(Dis)Information Wars*†

Adrian Casillas‡ Maryam Farboodi§ Layla Hashemi¶

Maryam Saeedi‖ Steven Wilson**

February 23, 2024

Abstract

With the unprecedented rise of internet access across the globe, social media platforms have emerged as prominent vehicles for displaying dissent. In response, numerous entities engage in spreading fake news on these platforms. We focus on a specific form of disinformation supply on social media—*disinformation wars*: the intentional spread of fake news while pretending to be an ordinary account. We demonstrate that this new form of disinformation supply is considerably more effective in spreading fake news on X, formerly known as Twitter, compared to traditional propaganda. We then propose a novel approach to preempt the spread by assigning a *disinformation score* to accounts and assess the effectiveness of the 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 Isaiah Andrews, Azam Heydari, Andrea Prat, Maryam Nazari, Jesse Shapiro, Ali Shourideh, Steven Tadelis, David Yang, participants of the Brookings 2023 AI Authors' Conference and European Summer Symposium in Economic Theory (ESSET) 2023, and many others who have helped us throughout this project. All the remaining errors are ours.

†All figures and tables in this paper are prepared by the authors using API v1 of X, formerly known as Twitter.

‡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 that 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 (Hynes, 2021).

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. Therefore, they derange the opposition movement without apparent aggression.[1] Furthermore, it is difficult to identify the pieces of fake news that have originated from 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 to derail and discredit the protest movement. 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 pressing limitations. Real-time fact-checking or content moderation is time-consuming, allowing disinformation to go viral before it can be addressed;

[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).

moreover, ex-post debunking shows limited impact on debiasing the public.[2] In this paper, we propose an alternative approach to restrict the supply of disinformation—ex-ante content moderation. 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-X, formerly known as 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 per 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-X, where the hashtag #MahsaAmini was created and widely used.[3] Concurrently, a surge in disinformation supply across Farsi-X 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.[4] There is a variety of entities who engage in spreading disinformation including the government and certain opposition groups which further complicates identifying them.

We group all the accounts on Farsi-X into three groups: propaganda, ordinary and unsafe. Propaganda accounts openly engage in spreading pro-government propaganda. Ordinary accounts are those who generally do not engage in either propaganda or disinformation. Unsafe accounts constitute two groups. The first group is imposter accounts, those who start by pretending to be dissidents, posting pro-dissidence content and hashtags to build a network among the protesters. Subsequently, they *intentionally* start posting disinformation. This disinformation then spreads by other imposter accounts as well as some of the ordinary X accounts. The second group consists of normal X accounts who actively engaged in the spread of disinformation, albeit unintentionally.

We create a unique dataset comprising all posts, formerly called tweets, in Farsi from September 16, 2019, to March 14, 2023.[5] We augment this data with a network of user re-

[2] Caplan et al. (2018) argues that the speed of disinformation spread is higher than content moderation. Chan et al. (2017); Nyhan and Reifler (2010); Ecker et al. (2022) study the impact of debunking and rebuttal.

[3] Fury grows in Iran over woman who died after hijab arrest. Accessed January 17, 2024.

[4] Meta removes Iran-based fake accounts targeting Instagram users in Scotland. Accessed January 17, 2024.

[5] The end date of data collection coincides with the cessation of the X v1 API.

lationships, including follower-following connections and re-posts, formerly called retweets, 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 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, to shed light on the supply and spread of disinformation, we devise a classifier to determine whether an account on Farsi-X 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-X accounts. The disinformation (propaganda) score of each X account reflects the probability of that account being an unsafe (propaganda) account. The disinformation score indicates the likelihood of an account acting as an unsafe account and actively engaging in the dissemination of disinformation, intentionally or not, even if it has not yet done so. The propaganda score indicates the likelihood that an account engages in explicit propaganda. We find that the unsafe 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. To address this concern, we propose an alternative classifier that only relies on network proximity measures and find that this limited classifier achieves 81.7% accuracy 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 posts 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 analyze 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 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., 2020; Angelucci et al., 2020; Levy, 2021).

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 three-fifths of adults in the United States access their news through these platforms. Cagé et al. (2020) show that even mainstream media are 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 X.

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 X until the end of 2016. However, they also discovered a sharp decline in interactions with false content on Facebook since then, while interactions on X continued to rise. Additionally, Bradshaw and Howard (2018) conducted an examination of organized social media manipulation campaigns in 28 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 X 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-X 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.

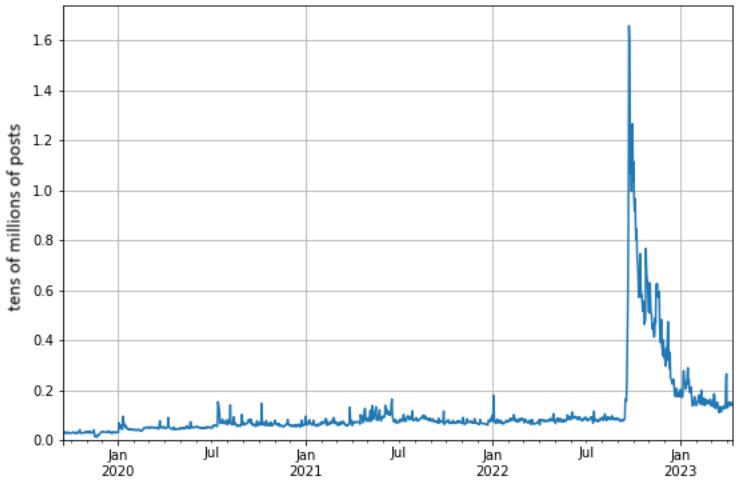


Figure 1: Daily volume of Farsi-language posts throughout the sample period.

2 X 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 X, 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.

Source of the data that underlies all the figures, tables and calculations in the paper is X's API v1 accessed directly or stored in an archive of all Farsi-language posts made after September 16, 2019.[6]

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 X API's v1 streaming service on March 14, 2023. As shown in Figure 1, we observe a substantial jump in the activity on Farsi-language X after the start of the protests in Iran.

X's v1 API post stream allowed its users to collect a stream of all incoming posts filtered on a number of parameters provided that the stream made up less than 1% of X's total post volume at any given moment. By including a filter that only lets through posts tagged as being written in Farsi and given that Farsi language posts rarely made up more than 1% of X's volume, we were able to archive nearly all Farsi-language posts since the start of our

[6]The archive is hosted by Brandeis University. See Hashemi et al. (2022) for an overview of this dataset.

	Total	Monthly average pre 09/2022	Monthly average post 09/2022
Number of users who have posted in Farsi	9,525,800	711,600	1,240,600
Number of active Farsi-X users*	1,767,350	614,000	1,016,300
Number of posts sent	83,937,899	1,896,500	2,610,900
Number of posts sent*	62,212,685	1,351,200	2,261,600
Number of reposts sent	639,337,800	9,637,000	48,734,300
Number of reposts sent*	589,714,000	8,364,400	44,186,200
Number of quote posts sent	57,919,800	1,345,300	1,581,500
Number of quote posts 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

*Users with more than 10% of posts and engagements in Farsi who have posted at least ten posts since the start of the protests in Farsi.

Table 1: Aggregate and average monthly statistics of Farsi-X posts pre and post September 2022

stream on September 16, 2019.[7]

These post objects are rich in detail and vary in kind. All posts contain not just their post 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 post objects contain at least their unique post ID, text, timestamp, public metrics (repost, quote post, 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 posts) at the time of the post's creation. X'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 posts, and reposts) will also include references to the post they are engaging with.

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

Moreover, if we swap each post's ID for that of their creator, then we have recounted a history of every engagement between any two Farsi-X users since the start of the stream.

[7]The 1% limit was set by X for streaming posts using their v1 API.

