(Dis)Information Wars*

Adrian Casillas† Maryam Farboodi‡ Layla Hashemi§
Maryam Saeedi¶ Steven Wilson‖

August 17, 2025

Abstract

Digitization has radically transformed news consumption. While social media are now essential information channels, their lack of editorial oversight allows sophisticated disinformation campaigns that threaten democratic discourse. We leverage endogenous patterns of disinformation supply to significantly reduce disinformation consumption without impacting legitimate news spread. Our Network-based Disinformation Labeling (NDL) approach identifies false information through network structure analysis rather than content review. Using comprehensive data, we show that disinformation originates from accounts with distinct network characteristics that coordinate on news origination. Simulations show that NDL reduces disinformation posts by 70% and maximum engagement by 50%, correctly identifying 85% of disinformation events.

JEL codes: A13, D72, L82, P0, Z13

*This paper was initially 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 and the NBER Digitization group for financial support. We are grateful to Isaiah Andrews, Hülya Eraslan, Thomas Fujiwara, Matthew Gentzkow, Azam Heydari, Andrea Prat, Maryam Nazari, Jesse Shapiro, Ali Shourideh, Steven Tadelis, Francesco Trebbi, David Yang, participants of NBER Political Economy program meeting Fall 2023, 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.

†NYU Stern; ac12001@stern.nyu.edu
‡MIT Sloan, NBER and CEPR; farboodi@mit.edu
§George Mason University; lhashem2@gmu.edu
¶Carnegie Mellon University; msaeedi@andrew.cmu.edu
‖Brandeis University; stevenwilson@brandeis.edu

1 Introduction

The digital transformation of news consumption has fundamentally altered how information spreads through society. With more than half of U.S. adults now receiving news from social media platforms that lack traditional editorial oversight, the media landscape has become increasingly susceptible to deliberate manipulation.[1] This vulnerability is especially acute given the speed, scale, and opacity with which content spreads on these platforms, allowing false narratives to go viral before fact-checkers or platform interventions can respond (Zhuravskaya et al., 2020).[2]

These "disinformation wars" represent deliberate efforts by concerted actors to manipulate information supply chains, intentionally spreading false information to influence public beliefs for political, economic, or reputational gain. A growing body of evidence shows that many such campaigns are organized or supported by state actors, often through *troll farms*, that industrialize the production and amplification of false narratives (Bradshaw and Howard, 2018; Denter and Ginzburg, 2024). These operations typically involve paid operatives managing large clusters of inauthentic accounts to flood networks with coordinated disinformation. Investigations have revealed how authoritarian regimes use such troll farms to silence dissent and distort political discourse.[3] The scale and sophistication of such efforts further exacerbate the limitations of conventional interventions like content moderation or fact-checking, which are often too slow to counteract the viral spread of these campaigns.

The economic literature has extensively documented how information provision strategies affect the consumption of news and political outcomes, from media capture by governments (Besley and Prat, 2006) to the rise of "informational autocrats" who manipulate information rather than rely on coercion (Guriev and Treisman, 2019). More recently, Guriev et al. (2021) show how internet access affects political attitudes and trust in institutions across countries. However, despite this sophisticated understanding of information manipulation, existing policy tools have proven inadequate to counter modern disinformation threats.

In this paper, we consider an alternative strategy to combat the spread and consumption of disinformation. Rather than fact-checking or content moderation, we propose *Network-*

[1] Pew Research Center, "Social Media and News Fact Sheet," 2024. Available at: https://www.pewresearch.org/journalism/fact-sheet/social-media-and-news-fact-sheet/. Accessed: July 2, 2025.

[2] For example, Allcott and Gentzkow (2017) document the prevalence of disinformation and misinformation regarding the presidential election in 2016.

[3] Human Rights Foundation, Like, Tweet, & Torment: Authoritarian Troll Farms, 2023; available at: https://hrf.org/latest/like-tweet-torment-authoritarian-troll-farms/, Accessed July 2, 2025.

based Disinformation Labeling (NDL) to identify inorganic distribution patterns coordinated by adversarial agents using network structure analysis. The NDL approach is grounded in two key observed insights about disinformation networks: First, disinformation originates from accounts with distinct network characteristics that cluster together and follow specific engagement patterns. Second, disinformation campaigns typically begin with coordinated, near-simultaneous sharing by multiple accounts, while legitimate news spreads through more diverse organic patterns.

Using comprehensive data from X (formerly Twitter), we demonstrate that NDL can effectively detect disinformation and disrupt its supply chain before it achieves widespread impact. We use simulations to show that NDL reduces disinformation posts by 70% and decreases both maximum engagement and effective lifespan of disinformation events by 50%. Critically, NDL accomplishes this early detection without disrupting the flow of legitimate news.

Intuitively, we draw on the theoretical foundations in strategic information transmission and network economics to leverage users' behavioral responses, rather than working against them, to stop the spread of disinformation.[4] Consistent with recent findings by Guriev et al. (2023), we show that early interventions through information labeling—even without full verification—can substantially reduce user engagement with disinformation.[5]

Our methodology reduces the consumption of disinformation more effectively than traditional countermeasures, which have proven insufficient due to three key limitations: Manual content verification is time-consuming and costly which creates intervention delays; disinformation spreads exponentially on social networks, implying that damage occurs before verification is complete;[6] and psychological research shows that corrections have limited effectiveness once false beliefs are internalized.[7] Content-based detection approaches face additional challenges as sophisticated language models both undermine existing detection algorithms and enable rapid production of multilingual disinformation.

Although recent advances in natural language processing can considerably improve the precision of content-based approaches, they still suffer from major delays in detecting false news. The critical advantage of our network-based approach is rapid detection of disinformation with satisfactory precision, which provides a robust solution to the aforementioned problems. The effectiveness of NDL stems from its ability to exploit the inherent constraints faced by the disinformation supply network. While disinformant accounts can strategically

[4]See Crawford and Sobel (1982) and Acemoglu et al. (2010).

[5]Our recommendation differs from the recommendation in Guriev et al. (2023) to some extent, but shares immediate warning aspects without much delay.

[6]Caplan et al. (2018) argues that the speed of disinformation spread is higher than content moderation.

[7]See Chan et al. (2017), Nyhan and Reifler (2010) and Ecker et al. (2022)

alter their posting behavior to mimic ordinary users, they face fundamental limitations in building authentic network relationships. This creates an asymmetry that yields distinctive network signatures, enabling reliable detection even as content generation technology evolves. Using insights from Akbarpour et al. (2020) who demonstrate the importance of network structure in determining information diffusion patterns and optimal seeding strategies, we believe the network-based approach remains effective as content generation technology evolves.

To validate NDL methodology and estimate its impact on the spread of disinformation on social media platforms, we apply it to a real-world scenario. We use data from X (formerly Twitter) during Iran's "Woman, Life, Freedom" movement, which began in September 2022.[8] This setting provides an ideal laboratory to study disinformation wars, as it features both organized disinformation campaigns and grassroots efforts to counter false narratives, allowing us to observe the full spectrum of information warfare tactics.

We construct a comprehensive dataset, comprising all Farsi-language posts on X from September 16, 2019, to March 14, 2023, augmented with network-based and non-network-based account characteristics. We supplement our dataset with hand-collected instances of disinformation and legitimate news events that occurred during this period on platform X, focusing on very similar news types.[9] Finally, we include organic rebuttals to these disinformation events. This rich data environment enables us to trace the complete life-cycle of information campaigns and measure intervention effectiveness across multiple dimensions.

In the first stage of our approach, we devise a multinomial logit model with elastic net regularization, based on network and non-network account characteristics, to categorize accounts that post in Farsi on X into three categories: "unsafe" accounts that actively participate in spreading disinformation, "ordinary" accounts that primarily share legitimate information, and "pro-regime" accounts that openly support state propaganda. We utilize a hand-collected labeled dataset of unsafe, ordinary, and pro-regime accounts to train and test our model. For each account, we construct various variables related to their characteristics, such as age, number of followers, number of following, and rate of postings. In addition to these variables, we construct *network proximity scores*. These network-based variables measure how close an account is to known unsafe and pro-regime accounts, from the training data, based on their following and engagement patterns. Our algorithm achieves 95% accuracy in identifying account categories by highly relying on the network proximity scores.

[8] This movement started after a 22-year-old Iranian woman, Mahsa Zhina Amini, died in a Tehran hospital after her arrest by Iran's morality police for alleged hijab violations. Fury grows in Iran over woman who died after hijab arrest. Accessed July 2, 2025.

[9] When choosing legitimate news, we selected a set of political news that was related to the women-life-freedom movement, as did the set of disinformation news we incorporated.

The second stage aims to identify disinformation events on X as soon as they begin. We employ the first stage's categorization of accounts and the initiators' category to label news events. As unsafe accounts try to mimic the behavior of ordinary accounts, not every piece of news that an unsafe account spreads is disinformation. However, unsafe accounts coordinate to initiate the spread of disinformation. As such, if many of the first accounts that initiate a piece of news are classified as unsafe, we label that news as "disinformation."

We find that using data up to four months prior to the disinformation date, our model identifies at least 85% of ex-post verified disinformation events in our sample without mistakenly labeling any legitimate news as disinformation. The accuracy increases as we expand the training set or the length of data used to construct the variables. Both result in even higher accuracy in identifying disinformation instances.

To estimate the effectiveness of employing NDL to curb the spread of disinformation, we simulate the impact of its implementation on the spread of false news on the social media platform. To perform these simulations, we need a measure of the impact of information provision on disinformation consumption on social media. In other words, we have to estimate how much informing social media users about the false nature of news impacts their behavior in spreading it. In order to measure this value, we use the organic rebuttals by political activists on X during the same time period to inform. We estimate the causal impact of these rebuttals on different measures of disinformation spread on X and use these estimates to predict the impact of the adoption of NDL by X on the extent of disinformation spread.

Our simulations in Section 4, informed by these estimates, illustrate that implementing our approach leads to a three-fold reduction in the number of posts in disinformation campaigns. Furthermore, the maximum user engagement rate and the effective lifespan of disinformation are reduced by at least half.

These findings suggest that our approach can be an effective mechanism to combat the spread of disinformation and significantly reduce its consumption. It is worth noting that the key to successfully limiting the spread is early detection of disinformation before it goes viral. Our approach achieves this goal by relying only on the type of the first few initiators of news events to identify disinformation.

We also highlight the importance of network-based characteristics in the high performance of our method for identifying both unsafe accounts and disinformation events. Removing network proximity scores reduces algorithm accuracy from 95% to 74.7%, with many instances of unsafe and ordinary accounts being misclassified as one another.[10] In

[10]Importantly, this is the case despite using the common centrality measures for classification.

contrast, restricting our model to only network proximity scores minimally reduces the accuracy to 93.7%. Furthermore, without network proximity scores, the accuracy of labeling disinformation drops dramatically. Using data up to four months prior to disinformation events results in only 20-60% accuracy, with a concerning 20% chance of misidentifying legitimate news as disinformation. In contrast, using only network-based characteristics maintains substantial accuracy in identifying disinformation events, further emphasizing the importance of these network proximity scores and the use of past relationships of users with one another.

Our research demonstrates that NDL is inherently resistant to manipulation attempts by disinformation spreaders. While unsafe accounts may strategically alter their behavior to evade detection, they face a fundamental constraint: They cannot compel ordinary users to follow them or share their content, creating an asymmetry that protects distinctive network signatures. As mentioned above, focusing on only network-based variables does not significantly decrease the accuracy of the model, either in identifying accounts or disinformation events. Another feature of our proposed algorithm is its disinformation labeling approach, which aligns with theoretical insights on optimal information signals under manipulation. We employ an opaque scoring scheme where unsafe accounts remain unaware of which accounts have been flagged and why, creating substantial obstacles to strategic adaptation.[11] Even potential long-term evasion strategies, such as creating entirely new account networks, would require overcoming significant barriers in establishing convincing network positions among legitimate users while generating new detectable network patterns.

Beyond its immediate application to Iran's social media landscape, our methodology has broad implications for combating disinformation globally. NDL shifts the focus from content-based approaches to supply-network detection. NDL is particularly valuable because it: (1) enables early intervention before false information spreads widely, (2) does not require resource-intensive content moderation, and (3) remains effective even when disinformation spreaders attempt to evade detection. As social media platforms grapple with escalating disinformation campaigns worldwide, our findings provide a practical framework for identifying and limiting the spread of false narratives while protecting legitimate information flow.

[11] This feature has been proposed by Frankel and Kartik (2019); Perez-Richet and Skreta (2022); Saeedi and Shourideh (2023)

1.1 Literature Review

Our paper contributes to a large and interdisciplinary literature on the economics of media, information, and disinformation, with particular relevance to the design of interventions under strategic manipulation.

One key strand of literature examines media capture and the political economy of information control, particularly by governments. Seminal work by Besley and Prat (2006) models how media capture weakens government accountability, while Guriev and Treisman (2019), and Guriev et al. (2021) study how "informational autocrats" influence beliefs and trust by controlling the information environment rather than relying on coercion. Schedler (2010) and Shadmehr and Bernhardt (2015) analyze state censorship and propaganda, while Chen and Yang (2019) and Simonov and Rao (2022) focus on public demand for uncensored or ideologically neutral information. Our work complements these studies by focusing not on the content of propaganda itself, but on the network structure through which it spreads, offering a tool for limiting disinformation by targeting the strategic actors who supply it.

A second literature explores the effects of digital technology on information production and consumption. Cagé et al. (2020) and Angelucci et al. (2020) document how online technologies alter the news supply chain, while Falck et al. (2014), Gavazza et al. (2019), and Campante et al. (2018) investigate how internet expansion affects political participation and attitudes. Zhuravskaya et al. (2020) provide a comprehensive summary of how internet access reshapes the media landscape, often with consequences for misinformation. Our paper contributes to this literature by modeling how network topology constrains the dissemination of false information, and by proposing a scalable tool that harnesses these constraints to reduce disinformation spread.

There is also a growing body of work studying misinformation and disinformation on social media platforms. Allcott and Gentzkow (2017), Guess et al. (2019) and Grinberg et al. (2019) document the prevalence and diffusion of false news on platforms like Facebook and Twitter, especially around electoral events. Allcott and Gentzkow (2017) track user interactions with disinformation across time and platforms, finding divergence in platform effectiveness. Thomas et al. (2012) and Stukal et al. (2017) provide evidence for the extensive use of false information, particularly on Russian Twitter. Vosoughi et al. (2018) highlight that false information spreads more rapidly than truth online, while Bradshaw and Howard (2018) document how coordinated disinformation campaigns, often state-sponsored, are deployed across dozens of countries. Our empirical focus on disinformation in the Iranian context during the "Woman, Life, Freedom" movement contributes new evidence on these dynamics in a non-Western setting and demonstrates how network structure reflects the

strategic behavior of disinformants.

