Troll Farms and Voter Disinformation

Denter, Philipp and Ginzburg, Boris

3 September 2021

Troll Farms and Voter Disinformation*

Philipp Denter† Boris Ginzburg‡

September 3, 2021

Abstract

Political agents often attempt to influence elections through "troll farms" – groups of fake social media accounts that flood social media platforms with messages emulating genuine information. We study the ability of troll farms to manipulate elections. We show that such disinformation tactics is more effective when voters are otherwise well-informed. Thus, for example, societies with high-quality media are more vulnerable to electoral manipulation.

Key Words: Fake News, Disinformation, Troll Farms, Elections, Social Media, Information Aggregation, Fact-Checking
JEL Codes: D72, D83, D85, D91

*We gratefully acknowledge the support from the Ministerio Economia y Competitividad (Spain) through grant PGC2018-098510-B-I00 and of the Comunidad de Madrid (Spain) through grants EPUC3M11 (V PRICIT) and H2019/HUM-5891.

†Universidad Carlos III de Madrid, Department of Economics, Calle de Madrid 126, 29803 Getafe, Spain. E-Mail: pdenter@eco.uc3m.es.

‡Universidad Carlos III de Madrid, Department of Economics, Calle de Madrid 126, 29803 Getafe, Spain. E-Mail: bginzbur@eco.uc3m.es.

1 Introduction

The rise of social media has enabled voters to acquire information from many sources, but it has also made them exposed to disinformation. One concern are the so-called keyboard armies, or troll farms – coordinated groups of ficticious social media accounts that emulate real individuals and spread messages favouring particular political positions. Such troll farms are increasingly common – one report estimates that in 38 out of 65 surveyed countries, political leaders use them to manipulate elections and other political outcomes (Freedom House, 2019).[1] Since multiple fake accounts can be controlled by a single user, or even by algorithms that do not require human intervention, a large number of trolls or automated "bots" can be deployed at negligible cost, and the messages sent by these trolls can drown out other messages.[2]

In this paper, we analyse the impact of troll farms on voting outcomes. We develop a model in which a continuum of voters need to make a choice between two policy outcomes, such as reelecting or not reelecting the government. There is a binary state of the world which indicates, for example, whether the government is competent. All voters prefer to reelect the government in the high state, but not in the low state. Voters receive imperfect binary signals about the state. There is a political operator, who wants the government to be reelected. She can organise a troll farm, which sends messages mimicking the informative signals. Each voter receives exactly one message, and does not know whether it is an informative signal, or comes from the troll farm. The political operator can choose the number of trolls – that is, the share of voters that receive messages from the troll farm – and the share of each signal realisation that trolls send.

Because signals are informative, in the absence of the troll farm the majority of voters would receive the correct signal. If signals are sufficiently precise, voters would then make the correct decision – that is, reelect the government if and only if the state is high. Thus, the election would aggregate information.

The presence of the troll farm changes this picture. If voters' preferences are such that they are ex ante willing to reelect the government, the sender can set the number of trolls to be so large that any message that a voter receives is almost surely coming from trolls. This would prevent voters from updating their beliefs, ensuring that the government wins in both

[1] Suspected use of a troll farm by Russia's Internet Research Agency to influence the US 2016 presidential election is one of the most well-known examples, but troll farms have also been used by governments and other agents in Iran, Philippines, Thailand, and other countries (The New Republic, 2017; Reuters, 2020).

[2] Existing software allows a single user to manage multiple "sock puppet" accounts (The Guardian, 2011). Furthermore, in at least 20 countries, fully automated bots appear to be used to manipulate online opinions (Freedom House, 2017).

states.

A more interesting case is when voters are ex ante unwilling to vote for the government. In that case, we show that if the precision of voters' informative signals is moderate, the election aggregates information even in the presence of the troll farm. However, if signals are very precise, the political operator can ensure that the government is reelected in both states, unless voters are heavily leaning against the government ex ante.

The reason for this result is that the troll farm needs to ensure that sufficiently many voters receive a favourable message, and that this message is able to overcome voters' initial unwillingness to vote for the government. To achieve the first, the number of trolls must be sufficiently large. However, increasing the number of trolls makes the favourable message weaker, as voters understand that with a high probability the message comes from trolls. Because a message from trolls emulates informative signals, its strength is greater when these signals are more precise. Hence, greater precision of informative signals helps the troll farm to manipulate the election.

This suggests that when technology allows troll farms to exist, an increase in signal precision can, paradoxically, make information aggregation harder to achieve. Thus, for example, societies with higher overall quality of the media are more vulnerable to manipulation by troll farms. In this, our paper differs from the standard literature on persuasion, in which receivers observe where signals originate, and hence more precise independent signals make it harder for the sender to manipulate their beliefs (see Denter, Dumav, and Ginzburg, 2021 on persuasion of voters; as well as Bergemann and Morris, 2016, and Matyskova, 2018).

In an extension, we show that the results remain fundamentally unchanged if some voters are naive – that is, unaware of the troll farm's existence – as long as the share of such voters is below $\frac{1}{2}$. We also study what happens when the troll farm has limited reach – for example, because there are capacity constraints, because some voters do not use social media, or because voters have access to fact-checking. When the share of such voters is low, the results are qualitatively unchanged. However, when the share of these voters is sufficiently large, information aggregation can be achieved when signals are very precise. Nevertheless, as long as a positive share of voters can be reached by the troll farm, there remains a parameter range over which an increase in signal precision hurts information aggregation.