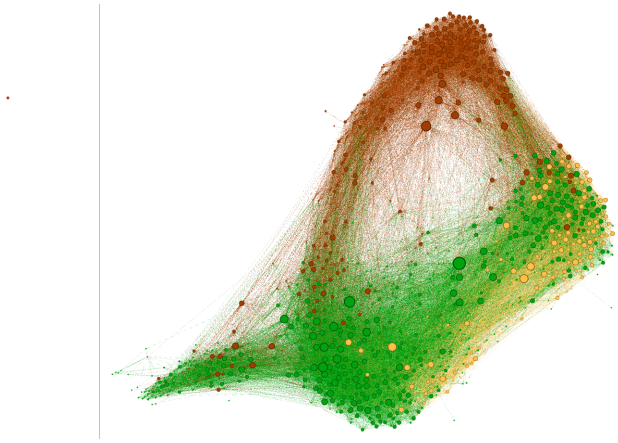


Figure 2: Network structure of followers and followings between the 1,000 most followed Farsi-X users as of December 2022. This is a directed graph with a link from node i to node j if i follows i on X. Yellow nodes are unsafe accounts, red nodes are propaganda accounts, and green nodes are ordinary accounts, as classified by the algorithm.

Finally, we supplement these engagement networks with the follower-following network of all active Farsi-X users, which is a directed friendship network.

We need to distinguish between what we call a Farsi-X user and a user who has posted in Farsi. We call users who have at least 10% of their posts or engagements in Farsi language and who have posted at least ten posts since the start of the protests in Farsi "active Farsi users," limit our sample to these users, and refer to this data as "Farsi-X".[8] Table 1 presents some monthly average statistics of all Farsi posts before and after the start of the protests in September 2022. The number of X users who have ever posted a post 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 posts and do not have any meaningful presence in the Farsi-X 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 since the start of the protests in Farsi. Table 1 also presents summary statistics limited to this sub-sample of Farsi-X accounts. Comparing the figures makes it clear that 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 repost engagement networks. Each figure is a directed graph that represents the 1,000 largest accounts in each network (i.e., the accounts with the most followers or reposts) and their connections to one

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

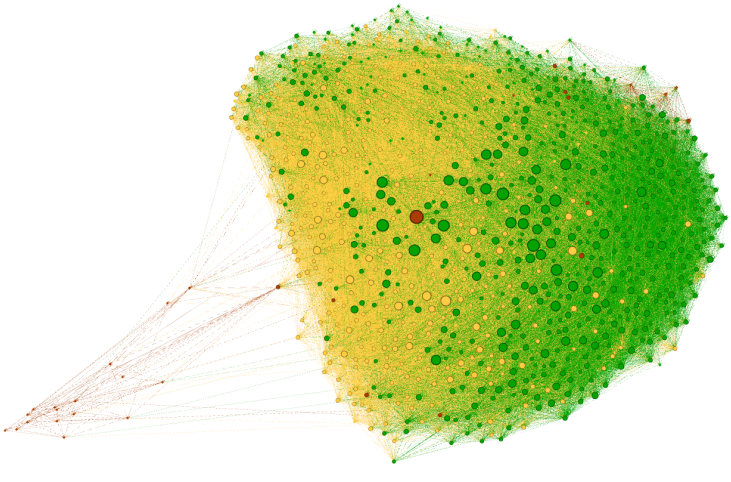


Figure 3: Network structure of reposts between the 1,000 most reposted Farsi-X accounts since the start of the sample on September 16, 2019. This is a directed graph with a link from node i to node j if i has reposted a post from j. Yellow nodes are unsafe accounts, red nodes are propaganda accounts, and green nodes are ordinary accounts, as classified by the algorithm.

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. In Figure 2, there is a directed edge from each account i to the accounts that i is following, whereas in Figure 3 there is a directed edge from account i to the accounts that i has reposted a post from. 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 reposts 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 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

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

X'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 reposts; and their activity was often implausible for human users (e.g., reposting 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 X 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, the imposter accounts. These measures are constructed from the reposts 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 imposters 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-X, we classify Farsi-X accounts into three disjoint sets: "unsafe", "propaganda", and "ordinary" accounts. Unsafe accounts consist of two groups: First, the imposter accounts who pretend to be part of the opposition

but are in fact operated by the government or other third parties; second, a group of normal Farsi-X accounts who actively participate in the spread of disinformation. We do not argue that these latter accounts intentionally spread fake news, rather, they might be fooled by the imposter accounts impersonating protesters.

Alternatively, propaganda accounts employ more traditional forms of state agitprop and openly support the Islamic Republic of Iran (IRI).[9] The third set 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 non-network variables that experienced X users employ to identify unsafe accounts anecdotally. Table 2 lists all of the characteristics of labeled Farsi-X 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 some other variables as well. For instance, unsafe accounts tend to be newer and more active on X.

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-X at some point during the protests but were later publicly refuted. These rumors are outlined in Table 20 in the Appendix. We use our extensive Farsi-X data to construct the diffusion network of each piece of fake news. The diffusion network consists of all the original posts, reposts, replies,

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

and quoted posts 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 created at least one piece of disinformation on Farsi-X 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 posts 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 posts related to the same fake news 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.[10] This procedure gives us a list of about 350 unsafe accounts that have initiated some verified piece of disinformation in the past.

We next label an initial set of "propaganda" accounts. In contrast with unsafe accounts, propaganda accounts do not hide their true allegiance, and, therefore, it is easy to distinguish and label them. We include two sets of accounts here. First, we include various I.R.I. leaders and other popular propaganda regime accounts with a high number of followers. Second, we choose random posts of the I.R.I leader and then randomly pick accounts who have liked those posts. We then manually check these accounts one by one to make sure that they are all indeed propaganda on Farsi-X and include them in our labeled dataset.

Given that the use of X is banned in Iran, ordinary users often operate under anonymity to protect themselves and their associates from political prosecution. For this reason, we label a set of "ordinary" X accounts from a diverse group of individuals, ranging from well-known opposition leaders to our friends and family, and their acquaintances whose identity can be attested.

We have a total of 929 labeled accounts: 319 ordinary accounts; 342 unsafe accounts; and 268 propaganda accounts. We split this set this set into a training set containing 60%

[10] 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.

of labeled accounts from each category and a testing set with the remaining 40%.

Figure 4 depicts one of the disinformation campaigns we used for training.[11] 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 reposting, quote posting, or replying to a post 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.[12] 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 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 several relationships. For instance, when an unsafe account

[11] Despite being disinformation, some less-reputable news organizations still report on such stories. Another Schoolgirl Beaten By Iranian Security In Critical Condition. Accessed on January 17, 2024.

[12] For the rumor depicted in Figure 4, there is a single user who was so influential in spreading the rumor, that their node graphically trumps all other participants by comparison.

	All		Ordinary		Unsafe		Propaganda	
	mean	stdev	mean	stdev	mean	stdev	mean	stdev
log(Unsafe Following Measure)	3.40	2.60	3.45	2.33	5.13	2.20	1.15	1.42
log(Unsafe Followers Measure)	3.25	2.70	1.53	1.55	5.77	2.25	2.09	1.74
log(Unsafe Reposts Measure)	3.58	3.25	3.78	3.06	5.66	2.77	0.70	1.36
log(Unsafe Reposted Measure)	3.23	3.22	2.29	2.18	6.39	2.43	0.32	0.68
log(Propaganda Following Measure)	2.53	3.00	1.37	1.38	0.37	1.02	6.68	1.65
log(Propaganda Followers Measure)	2.52	3.40	0.38	0.86	0.71	1.43	7.37	1.86
log(Propaganda Reposts Measure)	1.95	2.93	0.64	1.07	0.19	0.56	5.76	2.75
log(Propaganda Reposted 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 (days)	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 Reposts 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 Reposts 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 Posts Sent	422.39	4,735.54	319.60	490.23	753.68	7,782.55	121.97	240.60
New Quote Posts 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 Posts 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 Posts 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 Posts 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
Post Rate	8.66	15.13	6.60	7.77	12.25	20.99	6.53	11.60
Reposts Sent Rate	4.57	11.68	1.47	3.39	8.15	16.84	3.68	8.35
Reposts Received Rate	49.21	337.62	51.73	314.31	68.16	460.61	22.02	79.61
Quote Posts Sent Rate	0.61	3.42	0.57	0.83	0.93	5.55	0.25	0.50
Quote Posts 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 Reposts	0.40	0.32	0.25	0.24	0.52	0.33	0.41	0.31
Proportion Quote Posts	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 Reposts Sent Rate	2.11	10.17	0.28	2.80	5.07	15.62	0.51	4.57
New Account x Reposts Received Rate	4.16	80.75	0.49	3.78	10.56	132.83	0.37	5.42
New Account x Quote Posts Sent Rate	0.14	0.68	0.06	0.45	0.31	0.98	0.03	0.29
New Account x Quote Posts 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 Reposts	0.13	0.29	0.04	0.16	0.27	0.37	0.07	0.22
New Account x Proportion Quote Posts	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-X accounts with mean and standard deviation across labeled accounts. Table 21 in the Appendix provide descriptions of each characteristic.