Several methods have been proposed to combat online disinformation. Fact-checking and content moderation have been evaluated both theoretically and empirically, with mixed success. Some studies, such as Kunda (1990), Nyhan and Reifler (2010), Chan et al. (2017), Kahan et al. (2017), Barrera et al. (2020), and Ecker et al. (2022) show that ex-post rebuttals of misinformation have limited effectiveness due to belief entrenchment. Caplan et al. (2018) and Bak-Coleman et al. (2022) emphasize the speed advantage of disinformation over correction. Guriev et al. (2023) find that algorithmic labeling of false information can reduce engagement, even in the absence of full content verification. Our approach augments these insights by proposing a network-based detection method that enables early intervention, well before content can be verified or rebutted manually or before the content can go viral.

Finally, our paper is closely related to the theoretical literature on strategic information transmission and information diffusion in networks. Crawford and Sobel (1982) provide the foundational model of cheap talk between senders and receivers with misaligned interests. Acemoglu et al. (2010) study how network structure mediates the trade-off between information aggregation and misinformation. Akbarpour et al. (2020) explore optimal seeding strategies in information diffusion, while Budak et al. (2011) develop algorithms to limit misinformation spread. More recently, Wang et al. (2024) model reputational incentives in social media communication, and Frankel and Kartik (2019), Perez-Richet and Skreta (2022), and Saeedi and Shourideh (2023) study the design of opaque mechanisms that are robust to strategic manipulation.

Taken together, our paper contributes to the literature in four directions. First, we provide a novel approach for early detection of disinformation that does not rely on content but instead exploits the structural constraints of adversarial supply networks. Second, we offer empirical evidence, using X (formerly Twitter) data, on how disinformation actors coordinate and how their network signatures differ systematically from legitimate users. Third, we propose a policy-relevant mechanism that is scalable, cost-efficient, and resilient to evasion, expanding the toolkit for social media platforms and regulators. Fourth, we show that our method leads to significant reductions in disinformation spread, even when implemented under realistic conditions, offering a proof-of-concept for supply-side interventions grounded in network economics.

The rest of the paper is organized as follows. Section 2 provides details of the data that we use for the estimation. Section 3 describes our network-based disinformation labeling approach, NDL, for prompt identification of disinformation and the baseline estimation

results. Section 4 estimates the impact of NDL on the spread of disinformation on the social media platform. Section 5 compares our network-based approach with more traditional approaches and discusses its resilience towards manipulation attempts. Section 6 presents a number of robustness exercises. Section 7 concludes.

2 Data

We use data from X (formerly Twitter) spanning the period from the inception of the "Woman, Life, Freedom" movement in Iran–triggered by Mahsa Amini's death on September 16, 2022–to the discontinuation of the X API v1 streaming service on March 14, 2023, to construct a novel dataset.[12] We use this dataset to devise an algorithm to detect disinformation on X and measure its effectiveness in containing the spread of disinformation.

Our dataset has two main components. The first component consists of all the X accounts that have more than 10% of their interactions in Farsi and have posted at least ten times after September 2022. This component includes these accounts' characteristics, posts, engagements, and their social network. The second component consists of 14 instances of disinformation, augmented by 1,374 organic rebuttals of these disinformations in 924 distinct threads of posts, as well as 10 instances of relevant legitimate news that were spread on X during the time period of interest. We explain each component separately below.

Accounts, posts, and social network In order to construct a comprehensive set of accounts active on Farsi X during this period, we combine the data from X v1 API and v2 API. The X v1 API gives us an archive of all Farsi-language posts made after September 16, 2019.[13] The X v2 API allows us to get a complete network construction of active accounts in our data by finding all follower-following links among them.

X v1 API stream allowed its users to collect a stream of all new posts filtered on a number of parameters, provided that the stream made up less than 1% of new posts. By including a filter that only lets through posts written in Farsi and since Farsi language posts are almost never more than 1% of new posts, the combination includes almost all posts in

[12] See https://www.britannica.com/topic/Woman-Life-Freedom for more details about the "Woman, Life, Freedom" movement in Iran. Accessed 07/06/2024.

[13] The archive is hosted by Brandeis University. The data from X v1 API was originally collected by Leyla Hashemi and Steven Wilson; and the early part of the data, September 19, 2019 - January 28, 2021, underlies the paper Hashemi et al. (2022). The streamer continued to collect data until the termination of the X API v1 streaming service on March 14, 2023. See Hashemi et al. (2022) for an overview of this dataset.

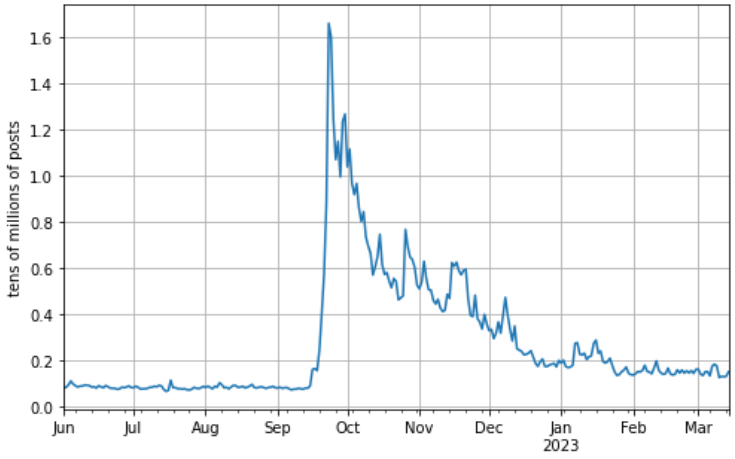


Figure 1: Daily volume of Farsi-language posts on X during 6/1/2022-3/14/2023

Farsi language since September 16, 2019. Figure 1 depicts a substantial increase in activity in the Farsi language X after the start of the protests in Iran.

A post object contains detailed information about its content and its creator. Each post is identified by a unique post ID and includes text, timestamp, and public metrics (such as the count of reposts, quote posts, replies, and likes), along with the creator's unique ID, username, and public metrics (including the number of followers, followings, and posts) at the time of posting.

As posts are stored immediately upon creation, they report no subsequent engagements. However, posts that interact with another post (such as reposts, quote posts, and replies) include a reference to the post they engage with. As such, by documenting nearly all Farsi posts after their creation, we can reconstruct a sequence of all subsequent engagements to every post. This allows us to map out the complete structure of Farsi-language X, including every reply thread and every chain of quote posts and reposts.

Furthermore, by replacing each post's ID with that of its creator, we can trace the history of follower/following interactions between any two accounts on X and construct the social network among them.

The number of X accounts that have ever posted in Farsi is very large, with more than nine million accounts. However, most of these accounts posted only a handful of times in Farsi.[14] To study accounts with a significant presence in the Farsi-speaking network of accounts, we restrict our sample to accounts that post in Farsi often. We define an *active*

[14]These are often non-Iranian users employing the popular "Woman, Life, Freedom" hashtags or sharing a friend's post. Furthermore, the algorithm that determines the language of posts has errors and sometimes misclassifies Arabic or Urdu as Farsi.

	Total	Monthly average pre 09/2022	Monthly average post 09/2022
Full Sample			
Number of accounts	9,525,800	711,600	1,240,600
Number of posts sent	83,937,899	1,896,500	2,610,900
Number of reposts sent	639,337,800	9,637,000	48,734,300
Number of quote posts sent	57,919,800	1,345,300	1,581,500
Number of replies sent	377,199,600	8,264,700	13,278,400
Active Accounts			
Number of active accounts	1,767,350	614,000	1,016,300
Number of posts sent	62,212,685	1,351,200	2,261,600
Number of reposts sent	589,714,000	8,364,400	44,186,200
Number of quote posts sent	52,452,100	1,198,945	1,248,300
Number of replies sent	349,949,200	7,840,700	12,080,400

Notes: Active accounts are those with more than 10% of posts and engagements in Farsi and at least ten posts since September 2022 in Farsi.

Table 1: Aggregate and monthly statistics pre- and post-September 2022

account as accounts that have at least 10% of their posts in Farsi and have also posted at least ten times in Farsi since the protests began. As shown in Table 1, while focusing on active accounts reduces the number of accounts by a factor of 5, the number of interactions does not drop significantly. Therefore, restricting the data to this subset does not result in a significant loss of generality.

Disinformation and legitimate news In order to test and train the model and measure its effectiveness in classifying disinformation on X and containing its spread, we need a sample of disinformation and legitimate news instances.

We hand-collected two sets of news spread on X during this time period. The first set consists of 14 ex-post verified disinformation campaigns that originated on X. These rumors are described in Table A.1 in Appendix A.1.[15] We then manually identify 1,374 organic rebuttals in 10 of the disinformation instances. These rebuttals are in the form of engagements with disinformation posts by activists to contest the validity of disinformation posts and have occurred on 924 independent threads of posts.[16] We augment our disinformation

[15]Three independent journalists provided assistance, verifying that these campaigns were indeed disinformation campaigns, and traced their origin to X. See Section 6.2 for more details.

[16]We review all the posts associated with each of these 10 disinformation campaigns and mark all posts that rebut the initial disinformation. The reason we did not use the others was the lack of apparent rebuttal before the rumor stopped spreading.

sample with these rebuttals. The second set of news consists of 10 verified ex post instances of relevant legitimate news that occurred during the same episode, described in Table A.2.

An important criterion for choosing the two news groups is the similarity in their content. As such, it is challenging to distinguish from the content alone whether each news item is true or false. Thus, these groups serve as a great test for validating our labeling methodology.

Using the above dataset, we collect and construct various characteristics for each account, as explained in Sections 2.2 and 2.3,[17] and use these characteristics to classify accounts in Section 3.1. The most important set of characteristics in our identification process is a set of network proximity to labeled accounts. In what follows, we first explain how we label accounts, and then we describe the construction of account characteristics.

2.1 Labeled Dataset

A "disinformation campaign" corresponds to a piece of news whose initiators are aware of its falsehood when they start spreading it on the social network. In our context, these false news were spread mainly to obfuscate what was happening during the protests and to undermine the opposition. Many of these disinformation campaigns originated on X and spread to other social media platforms, such as Instagram and Telegram. Section 3 provides details of our two-step approach to identify these disinformation campaigns. We first construct a model to identify Farsi-X accounts into three categories, and then use the category of accounts that initiate a piece of news to label the news as disinformation or not.

We label accounts in our dataset as one of the following three categories: "unsafe" accounts that actively participate in disseminating disinformation; "ordinary" accounts that do not; and "pro-regime" accounts that openly support the Islamic Republic of Iran and participate in its propaganda efforts.[18] One can further consider two groups of unsafe accounts. First, the "impostor" accounts that initiate the disinformation spread or spread it intentionally. In the context of the "Woman, Life, Freedom" protests, these accounts initially pretend to be from an opposition group, posting pro-dissidence content and hashtags to build a network among the protesters. Subsequently, they start posting disinformation at the right time.[19] The second group consists of "naive" normal X accounts who engage in

[17] Table A.3 in the Appendix A provides more detail of the construction.

[18] These accounts publicly engage with the propaganda of the Islamic Republic but do not hide their allegiance and therefore are easier to identify by both the public and our algorithm, as reported later on.

[19] 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

the spread of disinformation, albeit unintentionally. To identify disinformation news events, we do not need to distinguish between these two sets of accounts.

To execute and evaluate our methodology, it is necessary to begin with a collection of labeled accounts of these three groups. The most challenging category is unsafe accounts. We only use imposter-unsafe accounts to train our classifier to be conservative. We label the first 5% of the *initiators* of the verified disinformation campaigns that we have collected in our dataset as unsafe. As these accounts have created at least one instance of disinformation on X within our data, they are by definition unsafe. This procedure gives us a list of 476 unsafe accounts.

We next label an initial set of "pro-regime" accounts. In contrast to unsafe accounts, pro-regime 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 accounts of popular pro-regime agents. Second, we choose random posts of the I.R.I. leader and then randomly pick accounts that have liked these posts. We then manually check these accounts one by one to make sure that they are all pro-regime accounts. This procedure gives us a list of 470 pro-regime accounts.

An obstacle in labeling ordinary accounts for the case of Iran is that ordinary accounts often operate in anonymity to protect themselves and their associates from political prosecutions, making it difficult to distinguish them from unsafe accounts.[20] Thus, we label a set of accounts as ordinary accounts from a diverse group of individuals, ranging from well-known opposition leaders and celebrities to our friends and family and their acquaintances, whose identities can be attested. This procedure gives us a list of 489 ordinary accounts.

In summary, we have a total of 1,435 labeled accounts: 476 unsafe, 489 ordinary, and 470 pro-regime accounts. We divide this set into a training set containing 70% of labeled accounts from each category and a testing set with the remaining 30%.

2.2 Basic Characteristics

As a baseline, our classifier includes basic account features and activity statistics such as the number of followers and followings, follower-to-following ratio, account age, posting rate by post type, and post composition by type. However, we can expect these measures to become less predictive as disinformants strategically change their behavior to impede detection.

word choice.

[20]There have been numerous cases of prosecutions and even executions for online activities on X or other social media in Iran, see for example https://iranwire.com/en/politics/110844-young-iranian-woman-detained-for-46-days-over-tweet/. Accessed 06/21/2024.

In order to take advantage of the explosion of activity that took place on Farsi-X following the death of Mahsa Amini, we consider account features and activity before and after September 16, 2023. This includes whether an account was created before or after the start of the protests, changes in account activity throughout the protests, and the interaction of the two.[21]

2.3 Network-based Characteristics

The key distinguishing feature of NDL is the use of network characteristics. These characteristics indicate different proximity measures for each account to the accounts in the training set. We construct two sets of these metrics, one set in relation to known unsafe accounts and one in relation to known pro-regime accounts. The intuition behind these characteristics is that unsafe (pro-regime) accounts tend to follow and echo other accounts that are unsafe (pro-regime). In addition, this connectivity is a crucial factor for the success of a disinformation campaign. For a post to go viral and spread through the network, it must be widely shared and liked by numerous other accounts. In what follows, we explain the construction of these network proximity measures.

For any two accounts, u and v, we consider four types of relationships which we call "proximity scores:" 1) Following: u follows v, 2) Follower: u is followed by v, 3) Repost: u has reposted v, 4) Reposted: u has been reposted by v. We construct eight proximity scores for each account, four for each of the above engagements with unsafe accounts, and four for engagements with pro-regime accounts.

As an example, consider the construction of the proximity score for *unsafe following*. Start with the initialization step and give all the unsafe accounts in the training set a score of 1 and all other accounts a score of 0. Next, move to the iterative step. In this step, first, choose an account randomly among the ones with the highest score. Note that in this example, it will certainly be an unsafe account in the training set in the first round. Find all accounts that this account is *following* and add one to their score. That is, in this example, the "connected" unsafe accounts in the training set will get a score of 2, and the rest of the "connected" accounts will get a score of 1 in the first round. Repeat the iterative score until a terminal condition is satisfied.