The paper contributes to the growing literature on disinformation on social media (e.g., Del Vicario, Bessi, Zollo, Petroni, Scala, Caldarelli, Stanley, and Quattrociocchi, 2016, Allcott and Gentzkow, 2017, and Vosoughi, Roy, and Aral, 2018). Papanastasiou (2020) develops a model of how fake news spread when social media users choose which news to share. Kranton

and McAdams (2020) analyse how the sources of news producers' revenue and the structure of the network over which users share news together affect news veracity. Candogan and Drakopoulos (2020) study how a social media platform can design signals about content accuracy to maximise user engagement and minimise disinformation. Our paper contributes to this literature by studying decisions of a fake news provider who is interested in achieving a certain voting outcome.

The paper is also related to the literature that studies persuasion of privately informed voters (e.g., Denter, Dumav, and Ginzburg, 2021, and Heese and Lauermann, 2021). In that literature, a sender designs an experiment that sends signals conditional on the state, aiming to overcome voters' private signals. The key feature of our model is that the sender's messages mimic voters' independent signals. Because of this, increased precision of voters' private signals *helps* the sender to achieve her desired outcome.[3]

2 Model

A continuum of voters of mass one need to choose whether to reelect the government. There is an unknown state of the world $\theta \in \{0,1\}$, which indicates, for example, whether the government is competent. A voter who votes for the government receives a payoff of $1-\lambda$ if the state turns out to be 1, and a payoff of $-\lambda$ if it turns out to be 0, where $\lambda \in (0,1)$ measures the degree of voters' ex ante opposition to the government. The payoff of a voter who votes against the government is normalised to zero.[4] We assume that a voter who is indifferent votes for the government. The government is reelected if the share of voters who vote for it is at least $\frac{1}{2}$.

The prior belief that $\theta = 1$, i.e. that the government is competent, equals $q \in (0,1)$. Each voter receives a private signal $s \in \{0,1\}$ about the state. Without a troll farm, the realisation of the signal equals the state with probability $r \in \left(\frac{1}{2},1\right)$, and is distinct from the state with the complementary probability. Thus, r measures the quality of information in the absence of a troll farm.

A political operator, whom we will call the sender, is trying to make sure that the government is reelected. She can do it by setting up a troll farm, that is, by flooding the information

[3]Because the sender emulates informative signals and so decreases their informativeness, the paper is also loosely related to the literature on "signal jamming", e.g. Holmström (1999), Stone (2011), or Hermalin and Weisbach (2017).

[4]Thus, voters receive payoffs from their actions, and not from the outcome of the election. Since the set of voters is a continuum, each voter is pivotal with probability zero. Hence, allowing voters' payoffs to also depend on the voting outcome has no effect on their behaviour at an equilibrium.

environment with messages that imitate the informative signals but are not correlated with the true state. Specifically, a fraction $p \in [0,1]$ of trolls send signal 1 in each state, and a fraction $1-p$ send signal 0. The sender can choose p; additionally, by choosing the number of trolls, she can choose the probability $\alpha \in [0,1)$ that a given voter observes a signal from trolls instead of an informative signal. For example, $\alpha = 0$ means that no trolls are operating, and thus all signals are coming from informative sources. Similarly, $\alpha \to 1$ means that the number of trolls tends to infinity, and hence a signal is almost surely coming from a troll. Setting up any number of trolls is costless.

The timing of the game is as follows. First, the sender selects α and p. Then, nature draws the state θ. Each voter then receives either a signal from the troll farm or an informative signal, without being able to distinguish between the two. With probability α, a given voter observes a signal from a troll. Of these voters, fraction p observe signal 1, and fraction $1-p$ observe signal 0. With probability $1-\alpha$, a voter observes a signal from an informative source. That signal equals θ with probability r and $1-\theta$ with probability $1-r$, independently across voters. Voters then form their posterior beliefs and vote.

3 Election Outcomes

Let $\pi(s)$ be the probability that a voter assigns to the government being competent when she observes signal s. Her expected payoff is $\pi(s) - \lambda$ if the she votes for the government, and zero otherwise. Hence, she votes for the government if and only if $\pi(s) \geq \lambda$.

We will say that the election aggregates information if the government is reelected in state 1 but not in state 0.

No Troll Farms: A Benchmark. As a benchmark, consider the case when the troll farm is not operating. Then voters who observe signal $s = 1$ form a belief

$$\pi(s=1) = \frac{qr}{qr + (1-q)(1-r)},$$

while voters who observe signal $s = 0$ form a belief

$$\pi(s=0) = \frac{q(1-r)}{q(1-r) + (1-q)r}.$$

Because $r > \frac{1}{2}$, in each state a majority of voters receives the correct signal. Hence, the government wins the election in state θ if and only if $\pi(s=\theta) \geq \lambda$. The election

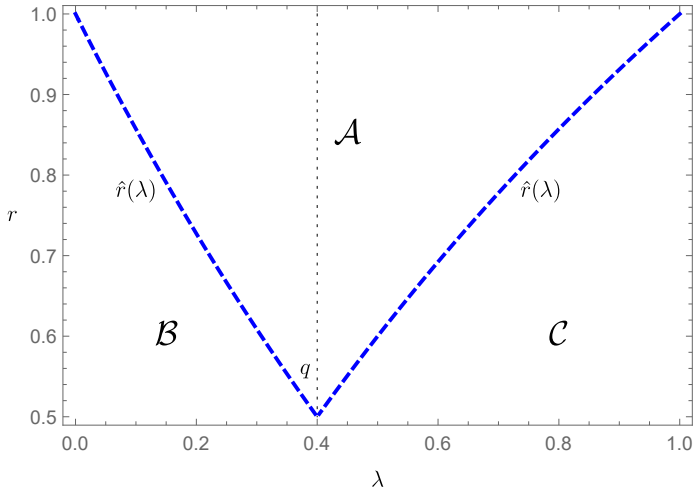

Figure 1: Election outcomes without a troll farm in both states when $q = \frac{4}{10}$ as a function of (λ, r).