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

starts its activity on X, it begins 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 reposting the posts of other unsafe accounts or by reposting their content verbatim.[13] The same goes for propagandists. We now explain the construction of the 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-X 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 reposted user v; or
4. User u has been reposted by user v.

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

We define our disinformation proximity measurement algorithm as follows. Given a random set of known unsafe and ordinary accounts, at $t = 0$, we initiate our set of disinformation proximity measures on relationship-R (either following, followers, reposts, or reposted) by assigning a score of one to known unsafe accounts 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, $u \in \text{argmax}_v\{L_t^R(v)\}$ (the set of users with the highest proximity measure at time t). If $u \in U_t$, go to $t + 1$.
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.

[13] To measure propensity for spreading disinformation, the only engagements we consider are reposts. This is because reposts are the only form of engagement whose unambiguous purpose is the amplification of the original post and are not context-dependent like quote posts or replies.

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 reposted them in the past. These users have a higher proximity measure for disinformation due to their 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 reposted 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 user can't spread disinformation if they never come across it. Likewise, even 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-X as the probability of belonging to the ordinary, unsafe, or propaganda class of users. For each account with attributes $x \in X$ we estimate $(p_s(x), p_u(x), p_p(x))$ where p_s is the probability of this account being an ordinary account,

p_u is the probability of being an unsafe account, and p_p is the probability of being a propaganda account, and $p_s + p_u + p_p = 1$. In the rest of the paper, we will refer to p_u 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

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

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 sixty percent 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. 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 forty percent 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_u, 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.

	Ordinary		Unsafe		Propaganda	
	coef	std err	coef	std err	coef	std err
log(Unsafe Following Measure)	0.18	0.55	0.33	0.88	-1.50	0.89
log(Unsafe Followers Measure)	-2.11	0.62	3.54	0.83	-0.44	0.96
log(Unsafe Reposts Measure)	0.74	0.42	0.00	0.57	-1.62	1.00
log(Unsafe Reposted Measure)	-0.06	0.48	2.56	0.70	-1.50	1.17
log(Propaganda Following Measure)	0.00	0.90	-1.57	0.98	2.49	1.18
log(Propaganda Followers Measure)	-1.59	1.01	0.00	0.82	2.26	1.23
log(Propaganda Reposts Measure)	-0.38	0.99	-0.65	1.52	2.02	0.55
log(Propaganda Reposted Measure)	-1.69	1.02	0.00	0.52	1.14	0.69
Account Age	0.42	0.52	0.00	0.65	0.00	0.73
Is New Account	-0.92	1.03	1.28	0.67	0.00	0.87
% Change Following	0.47	0.38	0.00	0.35	-0.05	0.42
New Replies Posts Sent	0.54	5.34	-0.13	4.60	0.00	8.28
Replies Sent Rate	0.38	7.06	-0.01	5.74	0.00	12.46
Proportion Reposts	0.09	0.62	-0.29	0.60	0.00	0.72
Proportion Quote Posts	0.75	1.53	-0.20	1.62	0.00	3.49
Proportion Replies	1.08	0.54	-1.95	0.92	0.00	0.73
Followers to Following Near 1	0.15	0.57	0.00	0.55	0.00	0.35
New Account x Quote Posts Sent Rate	0.51	5.06	-0.44	3.20	0.00	14.48
New Account x Proportion Reposts	0.83	1.27	-0.26	0.73	0.00	0.98
Propaganda Hashtag Measure	-0.68	1.15	-0.27	1.26	0.00	0.97
No. Observations:	556					
Log-Likelihood:	-0.15					
Pseudo R-squ.:	0.92					

Table 3: Baseline classifier, a multinomial logistic regression using all variables.
To choose the explanatory variables, we apply an elastic net with equal L1 and L2 penalties. Factors with point estimates of zero were pushed to zero by the elastic net.

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 sixty percent 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. Note that in a multinomial logistic regression, the classification is determined by the difference between the coefficients across different categories. In other words, the coefficients are best interpreted relative to a base category. Table 10 in the Appendix reports the normalized coefficients from Table 3.

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, almost all the non-network variables chosen, i.e. those that do not depend on the structure of the social networks and/or interactions therein, by the elastic net for both ordinary and unsafe categories have opposite signs.[14] Although it is in the interest of unsafe accounts to mimic ordinary accounts to avoid suspicion, either their agenda or technical shortcomings lead them to behave in a distinct and identifiable way according to our model.

To assess the performance of our model, we utilize the remaining forty percent of the labeled accounts, not used in training, as the test data to calculate the probability of account misclassification.

These probabilities can be computed using the confusion matrix. Figure 5 represents the general structure of the confusion matrix as it relates to type I and type II error rates. We use the confusion matrix to construct three measures of the performance of each classier:

$$\text{Total accuracy} = \frac{\text{sum of the diagonal entries}}{\text{sum of all entries}}$$

$$\text{Ordinary type I error} = 1 - \frac{(1,1)}{\text{sum of first row entries}} = \frac{(1,2)+(1,3)}{\text{sum of first row entries}}$$

$$\text{Unsafe type I error} = 1 - \frac{(2,2)}{\text{sum of second row entries}} = \frac{(2,1)+(2,3)}{\text{sum of second row entries}}$$

Two other well-known measures are

$$\text{Ordinary type II error} = \frac{(2,1)+(3,1)}{\text{sum of second and third row entries}}$$

$$\text{Unsafe type II error} = \frac{(2,1)+(2,3)}{\text{sum of first and third row entries}}$$

As we argue below, the last two measures are not well suited to our framework. As such, we only report them in the appendix.

[14] The only exception is the propaganda hashtag measure which measures how similar the frequency of a user's hashtags is to the 3,000 hashtags most used by the propaganda accounts in the training set.

	classified Ordinary	classified Unsafe	classified Propaganda
Ordinary accounts	correct	type I error (O) type II error (U)	type I error (O) type II error (P)
Unsafe accounts	type I error (U) type II error (O)	correct	type I error (U) type II error (P)
Propaganda accounts	type I error (P) type II error (O)	type I error (P) type II error (U)	correct

Figure 5: General structure of confusion matrix

Table 4 presents the confusion matrix of the baseline model in terms of the number of testing 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 two are misclassified, as an ordinary account. In the third column, no ordinary or unsafe accounts are classified as propaganda. Thus, the type I and type II errors in identifying propaganda accounts are virtually zero.

However, differentiating between the unsafe and ordinary accounts accurately is more challenging as unsafe accounts may 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.

Note that as the classifier does an almost perfect job of partitioning the propaganda accounts, the only place that they affect the type I and type II errors of ordinary and unsafe account classification is in the denominator of type II error: Having many propaganda accounts mechanically reduces the type II error of classifying both ordinary and safe accounts without improving the classification in a meaningful way. As such, in the main body of the paper we will restrict attention to comparing ordinary and unsafe type I errors.

We next use the multinomial logit to classify all the Farsi-X accounts into ordinary, unsafe, and propaganda accounts and report the results in Table 5. We classify about

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

Table 4: Confusion matrix of the baseline classifier with a total accuracy of 94.64%.

one-fifth of active Farsi users as unsafe accounts, i.e. accounts that participate in supplying disinformation to social media users, intentionally or unintentionally.

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

Cross-Validation. We use cross-validation to evaluate the performance of our classifier. We tailor the cross-validation procedure to our specific setting. In particular, one could argue that imposter participants in each disinformation campaign are correlated with each other. Thus, if they are present in both the training and testing set, they mechanically increase the classification accuracy due to an omitted variable bias, and such accuracy cannot be achieved when classifying imposter accounts that do not participate in that disinformation campaign.