As the score construction algorithm is random, we simulate it ten times and set the average of the scores to be the final score for each account. The set of final scores constitutes the corresponding proximity scores. Appendix A.3 provides details of the algorithm for the construction of the proximity score for *unsafe following*. The other seven proximity scores

[21] We constructed this dataset using input from multiple activists and journalists.

are defined similarly.

3 NDL: Network-based Disinformation Labeling

The spread of disinformation on social networks follows dynamics similar to those of infectious disease: It starts exponentially, reaches a peak, and dies out rapidly. As such, it is crucial to flag disinformation promptly after its origination and before the spread becomes pervasive.

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, thus allowing disinformation to go viral before it can be addressed. Furthermore, ex-post debunking has shown limited impact in debiasing the public.[22] Motivated by these deficiencies, we consider an alternative approach to restrict the "supply" of disinformation—a network-based disinformation labeling approach to enable ex-ante content moderation.

NDL has two steps. First, we utilize network-based account characteristics to predict users likely to engage in the disinformation supply network, even before they do so, and assign a category to them. Second, we employ the category of initiators of the news spread on social media to label them as disinformation or truth, shortly after their initiation.

3.1 Determination of Account Category

This section outlines our baseline methodology for identifying account categories and subsequently presents the estimation results.

Estimation methodology We calculate a triplet propensity score for each account in our data as the probability of belonging to the ordinary, unsafe, or pro-regime group. For each account with attributes $y \in Y$ we estimate $(p_o(y), p_u(y), p_p(y))$ where p_o is the probability that this account is an ordinary account, p_u is the probability of being an unsafe account, and p_p is the probability of being a pro-regime account, and $p_o + p_u + p_p = 1$. In the remainder of the paper, we will refer to p_u and p_p as *disinformation score* and *pro-regime score*, respectively.

[22] 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 and find limited, if any, impact.

Our estimation methodology is analogous to propensity score matching, a widely adopted method to estimate treatment effects (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 $y \in Y$, the propensity score of being an ordinary account, i.e., the likelihood of an account being ordinary, is given by

$$p_o(y) = \frac{exp(\beta_o y)}{exp(\beta_o y) + exp(\beta_u y) + exp(\beta_p y}.$$

Where $\beta_o, \beta_u, \beta_p$ are the coefficients corresponding to the ordinary, unsafe, and pro-regime groups, respectively. The likelihood of an account being classified as unsafe or pro-regime is described analogously. As with other choice models, we can only identify these coefficients up to a constant; therefore, we normalize the coefficients of ordinary accounts, β_o, to zero and report the other two sets of coefficients.

To estimate the logistic regression, we use seventy percent of the labeled accounts and the explanatory variables described in Section 2.[23] To avoid overfitting, we apply a regularized fit. This fit is an elastic net with equal weights on L1 and L2 penalties. The resulting set of nonzero regressors is then passed into logistic regression.

The estimated Logistic regression provides us with three propensity scores (p_o, p_u, p_p) for each account. We categorize each account into the group with the highest propensity score. For instance, if an account's scores are represented as $p = (0.5, 0.2, 0.3)$, it would be categorized as an ordinary account. The results and accuracy rate of this method are reported in the next section.

Estimation results Table 2 reports our baseline estimation results. Recall that these coefficients should be interpreted relative to ordinary accounts, for which the coefficients are normalized to 0. Also, note that the reported variables in the table are those selected by the elastic net. Several key findings warrant additional emphasis. First, all of the network proximity measures defined in Section A.3 are selected by the elastic net which underscores their importance in identifying account categories. Next, when examining the coefficients for unsafe accounts, two variables within the unsafe network proximity measures are positive and highly significant: *unsafe follower* and *unsafe reposted*. In contrast, the other two variables, *unsafe following* and *unsafe reposts*, are not significant. This is a notable observation as it implies that unsafe accounts are not easily differentiated from ordinary

[23] Details of the explanatory variables are provided in Appendix A.2.

	Unsafe		Pro-regime	
	coef	std err	coef	std err
log(Unsafe Following Score)	-0.72	0.46	-3.10	0.54
log(Unsafe Followers Score)	4.65	0.44	-0.46	0.53
log(Unsafe Reposts Score)	0.15	0.38	-1.79	0.69
log(Unsafe Reposted Score)	3.45	0.38	-0.41	0.85
log(pro-regime Following Score)	-2.39	0.72	3.62	0.88
log(pro-regime Followers Score)	1.02	0.61	4.22	0.72
log(pro-regime Reposts Score)	-0.32	1.70	0.25	0.26
log(pro-regime Reposted Score)	2.31	0.49	4.01	0.36
Eigen Followers Centrality	0.38	4.51	0.75	0.49
Followers to Following Ratio	0.06	109.82	0.12	1.91
Account Age	-1.56	0.55	-0.59	0.49
Is New Account	1.30	0.29	0.93	0.28
% Change Following	-1.34	0.27	-0.70	0.24
Replies Sent Rate	-0.11	0.71	-0.11	0.74
Proportion Reposts	-1.35	0.32	-0.24	0.34
Proportion Quote Posts	-2.75	0.96	-1.48	1.10
Proportion Replies	-2.36	0.36	-1.07	0.35
Followers to Following Near 1	0.40	0.47	0.38	0.22
New Account x Quote Posts Sent Rate	-1.24	0.72	-0.65	2.23
New Account x Proportion Reposts	-1.23	0.38	-0.77	0.46
New Account x Proportion Quote Posts	-0.57	1.98	-0.06	1.36
pro-regime Hashtag Score	0.86	1.08	0.65	0.66
No. Observations:	1,004			
Log-Likelihood:	-0.11			
Pseudo R-squ.:	0.94			

Notes: This table shows the multinomial logistic regression coefficients for the baseline model, regularized with an elastic net with equal L1 and L2 penalties. Factors with point estimates of zero were pushed to zero by the elastic net. The coefficients for ordinary category is normalized to zero.

Table 2: Baseline model

accounts based on their own following or reposting behavior, as they attempt to mimic the patterns of ordinary accounts by following and reposting ordinary accounts. However, they are unable to compel ordinary accounts to follow them or share their content, resulting in a higher number of similar unsafe accounts following and reposting their content, which enables us to identify them.

Furthermore, several non-network characteristics are also significant. In particular, unsafe accounts tend to be newer, which is likely caused by X's policy of shutting down accounts once they realize that an account is suspicious.[24] Additionally, unsafe accounts' engagements involve a larger proportion of original posts as opposed to reposting or replying

[24]Although there is a disinformation policy in place in X, our analysis indicates that this policy only addresses a small portion of these accounts, as the majority are still active.

Category	Precision	Sensitivity
Ordinary	92.6%	93.9%
Unsafe	93.6%	92.3%
pro-regime	99.3%	99.3%

Notes: Total accuracy: 95.13%

Table 3: Precision and sensitivity for the baseline model

to other accounts' posts, when compared to ordinary accounts.

Let us now examine the characteristics of the pro-regime accounts compared to the ordinary accounts, the second block of Table 2. The coefficients of network proximity to pro-regime accounts are all positive and highly statistically significant, confirming that these accounts tend to be very connected and echo each other's message. Additionally, the coefficients of network proximity to unsafe accounts for these accounts are all negative, and *unsafe following* and *unsafe reposts* are significant. The latter finding illustrates that pro-regime accounts try to distance themselves from unsafe accounts.

We utilize the remaining 30 percent of labeled accounts, which were excluded from the training set, as test data to measure the performance of the model. Table 3 reports the precision and sensitivity of the model classification for the three categories of accounts, as well as its total accuracy.[25,26] The total accuracy of the model is reassuringly high, 95.13%. The precision and sensitivity rates are also high for all three groups. In particular, for pro-regime accounts, both precision and sensitivity are nearly 100%, with a single false positive and a single false negative, ensuring that we can accurately identify almost all such accounts. However, differentiating unsafe and ordinary accounts is more challenging as unsafe accounts try to mimic ordinary accounts actively to avoid detection. Table 3 indicates that we misclassify between 6% and 7% of ordinary and unsafe accounts. There is a tradeoff between the precision of unsafe and ordinary accounts. Depending on the policy one wishes to implement, it may be necessary to prioritize reducing false negatives in one group over the other. We explore this tradeoff further in Section 6.3.

Next, we use the multinomial logit model to categorize all of the accounts in our sample into ordinary, unsafe, and pro-regime groups by assigning them to the category with the highest propensity score. The results are reported in Table 4. We classify about 16% of active Farsi accounts as unsafe accounts, i.e., accounts that participate in disseminating disinformation, intentionally or unintentionally. We further classify 8% of accounts as pro-

[25]For each account category, precision is calculated as $\frac{tp}{tp+fp}$, and sensitivity (also known as recall) is calculated as $\frac{tp}{tp+fn}$, where tp represents true positives, fp represents false positives, and fn represents false negatives. Total accuracy is defined as the percentage of correctly classified instances.

[26]Appendix Table B.6 reports the confusion matrix for this model.

Account Category	Count
Ordinary	1,339,454
Unsafe	285,549
Pro-Regime	142,347

Table 4: Account classification

regime accounts.

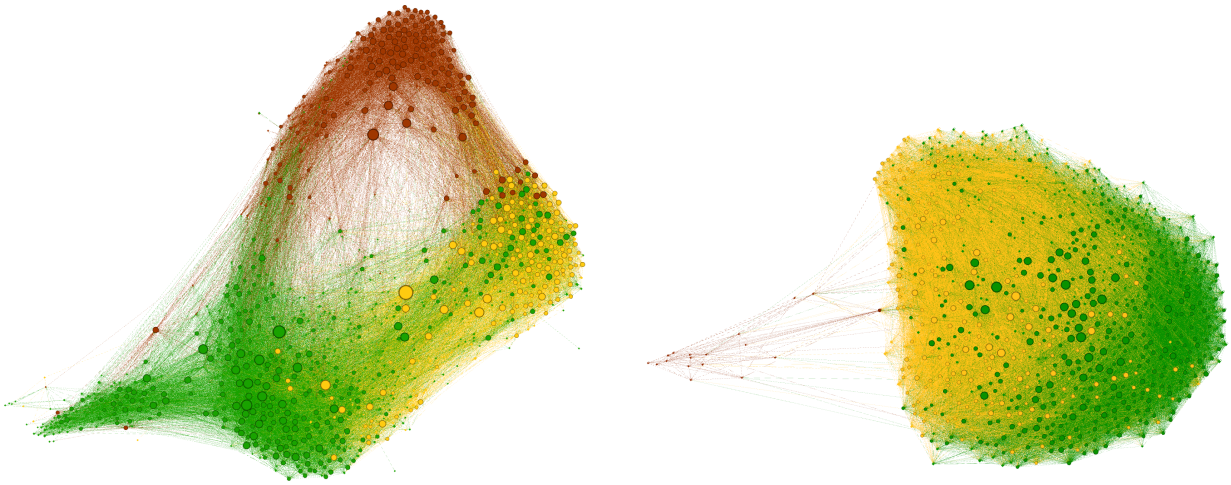
Our account classification provides interesting insights into the network structure of the social media platform. Since the number of active accounts on X is prohibitively high for meaningful visualization, we limit our analysis to the networks of the 1,000 most active accounts, displayed in Figure 2. Figure 2a illustrates the follower-following network among the 1,000 accounts with the highest number of followers, while Figure 2b depicts the tweet-reposting activity network of the 1,000 accounts with the highest number of reposts. The size of each node is proportional to the account's centrality in the network, and its color is based on the classification of accounts: unsafe accounts are shown in yellow, pro-regime accounts in red, and ordinary accounts in green.

There are a few noteworthy observations. First, as illustrated in Figure 2a, accounts within each category are highly interconnected. However, unsafe accounts have managed to infiltrate ordinary account networks. This integrated frontier is where disinformation permeates ordinary accounts. In contrast, most ordinary accounts steer clear of pro-regime accounts. Moreover, comparing the two figures, it is evident that unsafe accounts are significantly more active in posting and reposting content, as they have a greater presence in the repost network of the top 1,000 accounts compared to the most followed accounts. On the other hand, pro-regime accounts fail to get much engagement from either ordinary or unsafe accounts.

3.2 Labeling Based on Initiators' Category

In Section 3.1, we explain how NDL uses existing data, mainly from the network structure of the social media platform, to determine accounts that actively engage in the spread of disinformation already or are likely to do so in the future, even if they have not yet. In this section, we leverage this information to identify the disinformation itself as quickly as possible to contain its spread.

Our algorithm indicates two consistent patterns in supply of different pieces of news on X. First, each instance of disinformation is originated predominantly from unsafe accounts.



(a) Follower-following network of most followed accounts

(b) Repost-reposted network of most reposted accounts

Figure 2: Network of the 1,000 most active accounts

Notes: Yellow nodes represent unsafe accounts that are likely to spread disinformation, red nodes represent pro-regime accounts that are likely to spread propaganda, and green nodes represent ordinary accounts that are less likely to engage with either activity.

These accounts act in a highly concerted manner, engaging in temporally synchronized posting of nearly identical content to manufacture the illusion of grassroots consensus. This tactic mirrors the operational logic of troll farms and coordinated influence networks, where multiple actors push aligned narratives in lockstep. Second, although unsafe accounts do participate in the spread of legitimate news, ordinary accounts are also active originators of such news.

The distinctively non-organic signature of supply network of disinformation is what enables our network-based labeling approach to detect disinformation early. In particular, we use the above two observations to discern disinformation from legitimate news by identifying the composition of their initiators on social media. We posit that if m out of the first n initiators of a news item on the platform are unsafe accounts, then it is highly likely that the piece of news is disinformation, and suggest the following disinformation labeling rule:

Definition 1 (**Network-based Disinformation Labeling NDL-(n,m)**). *A piece of news is labeled as disinformation if and only if the baseline model classifies m out of its first n initiators as unsafe accounts.*

In order to choose the optimal values of (n, m) for our application, we perform the following procedure to select parameters that minimize type I and type II errors while being robust to the training sample and keep n low to ensure early detection of disinformation.

We train the baseline model using the first six ex-post verified disinformation events in our dataset (occurring between early September 2022 and mid-October 2022), and test the trained classifier on two sets of data, the remaining eight ex-post verified disinformation instances (from mid-October 2022 to end of February 2023, see Table A.1), and ten ex-post verified legitimate news events from the same period (see Table A.2). We use different pairs of parameters for testing to find (n, m) that maximize the percentage of disinformation correctly classified as disinformation and minimize the percentage of legitimate news misclassified as such, while keeping the performance of the model robust to the training sample and n low.