then aggregates information whenever beliefs are such that $\pi(s=0) < \lambda \leq \pi(s=1)$. To achieve this, individual signals need to be sufficiently precise. The following result states this formally:

Proposition 1. *Let*

$$\hat{r}(\lambda) := \max\left\{\frac{q(1-\lambda)}{(1-q)\lambda + q(1-\lambda)}, \frac{(1-q)\lambda}{(1-q)\lambda + q(1-\lambda)}\right\} \in \left[\frac{1}{2}, 1\right).$$

If $r \leq \hat{r}(\lambda)$ and $\lambda \leq q$, the government wins the election in both states. If $r < \hat{r}(\lambda)$ and $\lambda > q$, the government loses the election in both states. If $r > \hat{r}(\lambda)$, or if $r = \hat{r}(\lambda)$ and $\lambda > q$, the election aggregates information.

Intuitively, $\hat{r}(\lambda)$ is the minimum level of signal precision at which a signal can induce a voter to vote differently from her ex ante choice. When signals are more informative than $\hat{r}(\lambda)$, voters vote according to their signals, so the election aggregates information. Otherwise, signals do not change voters' ex ante choices, and the government either wins or loses in both states. Figure 1 illustrates the set of (λ, r) pairs for which the election aggregates information ($\mathcal{A}$). It also shows the values of (λ, r) for which the government always wins ($\mathcal{B}$), and always loses ($\mathcal{C}$).

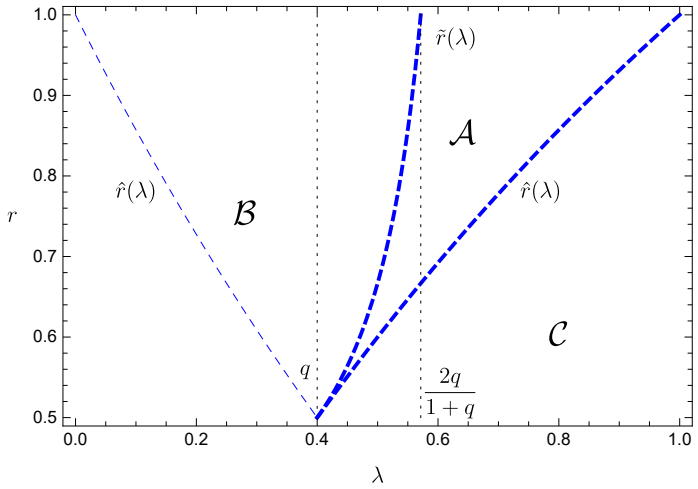

Figure 2: Election outcomes with a troll farm for different values of (λ, r).

Troll Farms. Now consider the full model when trolls are available. Given the sender's choice of α and p, a voter who observes a given signal forms the following posterior beliefs:

$$\pi(s=1) = \frac{q\left[(1-\alpha)\, r + \alpha p\right]}{q\left[(1-\alpha)\, r + \alpha p\right] + (1-q)\left[(1-\alpha)\,(1-r) + \alpha p\right]},$$

and

$$\pi(s=0) = \frac{q\left[(1-\alpha)\,(1-r) + \alpha\,(1-p)\right]}{q\left[(1-\alpha)\,(1-r) + \alpha\,(1-p)\right] + (1-q)\left[(1-\alpha)\, r + \alpha\,(1-p)\right]}.$$

The sender chooses α and p, aiming to ensure that the government wins the election. The following proposition characterises the resulting political outcomes:

Proposition 2. *Let*

$$\tilde{r}(\lambda) := \begin{cases} \frac{q(1-\lambda)}{3q-2\lambda q-\lambda} & \text{if } \lambda < \frac{2q}{1+q}, \\ 1 & \text{else.} \end{cases}$$

Then the following is true:

(i) If $\lambda > q$ and $r < \hat{r}(\lambda)$, at the equilibrium the government loses the election in both states.

(ii) If $\lambda > q$ and $\hat{r}(\lambda) \leq r < \tilde{r}(\lambda)$, at the equilibrium the government wins the election if and only if $\theta = 1$, and the election aggregates information.

(iii) If $\lambda \leq q$, or if $r \geq \tilde{r}(\lambda)$, at the equilibrium the government wins the election in both states.

Figure 2 illustrates this result. In words, Proposition 2 says that when voters are ex ante opposed to the government and signals are weak (area $\mathcal{C}$ in Figure 2), the government loses the election in both states, as in the case without the troll farm. When voters are ex ante moderately opposed to the government and signals are moderately strong, or if voters are very opposed to the government and signals are strong (area $\mathcal{A}$), then the election aggregates information, again as in the case without the troll farm.

However, in the remaining cases (area $\mathcal{B}$), the sender is able to ensure government victory in both states. If voters are ex ante willing to vote for the government, $q \geq \lambda$, the sender can set α to be arbitrarily close to one. This drowns out informative signals, preventing voters from updating their beliefs. If voters ex ante oppose the government, $\lambda > q$, the sender needs to set α to be high enough that the majority of voters receive signal $s = 1$. If informative signals are very precise, there exist values of α that achieve this while ensuring that the signal remains sufficiently strong to make them vote for the government.

The key insight of Proposition 2 is the connection between signal precision and the ability of the election to aggregate information. Recall that in an environment without the troll farm, Proposition 1 shows that information aggregation happens when signals are sufficiently precise. However, Proposition 2 implies that when the troll farm is present, an increase in signal precision r can move the outcome *away from information aggregation*, that is, from area $\mathcal{A}$ to area $\mathcal{B}$.