To address this potential bias, we use a procedure similar to k-fold cross-validation. Recall that we have 8 verified disinformation campaigns. To replicate 8-fold cross-validation, we first divide the labeled ordinary and propaganda accounts randomly into 8 non-overlapping subsets.[15] Then we create 8 subsets of labeled unsafe accounts where subset i consists of imposter accounts of disinformation campaigns i. This approach differs from the traditional

[15]We use two different ways to determine the size of the 8 subsets. In one, we use 8 equally sized subsets for labeled ordinary and propaganda accounts. In the other one, we determine the size of the subsets proportional to the size of the labeled imposter accounts in the corresponding disinformation campaign. Thus in the second method, the size of the labeled set of accounts in the i^{th}-fold is larger if the disinformation campaign i is larger. The two methods give virtually identical overall accuracy and type I errors for ordinary and unsafe account classification. We report the numbers for the equally sized subsets of ordinary and propaganda accounts.

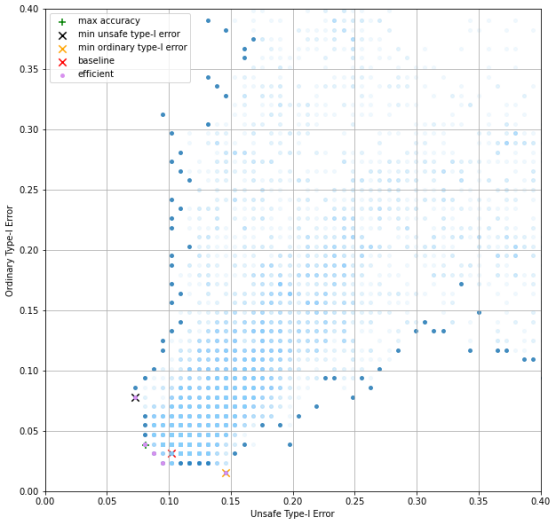


Figure 6: Trade-off between ordinary and unsafe type I errors for classifiers that use different combinations of all account characteristics as explanatory variables.
Dark blue points are the classifiers on the lower envelope of the trade-off. Purple points are the classifiers on the Pareto efficient frontier of ordinary-unsafe type I error trade-off. The black (yellow) cross is the classifier that minimizes type I error of unsafe (ordinary) account classification. The green cross is the classifier with the highest total accuracy. The red cross is the baseline classifier in Table 3.

k-fold cross-validation in that the unsafe accounts that are used for training and validation in each fold are note chosen at random. In particular, the unsafe accounts in validation set (fold) i are the imposter accounts specific to the i^{th} disinformation campaign.

The 8-fold cross-validation achieves a total accuracy of 95.1% on average. The ordinary and unsafe type I errors are 5.7% and 6.3%, respectively. This procedure mimics an authority who does not necessarily know which accounts will purvey a new piece of disinformation, but who knows which accounts have started spreading previous pieces of disinformation, and so can use information on network distance to prior imposter accounts to form a probabilistic prediction.

The high degree of accuracy in identifying out-of-sample imposter accounts shows that the classifier successfully curbs the supply of disinformation by identifying the accounts that engage in it, even if they have never done so before.

4.2 Ordinary and Unsafe Type I Errors

Using different explanatory variables to categorize the Farsi-X accounts affects the performance of the model in classifying ordinary and unsafe accounts differently, despite being relatively successful in segmenting the propaganda accounts. The classification process

	Baseline	0.5	0.6	0.7	0.8	0.9
Ordinary Type I	3.1%	1.6%	1.6%	2.3%	3.1%	3.1%
Unsafe Type I	10.2%	13.9%	12.4%	11.6%	10.9%	10.2%

Table 6: Effect of reclassifying the accounts classified as unsafe to ordinary based on the share of their initial ordinary followers on ordinary and unsafe type I errors.

always has a trade-off between minimizing ordinary type I versus unsafe type I errors. For instance, if our objective is to accurately identify all unsafe accounts, it is inevitable to classify some ordinary accounts mistakenly as unsafe. Conversely, we might prioritize avoiding misclassification of 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 train the classifier, i.e. the Logistic regression repeatedly, using different subsets of account characteristics listed in Table 2 as the explanatory variables. We then report the ordinary and unsafe type I errors associated with classifying the test data by the estimated classifier. Figure 6 illustrates the result of this exercise. Each point indicates the pair (unsafe type I, ordinary type I) error. The translucency of the points represents where the classifiers with the same combination of errors are concentrated.

The classifiers that minimize each of ordinary and unsafe type I errors are at the two end points of the efficient frontier in Figure 6. Furthermore, their position implies that attempting to reduce one type of error, without any constraints, leads to a 2-3 orders of magnitude increase in the other type of error. Interestingly, both the baseline classifier and the classifier that achieves maximum total accuracy take a middle ground in trading off the two types of error.

Unsafe accounts that are misclassified as ordinary (unsafe type I error) are mostly older accounts that have done a good job of staying out of the nexus of disinformation spread. They are relatively inactive, and when they are active, they usually avoid reposting. They are mostly incendiary, being used to initiate disinformation rather than engaging with it after the fact.

Ordinary accounts that are classified as unsafe (ordinary type I error) are typically run by users who joined in response to the most recent wave of protests. Many of their non-network account characteristics resemble those of imposter accounts which were created around the same time. In the interest of free speech, it is reasonable to err on the side of caution when classifying any account as unsafe. Given this motivation, we suggest the following augmentation to the baseline classifier in order to reduce the ordinary type I error.

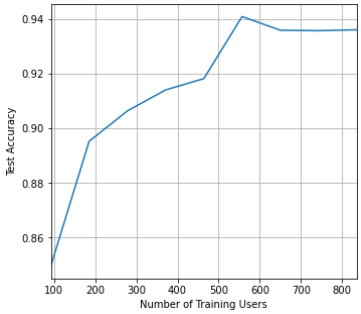


Figure 7: Total accuracy as a function of the size of the training set.

Naive Ordinary versus Imposter Accounts. We propose to solve the problem of misclassifying ordinary accounts as unsafe by looking at each account's set of first followers. We hypothesize that imposter accounts, due to their lack of genuine identity, would have a hard time collecting ordinary accounts as followers. As such, having many initial ordinary followers is a sign of being an ordinary account.

We take the first followers of each account and observe how our algorithm classifies them. To decrease type I error, we examine all accounts classified as unsafe. If more than $x\%$ of their first 10 followers are ordinary, we reclassify them as ordinary. Table 6 presents the results. The first column displays the type I errors in our baseline model for both ordinary and unsafe accounts. When we reclassify accounts as ordinary if over 50% of their first followers are ordinary, type I errors for ordinary accounts decrease to just over half the initial amount. However, this leads to higher type I errors for unsafe accounts, increasing it from the initial 10.2% to 13.9%. Alternatively, if we focus on at least 70% of the initial followers being ordinary, there is less improvement in type I error for ordinary accounts, but the increase in type I error for unsafe accounts is also smaller. Using a higher threshold does not improve the performance of the baseline classifier. Thus, one can use an appropriate threshold for ordinary initial follower to balance different types of errors.

Increasing the Size of the Training Set. One approach to simultaneously reducing 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 sub-samples of users in our training set. For each training subset, we repeat the entire procedure, including the calculation of network proximity measures. Figure 7 illustrates that as expected, the test accuracy initially increases with the size of the training set, but plateaus at around 94% once the size of the training set reaches

	Ordinary		Unsafe		Propaganda	
	coef	std err	coef	std err	coef	std err
log(Unsafe Following Measure)	1.84	0.50	0.00	0.68	-1.23	0.71
log(Unsafe Followers Measure)	-2.64	0.54	4.06	0.67	-0.42	0.73
log(Unsafe Reposts Measure)	1.02	0.38	0.00	0.46	-1.61	0.69
log(Unsafe Reposted Measure)	-2.02E-03	0.40	2.52	0.53	-1.53	0.96
log(Propaganda Following Measure)	0.00	0.76	-2.81	0.83	2.46	0.90
log(Propaganda Followers Measure)	-0.95	0.95	-0.35	0.71	2.29	0.95
log(Propaganda Reposts Measure)	-0.70	0.81	-0.41	1.32	2.12	0.47
log(Propaganda Reposted Measure)	-1.48	0.96	0.00	0.43	1.25	0.57
Degree Followers Centrality	0.76	4.51	0.00	6.23	-0.17	4.43
Eigen Followers Centrality	-3.02	2.17	-1.90	2.30	5.91	0.91
No. Observations:	556					
Log-Likelihood:	-0.23					
Pseudo R-squ.:	0.85					

Table 7: Non-manipulable classifier, a Multinomial logistic regression using network-based characteristics only.
To choose the explanatory variables, we apply an elastic net with equal L1 and L2 penalties. Factors with point estimates of zero were pushed to zero by the elastic net.

approximately 550 account.[16] This finding has two implications: On one hand, it suggests a minimal benefit for further increasing the size of the training sample. On the other hand, it indicates that we achieve a good level of accuracy without relying on an excessively large training sample. This observation is particularly valuable when applying this method to similar applications as it suggests that a high degree of classification accuracy can be obtained without requiring a prohibitively large amount of training data.