Table B.9 demonstrates how the success rate of labeling disinformation events varies with different (n, m) choices. We observe that as we increase the required share of initial accounts deemed unsafe to classify news as false, we risk missing some disinformation events. Alternatively, Table B.10 shows how the failure rate of labeling legitimate news events as disinformation varies with (n, m). It depicts the reverse pattern: As we increase the required share of initial accounts deemed unsafe to classify news as false, we reduce the likelihood of misclassifying legitimate news as disinformation. Comparing these two observations illustrates the main trade-off: Decreasing m relative to n leads to capturing more disinformation events, at the cost of misclassifying real news. Finally, Table B.11 reports the sensitivity of the error rates with $n = 10$ for different choices of m, as we vary the time intervals of training data.

We find that for our case, $(n, m) = (10, 7)$ provides a high detection rate without misclassifying any real news while being robust to the training sample. Figure 3 illustrates the performance of NDL in correctly detecting disinformation events as we change the length of the data used for training. The solid blue line represents the percentage of legitimate news events labeled as disinformation, which remains at zero regardless of the length of the training data. This minimal false-positive error rate ensures that our disinformation labeling approach does not impede the flow of legitimate news on social media. Conversely, the dashed red line indicates the percentage of disinformation events labeled as such. The model performs well in correctly detecting disinformation, although there is a non-zero false-negative error rate. As expected, the dashed red line increases modestly with the length of the training data period, ultimately reaching 100% correct detection of disinformation by the midpoint of our sample time interval. The figure also confirms that the model performance is robust to changes in the training sample.

The distinct advantage of NDL is that it enables swift discovery of disinformation in online platforms, through two distinct channels. First, we suggest detecting a disinforma-

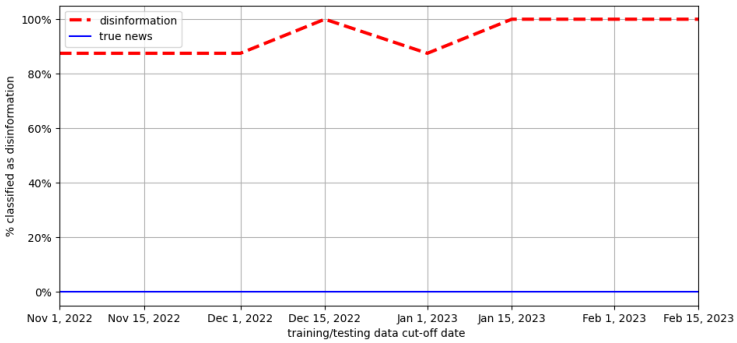


Figure 3: Labeling disinformation and legitimate news as disinformation

Notes: This figure illustrates the accuracy of labeling eight instances of disinformation (last 8 rows of Table A.1) and ten legitimate news events (Table A.2) for different lengths of training data period, as a function of duration of data used for training. If the algorithm classifies at least seven of the first 10 accounts that initiate the corresponding thread as unsafe, we label the news as disinformation.

tion event after the participation of only ten accounts. Due to the exponential spread of news on social media, this is crucial for the effective disruption of disinformation. Second, we are able to correctly detect disinformation events using account data last updated four months prior. Specifically, using account data up to the end of October 2022, which only includes the six earliest disinformation campaigns in our sample, we successfully identify disinformation events as of the end of February 2023 as illustrated in Figure 3.[27] As such, adopting this approach significantly reduces the need for the platform to engage in extensive content moderation ex-post.[28] Section A.4 in the appendix provides suggestive guidelines for implementing this approach by social media platforms.

4 NDL: Restricting the Spread of Disinformation

The US Supreme Court ruling on June 26, 2024, allows the White House and federal agencies to continue urging social media platforms to take down disinformation without violating the First Amendment. This ruling underscores the importance of prompt identification of disinformation to be able to prevent the proliferation of disinformation among social network accounts.

As such, we would like to estimate the effectiveness of NDL in mitigating the spread

[27]While expanding the training set and data can enhance the performance of the algorithm, continuous updates are not necessary to maintain its effectiveness.

[28]After facing a class action lawsuit from content moderation workers, Facebook began hiring external contractors, with Accenture being the largest, reportedly receiving over $500 million in 2021. https://www.nytimes.com/2021/08/31/technology/facebook-accenture-content-moderation.html, accessed 08/10/2024.

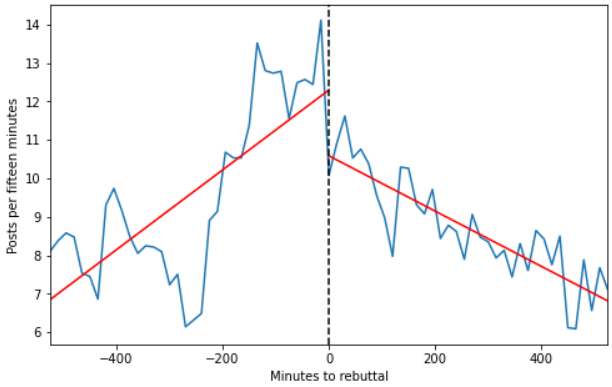


Figure 4: Effect of rebuttal on disinformation spread

Notes: This figure exhibits the average number of posts per fifteen minutes across all threads in which a rebuttal has occurred over time. Time zero is the time of the first rebuttal of the thread. The best-fit line is in red.

of disinformation by disrupting the supply chain originating from concerted disinformation efforts. To do so, we first measure the effectiveness of flagging a piece of news as disinformation on its social media spread. We then use these estimates to quantify the efficiency of NDL in restricting the spread of disinformation. We find that this approach reduces the number of posts by a factor of three, and decreases the maximum user engagement and the lifetime of a rumor by at least a factor of two.

4.1 Decentralized Rebuttals

During the Woman, Life, Freedom movement in Iran, the most common interruption to the spread of disinformation on X was through rebuttals of disinformation posts by political activists, in a fully organic, decentralized fashion. Recall that we have manually identified 1,374 such rebuttals in 10 of the 14 ex-post-verified disinformation campaigns in our sample.

Let an *original post* be one which is not a comment within another post, a repost, or a quote-repost. Let a *thread* denote a tree of posts, comments, reposts, and quote reposts whose root is an original post. Each of the disinformation campaigns we study consists of many threads of different lengths. Of all the rebuttals, 924 of them occur on different threads. When multiple rebuttals occur on a single thread, we restrict attention to the first rebuttal.

Figure 4 illustrates the average number of posts per 15 minutes across all threads where a rebuttal has occurred. Time zero represents when the first rebuttal happens, and positive (negative) numbers are the time after (before) the rebuttal. There is a clear discontinuity at time zero when the first rebuttal occurs. More importantly, rebuttals disrupt the

exponential spread of the rumor and revert the slope of the number of posts to negative.

In order to model the impact of rebuttals on the spread of disinformation, we need to make an assumption about who can see each rebuttal, as we do not have information on impressions. We assume that the rebuttals in each thread are observed only by accounts who participate in that particular thread and not the rest of the disinformation campaign.[29]

We then conduct a difference-in-differences analysis to determine the effect of rebuttals. We model the number of posts within each thread during each 15-minute time interval as a Poisson process. As the number of posts changes even in the absence of rebuttals, we assume that the process is mean-varying. We estimate the mean using various time controls, as well as other fixed effects which depend on the history, and also include disinformation campaign fixed effects.

We assume that the number of posts in thread i of a news event j at time t follows a generalized Poisson linear model, $Y_{i,j,t} \sim Poisson(\mu_{i,j,t})$, where $\mu_{i,j,t}$ represents the mean of the Poisson process. For each thread i of the news event j, we denote the time of the first rebuttal as $\tau_{i,j}$. We estimate β_1, the effect of a rebuttal that takes place at time $\tau_{i,j}$, on $Y_{i,j,t}$, the number of posts at time $t \geq \tau_{i,j}$, using the following regression discontinuity in time model:

$$\mu_{i,j,t} = \alpha + \beta_1 \mathbb{1}_{t \geq \tau_{i,j}} + \beta_2 Y_{i,j,t-1} + \beta_3 Y_{i,j,t-1} \times \mathbb{1}_{t \geq \tau_{i,j}} + \text{Fixed-Effects}_{i,j,t}. \tag{1}$$

We include three sets of fixed effects. The first set is disinformation-specific fixed effects, which controls for the varying levels of engagement with different disinformation campaigns. The second set involves time-specific fixed effects, such as the day of the week and the hour of the day. The third set pertains to the rumor's history, such as the count of different types of posts in the campaign up to time t.

Table 5 presents the regression results from Equation (1) with various fixed effects. All coefficients across specifications are significant at the 1% level. We found that previous posts increasingly drive new posting activity, while the passing of time gradually dampens engagement.

When a rebuttal occurs, we observe two notable effects. First, there's an immediate drop in posting activity, the number of posts per 15 minutes drops by 2, from an average of 13 to

[29]This assumption has a couple of implications. First, it implies that threads are independent of one another. As such, we ignore the spillover that rebuttals have on the rest of the campaign and can lead to earlier termination of the spread of false news. On the other hand, it implies that when a rebuttal appears, it is seen by all participants of that thread going forward. Neither of these simplifying assumptions is precise. However, we believe that the aggregate effect of this simplifying assumption likely underestimates the impact of rebuttals.

constant	4.78	4.69	1.84	5.28	2.69
	(2.0E-2)	(8.0E-3)	(0.04)	(3.0E-3)	(0.07)
rebuttal	-2.08	-1.98	-1.97	-2.35	-2.15
	(0.03)	(0.03)	(0.03)	(0.03)	(0.03)
lagged posts	7.0E-4	8.0E-4	9.0E-4	8.0E-4	9.0E-4
	(5.5E-7)	(8.3E-7)	(1.1E-6)	(1.2E-6)	(1.9E-6)
lagged posts × rebuttal	0.01	0.01	6.7E-3	6.7E-3	6.6E-3
	(7.9E-5)	(8.15E-5)	(8.9E-5)	(8.6E-5)	(8.9E-5)
time elapsed	-1.1E-7	-1.0E-7	-3.1E-7	-1.7E-7	-2.7E-7
	(5.2E-10)	(5.2E-10)	(1.7E-9)	(6.3E-10)	(1.9E-9)
time elapsed × rebuttal	-2.4E-6	-2.7E-6	-2.6E-6	-6.8E-7	-1.3 E-6
	(1.1E-7)	(1.2E-7)	(1.3E-7)	(9.8E-8)	(1.1E-7)
time fixed-effects	X	✓	X	X	✓
history fixed-effects	X	X	✓	X	✓
campaign fixed-effects	X	X	X	✓	✓
pseudo R-squared (McFadden)	0.84	0.89	0.88	0.86	0.91
N	4,851	4,851	4,851	4,851	4,851

Notes: In this regression, rebuttal is a dummy variable equal to 1 if the time period is after a rebuttal occurred.

Table 5: Results of regression discontinuity in time

11 posts.[30] More interestingly, rebuttals transform engagement patterns by altering both the impact of previous posts and time effects. Rebuttals amplify the influence of previous posts, likely due to the short-term discussions they generate. At the same time, they strengthen the negative effect of elapsed time, reversing posting trends altogether. These findings align with recent studies by Henry et al. (2022) and Martel and Rand (2023), which demonstrate how warning labels can significantly reduce the spread of social media content. Another recent paper, Guriev et al. (2023), shows that fast algorithmic information labeling can be one of the most effective tools in curtailing disinformation on social media.

For our subsequent simulation and prediction analyses, we use the final column—which includes all fixed effects—as our reference specification.

4.2 How Much Does NDL Restrict Disinformation Spread?

In this section, we use the estimated impact of decentralized rebuttals —a 2.15 decline in the average number of posts in each 15-minute interval as well as the change in trends—to estimate the potential impact of NDL if implemented platform-wide by social media com-

[30]This is the average right before the occurrence of rebuttal for threats. The mean number of posts per 15 minutes of all debunked threats is 11 and the median is 9.

	Realized	Baseline Simulation	NDL Simulation
Total # of Posts	9,668	11,204	3,238
Maximum Rate of Posts	629	552	342
Time to $<$ 42 Posts	168 hours	104 hours	26 hours
Time to $<$ 28 Posts	170 hours	125 hours	55 hours

Notes: This table reports average measures of spread of disinformation on platform X for the 10 realized disinformation events that have been rebutted, as well as their corresponding baseline and NDL simulations. We report three types of diffusion statistics: 1) number of posts in a disinformation event, 2) the maximum number of posts per hour, 3) the time until it falls and stays below 42 and 28 posts per hour (these threshold are the average number of posts at the 90th and 95th percentile across all the disinformation events). The three columns are: values from the data, the corresponding simulations using Equation (1), and a counterfactual where the disinformation was labeled by NDL using the estimates for the impact of rebuttal reported in Table 5.

Table 6: NDL: Restricting the spread of disinformation

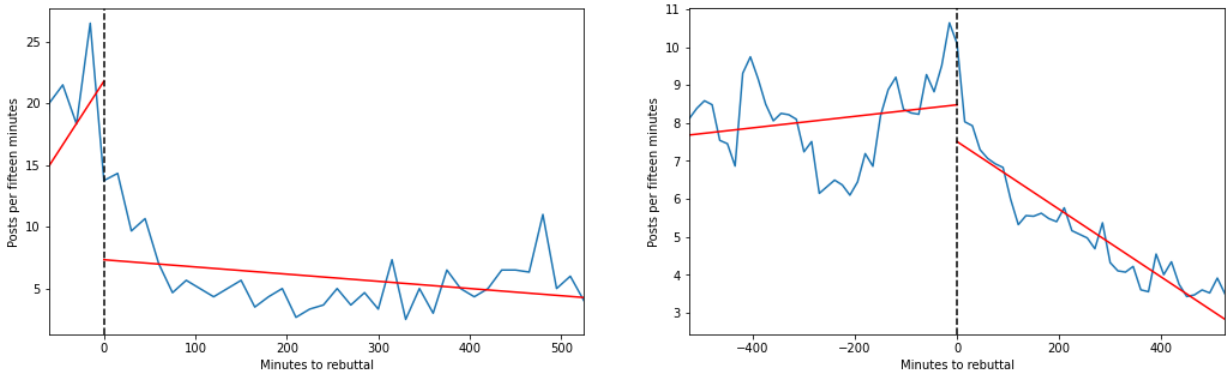
panies. This analysis helps us predict how centralized adoption might affect the spread of disinformation.

We perform two distinct sets of simulations for each thread of each disinformation event, according to Equation (1). The first set, the "baseline" simulations, follows the history of realized disinformation events by setting the rebuttal indicator of a rebutted thread to the time its corresponding thread was rebutted.[31] As such, these simulations replicate the average statistics of the realized disinformation events.[32]

The second set, the NDL simulations, measures the impact of NDL on the spread of disinformation, if it is adopted by the social media platform in a decentralized fashion and is used to label news events as disinformation early in their lifespan. We simulate each disinformation event thread-by-thread, assuming that the number of posts in a thread at time t follows a Poisson distribution with mean given by Equation (1), using estimates reported in the final column of Table 5. Each thread begins at the time indicated in the data but evolves based on the corresponding simulation. For the "baseline" simulations, we use the time of the first rebuttal within each thread in the data as $\tau_{i,j}$. For the NDL simulations, we assume that all existing threads of the disinformation event are rebutted simultaneously at the time of the 10^{th} original post of the news event in the data, which is when our approach labels a piece of disinformation as such. The simulation method is explained in detail in Appendix A.5.