Using this analysis, we can characterise settings in which the troll farm has an effect on elections – that is, settings in which information aggregates when the troll farm is absent but not when it is present. Comparing the result of Proposition 2 to that of Proposition 1, we find the following:

Corollary 1. *The troll farm prevents information aggregation if and only if* $r > \hat{r}(\lambda)$ *and* $r \geq \tilde{r}(\lambda)$*, that is, if and only if signals are sufficiently precise.*

Figure 3 illustrates this result. In words, Corollary 1 shows that troll farms tend to have an effect in societies in which voters receive precise information – for example, in societies with trustworthy media. Intuitively, the persuasive power of the troll farm comes from its ability to emulate informative signals. Hence, it is larger when the informational content of these signals is higher.

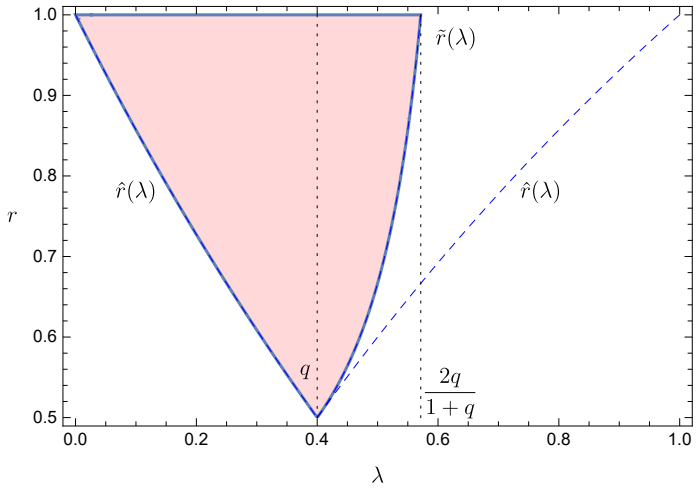

Figure 3: The impact of the troll farm. The shaded area corresponds to (λ, r) pairs for which the troll farm prevents information aggregation.

4 Discussion

Naive Voters. Suppose that some of the voters are not aware of the existence of troll farms. Specifically, suppose that a fraction $\phi \in [0,1]$ are naive, that is, update their beliefs assuming that $\alpha = 0$.[5] When $\phi = 0$, the setting is equivalent to our baseline model. The next result shows that for any ϕ, the general message of Corollary 1 holds: the sender is able to manipulate the election whenever voters' signals are sufficiently precise.

Proposition 3. *If $\phi < \frac{1}{2}$, then the troll farm prevents information aggregation if and only if $r > \hat{r}(\lambda)$ and $r \geq \tilde{r}(\lambda)$. If $\phi \geq \frac{1}{2}$, then the troll farm prevents information aggregation if and only if either $r > \hat{r}(\lambda)$, or $r = \hat{r}(\lambda)$ and $\lambda > q$.*

When not too many voters are naive, $\phi < \frac{1}{2}$, then results are identical to those from our earlier analysis. The sender still needs to convince some non-naive voters to vote for the government, and still needs to guarantee that at least half of voters receive message $s = 1$. This means that Corollary 1 remains valid.

When a majority of voters is naive, $\phi \geq \frac{1}{2}$, the sender only needs to persuade naive voters to ensure that the government wins. If signals are imprecise, Proposition 1 shows that information does not aggregate even without the troll farm. However, when signals are sufficiently accurate, information aggregates in the absence of a troll farm. This happens over region $\mathcal{A}$ in Figure 1. Then, with the troll farm, over this region the sender prevents

[5] Other papers studying naive receivers are, for example, Ottaviani and Squintani (2006), Kartik, Ottaviani, and Squintani (2007), and Little (2017).

information aggregation by setting $\alpha \to 1$, which ensures that almost all naive voters receive a favourable signal and vote for the government.

Limited Reach and Fact-Checking. It is possible that not all voters can be influenced by the sender through the troll farm. For example, some voters do not use social media, and hence only receive messages that do not come from trolls. The troll farm may also have capacity constraints, and be unable to reach the entire electorate. Alternatively, some voters fact-check the signal that they receive, and hence ignore the message that comes from the troll farm.[6] Voters that cannot be reached by the troll farm, whom we will refer to as *sceptical voters*, will only receive informative signals.

Suppose that a fraction $\mu \in [0,1]$ of voters are sceptical. When $\mu = 1$, the setting is identical to the benchmark without a troll farm, while $\mu = 0$ is equivalent to our model of a troll farm with no fact-checking. The next result describes how the impact of the troll farm depends on μ:

Proposition 4. *Define* $\bar{\mu} := \frac{1}{4}\left(3 - \sqrt{5}\right) \approx 0.191$.

(i) If $\mu \leq \bar{\mu}$, *then there exists a function* $\check{r}(\lambda)$ *such that the troll farm prevents information aggregation if and only if* $r > \hat{r}(\lambda)$ *and* $r \geq \check{r}(\lambda)$.

(ii) If $\mu \in \left(\bar{\mu}, \frac{1}{2r}\right]$, *then there exist* $\lambda^* > q$, *and functions* $\underline{r}(\lambda)$ *and* $\overline{r}(\lambda)$ *such that* $\hat{r}(\lambda) \leq \underline{r}(\lambda) \leq \overline{r}(\lambda) \leq 1$ *for all* $\lambda \in [q, \lambda^*]$; *and the troll farm prevents information aggregation if and only if either* $\lambda < q$ *and* $r \in \left(\hat{r}(\lambda), \min\left\{1, \frac{1}{2\mu}\right\}\right]$, *or* $\lambda \in [q, \lambda^*]$ *and* $r \in [\underline{r}(\lambda), \overline{r}(\lambda)]$.