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-X users or using the scoring mechanism to take punitive action is factor manipulation. Unsafe accounts would have the incentive to change their strategies 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 difficult to manipulate. These characteristics

[16]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.

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

Table 8: Confusion matrix for the non-manipulable classifier with a total accuracy of 89.28%.

depend on the structure of the whole network which includes the ordinary accounts. As such, collective action by unsafe account cannot effectively manipulate these measures. For instance, the unsafe accounts can follow and repost ordinary accounts, but cannot force the ordinary accounts to follow and repost them. Furthermore, for each account i, these network measures depend on behavior of accounts many links away from i, which makes effective manipulation even less likely.

Table 7 reports the results of estimating our baseline model with network-based account characteristics only.[17] The coefficients of these variables tend to be similar to the ones in the baseline model which included a wider set of variables. Table 11 in the Appendix reports the normalized coefficients from Table 7 relative to a base category.

To assess the impact of restricting the set of explanatory variables on the classification performance, Table 8 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.[18]

We additionally conduct the same analysis as that of Section 4.2 to analyze the trade-off between ordinary and unsafe type I classification errors when network-based characteristics are used only. Figure 8 illustrates the results.

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 to manipulate, as introduced in Saeedi and Shourideh (2023). That would enable us to

[17] 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.

[18] Table 19 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 to 67.83%. The substantial decline in test accuracy when excluding these characteristics emphasizes their crucial role in classifying accounts.

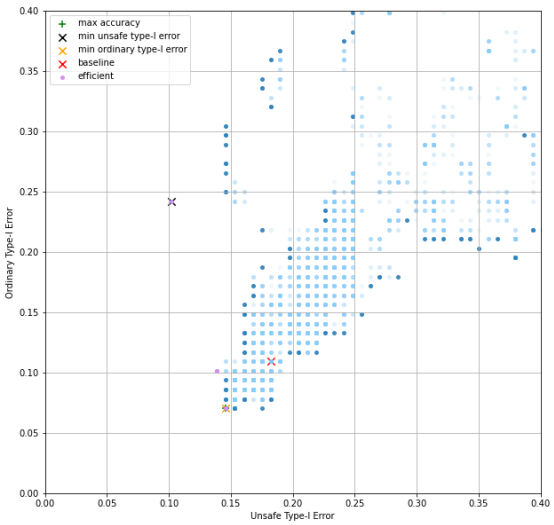


Figure 8: Trade-off between ordinary and unsafe type I errors for classifiers that omit all the non-network characteristics and use different combinations of network characteristics as explanatory variables.
Dark blue points are the classifiers on the lower envelope of the trade-off. Purple points are the classifiers on the Pareto efficient frontier of ordinary-unsafe type I error trade-off. The black (yellow) cross is the classifier that minimizes type I error of unsafe (ordinary) account classification. The green cross is the classifier with the highest total accuracy. The red cross is the baseline classifier in Table 7.

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 the posts 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-X.

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.

We use a procedure similar to 8-fold cross-validation to avoid introducing a bias into our estimation from our training set. For each confirmed disinformation campaign, we remove the corresponding labeled imposter accounts from the training data and train the baseline classifier with the remaining labeled imposter account. We then run the three variations of each policy on the left-out disinformation campaign and measure the impact of each policy on the spread of the rumor using the different measures reported in Table 9.

The left and right blocks of Table 9 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 disinformation 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 post 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 posts or engages with (replies, reposts, or quote posts) 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.[19] 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 are restricted to posts 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, Consider block "Information Disclosure," row "Approximate Ordinary Views via Ordinary" and column "Unsafe". It reports that merely disclosing the disinformation score of unsafe accounts to users leads to a 36.9% decline in the number of ordinary accounts who view a post that contains a disinforma-

[19]X started reporting the number of views each post has received as a part of the post 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 post for a small random subsample of recent posts on Farsi-X, and we found a 92% correlation.

	Block Users			Information Disclose		
	Unsafe	Propaganda	Both	Unsafe	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%
Approximate Ordinary Views	-84.1%	-1.8%	-85.5%	-82.2%	-1.5%	-83.6%
Approximate Unsafe Views	-59.5%	-0.7%	-60.1%	-2.6%	-0.2%	-2.8%
Approximate Propaganda Views	-49.3%	-18.7%	-66.9%	-5.0%	-0.2%	-5.2%
Approximate Ordinary Views via Ordinary	-42.0%	-1.5%	-43.5%	-36.9%	-1.4%	-38.3%
Approximate Unsafe Views via Ordinary	-44.1%	-1.6%	-45.7%	-37.6%	-1.4%	-39.0%
Approximate Propaganda Views via Ordinary	-38.0%	-1.3%	-39.3%	-33.2%	-1.2%	-34.4%

Table 9: Average impact of policy treatments on network diffusion statistics of disinformation campaigns.

tion because they are following another ordinary account who (unintentionally) spread that piece of fake news.

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.

Blocking Unsafe and Propaganda Accounts. One straightforward yet strict policy intervention is to block the posts (post, repost, 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 9 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 9 is that blocking propaganda accounts has minimal impact on the diffusion of disinformation. This is largely because 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

(a) (b) (c)

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

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 post by using the disinformation score, the propaganda score, or both scores of the account that posts it. To measure the impact of this policy on the spread of disinformation, we need to make a behavioral assumption about Farsi-X users. In particular, unlike in the case of blocking, the effectiveness of information disclosure relies heavily on the behavior of the (ordinary) accounts on X.

Intuitively, an ordinary user would not want to spread disinformation. When ordinary users are better able to deem posts 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 post is accompanied by the disinformation score p_u, the propaganda score p_p, or both disinformation and propaganda scores (p_u, 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 repost the content of unsafe, propaganda, or both accounts because they are concerned that it might be a piece of disinformation. Consider account k that has reposted a post of account j, with disinformation and propaganda scores (p_u^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 repost account j's post in the policy experiment. Alternatively, if k is an ordinary account, it will repost the original post with probability $1 - p_u^j$, $1 - p_p^j$, and $1 - (p_u^j + p_p^j)$ corresponding to the relevant version of the information disclosure policy.

The second block of columns in Table 9 reports the average impact of each information disclosure policy across the identified disinformation campaigns. The impact of the information disclosure policy on exposure of ordinary accounts to disinformation and disinformation views through ordinary accounts is similar to the impact of the blocking policy. 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 post. However, these are propaganda and unsafe accounts that follow ordinary accounts, and ordinary accounts do reduce their engagements 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 9, 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-X 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-X accounts.

We propose two interventions to limit the spread of disinformation on social networks. These interventions represent active measures that social media platforms can implement themselves or can be imposed by policymakers on these platforms in order to prevent the spread of disinformation. 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.[20]

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

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

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.

References

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

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

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

Allcott, H., M. Gentzkow, and C. Yu (2019). Trends in the diffusion of misinformation on social media. *Research & Politics 6*(2), 1–8.

Angelucci, C., J. Cagé, and M. Sinkinson (2020). Media competition and news diets. Working Paper 26782, National Bureau of Economic Research.

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

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

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

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

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

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

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

Caplan, R., L. Hanson, and J. Donovan (2018). Dead reckoning: Navigating content moderation after" fake news".

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

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

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

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

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

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

Gottfried, J. and E. Shearer (2016). News use across social media platforms 2016. Pew Research Center Poll. https://www.pewresearch.org/journalism/2016/05/26/news-use-across-social-media-platforms-2016/.

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

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

Hashemi, L., S. Wilson, and C. Sanhueza (2022). Five hundred days of farsi twitter: An overview of what farsi twitter looks like, what we know about it, and why it matters. *Journal of Quantitative Description: Digital Media 2*.

Heckman, J. J., H. Ichimura, and P. E. Todd (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., G. W. Imbens, and G. Ridder (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica 71*(4), 1161–1189.