[31] Some of the disinformation campaigns never experience a rebuttal.

[32] This approach ensures a more robust comparison set, as we can verify that results are not driven by differences between the simulations and the actual data.



(a) Rebuttal at most 1 hour after the first tweet of the thread

(b) Rebuttal at least 4 hours after the first tweet of the thread

Figure 5: Effect of early versus late rebuttals on disinformation spread

Notes: This figure exhibits the average number of posts per fifteen minutes across threads in which a rebuttal has occurred over time, grouped by time of the first rebuttal relative to start of the thread, early and late. Time-zero is the time of the first rebuttal of the thread. The best fit line is in red.

We measure the extent of spread of disinformation using three statistics. The number of posts in the disinformation campaign, the maximum user engagement rate, and the duration of the disinformation campaign. Table 6 reports these statistics, averaged over the disinformation events in our sample, for the realized events and the two sets of simulations.

The number of posts related to the disinformation campaign decreases significantly, by a factor of three. The maximum user engagement rate, measured as the maximum number of posts per hour on a given campaign, decreases to approximately half. Finally, the effective lifespan of disinformation has notably shortened, by half or a quarter, depending on the measure used. To estimate the effective lifetime of a disinformation event, we calculate the number of posts per hour at the 90th and 95th percentiles of all instances of disinformation. We then determine the duration required for a disinformation event to reach at most that many posts per hour using both the actual data and the two sets of simulations.

While the estimates reported in Table 6 already indicate that NDL significantly reduces engagement with disinformation, they are likely a lower bound on its effectiveness in containing the spread of disinformation. That is because a key advantage of NDL is to promptly label disinformation before it spreads widely and becomes viral, while organic rebuttals may occur at various stages throughout the spread of false news, either early or late. Due to the exponential nature of news dissemination on social platforms, early detection is crucial for curtailing the spread of disinformation.

To explore the differential impact of early versus late disinformation detection on its spread, we consider the impact of rebuttals on the number of posts in the disinformation

event as a function of timing of the rebuttal relative to the original post that started the corresponding thread. We examine two subsets of rebuttals: early, corresponding to threads that are rebutted within the first hour of the original post, versus late, those rebutted after four hours. Figure 5 illustrates our findings. Comparing the early rebuttal panel 5a with the late rebuttal panel Figure 5b exhibits a substantial impact for early rebuttals. As expected, the average impact of rebuttals shown in Figure 4 falls between these two extremes.[33]

A potential concern is that disinformation spreaders might alter their behavior to evade detection if a platform adopts NDL. The estimates presented in Table 6 are based on two assumptions. First, to successfully spread disinformation, unsafe accounts must maintain high connectivity within the social network while avoiding detection by ordinary users. This necessitates not only strong connections among themselves but also strategic positioning within the broader network through interactions with ordinary accounts. As depicted in Figure 2, this involves engaging in the spread of legitimate news and establishing follower and following relationships with ordinary accounts. Given these requirements, the network-based characteristics are not susceptible to rapid manipulation. Section 5.2 demonstrates that a non-manipulable classifier that relies solely on network-based characteristics performs nearly as well as the baseline model in labeling disinformation. Second, unsafe accounts behave in a coordinated manner to initiate disinformation, while they behave independently in spreading legitimate news. With these two assumptions, the estimates in Table 6 provide reasonable approximations of the approach's impact in the short term, even when considering potential manipulative behaviors by disinformation spreaders.

So far, we have focused on the benefits of adopting NDL in restricting the spread of disinformation on social media. At the same time, we believe this approach can potentially reduce the cost of detecting disinformation for platforms substantially through process automation and reduction of manual content moderation labor costs. NDL involves two labor-intensive steps: first, verifying a number of disinformation events using news content to initialize the process, and second, creating a labeled dataset for training and testing the algorithm. The rest of the algorithm can be fully automated. While these steps need to be repeated intermittently to ensure the labeling algorithm's performance, the rest of the algorithm needs minimal if any updates. The automation of the majority of the process is likely to provide a significant cost-saving opportunity for social media platforms.

[33]We have decided to use all rebuttals in the estimates reported in Table 6 to avoid small sample concerns.

5 Proximity Scores and Manipulation Resistance

In this section, we highlight the importance of network-based characteristics in two dimensions. First, we show that they are critical to correctly classify accounts, and subsequently, disinformation events and legitimate news. In particular, we show that dropping the network characteristics from the classifier leads to a significant decline in its performance.

Second, we illustrate that a model that only relies on network-based characteristics maintains a high classification accuracy while being less susceptible to manipulation by disinformants.

5.1 Traditional Disinformation Labeling

Although traditional approaches to detecting disinformation might employ different network centrality measures of accounts, they do not use the proximity scores that we have introduced in Section 2.3. To test the performance of the traditional approach, we revise our baseline algorithm by removing the 8 proximity scores, while keeping the common network centrality measures. The resulting algorithm is presented in Appendix 5.2.[34]

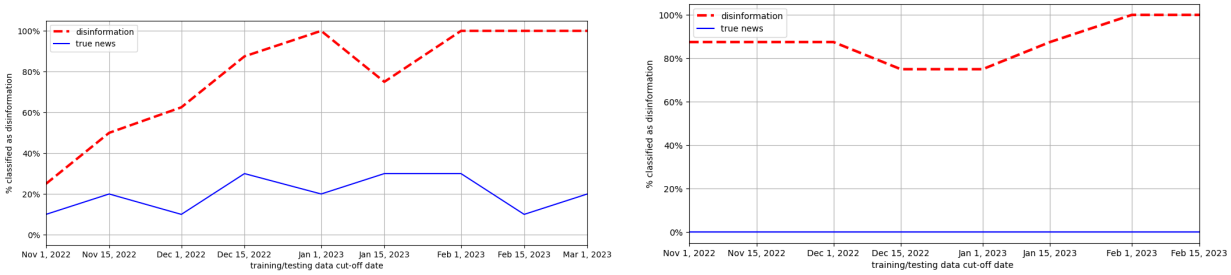
Removing the network proximity scores results in a sharp decline in accuracy to 74.7%. This decline is primarily due to the algorithm's inability to successfully identify ordinary and unsafe accounts. Interestingly, including the network centrality measures is not helpful in distinguishing unsafe accounts from ordinary accounts, in fact, their coefficients remain insignificant, even those that survive the elastic net filtering process.

This 20% drop in accuracy has dire consequences for identifying disinformation campaigns. When testing with data up to three to four months prior to a campaign's launch, the algorithm correctly identifies only 25-50% of disinformation campaigns, while misclassifying roughly 20% of legitimate news as disinformation. The poor performance of the model in correctly classifying disinformation campaigns, even with network centrality measures, reinforces the importance of *proximity scores* in disinformation labeling.

5.2 Manipulating Account Characteristics

To successfully spread disinformation, disinformants, especially concerted agents, often behave as ordinary accounts, making it appear that their posts are legitimate news. Put differently, unsafe accounts mimic ordinary accounts to blend in, which in turn makes their

[34]Using NLP algorithms to process the content of messages can be useful for both network-based and traditional disinformation labeling. However, we have decided to abstract away from text processing as content verification is time consuming step and our focus is on early detection of disinformation campaigns.



(a) Model without proximity scores

(b) Model with only network variables

Figure 6: Labeling disinformation and legitimate news as disinformation

Notes: This figure illustrates the accuracy of labeling eight instances of disinformation and ten legitimate news events for different sizes of training data, as a function of the duration of data used for training. If the algorithm classifies at least seven of the first 10 accounts that initiate the corresponding thread as unsafe using only network variables, we label the news as disinformation.

detection challenging. Identifying disinformation events based on the identity of their initiator accounts can further encourage users to alter their behavior to manipulate the scores and avoid punitive measures, which would in turn decrease the model's accuracy.

In order to counter possible manipulation from unsafe accounts, we restrict attention to characteristics that are hard to manipulate by unsafe accounts. We posit that network characteristics are particularly resistant to manipulation as they depend on the entire network structure, including ordinary accounts. As such, this makes coordinated manipulation by unsafe accounts challenging. This is especially relevant for *proximity scores*. For example, while unsafe accounts can follow and share posts from ordinary accounts, they cannot compel ordinary accounts to reciprocate. Additionally, network metrics for each account are influenced by accounts several connections away, further reducing manipulation potential.

To address this concern, we revise the baseline model of Section 3.1 by using exclusively network characteristics. Tables B.16 and B.17 present the updated estimation results.[35] Table B.16 illustrates that the estimated coefficients are generally aligned with the more comprehensive baseline model. Restricting to network characteristics leads to a small reduction in the model's accuracy, from 95.1% to 93.7%, as reported in Table B.17.[36]

Despite this slight decrease in accuracy when using only the network characteristics that are less susceptible to manipulation, the impact on identifying disinformation campaigns remains minimal, as illustrated in Figure 6b. There is a slight reduction in the probability of identifying disinformation campaigns, while still no instances of legitimate news are

[35]There are other hard-to-manipulate variables from the baseline model. To err on the side of caution, we opted to only use the network-based characteristics.

[36]These results are in contrast to ones in the previous section, where removing network scores caused the accuracy to drop to 74.7%.

misidentified as disinformation.

The key feature of our proposed intervention to fend off manipulation is to flag disinformation news rather than individual accounts. This approach aligns with theoretical insights from studies on optimal information signals in the presence of manipulation, as discussed in Frankel and Kartik (2019); Perez-Richet and Skreta (2022); Saeedi and Shourideh (2023). These studies recommend employing opaque scoring schemes to hinder manipulation attempts. Our labeling mechanism creates significant obstacles for unsafe accounts' attempt to evade detection. Firstly, they are unaware of which accounts under their control have been flagged as unsafe. Secondly, they lack precise information on the reasons for the flagging, making it harder to evade.

A potential longer-term strategy to spread disinformation could involve creating entirely new accounts and building new relationships. However, positioning these new accounts within the social network of ordinary users is a time-consuming process. Furthermore, to evade detection based on network characteristics, these accounts would need to establish connections within clusters of other new unsafe accounts while avoiding links to previously identified unsafe accounts. To counter these methods, we could add new variables to our algorithm, such as the percentage of an account's followers that are new, or the proportion of its first ten followers created within k-days of the account's own creation. Although implementing such enhancements creates the need to update the algorithm over time, they could significantly mitigate potential long-term manipulation efforts by disinformation spreaders.

6 Robustness

In this section, we conduct several exercises to condfirm the robustness of the model and its external validity.

6.1 Value of Training Data

Our model relies on a classifier for detecting disinformation, where the training data plays a crucial role. There are two dimensions through which the training data can be improved: 1) increasing the number of disinformation campaigns included in the training data, hence enlarging the training set, and 2) extending the duration of data used for training while keeping the training set constant. The former dimension allows the model to identify unsafe accounts based on their network interactions with a broader set of verified unsafe accounts, while the latter dimension implies that the model considers longer-term interactions of the same set of accounts.

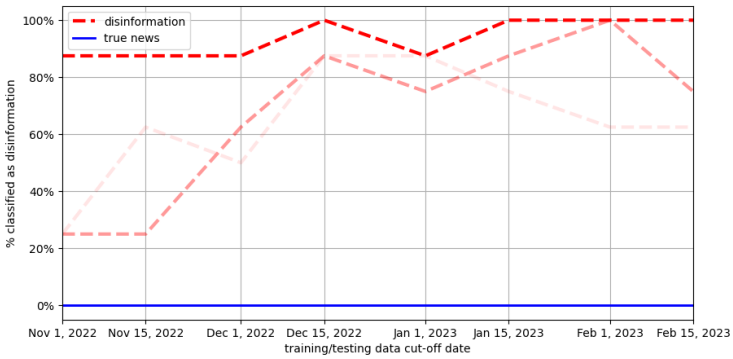


Figure 7: Labeling disinformation and legitimate news as disinformation: Training data

Notes: This figure illustrates the accuracy of labeling eight instances of disinformation and ten legitimate news events for different sizes of training data, as a function of the end date of data used for training. If the algorithm classifies at least seven of the first 10 users who initiate the corresponding thread as unsafe, we label the news as disinformation. The blue (red) lines correspond to disinformation (real news). Darker lines represent classifiers with a larger training set, measured by the number of disinformation campaigns used for training (1, 3 or 6).

We have already shown that when the training dataset includes six disinformation campaigns, the model accuracy starts high even with a limited duration of the training data. As such, although increasing the length of the data used for training results in higher accuracy, the improvement is marginal, as illustrated in Figure 3.

Figure 7 explores both dimensions of improvement in training data in a more general way. In particular, it compares the accuracy of the baseline model trained with 1, 3, or 6 disinformation campaigns, while the end of the training period varies from mid-September 2022 to March 2023.[37] Similar to Figure 3, blue curves correspond to the probability that a legitimate news is mistakenly classified as disinformation, while the red curves are the probability of correct detection of disinformation.

A significant finding in Figure 7 is that utilizing a single disinformation campaign and data up to mid-September 2022 to train the model is sufficient to discern that all legitimate news is not disinformation. However, increasing the number of disinformation events in the training sample significantly improves the accuracy of detecting disinformation, particularly with a limited duration of training data.

As such, while increasing the duration of training data enhances the algorithm's accuracy, the impact of enlarging the training set is notably more pronounced. We believe that this is due to the fact that unsafe accounts put a lot of effort to act as ordinary accounts and keep distance from some other unsafe accounts, to avoid all being detected. Thus, by

[37]With fewer disinformation events in the training set, one can consider an earlier final date for availability of training data.

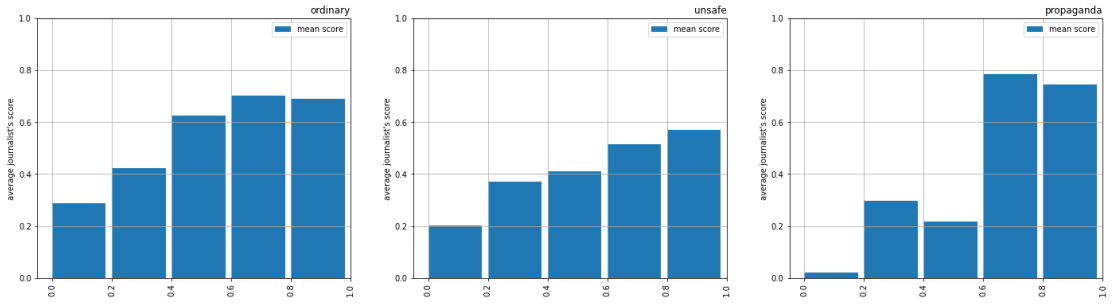


Figure 8: Journalists' average propensity score vs. algorithm's propensity score
Notes: Each figure illustrates the propensity score assigned to accounts, averaged over the three journalists, for quantiles of the algorithm-generated propensity scores, for one category of accounts.

including sufficiently many unsafe accounts in the training set, the model is able to identify new clusters of unsafe accounts who have not participated in the first few disinformation campaigns and thus identify new disinformation more successfully.