(iii) If $\mu > \frac{1}{2r}$, *then the troll farm cannot prevent information aggregation.*

When the share μ of sceptical voters is smaller than $\bar{\mu}$, the result is similar to the one described in Corollary 1: information aggregation is prevented if and only if voters' signals are sufficiently precise. When the share of sceptical voters is greater than $\bar{\mu}$ but smaller than $\frac{1}{2r}$, an increase in signal precision can still hurt information aggregation when r increases and crosses $\underline{r}(\lambda)$. However, a further increase in r that moves it above $\overline{r}(\lambda)$ restores information aggregation. Note that, as before, when λ is sufficiently large, that is, greater than λ^*, the troll farm cannot prevent information aggregation. Finally, when $\mu \geq \frac{1}{2r}$, that is, when μ or r is large, the troll farm cannot prevent information aggregation, either. Figure 4 illustrates

[6] Reporters' Lab at Duke University identified more than 300 fact-checking sites globally in 2020, see `https://reporterslab.org/fact-checking-count-tops-300-for-the-first-time/` (accessed on May 19, 2021). Recent studies show that fact-checking indeed has a positive impact on beliefs, for example Barrera, Guriev, Henry, and Zhuravskaya (2020), Walter, Cohen, Holbert, and Morag (2020), or Brashier, Pennycook, Berinsky, and Rand (2021).

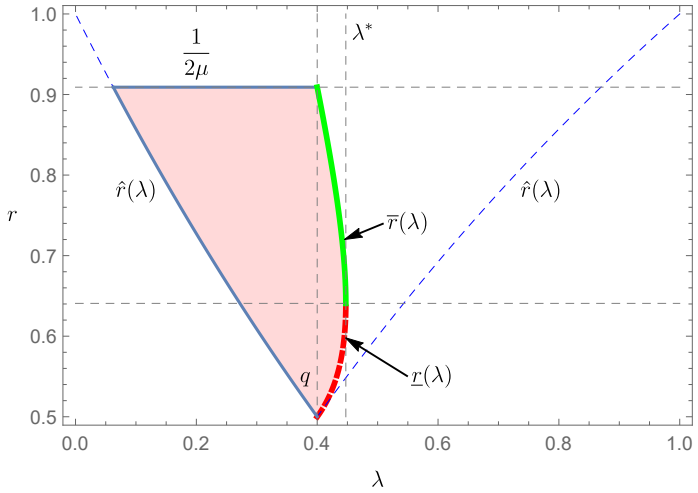

Figure 4: Outcomes for $q = 0.4$ and $\mu = 0.55 > \bar{\mu}$. The shaded area corresponds to (λ, r) pairs for which the troll farm prevents information aggregation. The solid (green) curve bordering the shaded area to the right represents $\overline{r}(\lambda)$, the dashed (red) curve represents $\underline{r}(\lambda)$.

the outcomes when the share of sceptical voters is greater than $\bar{\mu}$. The troll farm prevents information aggregation in the red shaded area, which is a subset of area $\mathcal{A}$ in Figure 1.

Taken together, Proposition 4 suggests that the availability of fact-checking (or the sender's limited reach) does not change the basic conclusions of the model when the share of voters using it is small. However, when fact-checking is available to sufficiently many voters, more precise signals can restore information aggregation. In particular, when fact checking is widely available and signals are very precise, the election aggregates information.

Mathematical Appendix

Proof of Proposition 1

Observe that $\pi(s=1) = \frac{qr}{qr+(1-q)(1-r)} \geq \lambda$ iff $r \geq \frac{(1-q)\lambda}{(1-q)\lambda+q(1-\lambda)}$. Furthermore, $\pi(s=0) = \frac{q(1-r)}{q(1-r)+(1-q)r} < \lambda$ iff $r > \frac{q(1-\lambda)}{(1-q)\lambda+q(1-\lambda)}$. Note also that $\frac{(1-q)\lambda}{(1-q)\lambda+q(1-\lambda)} \in (0,1)$ is monotone increasing in λ, while $\frac{q(1-\lambda)}{(1-q)\lambda+q(1-\lambda)} \in (0,1)$ is monotone decreasing in λ, and the two functions intersect at $\lambda = q$, reaching the value of $\frac{1}{2}$. Therefore, if $\lambda \leq q$, then $\frac{q(1-\lambda)}{(1-q)\lambda+q(1-\lambda)} \leq \frac{1}{2} \leq \frac{(1-q)\lambda}{(1-q)\lambda+q(1-\lambda)}$, so the government wins the election in both states if $r \leq \frac{(1-q)\lambda}{(1-q)\lambda+q(1-\lambda)} = \hat{r}(\lambda)$. Similarly, if $\lambda > q$, then $\frac{q(1-\lambda)}{(1-q)\lambda+q(1-\lambda)} > \frac{1}{2} > \frac{(1-q)\lambda}{(1-q)\lambda+q(1-\lambda)}$, so the government loses the election in both states if $r < \frac{q(1-\lambda)}{(1-q)\lambda+q(1-\lambda)} = \hat{r}(\lambda)$. In all other cases, the government wins the election

in state 1 and loses the election in state 0. □

Proof of Proposition 2

We will prove the three statements of the proposition in turn.