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

Hynes, M. (2021). *The Social, Cultural and Environmental Costs of Hyper-Connectivity: Sleeping Through the Revolution*, Chapter 9, pp. 137–153. Leeds: Emerald Publishing Limited.

Kahan, D. M., E. Peters, E. C. Dawson, and P. Slovic (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. https://www.theguardian.com/technology/2017/may/16/facebook-fake-news-tools-not-workin. The Guardian 16.

Levy, R. (2021). Social media, news consumption, and polarization: Evidence from a field experiment. *American economic review 111*(3), 831–870.

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

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

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

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

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

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

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

Appendix

A Additional Regressions

	Relative to Ordinary		Relative to Propaganda	
	Unsafe	Propaganda	Ordinary	Unsafe
log(Unsafe Following Measure)	0.14	-1.69	1.69	1.83
log(Unsafe Followers Measure)	5.66	1.67	-1.67	3.98
log(Unsafe Reposts Measure)	-0.75	-2.37	2.37	1.62
log(Unsafe Reposted Measure)	2.63	-1.44	1.44	4.07
log(Propaganda Following Measure)	-1.58	2.49	-2.49	-4.07
log(Propaganda Followers Measure)	1.59	3.86	-3.86	-2.27
log(Propaganda Reposts Measure)	-0.27	2.40	-2.40	-2.67
log(Propaganda Reposted Measure)	1.69	2.84	-2.84	-1.15
Account Age	-0.42	-0.42	0.42	0.00
Is New Account	2.22	0.93	-0.93	1.29
% Change Following	-0.47	-0.53	0.53	0.06
New Replies Posts Sent	-0.67	-0.54	0.54	-0.13
Replies Sent Rate	-0.41	-0.39	0.39	-0.02
Proportion Reposts	-0.39	-0.10	0.10	-0.29
Proportion Quote Posts	-0.96	-0.76	0.76	-0.20
Proportion Replies	-3.05	-1.09	1.09	-1.96
Followers to Following Near 1	-0.16	-0.16	0.16	0.00
New Account x Quote Posts Sent Rate	-0.97	-0.52	0.52	-0.45
New Account x Proportion Reposts	-1.10	-0.83	0.83	-0.26
Propaganda Hashtag Score	0.95	0.68	-0.68	0.27

Table 10: Baseline classifier coefficients from Table 3 relative to a base class.

	Relative to Ordinary		Relative to Propaganda	
	Unsafe	Propaganda	Ordinary	Unsafe
log(Unsafe Following Measure)	-1.85	-3.08	3.08	1.24
log(Unsafe Followers Measure)	6.70	2.22	-2.22	4.49
log(Unsafe Reposts Measure)	-1.03	-2.64	2.64	1.61
log(Unsafe Reposted Measure)	2.53	-1.53	1.53	4.05
log(Propaganda Following Measure)	-2.82	2.46	-2.46	-5.28
log(Propaganda Followers Measure)	0.60	3.25	-3.25	-2.65
log(Propaganda Reposts Measure)	0.29	2.83	-2.83	-2.54
log(Propaganda Reposted Measure)	1.48	2.73	-2.73	-1.25
Degree Followers Centrality	-0.77	-0.94	0.94	0.18
Eigen Followers Centrality	1.13	8.94	-8.94	-7.81

Table 11: Non-manipulable classifier coefficients from Table 7 relative to a base class.

Table 12 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 Following Measure)	7.61	0.55	10.51	0.88	-18.12	0.89
log(Unsafe Followers Measure)	-8.73	0.63	3.46	0.84	5.27	0.96
log(Unsafe Reposts Measure)	16.24	0.42	14.81	0.58	-31.04	1.00
log(Unsafe Reposted Measure)	9.08	0.49	17.70	0.70	-26.78	1.18
log(Propaganda Following Measure)	-9.33	0.90	-6.97	0.99	16.30	1.19
log(Propaganda Followers Measure)	-10.45	1.02	-9.86	0.82	20.30	1.24
log(Propaganda Reposts Measure)	-4.58	0.99	-21.13	1.52	25.71	0.55
log(Propaganda Reposted 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 Reposts Sent	-10.39	16.88	10.74	8.74	-0.35	20.93
New Reposts Received	-4.83	8.37	5.12	11.08	-0.29	30.35
New Quote Posts Sent	0.32	83.72	-0.22	104.17	-0.10	155.38
New Quote Posts Received	17.50	25.54	-17.12	36.12	-0.38	48.82
New Replies Posts Sent	6.81	5.40	-6.35	4.62	-0.46	8.34
New Replies Posts Received	-1.96	7.14	1.17	12.31	0.79	7.85
New Posts Sent	-5.07	36.42	5.47	20.06	-0.39	59.85
Post Rate	-7.48	21.93	8.88	10.52	-1.39	40.76
Reposts Sent Rate	9.83	23.68	-9.45	11.71	-0.38	29.11
Reposts Received Rate	-13.56	12.88	13.64	14.97	-0.08	42.22
Quote Posts Sent Rate	-4.92	38.97	5.19	51.82	-0.28	70.28
Quote Posts 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 Reposts	-3.58	0.62	-10.67	0.61	14.24	0.72
Proportion Quote Posts	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 Reposts Sent Rate	-3.01	11.82	3.30	6.07	-0.28	10.28
New Account x Reposts Received Rate	-24.95	97.13	25.05	46.47	-0.10	534.07
New Account x Quote Posts Sent Rate	7.39	5.03	-6.98	3.15	-0.41	14.46
New Account x Quote Posts 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 Reposts	8.11	1.27	6.50	0.73	-14.61	0.99
New Account x Proportion Quote Posts	-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:	-0.06					
Pseudo R-squ.:	0.96					

Table 12: Non-regularized multinomial logistic regression with all account characteristics

Table 13 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 13: Confusion matrix of non-regularized classifier with all account characteristics with a total test accuracy of 94.10%.

Table 14 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 Following Measure)	44.01	0.50	41.15	0.69	-85.16	0.72
log(Unsafe Followers Measure)	-35.79	0.54	-25.78	0.68	61.57	0.73
log(Unsafe Reposts Measure)	189.78	0.38	188.75	0.47	-378.53	0.70
log(Unsafe Reposted Measure)	152.14	0.41	154.29	0.53	-306.43	0.96
log(Propaganda Following Measure)	25.14	0.77	19.49	0.83	-44.63	0.91
log(Propaganda Followers Measure)	-76.44	0.95	-77.03	0.71	153.46	0.95
log(Propaganda Reposts Measure)	-122.27	0.81	-122.45	1.33	244.73	0.47
log(Propaganda Reposted 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:	-0.18					
Pseudo R-squ.:	0.87					

Table 14: Non-regularized multinomial logistic regression with network-based account characteristics

Table 15 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 15: Confusion matrix of non-regularized classifier with network-based account characteristics with a total test accuracy of 89.54%.

Table 16 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 Reposts Sent	-1.09	11.99	5.04	9.17	-3.95	14.07
New Reposts Received	-1.15	6.21	12.26	9.47	-11.11	7.06
New Quote Posts Sent	2.02	80.64	-1.56	92.00	-0.46	126.83
New Quote Posts Received	10.64	25.11	-6.03	34.87	-4.61	24.76
New Replies Posts Sent	3.25	3.94	1.48	4.57	-4.73	5.76
New Replies Posts Received	-1.91	5.79	2.92	10.64	-1.01	6.24
New Posts Sent	-5.21	26.31	9.13	22.12	-3.93	39.37
Post Rate	-6.95	21.22	5.31	11.03	1.64	25.85
Reposts Sent Rate	-5.70	19.56	-3.36	11.42	9.06	19.02
Reposts Received Rate	-3.04	9.16	10.24	12.33	-7.20	8.35
Quote Posts Sent Rate	3.51	38.17	-3.53	46.33	0.02	61.19
Quote Posts 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 Reposts	-0.92	0.60	-0.36	0.51	1.28	0.63
Proportion Quote Posts	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 Reposts Sent Rate	-4.93	9.56	2.17	5.66	2.76	7.70
New Account x Reposts Received Rate	-6.27	102.20	7.90	49.00	-1.62	90.99
New Account x Quote Posts Sent Rate	9.29	4.04	-1.82	2.90	-7.47	5.74
New Account x Quote Posts 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 Reposts	2.52	1.27	-0.91	0.64	-1.62	1.09
New Account x Proportion Quote Posts	-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:	-0.44					
Pseudo R-squ.:	0.62					

Table 16: Non-regularized multinomial Logistic regressions with non-network-based account characteristics

Table 17 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 17: Confusion matrix of non-regularized classifier with non-network-based account characteristics with a total test accuracy of 73.73%.

