate	7.96	27.97	-3.81	36.95	-4.15	26.72
Replies Sent Rate	10.81	6.57	2.64	6.03	-13.45	8.23
Replies Received Rate	5.02	7.12	-1.06	12.38	-3.96	7.85
Proportion Retweets	-0.92	0.60	-0.36	0.51	1.28	0.63
Proportion Quote Tweets	8.23	1.34	2.24	1.74	-10.46	2.24
Proportion Replies	1.33	0.56	-2.67	0.71	1.33	0.72
Followers to Following Near 1	-0.94	0.48	-0.06	0.52	1.00	0.36
New Account x Followers Rate	2.42	18.11	0.83	6.94	-3.25	14.14
New Account x Retweets Sent Rate	-4.93	9.56	2.17	5.66	2.76	7.70
New Account x Retweets Received Rate	-6.27	102.20	7.90	49.00	-1.62	90.99
New Account x Quote Tweets Sent Rate	9.29	4.04	-1.82	2.90	-7.47	5.74
New Account x Quote Tweets Received Rate	-5.46	93.76	7.07	45.90	-1.61	84.28
New Account x Replies Sent Rate	3.13	7.37	-0.87	3.47	-2.26	7.47
New Account x Replies Received Rate	5.36	14.15	-0.78	6.96	-4.58	11.71
New Account x Proportion Retweets	2.52	1.27	-0.91	0.64	-1.62	1.09
New Account x Proportion Quote Tweets	-6.70	2.81	-0.86	1.82	7.56	2.77
New Account x Proportion Replies	-1.64	1.40	0.26	0.94	1.38	1.10
Propaganda Hashtag Measure	-2.74	1.08	0.88	0.95	1.85	0.81
No. Observations:	556					
Log-Likelihood:	-0.44876					
Pseudo R-squ.:	0.62207					

Table 13: Non-regularized multinomial Logistic regressions with non-network-based account characteristics

Table 14 reports the confusion matrix of the non-regularized classifier with non-network-based account characteristics.

$$\begin{bmatrix} 97 & 25 & 6 \\ 31 & 94 & 12 \\ 16 & 8 & 84 \end{bmatrix}$$

Table 14: Confusion matrix of non-regularized classifier with non-network-based account characteristics
Test accuracy: 73.73%

Table 15 reports the results of a regularized classifier (multinomial Logistic regression) with only non-network-based account characteristics from Table 2

	Ordinary		Unsafe		Propaganda	
	coef	std err	coef	std err	coef	std err
Account Age	1.053713	0.543106	-	0.551627	-0.28843	0.61607
Is New Account	-0.1008	1.034961	0.627458	0.523488	-	0.924317
% Change Followers	-0.6897	2.21887	0.02598	2.93783	-	2.289221
% Change Following	0.54053	0.41056	-0.60367	0.33645	-	0.36835
New Followers Rate	1.110967	7.015551	-0.16318	9.379481	-	4.81201
New Following Rate	-0.41131	1.978515	-	1.922103	-	0.949311
New Retweets Sent	-0.12411	11.92311	1.599152	9.078003	-0.48843	13.96738
New Retweets Received	-	6.181782	-	9.436224	-0.38622	7.023378
New Quote Tweets Received	0.136584	25.06158	-	34.70702	-	24.71143
New Replies Tweets Sent	-	3.918706	0.06953	4.543436	-0.90799	5.71371
New Tweets Sent	-	26.10347	0.524033	21.87756	-0.03412	39.04303
Tweet Rate	-	21.28639	0.564041	10.91027	-0.27367	25.77769
Retweets Received Rate	-	9.189183	-	12.409	-0.04948	8.377934
Replies Sent Rate	0.541699	6.613083	-	6.056521	-1.04943	8.245923
Proportion Retweets	-1.36402	0.601262	-	0.506159	1.372054	0.629105
Proportion Quote Tweets	2.895083	1.334992	-	1.736023	-2.98703	2.237432
Proportion Replies	0.599447	0.556767	-2.40937	0.704888	0.823392	0.716553
Followers to Following Near 1	-0.30346	0.480375	-	0.515099	0.761134	0.362297
New Account x Quote Tweets Sent Rate	0.471716	4.069173	-	2.955171	-0.08003	5.794803
New Account x Proportion Retweets	1.605316	1.275881	-0.0192	0.635128	-0.6005	1.094229
New Account x Proportion Quote Tweets	-	2.813417	-	1.820554	-0.05524	2.76514
New Account x Proportion Replies	-0.61567	1.402751	-	0.9378	1.019305	1.096878
Propaganda Hashtag Measure	-0.17956	1.076841	-0.7493	0.954723	1.91544	0.811938
No. Observations:	556					
Log-Likelihood:	-0.67356					
Pseudo R-squ.:	0.27563					

Table 15: Regularized multinomial Logistic regressions with non-network-based account characteristics
To choose the explanatory variables, we apply an elastic net with equal L1 and L2 penalties.