Part (i): Government Always Loses. Suppose that $\lambda > q$ and $r < \frac{\lambda(1-q)}{(1-\lambda)q+\lambda(1-q)}$. The first inequality, together with the fact that $r > 1-r$, implies that

$$\begin{aligned}\pi(s=0) &= \frac{q(1-r)+q\frac{\alpha}{1-\alpha}(1-p)}{q(1-r)+(1-q)r+\frac{\alpha}{1-\alpha}(1-p)} \\ &< \frac{q(1-r)+q\frac{\alpha}{1-\alpha}(1-p)}{q(1-r)+(1-q)(1-r)+\frac{\alpha}{1-\alpha}(1-p)} \\ &= q \leq \lambda.\end{aligned}$$

At the same time, note that $\pi(s=1)$ is decreasing in p. To see this, observe that

$$\begin{aligned}\frac{\partial\pi(s=1)}{\partial p} &= \frac{\left[qr+(1-q)(1-r)+\frac{\alpha}{1-\alpha}p\right]q\frac{\alpha}{1-\alpha}-\left[qr+q\frac{\alpha}{1-\alpha}p\right]\frac{\alpha}{1-\alpha}}{\left[qr+(1-q)(1-r)+\frac{\alpha}{1-\alpha}p\right]^2} \\ &= \frac{q\alpha}{1-\alpha}\frac{(1-q)(1-2r)}{\left[qr+(1-q)(1-r)+\frac{\alpha}{1-\alpha}p\right]^2} < 0,\end{aligned}$$

as $r > \frac{1}{2}$. Hence, $\pi(s=1)$ takes the largest value when $p=0$. At that value of p, we have $\pi(s=1) = \frac{qr}{qr+(1-q)(1-r)}$, which is smaller than λ if $r < \hat{r}(\lambda) = \frac{(1-q)\lambda}{q(1-\lambda)+(1-q)\lambda}$.

Hence, if $\lambda > q$ and $r < \hat{r}(\lambda)$, all voters vote against the government for any p and α.

Part (ii): Social Optimum. Suppose that $\lambda > q$ and $\hat{r}(\lambda) \leq r < \tilde{r}(\lambda)$. The first condition implies that $\pi(s=0) < \lambda$. Hence, voters who receive signal 0 vote against the government, and only voters who receive signal 1 may vote for it. Therefore, the government wins the election in state $\theta \in \{0,1\}$ if and only if (i) $\pi(s=1) \geq \lambda$, and (ii) the share of voters who receive signal 1 is state θ is at least $\frac{1}{2}$. Let the share of voters who receive signal 1 in state θ be m_θ, where

$$m_1 = (1-\alpha)r + \alpha p$$

and

$$m_0 = (1-\alpha)(1-r) + \alpha p.$$

Since $r > \frac{1}{2}$, we have $m_1 > m_0$.

We can show that the sender cannot ensure government victory in state 0. To see this, note that for the government to win in state 0 we need $m_0 \geq \frac{1}{2}$ and $\pi(s=1) \geq \lambda$. The former is equivalent to

$$\frac{\alpha}{1-\alpha}p \geq \frac{1}{2(1-\alpha)} - (1-r). \tag{1}$$

The latter is equivalent to

$$\frac{\alpha}{1-\alpha}p \leq \frac{q(1-\lambda)r - (1-q)\lambda(1-r)}{\lambda - q}, \tag{2}$$

which uses the expression for $\pi(s=1)$ derived in Section 3.

Next, observe that if (1) and (2) hold for some (α, p) such that $p < 1$, we can increase p and decrease α such that $\frac{\alpha}{1-\alpha}p$ remains unchanged. Then (2) remains unchanged, while (1) continues to hold because its right-hand side decreases. Hence, (1) and (2) can simultaneously hold for some (α, p) if and only if they can simultaneously for some α when $p = 1$.

When $p = 1$, (1) becomes $\frac{\alpha}{1-\alpha} \geq \frac{1}{2(1-\alpha)} - (1-r) \Leftrightarrow \alpha \geq 1 - \frac{1}{2r}$, which is equivalent to

$$\frac{\alpha}{1-\alpha} \geq 2r - 1, \tag{3}$$

while (2) becomes

$$\frac{\alpha}{1-\alpha} \leq \frac{q(1-\lambda)r - (1-q)\lambda(1-r)}{\lambda - q}. \tag{4}$$

Note that $\dfrac{\alpha}{1-\alpha}$ can take any values on $[0, \infty)$. Therefore, (3) and (4) can jointly hold if and only if

$$2r - 1 \leq \frac{q(1-\lambda)r - (1-q)\lambda(1-r)}{\lambda - q} \Leftrightarrow r[3q - 2\lambda q - \lambda] \geq q(1-\lambda). \tag{5}$$

If $\lambda \geq \frac{3q}{2q+1}$, then $3q - 2\lambda q - \lambda \leq 0$, which, together with the fact that $q(1-\lambda) > 0$, implies that (5) cannot hold. If $\lambda < \frac{3q}{2q+1}$, then for (5) to hold we must have $r \geq \frac{q(1-\lambda)}{3q-2\lambda q-\lambda} = \tilde{r}(\lambda)$, which contradicts the initial assumption. Hence, (1) and (2) cannot simultaneously hold, and the government cannot win when $\theta = 0$.

At the same time, in state 1 the sender can ensure the government's victory by setting $\alpha = 0$. In that case, we have $m_1 = r > \frac{1}{2}$, and $\pi(s=1) = \frac{qr}{qr+(1-q)(1-r)} \geq \lambda$, where the inequality follows from the fact that $r \geq \hat{r}(\lambda) \geq \frac{\lambda(1-q)}{(1-\lambda)q+\lambda(1-q)}$.

Part (iii): Government Always Wins. Suppose first that $\lambda < q$. Note that

$$\lim_{\alpha \to 1} \pi(s=1) = \lim_{\alpha \to 1} \pi(s=0) = q.$$

Hence, by selecting a sufficiently high α, the sender can set both $\pi(s=0)$ and $\pi(s=1)$ to be weakly greater than λ, and thus ensure that all voters vote for the government in both states.