6.2 Expert Validation

To validate the results of our model, we conducted an external verification by hiring three independent journalists who are active on X. We provided them with a random list of accounts, including both accounts from our labeled set and other accounts identified solely based on our algorithm. The journalists were not informed which accounts belonged to the labeled set. We asked them to assign probabilities to each account, indicating the likelihood that an account falls into one of three categories: ordinary, unsafe, or pro-regime. For example, for a specific account, a journalist might assess an 80% probability of being ordinary, a 20% probability of being unsafe, and a 0% probability of being pro-regime. They could also assign a 100% probability to a single category if they were certain.

We first consider the propensity scores provided by the journalists for the set of labeled accounts they had received.[38] Several notable observations emerge. First, the journalists often expressed uncertainty in their classifications; in 65% of the cases, they did not assign an account to a category with 100% probability. Second, the scores assigned by different journalists varied significantly, especially for ordinary and unsafe categories.[39] Third, NDL demonstrated higher accuracy than the most accurate journalist, achieving 93% accuracy compared to 87. These observations underscore the value of incorporating a model capable of tracing network relationships which can achieve a more efficient and accurate outcome

[38]These accounts are ones for which we are certain of their category.
[39]The standard deviations for ordinary and unsafe scores provided by journalists were 0.26 and 0.24 respectively, while the standard deviation for pro-regime scores was 0.03.

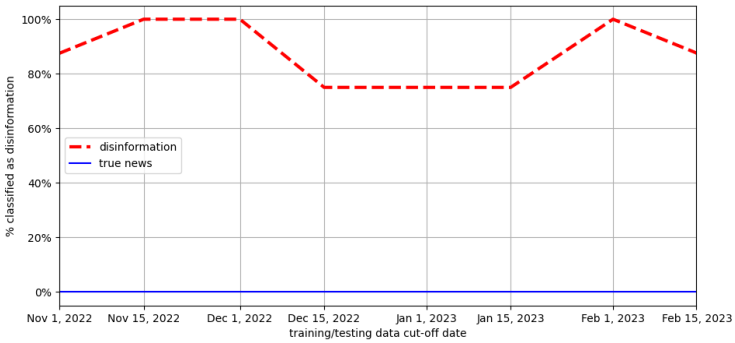


Figure 9: Labeling disinformation and legitimate news as disinformation: Estimation using ordinary accounts provided by independent journalists

Notes: This figure illustrates the accuracy of labeling eight instances of disinformation and ten legitimate news events for different sizes of training data and training using only ordinary accounts provided to us by independent journalists. If the algorithm classifies at least seven of the first 10 accounts who initiate the corresponding thread as unsafe using only network variables, we label the news as disinformation.

than that of experts.

Next, we compare the journalists' propensity scores with those of our algorithm for accounts not included in the labeled set. We calculate a weighted average score for each account by assigning a weight to each journalist based on their accuracy rate for labeled accounts.[40] Figure 8 compares their weighted propensity scores to the algorithm's propensity scores. This figure presents the average scores for all accounts within each quantile for the algorithm propensity scores. A positive correlation is observed for both ordinary and unsafe accounts, as shown in the first two panels. Accounts that our algorithm identified as highly likely to be ordinary (or unsafe) also received higher probabilities of being ordinary (or unsafe) from the journalists. Additionally, the journalists' predictions for pro-regime accounts closely align with that of the algorithm's, which is expected given that we defined pro-regime accounts as those overtly supporting the regime, which are thus easier to identify.

Finally, Figure 9 depicts the accuracy of the baseline model in categorizing disinformation and legitimate news using the ordinary account provided by the independent journalists to estimate the logistic regression. The figure clearly illustrates that the performance of the model is robust to the training data.

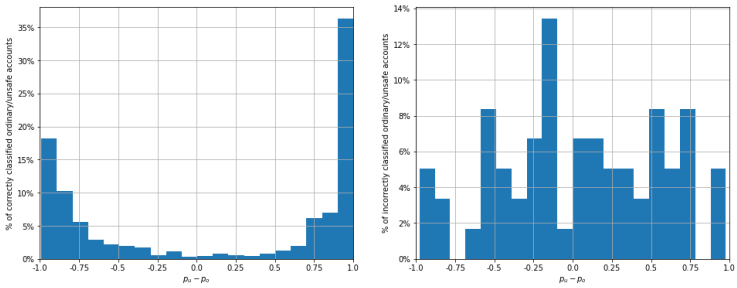


Figure 10: Histogram of differences between ordinary and unsafe propensity score

Notes: The left histogram shows the distribution of $p_u - p_o$ for accounts that are correctly labeled by the baseline model as ordinary or unsafe. The right histogram shows the distribution for accounts that are misclassified.

6.3 Precision and Sensitivity Tradeoff

In this section, we consider the precision and sensitivity of the model in labeling accounts, which is the crucial input to identifying disinformation campaigns. Similarly to any other classifier, our model exhibits a tradeoff between precision and sensitivity of labeling. However, in certain circumstances, one might be specifically concerned about the performance of the model for one group of accounts. For instance, one might seek to maximize the number of unsafe accounts that are correctly labeled as unsafe, or alternatively, one might aim to guarantee that every ordinary account is correctly labeled as ordinary. Either of these goals can be achieved through a slight adjustment of the algorithm, which we describe next.

Given the high precision and sensitivity in classifying pro-regime accounts, we exclude them from the subsequent analysis. We will continue to classify accounts with the highest pro-regime score into the pro-regime group as before. Our focus will be on accounts identified by our algorithm as either unsafe or ordinary. For each account, we calculate $p_u - p_o$, which represents the difference between their unsafe and ordinary propensity scores. Figure 10 illustrates the distribution of $p_u - p_o$ for the unsafe and ordinary test accounts labeled by the model. The left panel depicts the distribution of correctly labeled account,s while the right panel depicts that of the incorrectly labeled accounts. The two panels clearly show that for the accounts that are incorrectly labeled by the model, the propensity scores to be an unsafe versus ordinary are a lot more similar.

In order to control the precision and sensitivity of account classification directly, we introduce a "propensity score threshold," ϵ, and consider its comparison with $p_u - p_o$. By changing the propensity score threshold one can achieve a specific precision or sensitivity. If

[40]The weight for journalist i is $\text{accuracy}_i/(\sum_j \text{accuracy}_j)$.

	Baseline	0.05	0.1	0.15	0.2	0.25	0.3
Ordinary Precision	92.6%	92.6%	92.0%	91.4%	90.8%	90.8%	89.7%
Ordinary Sensitivity	93.9%	93.9%	93.9%	93.9%	93.9%	93.9%	95.2%
Unsafe Precision	93.6%	93.6%	93.6%	93.6%	93.6%	93.6%	94.8%
Unsafe Sensitivity	92.3%	92.3%	91.6%	90.9%	90.2%	90.2%	90.2%

Notes: Effect of reclassifying the accounts classified as unsafe as ordinary by the baseline model using $p_u - p_o < \epsilon$ instead of $p_u - p_o > 0$, where $\epsilon = 0.05, 0.1, 0.15, 0.2, 0.25, 0.3$

Table 7: Precision and sensitivity for the baseline model using the $p_u - p_o < \epsilon$ criteria

$p_u - p_o < \epsilon$, we reclassify this account as ordinary. By increasing ϵ, we label fewer accounts as unsafe. In other words, relabeling accounts using a higher propensity score threshold increases sensitivity for ordinary accounts and precision for unsafe accounts by as it is more stringent about calling an account unsafe. Note that so far we identify an account as the category with the highest propensity score, i.e., $\epsilon = 0$.

Table 7 reports the precision and sensitivity of labeling ordinary and unsafe accounts as we change the threshold ϵ form 0.05 to 0.3. Consistent with the contrast between the two histograms in Figure 10, by increasing ϵ we reduce how often ordinary accounts are misclassified as unsafe, albeit at the expense of increasing the misclassification of unsafe accounts as ordinary.

7 Conclusion

In this paper, we analyze a novel method to curb consumption of disinformation on social media, particularly those driven by troll farms and other concerted actors —the network-based disinformation labeling approach, NDL. Our methodology relies on identifying disinformation events on social media using the characteristics of their first few initiators to contain the supply network of disinformation before it goes viral.

We take advantage of data from X during the recent wave of social unrest in Iran, the "Woman, Life, Freedom" movement, to estimate the performance of our proposed approach in identifying disinformation campaigns on X during this episode. We find that using data only up to four months ahead of the disinformation date, our model is able to identify at least 85% of ex-post verified disinformation instances without misclassifying any legitimate news as disinformation.

To quantify the impact of the NDL approach on containing the supply of disinformation, we first causally estimate the effect of organic, decentralized rebuttals of disinformation on X, during the same episode. We then employ these estimates to measure a lower bound on

the impact of the centralized implementation of our labeling approach by the platform and find that it leads to a three-fold reduction in the number of posts and reduces the maximum user engagement rate and the effective lifespan of disinformation by at least a factor of two.

These results suggest that, unlike live moderation and fact checking or ex-post debunking, a network-based detection approach that targets concerted efforts can significantly mitigate the spread of disinformation on social media. We believe that the substantial impact of this method makes it worth considering for a wide range of scenarios where the spread of disinformation is a problem, thereby bolstering our ability to counteract the negative impact of disinformation proliferation.

References

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

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

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

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

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

Barrera, O., Guriev, S., Henry, E., and Zhuravskaya, E. (2020). Facts, alternative facts, and fact checking in times of post-truth politics. *Journal of public economics*, 182:104123.

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

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

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

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

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

Campante, F., Durante, R., and Sobbrio, F. (2018). Politics 2.0: The multifaceted effect of broadband internet on political participation. *Journal of the European Economic Association*, 16(4):1094–1136.

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

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

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

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

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

Denter, P. and Ginzburg, B. (2024). Troll farms. *arXiv preprint arXiv:2411.03241*.

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

Falck, O., Gold, R., and Heblich, S. (2014). E-lections: Voting behavior and the internet. *American Economic Review*, 104(7):2238–2265.

Frankel, A. and Kartik, N. (2019). Muddled information. *Journal of Political Economy*, 127(4):1739–1776.

Gavazza, A., Nardotto, M., and Valletti, T. (2019). Internet and politics: Evidence from uk local elections and local government policies. *The Review of Economic Studies*, 86(5):2092–2135.

Grinberg, N., Joseph, K., Friedland, L., Swire-Thompson, B., and Lazer, D. (2019). Fake news on twitter during the 2016 us presidential election. *Science*, 363(6425):374–378.

Guess, A., Nagler, J., and Tucker, J. (2019). Less than you think: Prevalence and predictors of fake news dissemination on facebook. *Science advances*, 5(1):eaau4586.

Guriev, S., Henry, E., Marquis, T., and Zhuravskaya, E. (2023). Curtailing false news, amplifying truth. *Working Paper*.

Guriev, S., Melnikov, N., and Zhuravskaya, E. (2021). 3g internet and confidence in government. *The Quarterly Journal of Economics*, 136(4):2533–2613.

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

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

Henry, E., Zhuravskaya, E., and Guriev, S. (2022). Checking and sharing alt-facts. *American Economic Journal: Economic Policy*, 14(3):55–86.

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

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

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

Martel, C. and Rand, D. G. (2023). Misinformation warning labels are widely effective: A review of warning effects and their moderating features. *Current Opinion in Psychology*, page 101710.

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

Perez-Richet, E. and Skreta, V. (2022). Test design under falsification. *Econometrica*, 90(3):1109–1142.

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

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

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

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

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

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

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

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

Wang, C.-Y., Chen, H.-N., and Öry, A. (2024). A disclosure game with (non-)experts and the role of information-sharing incentives. Unpublished manuscript.

Zhuravskaya, E., Petrova, M., and Enikolopov, R. (2020). Political effects of the internet and social media. *Annual review of economics*, 12(1):415–438.

Appendix

A X Data: Supplemental Information

A.1 Ex-post Verified Disinformation and Legitimate News

Tables A.1 and A.2 outline the 14 pieces of disinformation and the 10 legitimate news, all ex-post verified, used in our analysis.

rumor	start date	description
Health of Khamenei	September 1, 2022	Prior to and through the start of the Women, Life, Freedom protests, rumors of Khamenei's declining health and death began being circulated.
Account balance of Reza Ostadi	September 10, 2022	Alleged screenshots of Reza Ostadi's bank account balances began circulating showing an exhorbitant balance.
Komle's arrival in Kurdistan	September 19, 2022	Word began to spread which indicated that separatist forces were beginning to amass in Kurdistan in historic opposition to the IRI.
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.
Murder of Asra Panahi	October 12, 2022	Asra Panahi died after being hospitalized following a clash with security forces at her school. The role of security forces in her death was quickly called into question.
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.
Murder of Hana Duzduzani	October 14, 2022	One of several false names reported to have died in conjunction with the real death of Asra Panahi, whose death was falsely reported to have been by suicide.
Assault of Armita Abbasi	October 23, 2022	Rumor began to circulate that the protestor Armita Abbasi had been hospitalized due to multiple sexual assault following her arrest.
Venezuelan political refuge	November 1, 2022	Rumor began to spread of high-level officials seeking political asylum in Venezuela over pressure from the protest movement.
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.
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.
Death of Judge Salavati	January 5, 2023	False reports of Judge Salavati's death began to circulate at the beginning of the year.
Assault of Sara Shirazi	February 21, 2023	An azreshi woman was reported to have assaulted a school girl in Isfahan for improperly wearing her religious garb.
Murder of Fatemeh Rezaee	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.

Table A.1: List of the Collected ex-Post Verified Disinformation

news event	start date	description
Arrest of Majid Tavakoli	September 22, 2022	The political activist Majid Tavakoli was arrested in the early days of the Women, Life, Freedom protest movement.
Execution of Mohsen Shekari	December 8, 2022	Mohsen Shekari was executed by hanging having been convicted of the assault of a paramilitary militia member.
Death of Rostam Ghasemi	December 8, 2022	Rostam Ghasem, Iran's former Minister of Urban Development, died as a result of chronic illness.
Release of Majid Tavakoli	December 19, 2022	The political activist Majid Tavakoli was released after having been detained for three months following his participation in the early Woman, Life, Freedom protest movement.
Torture and Death of Mahdi Zare	January 2, 2023	The protestor Mahdi Zare fell into a coma and died following his release from custody. His death is thought to have resulted from torture while he was in custody.
Arrest of Navab Ebrahimi	January 5, 2023	The social media influencer Navab Ebrahimi was arrested without pretense but is thought to be related to the taunting of general Qassem Suleimani.
Trial of Yalda Moeeri	January 6, 2023	The photojournalist Yalda Moeeri underwent sentencing for spreading anti-IRI propaganda.
Ahmad-Reza Radan Appointment	January 7, 2023	The general Ahmad-Reza Radan was appointed police chief likely in response to the ongoing protests.
Execution of Mohammad Mahdi Karami	January 7, 2023	Mohammad Mahdi Karami was executed by hanging having been convicted for the murder of a paramilitary militia member.
Black Reward Hack of Imam Sadegh University	January 20, 2023	A group of hackers hacked into the network of Imam Sadeq University and threatened to release sensitive documents and information.