Table 18 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	1.05	0.54	0.00	0.55	-0.28	0.61
Is New Account	-0.10	1.03	0.62	0.52	0.00	0.92
% Change Followers	-0.68	2.21	0.02	2.93	0.00	2.28
% Change Following	0.54	0.41	-0.60	0.33	0.00	0.36
New Followers Rate	1.11	7.01	-0.16	9.37	0.00	4.81
New Following Rate	-0.41	1.97	0.00	1.92	0.00	0.94
New Reposts Sent	-0.12	11.92	1.59	9.07	-0.48	13.96
New Reposts Received	0.00	6.18	0.00	9.43	-0.38	7.02
New Quote Posts Received	0.13	25.06	0.00	34.70	0.00	24.71
New Replies Posts Sent	0.00	3.91	0.06	4.54	-0.90	5.71
New Posts Sent	0.00	26.10	0.52	21.87	-0.03	39.04
Post Rate	0.00	21.28	0.56	10.91	-0.27	25.77
Reposts Received Rate	0.00	9.18	0.00	12.40	-0.04	8.37
Replies Sent Rate	0.54	6.61	0.00	6.05	-1.04	8.24
Proportion Reposts	-1.36	0.60	0.00	0.50	1.37	0.62
Proportion Quote Posts	2.89	1.33	0.00	1.73	-2.98	2.23
Proportion Replies	0.59	0.55	-2.40	0.70	0.82	0.71
Followers to Following Near 1	-0.30	0.48	0.00	0.51	0.76	0.36
New Account x Quote Posts Sent Rate	0.47	4.06	0.00	2.95	-0.08	5.79
New Account x Proportion Reposts	1.60	1.27	-0.01	0.63	-0.60	1.09
New Account x Proportion Quote Posts	0.00	2.81	0.00	1.82	-0.05	2.76
New Account x Proportion Replies	-0.61	1.40	0.00	0.93	1.01	1.09
Propaganda Hashtag Measure	-0.17	1.07	-0.74	0.95	1.91544	0.81
No. Observations:	556					
Log-Likelihood:	-0.67					
Pseudo R-squ.:	0.27					

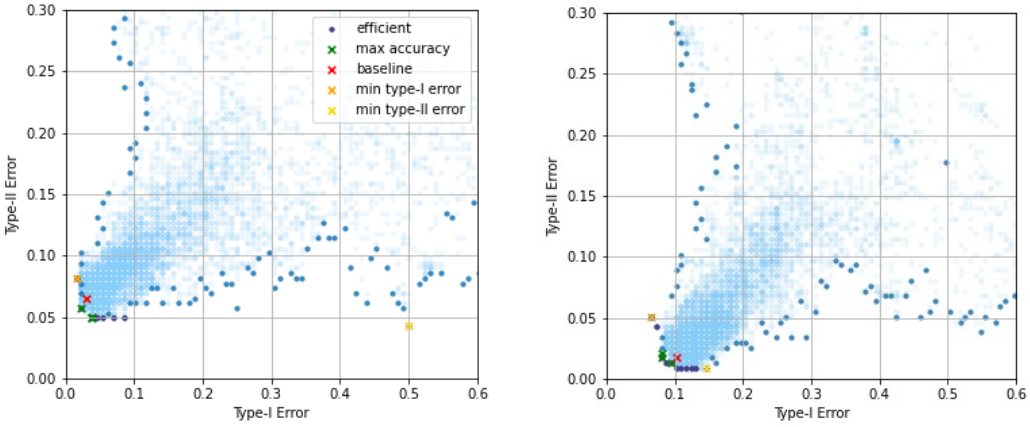
Table 18: Regularized multinomial logistic regressions with non-network-based account characteristics
To choose the explanatory variables, we apply an elastic net with equal L1 and L2 penalties. Factors with point estimates of zero were pushed to zero by the elastic net.

Table 19 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 19: Confusion matrix of regularized classifier with non-network-based account characteristics with a total test accuracy of 67.83%.

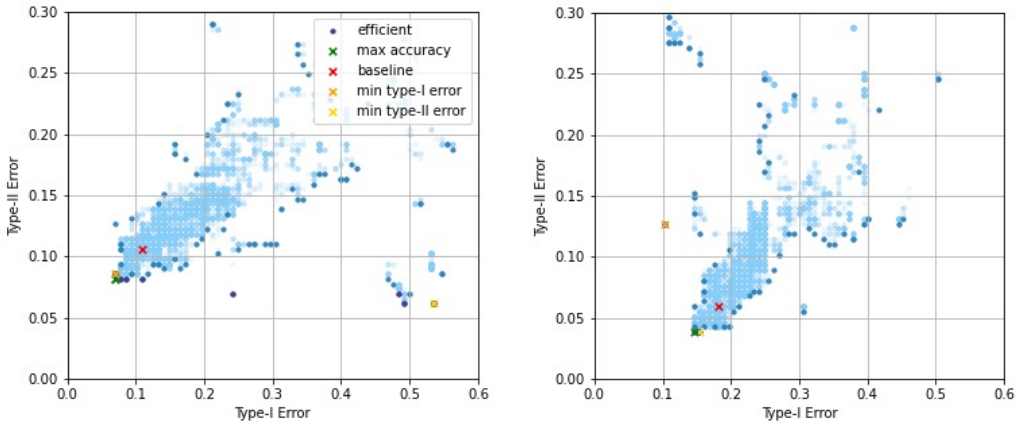
B Additional Charts and Figures



Null Hypothesis: Account is Ordinary

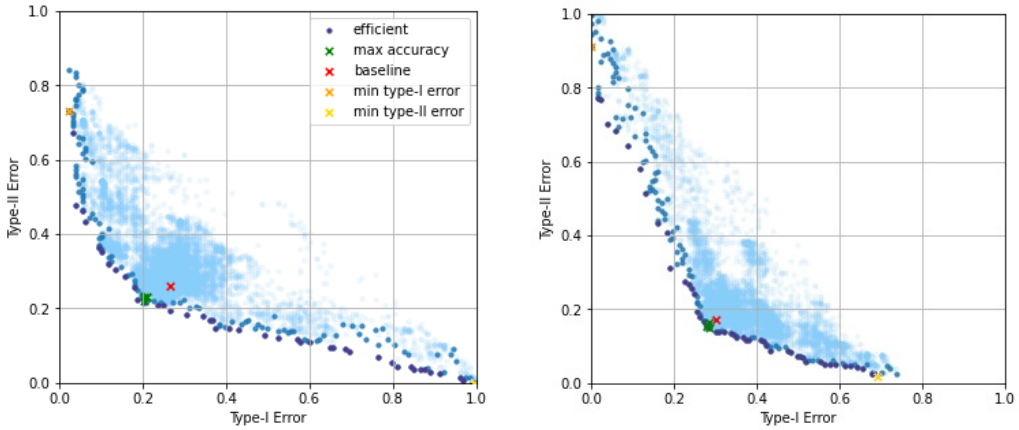
Null Hypothesis: Account is Unsafe

Figure 10: The type I-type II error trade-off using different combinations of all account characteristics.
Dark blue points are models that are on the lower envelope of the trade-off. Purple points are models that are on the Pareto efficient frontier of type I-type II error trade-off. The orange cross is the model that minimizes type I error. The yellow cross is the model that minimizes type II error. Green crosses are models with highest overall test accuracy. The red cross is the baseline classifier in Table 3.



Null Hypothesis = Account is Ordinary Null Hypothesis = Account is Unsafe

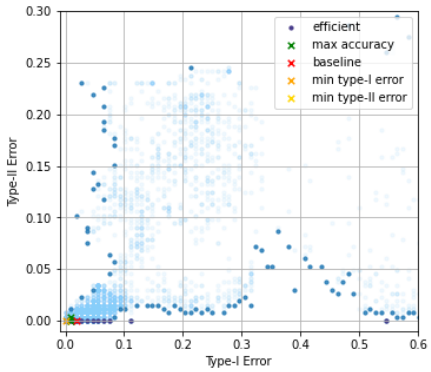
Figure 11: The type I-type II error trade-off using different combinations of non-manipulable account characteristics.
Dark blue points represent the lower envelope of the trade-off. Purple points represent models that are on the Pareto efficient frontier of type I-type II error trade-off. The orange cross is the model that minimizes type I error. The yellow cross is the model that minimizes type II error. Green crosses are models with highest overall test accuracy. The red cross represents the baseline classifier in Table 7.