Now suppose that $\lambda \geq q$ and $r \geq \tilde{r}(\lambda)$. Note that the latter condition can only hold when $\lambda < \frac{2q}{1+q}$, as otherwise $\tilde{r}(\lambda) = 1$. Hence, $r \geq \tilde{r}(\lambda)$ is equivalent to $r \geq \frac{q(1-\lambda)}{3q-2\lambda q-\lambda}$. At the same time $\lambda \geq q$ implies that $3q - 2\lambda q - \lambda > 0$. Consequently, $r \geq \tilde{r}(\lambda)$ implies that (5) holds, which by earlier reasoning means that there exists a pair (α, p) at which the government wins the election in both states. Finally, note that if $\lambda = q$, then $\tilde{r}(\lambda) = \frac{1}{2}$, so the condition $r \geq \tilde{r}(\lambda)$ is always satisfied. □

Proof of Corollary 1

Note that $\hat{r}(\lambda) > \tilde{r}(\lambda)$ if and only if $\lambda < q$. Hence, when $\lambda < q$, Propositions 1 and 2 imply that the troll farm changes the election outcome if and only if $r > \hat{r}(\lambda) > \tilde{r}(\lambda)$. When $\lambda \geq q$, Propositions 1 and 2 imply that the troll farm changes the election outcome if and only if $r \geq \tilde{r}(\lambda) \geq \hat{r}(\lambda)$. These facts together imply the result. □

Proof of Proposition 3

Note that $\pi(s=1) > \pi(s=0)$ for both naive and non-naive voters. Furthermore, $\pi(s=1)$ is greater for naive voters than for non-naive voters. Hence, if a non-naive voter votes for the government upon receiving message $s=1$, then so does a naive voter.

If $\phi < \frac{1}{2}$, ensuring the government's victory in a given state requires the sender to persuade some non-naive voters to vote for the government. Hence, she needs to ensure that (i) at least half of all voters receive message $s=1$; and (ii) $\pi(s=1) \geq \frac{1}{2}$. These conditions are identical to the conditions required for the government to win when naive voters are not present (see the proof of Proposition 2). Hence, the results are identical to those of Proposition 2 and Corollary 1.

If $\phi \geq \frac{1}{2}$, it is sufficient for the sender to persuade naive voters. If they do not vote for the government after receiving signal $s=1$, then the government cannot win the election. Otherwise, the sender can ensure that all naive voters vote for the government by setting $\alpha \to 1$. This prevents information aggregation if and only if information aggregates in the

absence of the troll farm, which happens under the condition specified in Proposition 1. □

Proof of Proposition 4

If $r < \hat{r}(\lambda)$, or if $r = \hat{r}(\lambda)$ and $\lambda \leq q$, by Proposition 1 information does not aggregate even without the troll farm. Hence, the troll farm can only prevent information aggregation when

$$r > \hat{r}(\lambda) \text{ or } [r = \hat{r}(\lambda) \text{ and } \lambda > q]. \tag{6}$$

The rest of the proof will focus on this case. Then, a sceptical voter votes for the government if and only if she receives signal $s = 1$, while a non-sceptical voter behaves exactly as in the baseline model.

When (6) holds, without a troll farm the government wins in state 1 but not in state 0. The sender can replicate this outcome by setting $\alpha = 0$. Therefore, information aggregation will be prevented if and only if the sender can guarantee a government victory in state 0.

In state 0, the mass of sceptical voters voting for the government is $\mu(1-r)$. Since this is smaller than $\frac{1}{2}$, the sender needs to induce non-sceptical voters to vote for the government. Let $m_0 = (1-\alpha)(1-r)+\alpha p$, as in the proof of Proposition 2, be the fraction of non-sceptical voters receiving message $s = 1$ in state 0. Additionally, let z_0 be the fraction of non-sceptical voters voting for the government in state 0. A necessary condition for the sender to be able to prevent information aggregation is

$$(1-\mu) z_0 + \mu(1-r) \geq \frac{1}{2}. \tag{7}$$

Note that the left-hand side of (7) is weakly smaller than $1 - \mu r$. If $\mu > \frac{1}{2r}$, then the left-hand side of (7) is strictly smaller than $\frac{1}{2}$, and hence information always aggregates. This proves part (iii) of the proposition.

Now consider the case when $\mu \leq \frac{1}{2r}$. If $\lambda < q$, the sender can, as in the baseline model, set $z_0 = 1$ by setting $\alpha \to 1$. Hence, the government's vote share is $1 - \mu r \geq \frac{1}{2}$, and information does not aggregate.

If $\lambda \geq q$, then voters who receive signal 0 do not vote for the government. For the government to win in state 0, the sender then needs to ensure that (i) non-sceptical voters who receive signal 1 vote for the government, and that (ii) $(1-\mu) m_0 + \mu(1-r) \geq \frac{1}{2}$. The first condition is equivalent to

$$\frac{\alpha}{1-\alpha} p \leq \frac{q(1-\lambda) r - (1-q)\lambda(1-r)}{\lambda - q}, \tag{8}$$

as in the proof of Proposition 2. The second condition is equivalent to

$$\begin{aligned} & (1-\mu)\left[(1-\alpha)(1-r)+\alpha p\right]+\mu(1-r)\geq\frac{1}{2} \\ \Leftrightarrow\quad & \frac{\alpha}{1-\alpha}p\geq\frac{1-2\mu(1-r)}{2(1-\mu)(1-\alpha)}-(1-r). \end{aligned} \tag{9}$$