Table A.2: List of Collected ex-Post Verified Legitimate News Events

A.2 Account Characteristics

account characteristic	description
log(Unsafe Following Measure)	log of unsafe following measure (how ingrained an account's followings are with imposter accounts)
log(Unsafe Followers Measure)	log of unsafe followers measure (how ingrained an account's followers are to imposter accounts)
log(Unsafe Retweets Measure)	log of unsafe retweets measure (how ingrained an account's retweets are to posts made by imposter accounts)
log(Unsafe Retweeted Measure)	log of unsafe retweeted measure (how ingrained an account's retweeted posts are to those retweeted by imposter accounts)
log(pro-regime Following Measure)	log of pro-regime following measure (how ingrained an account's followings are with pro-regime accounts)
log(pro-regime Followers Measure)	log of pro-regime followers measure (how ingrained an account's followers are to pro-regime accounts)
log(pro-regime Retweets Measure)	log of pro-regime retweets measure (how ingrained an account's retweets are to posts made by pro-regime accounts)
log(pro-regime Retweeted Measure)	log of pro-regime retweeted measure (how ingrained an account's retweeted posts are to those retweeted by pro-regime accounts)
Degree Followers Centrality	account's degree centrality in follower-following network (number of edges to or from an account)
Eigen Followers Centrality	account's eigenvector centrality in follower-following network (centrality weighted by an account's connection to highly-centered nodes)
Betweenness Followers Centrality	account's betweenness centrality in follower-following network (centrality measured by an account's presence in the shortest paths between all other accounts)
Followers to Following Ratio	ratio of an accounts 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 an account's number of followers since the start of the protests
% Change Following	percent change in an account's number of followings since the start of the protests
New Followers Rate	the rate at which an account has gained followers since the start of the protests
New Following Rate	the rate at which an account has followed accounts since the start of the protests
New Retweets Sent	the number of retweets an account has sent since the start of the protests
New Retweets Received	the number of time an account has been retweeted since the start of the protests
New Quote Tweets Sent	the number of quote tweets an account has sent since the start of the protests
New Quote Tweets Received	the number of time an account has been quote tweeted since the start of the protests
New Replies Tweets Sent	the number of replies an account has sent since the start of the protests
New Replies Tweets Received	the number of time an account has been replied to since the start of the protests
New Tweets Sent	the number of tweets an account has posted since the start of the protests
Tweet Rate	the rate at which an account tweets since its creation
Retweets Sent Rate	the rate at which an account retweets since its creation
Retweets Received Rate	the rate at which an account is retweeted since its creation
Quote Tweets Sent Rate	the rate at which an account quote tweets since its creation
Quote Tweets Received Rate	the rate at which an account is quote tweeted 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 Retweets	the proportion of an account's total activity which is retweets
Proportion Quote Tweets	the proportion of an account's total activity which is quote tweeting
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 Retweets Sent Rate	retweets sent rate interacted with an account being new
New Account x Retweets Received Rate	retweets received rate interacted with an account being new
New Account x Quote Tweets Sent Rate	quote tweets sent rate interacted with an account being new
New Account x Quote Tweets Received Rate	quote tweets 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 Retweets	proportion of retweets interacted with an account being new
New Account x Proportion Quote Tweets	proportion of quote tweets rate interacted with an account being new
New Account x Proportion Replies	proportion of replies rate interacted with an account being new
pro-regime Hashtag Measure	a measure of how often an account uses hashtags which are also used by propagandists

Table A.3: Description of Account Characteristics Included in the Classifier

	All		Ordinary		Unsafe		Pro-regime	
	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(Pro-regime Following Measure)	2.53	3.00	1.37	1.38	0.37	1.02	6.68	1.65
log(Pro-regime Followers Measure)	2.52	3.40	0.38	0.86	0.71	1.43	7.37	1.86
log(Pro-regime Reposts Measure)	1.95	2.93	0.64	1.07	0.19	0.56	5.76	2.75
log(Pro-regime 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
Pro-regime Hashtag Measure	0.10	0.08	0.10	0.07	0.09	0.06	1.13	0.09

Notes: This table reports the mean and standard deviation of the account characteristics listed in Table A.3, for all the labeled accounts and each category of labeled account separately.

Table A.4: Statistics of Account Characteristics Across Different Account Categories

A.3 Construction of Network-based Characteristics Algorithm

In this section, we 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 pro-regime accounts.

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

1. Account u currently follows account v;
2. Account u is currently being followed by account v;
3. Account u has reposted account v; or
4. Account u has been reposted by account v.

For 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 pro-regime 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 accounts 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 accounts and select the natural number constant-k as the exit threshold. At each iteration t, we loop through the following steps:

1. Randomly choose an account, $u \in \text{argmax}_v\{L_t^R(v)\}$ (the set of accounts with the highest proximity measure at time t). If $u \in U_t$, go to $t + 1$.
2. Get the set of accounts who share the relationship-R with u, R_u. These are all the accounts who share an edge with account-u in the graph defined on relationship-R.
3. Define L_{t+1}^R by starting with L_t^R and increment the measure of each account in R_u by one if they are already in the domain of L_t^R, or otherwise add them to the domain of L_t^R with a measure of one.
4. $U_{t+1} = U_t + u$.

The algorithm terminates after T-iterations when $\|L_t^R\| = \|L_{t+k}^R\|$. The domain of the disinformation proximity measure map has not changed after k-iterations. Account-u's final disinformation proximity measure on relationship-R is $L_T^R(u)$.

A.4 Implementation by Social Media Platforms

Implementing this approach has two stages:

1. Algorithm for classifying accounts
2. Labeling news events based on account classification

Stage I: Construction of the account classification algorithm

1. Classification design
    - Decide on the groups for account classification
        - For disinformation detection, the base groups are unsafe and ordinary[41]
        - Possibly augment the model with further groups of accounts who exhibit distinct behavior from the base groups and can thus contaminate the estimation process if not separately considered.
          In this study: pro-regime accounts
    - Define the characteristics of each group
2. Create a labeled dataset of platform accounts
    - For unsafe accounts: Trace verified disinformation events to their initiators
    - For ordinary accounts: Use verified public figures, independent journalists, seek advice from field experts
3. Construct the network-based characteristics
    - Construct network proximity measures using steps in Appendix A.3
4. Augment with non-network characteristics
    - Identify and collect relevant non-network characteristics specific to the social media platform

[41] This method can be extended to other contexts where adversarial users create social network connections and exhibit herding behavior, such as identity theft purposes and other similar malicious behavior.

- Refer to the list of characteristics used in this study (Table A.3)

5. Train and test the logistic classifier

   - Split the labeled dataset: 70% for training, 30% for testing
   - Train the (bi)multinomial logistic regression using both network and non-network characteristics
     - Apply regularization (e.g., elastic net) to avoid overfitting
     - Include only characteristics with non-zero estimated coefficients
   - Test the model using the remaining 30% of the labeled accounts and make adjustments if needed

6. Classify all accounts

   - Apply the trained model to classify all accounts active on the platform into the defined groups

Stage II: Disinformation Labeling

1. Define the labeling rule

   - Condition L: If m out of the first n accounts originating a news piece are classified as unsafe, label it as disinformation
   - Set parameters m and n
     - Optimize using a set of verified disinformation and legitimate news events to achieve the desired false positive and false negative rates
     - In this study: $m = 7$, $n = 10$

2. Implement the disinformation labeling

   - Monitor new news events on the platform
   - Identify the first n accounts sharing each news event
   - Use these accounts' classification to decide if condition L is satisfied
   - If so, label all the previous and subsequent posts related to this news as disinformation

To ensure that the algorithm stays up-to-the-date, these steps should be followed:

1. Periodic update and maintenance
    - Regularly update network and non-network characteristics of all accounts
    - Retrain the model periodically with fresh data to maintain accuracy
    - Add new accounts into the training set
2. Performance monitoring
    - Continuously evaluate the model's performance on verified disinformation and legitimate news events
    - Adjust parameters (m, n) if necessary based on the desired false positive and false negative rates

A.5 Rebuttal Simulations: Baseline and Network-based Labeling

This section details our simulation methodology for estimating the impact of disinformation labeling. To establish a robust comparison, we also simulate the baseline model using the regression results of Section 4.1.

Initialization Time and threads:

- Divide time into 15-minute intervals.
- Let t_0^j denote the time of the first post of disinformation event j.
- Partition each disinformation event j into threads i, where threads are sets of posts with a common origin.
- For each (thread, disinformation event) pair (i, j), let
    - $t_{i,j}^0$: time of origin post of the thread i.
    - $\tau_{i,j}$: the time of the first rebuttal if the thread (i, j) is ever rebutted.
    - $Y_{i,j,t_{i,j}^0-1} = 0$.

Baseline Model Simulation For each disinformation event j, we run many simulations until a termination criteria is satisfied. Table 6 reports the average statistics across all the simulations.

For each disinformation event j:

1. Start the simulation at $t = t_0^j$
2. For each thread i, if $t \geq t_{i,j}^0$
   (a) Calculate $\mu_{i,j,t}$ using Equation (1) and the estimates from Table 5's final column.
   (b) Simulate post count using a Poisson distribution with mean $\mu_{i,j,t}$ to obtain $Y_{i,j,t}$.
3. Continue to the next period: calculate the total number of posts across all threads at time t. If this total is
   - Less than or equal to one: go to step 4
   - Otherwise, set $t = t + 1$ and go back to step 2
4. Termination: Check the termination criteria. If it is satisfied, stop.[42] Otherwise, go back to step 1 and start a new simulation.
5. Output statistics: Return the average statistics over all of the performed simulations for disinformation event j.

Network-based Labeling Simulation These simulations follow the baseline simulations with one key difference: the value of $\tau_{i,j}$.

Let $\tau_{i,j} = \bar{\tau}_j$ be the time when labeling begins using NDL, i.e., the time of the 10^{th} original post of the disinformation event j, $t_{10,j}^0$. As we consider the impact of the adoption of this approach by the social media platform in a centralized fashion, we assume that all posts on all threads are labeled as disinformation after $\bar{\tau}_j$.

[42] For our termination criteria, we construct a moment (a moving average) from all the performed simulations. If incorporating the last simulation into the moment changes it less than threshold of $1e^{-3}$, we stop.

B Additional Results

B.1 Section 3.1

In this section, we first report the confusion matrix of our main model reported in Table 2. We then report the estimation results corresponding to a classifier which includes all the account characteristics and does not use elastic net to choose a subset of them.

	classified Ordinary	classified Unsafe	classified Pro-regime
Ordinary accounts	correct	incorrect	incorrect
Unsafe accounts	incorrect	correct	incorrect
Pro-regime accounts	incorrect	incorrect	correct

Table B.5: Structure of Confusion Matrix

$$\begin{bmatrix} 138 & 9 & 0 \\ 10 & 132 & 1 \\ 1 & 0 & 140 \end{bmatrix}$$

Notes: Total accuracy: 95.13%

Table B.6: Confusion Matrix of Baseline Model

	Unsafe		pro-regime	
	coef	std err	coef	std err
constant	-0.03	0.52	-2.37	0.60
log(Unsafe Following Score)	-0.70	0.58	-27.76	0.88
log(Unsafe Followers Score)	8.77	0.63	5.72	0.71
log(Unsafe Reposts Score)	-0.87	0.40	-7.31	1.47
log(Unsafe Reposted Score)	4.51	0.41	-7.43	0.94
log(pro-regime Following Score)	-4.28	0.78	19.88	1.01
log(pro-regime Followers Score)	-2.72	0.69	6.76	0.97
log(pro-regime Reposts Score)	-8.21	2.41	-5.39	0.26
log(pro-regime Reposted Score)	4.51	0.52	15.55	0.38
Degree Followers Centrality	-13.45	9.54	0.76	8.15
Eigen Followers Centrality	11.94	7.75	22.90	1.32
Betweenness Followers Centrality	-0.74	63.01	-0.09	58.43
Followers to Following Ratio	5.90	605.87	11.60	3.71
Account Age	-1.81	0.58	0.66	0.53
Is New Account	8.47	0.48	6.88	0.65
% Change Followers	16.79	2.50	8.17	2.05
% Change Following	-1.76	0.29	-1.38	0.25
New Followers Rate	-14.33	17.96	4.09	5.62
New Following Rate	10.58	1.75	3.11	0.92
New Reposts Sent	-5.70	14.26	-3.77	36.65
New Reposts Received	3.54	12.36	1.63	87.91
New Quote Posts Sent	3.37	30.75	1.41	145.35
New Quote Posts Received	1.70	41.83	0.76	61.97
New Replies Posts Sent	-1.05	4.01	-1.98	8.06
New Replies Posts Received	6.61	15.05	4.82	11.04
New Posts Sent	4.34	22.98	1.13	61.26
Post Rate	4.41	18.32	0.34	37.51
Reposts Sent Rate	1.59	17.21	-0.22	34.45
Reposts Received Rate	3.24	11.27	1.92	46.03
Quote Posts Sent Rate	2.86	24.37	0.80	65.16
Quote Posts Received Rate	-0.02	42.50	0.45	48.30
Replies Sent Rate	-1.66	6.05	-2.62	9.70
Replies Received Rate	1.99	14.11	5.73	10.83
Proportion Reposts	-4.10	0.42	3.87	0.44
Proportion Quote Posts	-7.36	1.71	-0.43	1.22
Proportion Replies	-5.24	0.58	-5.54	0.47
Followers to Following Near 1	1.37	0.49	1.27	0.22
New Account x Followers Rate	-8.23	7.98	-4.16	45.61
New Account x Reposts Sent Rate	4.87	2.03	0.89	15.83
New Account x Reposts Received Rate	4.59	10.82	2.26	1114.25
New Account x Quote Posts Sent Rate	-3.12	1.40	-7.10	16.34
New Account x Quote Posts Received Rate	6.04	12.24	2.99	1375.03
New Account x Replies Sent Rate	-9.82	2.01	-5.63	10.74
New Account x Replies Received Rate	-8.54	3.88	-4.27	68.26
New Account x Proportion Reposts	-7.52	0.55	-9.10	0.73
New Account x Proportion Quote Posts	-26.88	2.51	2.60	1.51
New Account x Proportion Replies	-4.31	0.81	-1.56	0.80
pro-regime Hashtag Score	3.41	1.19	-2.29	0.69
No. Observations:	1,004			
Log-Likelihood:	-0.10			
Pseudo R-squ.:	0.95			

Notes: This table shows the non-regularized multinomial logistic regression coefficients for the model with all baseline account characteristics. Coefficients are normalized with respect to those of the ordinary category.