Null Hypothesis = Account is Ordinary Null Hypothesis = Account is Unsafe

Figure 12: The type I-type II error trade-off using different combinations of manipulable account characteristics.
Dark blue points represent the lower envelope of the trade-off. Purple points represent models that are on the Pareto efficient frontier of type I-type II error trade-off. The orange cross is the model that minimizes type I error. The yellow cross is the model that minimizes type II error. Green crosses are models with highest overall test accuracy. The red cross represents the baseline classifier in Table 18.

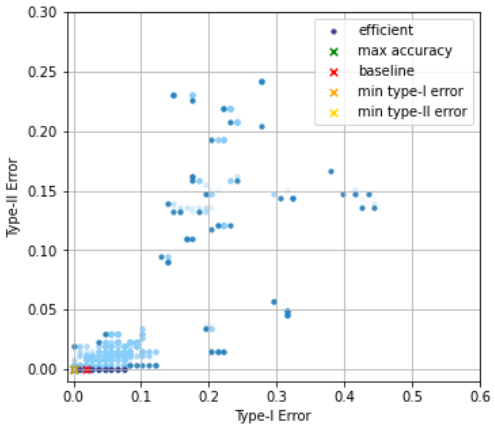
Figure 13 illustrates the trade-off between type I and type II errors in classifying propaganda accounts using different combinations of all account characteristics.



Null Hypothesis = Account is Propaganda

Figure 13: The type I-type II error trade-off using different combinations of all account characteristics.
Dark blue points represent the lower envelope of the trade-off. Purple points represent models that are on the Pareto efficient frontier of type I-type II error trade-off. The orange cross is the model that minimizes type I error. The yellow cross is the model that minimizes type II error. Green crosses are models with highest overall test accuracy. The red cross represents the baseline classifier in 3.

Figure 14 illustrates the trade-off between type I and type II errors in classifying propaganda accounts using different combinations of network-based account characteristics.



Null Hypothesis = Account is Propaganda

Figure 14: Trade-off between type I and type II errors in classifying propaganda accounts using different combinations of network-based account characteristics.
Purple points represent models that are on the Pareto efficient frontier of type I-type II error trade-off. The orange cross is the model that minimizes type I error. The yellow cross is the model that minimizes type II error. Green crosses are models with highest overall test accuracy. The red cross represents the baseline classifier in Table 7.

C Supplemental Information

Table 20 outlines the eight identified rumors used in our analysis.

rumor	start date	description	sample
Rape of Nika Shakarmi	October 4, 2022	Disappeared protestor Nika Shakarmi, whose death security guards had a suspected role in, was rumored to have also been raped before her death.	link
Murder of Pardis Javid	October 14, 2022	Reported death of a Kurdish student who was allegedly kidnapped by security forces during a protest just before her death.	link
Murder of Hana Duzduzani	October 14, 2022	One of several false names reported to have died in protests to cast doubt on real cases.	link
Massacre at Saadat-Abad Square	November 20, 2022	Reports of indiscriminate open fire by IRI security forces at civilians at Saadat-Abad Square after a soccer match.	link
Arrest and torture of Hasan Firoozi	December 8, 2022	Fictitious political prisoner who had a video of his circulate in which he pleads with IRI officials to let him see his newborn daughter before his scheduled execution.	link
Death of Judge Salavati	January 5, 2023	False reports of Judge Salavati's death began to circulate at the beginning of the year.	link
Assault of Sara Shirazi	February 21, 2023	A woman was reported to have assaulted a schoolgirl in Isfahan for improperly wearing her religious garb.	link
The murder of Fatemeh Rezaee	on February 26, 2023	Rumor began to spread that a protester in Qom had died due to overexposure to poisonous chemicals used by riot police. Qom officials had allegedly threatened anyone who knew about the death.	link

Table 20: List of false rumors collected and processed for use in our analysis.

Table 21 describes all of the factors we consider in creating our classifier.

factors	description
log(Unsafe Following Measure)	log of unsafe following measure (how ingrained a user's followings are with imposter accounts)
log(Unsafe Followers Measure)	log of unsafe followers measure (how ingrained a user's followers are to imposter accounts)
log(Unsafe Reposts Measure)	log of unsafe reposts measure (how ingrained a user's reposts are to posts made by imposter accounts)
log(Unsafe Reposted Measure)	log of unsafe reposted measure (how ingrained a user's reposted posts are to those reposted by imposter accounts)
log(Propaganda Following Measure)	log of propaganda following measure (how ingrained a user's followings are with propaganda accounts)
log(Propaganda Followers Measure)	log of propaganda followers measure (how ingrained a user's followers are to propaganda accounts)
log(Propaganda Reposts Measure)	log of propaganda reposts measure (how ingrained a user's reposts are to posts made by propaganda accounts)
log(Propaganda Reposted Measure)	log of propaganda reposted measure (how ingrained a user's reposted posts are to those reposted by propaganda accounts)
Degree Followers Centrality	user's degree centrality in follower-following network (number of edges to or from user)
Eigen Followers Centrality	user's eigenvector centrality in follower-following network (centrality weighted by a user's connection to highly-centered nodes)
Betweenness Followers Centrality	user's betweenness centrality in follower-following network (centrality measured by a user's presence in the shortest paths between all other users)
Followers to Following Ratio	ratio of a users number of followers to followings
Account Age (days)	days since account's creation
Is New Account	an indicator as to whether an account was created after the start of the protests on September 16, 2022
% Change Followers	percent change in a user's number of followers since the start of the protests
% Change Following	percent change in a user's number of followings since the start of the protests
New Followers Rate	the rate at which a user has gained followers since the start of the protests
New Following Rate	the rate at which a user has followed users since the start of the protests
New Reposts Sent	the number of reposts an account has sent since the start of the protests
New Reposts Received	the number of time an account has been reposted since the start of the protests
New Quote Posts Sent	the number of quote posts an account has sent since the start of the protests
New Quote Posts Received	the number of time an account has been quote posted since the start of the protests
New Replies Posts Sent	the number of replies an account has sent since the start of the protests
New Replies Posts Received	the number of time an account has been replied to since the start of the protests
New Posts Sent	the number of posts an account has posted since the start of the protests
Post Rate	the rate at which an account posts since its creation
Reposts Sent Rate	the rate at which an account reposts since its creation
Reposts Received Rate	the rate at which an account is reposted since its creation
Quote Posts Sent Rate	the rate at which an account quote posts since its creation
Quote Posts Received Rate	the rate at which an account is quote posted since its creation
Replies Sent Rate	the rate at which an account replies since its creation
Replies Received Rate	the rate at which an account is replied to since its creation
Proportion Reposts	the proportion of an account's total activity which is reposts
Proportion Quote Posts	the proportion of an account's total activity which is quote posting
Proportion Replies	the proportion of an account's total activity which is replying to content
Followers to Following Near 1	an indicator as to whether an account has as many followers as they do followings $\pm 5\%$
New Account x Followers Rate	followers rate interacted with an account being new
New Account x Reposts Sent Rate	reposts sent rate interacted with an account being new
New Account x Reposts Received Rate	reposts received rate interacted with an account being new
New Account x Quote Posts Sent Rate	quote posts sent rate interacted with an account being new
New Account x Quote Posts Received Rate	quote posts received rate interacted with an account being new
New Account x Replies Sent Rate	replies sent rate interacted with an account being new
New Account x Replies Received Rate	replies received rate interacted with an account being new
New Account x Proportion Reposts	proportion of reposts interacted with an account being new
New Account x Proportion Quote Posts	proportion of quote posts rate interacted with an account being new
New Account x Proportion Replies	proportion of replies rate interacted with an account being new
Propaganda Hashtag Measure	a measure of how often an account uses hashtags which are also used by propagandists

Table 21: Description of all factors considered in creating the classifiers.