Table 16 reports the confusion matrix of the regularized classifier with non-network-based account characteristics.

$$\begin{bmatrix} 94 & 25 & 9 \\ 36 & 96 & 5 \\ 28 & 17 & 63 \end{bmatrix}$$

Table 16: Confusion matrix of regularized classifier with non-network-based account characteristics
Total accuracy: 67.83%

B Additional Charts and Figures

Figure 10 illustrates the trade-off between Type I and Type II errors in classifying propaganda accounts using different combinations of all account characteristics.

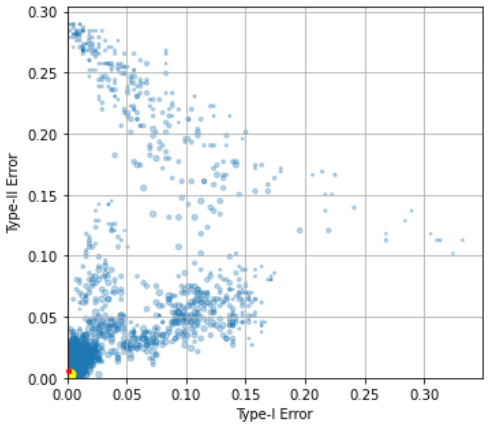


Figure 10: Trade-off between Type I and Type II errors in classifying propaganda accounts using different combinations of all account characteristics.
The red point represents the baseline classifier in Table 3. The yellow point represents the non-regularized classifier of all accounts characteristics in Table 9.

Figure 11 illustrates the trade-off between Type I and Type II errors in classifying propaganda accounts using different combinations of network-based account characteristics.

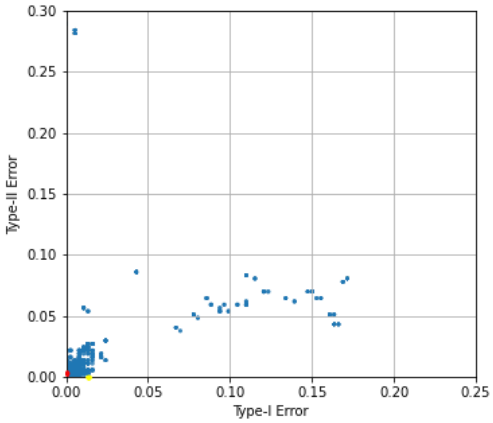


Figure 11: Trade-off between Type I and Type II errors in classifying propaganda accounts using different combinations of network-based account characteristics.
The red point represents the regularized network-based classifier in Table 6. The yellow point represents the non-regularized classifier of network-based account classifiers in Table 11.

Figure 12 depicts the the trade-off between Type I and Type II errors in classifying the propaganda accounts (left panel) and the unsafe accounts (right panel) using different combinations of non-network-based account characteristics.

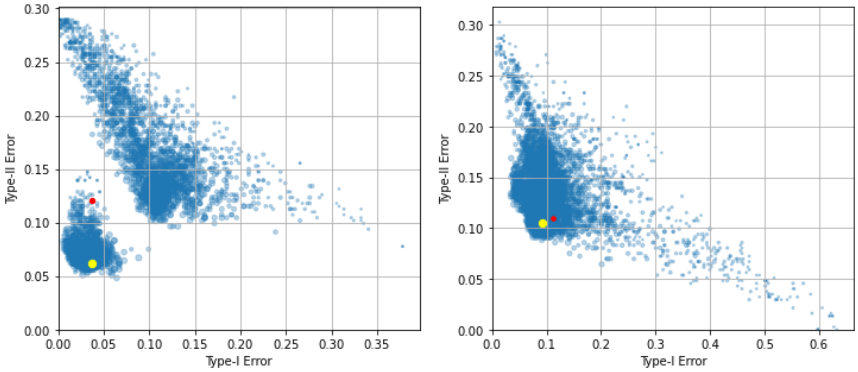


Figure 12: Trade-off between Type I and Type II errors in classifying propaganda accounts (left) and unsafe accounts (right) using different combinations of non-network-based account characteristics.
The red point represents the regularized non-network-based classifier in Table 15. The yellow point represents the non-regularized classifier of non-network-based account classifiers in Table 13.