Table B.7: Non-Regularized Baseline Model

Category	Precision	Sensitivity
Ordinary	90.8%	93.9%
Unsafe	93.5%	90.9%
pro-regime	100.0%	99.3%

(a) Precision and Sensitivity

$$\begin{bmatrix} 138 & 9 & 0 \\ 13 & 130 & 0 \\ 1 & 0 & 140 \end{bmatrix}$$

(b) Confusion Matrix

Notes: These tables report measures of performance of the non-regularized model with all baseline characteristics. The test accuracy is 94.7%.

Table B.8: Precision and Sensitivity, and Confusion Matrix; Non-Regularized Baseline Model

B.2 Section 3.2

		m									
		1	2	3	4	5	6	7	8	9	10
n	5	100%	100%	87%	87%	50%					
	6	100%	100%	100%	87%	63%	50%				
	7	100%	100%	100%	100%	87%	63%	37%			
	8	100%	100%	100%	100%	100%	75%	63%	37%		
	9	100%	100%	100%	100%	100%	100%	75%	63%	37%	
	10	100%	100%	100%	100%	100%	100%	100%	75%	50%	25%

Notes: In this table, n is # of initiator accounts, and m is # of classified unsafe accounts in Definition 1.

Table B.9: Success Rate in Labeling Disinformation Events

		m									
		1	2	3	4	5	6	7	8	9	10
n	5	70%	50%	20%	0%	0%					
	6	70%	60%	20%	10%	0%	0%				
	7	70%	60%	60%	20%	0%	0%	0%			
	8	70%	60%	60%	30%	10%	0%	0%	0%		
	9	70%	70%	60%	50%	20%	10%	0%	0%	0%	
	10	80%	70%	60%	50%	30%	10%	0%	0%	0%	0%

Notes: In this table, n is # of initiator accounts, and m is # of classified unsafe accounts in Definition 1.

Table B.10: Failure Rate in Labeling Legitimate News Events

	m			
End of training period	5	6	7	8
November 1, 2022	(100%,30%)	(100%,0%)	(87%,0%)	(75%,0%)
November 15, 2022	(100%,40%)	(87,10%)	(87%,0%)	(87%,0%)
December 1, 2022	(100%,30%)	(100%,20%)	(87%,0%)	(63%,0%)
December 15, 2022	(100%,30%)	(100%,0%)	(100%,0%)	(75%,0%)
January 1, 2023	(100%,30%)	(100%,10%)	(87%,0%)	(63%,0%)
January 15, 2023	(100%,30%)	(100%,10%)	(100%,0%)	(100%,0%)
February 1, 2023	(100%,10%)	(100%,0%)	(100%,0%)	(63%,0%)
February 15, 2023	(100%,10%)	(100%,10%)	(100%,0%)	(75%,0%)
March 1, 2023	(100%,0%)	(100%,0%)	(100%,0%)	(87%,0%)

Notes: This table reports the performance of the non-manipulable disinformation labeling-(10,m) in Definition 1 as a pair $(D_T\%, N_F\%)$, for different values of m and different lengths of testing and training sample. $D_T\%$ is the percentage of disinformation events that are correctly classified as disinformation while $N_F\%$ is the percentage of legitimate news that are falsely classified as disinformation by the model. The training set consists of the first 6 disinformation events in our sample. The test sample consists of 8 disinformation and 10 legitimate news events. Data used for both the training and testing sample comes from the same period of time. Each row corresponds to a duration for the test and training sample, reported as the last day of the sample, and each column corresponds to a value for m.

Table B.11: Percentage of disinformation & legitimate news events classified as disinformation when non-manipulable disinformation labeling is used with $n = 10$.

B.3 Section 5.2

This section reports estimation and model performance results for two other models. The first set, Tables B.12-B.15 are for the model with non-network characteristics augmented with commonly used centrality measures (not including network-based characteristics). Tables B.12-B.13 (B.14-B.15) correspond to regularized (non-regularized) results for the model with without network-based characteristics. The second set Tables B.16-B.19, are for the network only, non-manipulable model. Tables B.16-B.17 (B.18-B.19) correspond to regularized (non-regularized) results.

	Unsafe		Pro-regime	
	coef	std err	coef	std err
Degree Followers Centrality	-0.94	0.91	-1.61	9.52
Eigen Followers Centrality	1.10	10.17	12.42	187.61
Propaganda Hashtag Score	-2.45	9.36	3.31	161.55
Account Age	-0.42	0.14	-1.95	0.21
Is New Account	1.27	0.24	0.27	0.87
% Change Followers	0.56	0.28	0.25	0.74
% Change Following	-1.91	0.58	-0.65	0.73
New Followers Rate	-2.92	0.67	-3.15	0.83
New Following Rate	0.79	1.33	0.11	0.23
New Retweets Sent	1.50	0.50	0.24	0.34
New Retweets Received	0.25	43.59	-0.27	4.61
New Replies Tweets Sent	1.38	2.45	-1.44	1.98
New Tweets Sent	2.90	2.51	0.52	2.45
Tweet Rate	1.86	1.55	-0.02	0.66
Retweets Sent Rate	0.64	4.72	0.13	17.10
Retweets Received Rate	0.26	9.13	-0.22	43.90
Quote Tweets Received Rate	-0.10	3.89	-0.10	21.13
Replies Sent Rate	-1.08	0.99	-2.26	2.91
Replies Received Rate	-0.54	9.24	-0.54	6.27
Proportion Retweets	0.96	5.09	3.13	26.71
Proportion Quote Tweets	-3.47	7.51	-3.96	10.96
Proportion Replies	-2.00	8.25	1.54	12.18
Followers to Following Near 1	0.47	7.60	1.27	23.73
New Account x Quote Tweets Sent Rate	-0.07	8.89	-0.07	6.39
New Account x Proportion Retweets	-1.74	0.44	-2.44	0.21
New Account x Proportion Quote Tweets	-0.26	8.04	0.93	20.13
New Account x Proportion Replies	0.00	1.16	0.45	9.49
No. Observations:	1,004			
Log-Likelihood:	-0.30			
Pseudo R-squ.:	0.77			

Notes: This table shows the regularized multinomial logistic regression coefficients for the model with easily-manipulable account characteristics. Coefficients are normalized with respect to those of the ordinary category. It is regularized by an elastic net with equal L1 and L2 penalties. Factors with point estimates pushed to zero by the elastic net were dropped from the table.

Table B.12: Regularized Easily-Manipulable Model

Category	Precision	Sensitivity
Ordinary	70.4%	81.0%
Unsafe	75.9%	70.6%
pro-regime	79.1%	72.3%

(a) Precision and Sensitivity

$$\begin{bmatrix} 119 & 17 & 11 \\ 26 & 101 & 16 \\ 24 & 15 & 102 \end{bmatrix}$$

(b) Confusion Matrix

Notes: These tables report measures of performance of the regularized model with easily-manipulable characteristics. The test accuracy is 74.7%.

Table B.13: Precision and Sensitivity, and Confusion Matrix; Regularized Easily-Manipulable Model

	Unsafe		pro-regime	
	coef	std err	coef	std err
constant	1.51	0.41	-1.67	0.51
Degree Followers Centrality	-3.52	6.39	-52.94	8.81
Eigen Followers Centrality	31.90	9.65	201.05	1.14
Betweenness Followers Centrality	-0.80	85.44	-3.22	50.06
pro-Regime Hashtag Score	-1.20	1.01	5.96	0.67
Followers to Following Ratio	2.27	65.68	9.82	3.59
Account Age	0.12	0.56	-2.17	0.56
Is New Account	3.51	0.44	2.86	0.83
% Change Followers	19.95	2.05	-2.55	2.22
% Change Following	-2.24	0.27	-0.10	0.24
New Followers Rate	-65.17	14.38	-78.33	6.44
New Following Rate	9.82	2.14	-29.88	0.90
New Reposts Sent	0.19	16.11	-4.37	35.31
New Reposts Received	15.90	16.07	-12.34	71.32
New Quote Posts Sent	5.20	30.96	0.04	142.52
New Quote Posts Received	-19.36	46.21	-21.27	55.34
New Replies Posts Sent	0.59	4.23	-14.73	7.55
New Replies Posts Received	7.22	18.13	-3.98	9.93
New Posts Sent	29.75	25.58	4.39	58.22
Post Rate	18.47	21.30	15.26	36.27
Reposts Sent Rate	-3.69	20.34	8.76	33.99
Reposts Received Rate	20.18	12.95	-8.37	37.14
Quote Posts Sent Rate	2.20	24.33	-1.04	64.30
Quote Posts Received Rate	-20.94	45.01	-21.68	39.88
Replies Sent Rate	-9.28	6.87	-7.06	9.25
Replies Received Rate	-16.01	17.42	-14.68	9.45
Proportion Reposts	0.69	0.41	3.55	0.43
Proportion Quote Posts	-4.08	1.87	-2.86	1.45
Proportion Replies	-1.82	0.57	3.06	0.48
Followers to Following Near 1	0.34	0.52	0.78	0.22
New Account x Followers Rate	9.09	7.51	-6.14	19.14
New Account x Reposts Sent Rate	17.88	1.70	1.71	10.98
New Account x Reposts Received Rate	8.52	10.94	-0.07	56.78
New Account x Quote Posts Sent Rate	-15.32	1.23	-28.58	11.12
New Account x Quote Posts Received Rate	6.38	11.48	-1.18	50.21
New Account x Replies Sent Rate	6.48	1.91	10.89	8.43
New Account x Replies Received Rate	-5.59	3.95	-0.04	13.00
New Account x Proportion Reposts	-4.27	0.52	-5.15	0.92
New Account x Proportion Quote Posts	-1.42	2.40	4.47	1.97
New Account x Proportion Replies	-2.45	0.73	-2.61	0.97
No. Observations:	1,004			
Log-Likelihood:	-0.51			
Pseudo R-squ.:	0.94			

Notes: This table shows the non-regularized multinomial logistic regression coefficients for the model with easily-manipulable account characteristics. Coefficients are normalized with respect to those of the ordinary category.

Table B.14: Non-Regularized Easily-Manipulable Model

Category	Precision	Sensitivity
Ordinary	78.9%	83.7%
Unsafe	76.2%	76.2%
pro-regime	89.4%	83.7%

(a) Precision and Sensitivity

$$\begin{bmatrix} 123 & 20 & 4 \\ 24 & 109 & 10 \\ 9 & 14 & 118 \end{bmatrix}$$

(b) Confusion Matrix

Notes: These tables report measures of performance of the non-regularized model with easily-manipulable characteristics. The test accuracy is 81.2%.

Table B.15: Precision and Sensitivity, and Confusion Matrix; Non-Regularized Easily-Manipulable Model

	Unsafe		Pro-regime	
	coef	std err	coef	std err
log(Unsafe Following Score)	-1.74	0.44	-4.01	0.45
log(Unsafe Followers Score)	5.96	0.21	-0.19	0.38
log(Unsafe Reposts Score)	-0.23	0.36	-2.08	0.58
log(Unsafe Reposted Score)	2.75	0.34	-0.83	0.78
log(pro-regime Following Score)	-3.88	0.65	3.37	0.83
log(pro-regime Followers Score)	0.89	0.57	3.99	0.60
log(pro-regime Reposts Score)	-0.14	1.59	0.40	0.24
log(pro-regime Reposted Score)	2.32	0.44	3.94	0.33
Degree Followers Centrality	-0.91	0.97	35.44	36.35
Eigen Followers Centrality	1.08	9.30	1070.05	1065.89
No. Observations:	1,004			
Log-Likelihood:	-0.18			
Pseudo R-squ.:	0.89			

Notes: This table shows the regularized multinomial logistic regression coefficients for the model with non-manipulable account characteristics. Coefficients are normalized with respect to those of the ordinary category. It is regularized by an elastic net with equal L1 and L2 penalties. Factors with point estimates pushed to zero by the elastic net were dropped from the table.

Table B.16: Regularized Non-Manipulable Model

Category	Precision	Sensitivity
Ordinary	90.0%	91.8%
Unsafe	91.5%	90.2%
pro-regime	100.0%	99.3%

(a) Precision and Sensitivity

$$\begin{bmatrix} 135 & 12 & 0 \\ 14 & 129 & 0 \\ 1 & 0 & 140 \end{bmatrix}$$

(b) Confusion Matrix

Notes: These tables report measures of performance of the regularized model with non-manipulable characteristics. The total accuracy of the model is 93.7%.

Table B.17: Precision and Sensitivity, and Confusion Matrix; Regularized Non-Manipulable Model

	Unsafe		pro-regime	
	coef	std err	coef	std err
constant	-1.88	0.27	-1.51	0.38
log(Unsafe Following Score)	-2.00	0.55	-46.14	0.61
log(Unsafe Followers Score)	7.46	0.64	2.56	0.53
log(Unsafe Reposts Score)	-0.21	0.36	-1.79	0.60
log(Unsafe Reposted Score)	2.54	0.36	-8.43	0.79
log(pro-regime Following Score)	-5.82	0.67	41.02	0.90
log(pro-regime Followers Score)	0.63	0.60	-3.26	0.80
log(pro-regime Reposts Score)	-1.07	1.58	-6.15	0.24
log(pro-regime Reposted Score)	5.18	0.44	16.00	0.33
Degree Followers Centrality	-5.31	1.26	-37.36	21.99
Eigen Followers Centrality	20.59	11.70	152.45	1091.84
Betweenness Followers Centrality	-0.72	1.55	-2.39	8.31
No. Observations:	1,004			
Log-Likelihood:	-0.11			
Pseudo R-squ.:	0.94			

Notes: This table shows the non-regularized multinomial logistic regression coefficients for the model with non-manipulable account characteristics. Coefficients are normalized with respect to those of the ordinary category.

Table B.18: Non-Regularized Non-Manipulable Model

Category	Precision	Sensitivity
Ordinary	91.2%	91.8%
Unsafe	92.2%	90.9%
pro-regime	99.3%	100.0%

(a) Precision and Sensitivity

$$\begin{bmatrix} 135 & 11 & 1 \\ 13 & 130 & 0 \\ 0 & 0 & 141 \end{bmatrix}$$

(b) Confusion Matrix

Notes: These tables report measures of performance of the non-regularized model with non-manipulable characteristics. Total accuracy is 94.2%.

Table B.19: Precision and Sensitivity, and Confusion Matrix; Non-Regularized Non-Manipulable Model