Note that if (8) and (9) jointly hold for some α and some $p<1$, then we can increase p and reduce α such that $\frac{\alpha}{1-\alpha}p$ remains unchanged. Then (8) continues to hold because its right-hand side remains unchanged, while (9) continues to hold because its right-hand side is either negative or monotone increasing in α. Hence, the government can win in state 0 for some (α,p) if and only if it can win in state 0 for some α when $p=1$. We can thus without loss of generality restrict attention to the case when $p=1$. Then (9) becomes

$$\alpha\geq\frac{2r-1}{2(1-\mu)r}\Leftrightarrow\frac{\alpha}{1-\alpha}\geq\frac{2r-1}{1-2\mu r}, \tag{10}$$

while (8) becomes

$$\frac{\alpha}{1-\alpha}\leq\frac{q(1-\lambda)r-(1-q)\lambda(1-r)}{\lambda-q}. \tag{11}$$

Since $\frac{\alpha}{1-\alpha}\in[0,\infty)$, it is possible to find $\alpha\in[0,1)$ at which (10) and (11) hold if and only if

$$\begin{aligned} & \frac{2r-1}{1-2\mu r}\leq\frac{q(1-\lambda)r-(1-q)\lambda(1-r)}{\lambda-q} \\ \Leftrightarrow\quad & \lambda\leq\tilde{\lambda}(r):=\frac{q(r(2\mu r-3)+1)}{q(2r-1)(2\mu r-1)+r(2\mu(1-r)-1)}. \end{aligned} \tag{12}$$

Note that $\tilde{\lambda}(r)\Big|_{r=\frac{1}{2}}=\tilde{\lambda}(r)\Big|_{r=\frac{1}{2\mu}}=q$ and

$$\frac{\partial\tilde{\lambda}(r)}{\partial r}=\frac{(1-q)q(4\mu r((\mu-2)r+1)+1-2\mu)}{(q(2r-1)(2\mu r-1)+r(2\mu(1-r)-1))^2}.$$

If $\mu\leq\bar{\mu}:=\frac{1}{4}\left(3-\sqrt{5}\right)\approx 0.191$, then $\tilde{\lambda}(r)$ is monotone increasing in r on $r\in\left[\frac{1}{2},1\right]$. Therefore, the equation $\lambda=\tilde{\lambda}(r)$ has a unique solution $r(\lambda)\geq\hat{r}(\lambda)$, where the inequality follows from the fact that the government always wins at $r=r(\lambda)$, and always loses at $r<\hat{r}(\lambda)$ when $\lambda\geq q$. Defining $\check{r}(\lambda)$ to equal that solution when $\lambda\geq q$, and to equal $\hat{r}(\lambda)$ otherwise yields part (i) of the proposition.

If $\mu\in\left(\bar{\mu},\frac{1}{2r}\right]$, $\tilde{\lambda}(r)$ attains a unique interior maximum λ^* on $r\in\left[\frac{1}{2},1\right]$. This means that when $\lambda\in[q,\lambda^*)$, the equation $\lambda=\tilde{\lambda}(r)$ has two solutions. Define $\underline{r}(\lambda)$ to equal the

smaller of the two solutions when $\lambda \in [q, \lambda^*]$. Also define $\overline{r}(\lambda)$ to equal the greater of the two solutions when $\lambda \in [q, \lambda^*]$, and to equal $\min\left\{1, \frac{1}{2\mu}\right\}$ otherwise. Note that when $\lambda = \lambda^*$, we have $\underline{r}(\lambda) = \overline{r}(\lambda)$. Then when $\lambda \in [q, \lambda^*]$, the troll farm prevents information aggregation if and only if $r \in [\underline{r}(\lambda), \overline{r}(\lambda)]$. Together with the fact that when $\lambda < q$ the troll farm prevents information aggregation if and only if $r \in (\hat{r}(\lambda), \frac{1}{2\mu}]$, this implies part (ii) of the proposition. □

References

Allcott, H., and M. Gentzkow (2017): "Social media and fake news in the 2016 election," *Journal of Economic Perspectives*, 31(2), 211–36.

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

Bergemann, D., and S. Morris (2016): "Information Design, Bayesian Persuasion, and Bayes Correlated Equilibrium," *American Economic Review*, 106(5), 586–91.

Brashier, N. M., G. Pennycook, A. J. Berinsky, and D. G. Rand (2021): "Timing matters when correcting fake news," *Proceedings of the National Academy of Sciences*, 118(5).

Candogan, O., and K. Drakopoulos (2020): "Optimal signaling of content accuracy: Engagement vs. misinformation," *Operations Research*, 68(2), 497–515.

Del Vicario, M., A. Bessi, F. Zollo, F. Petroni, A. Scala, G. Caldarelli, H. E. Stanley, and W. Quattrociocchi (2016): "The spreading of misinformation online," *Proceedings of the National Academy of Sciences*, 113(3), 554–559.

Denter, P., M. Dumav, and B. Ginzburg (2021): "Social Connectivity, Media Bias, and Correlation Neglect," *The Economic Journal*, 131, 2033–2057.

Freedom House (2017): "Manipulating Social Media to Undermine Democracy," Accessed on 26 April 2021.

——— (2019): "The Crisis of Social Media," Accessed on 26 April 2021.

Heese, C., and S. Lauermann (2021): "Persuasion and Information Aggregation in Elections," .

Hermalin, B. E., and M. S. Weisbach (2017): "Assessing Managerial Ability: Implications for Corporate Governance," in *The Handbook of the Economics of Corporate Governance*, ed. by B. E. Hermalin, and M. S. Weisbach, vol. 1 of *The Handbook of the Economics of Corporate Governance*, pp. 93–176. North-Holland.

Holmström, B. (1999): "Managerial incentive problems: A dynamic perspective," *The Review of Economic Studies*, 66(1), 169–182.

Kartik, N., M. Ottaviani, and F. Squintani (2007): "Credulity, lies, and costly talk," *Journal of Economic Theory*, 134(1), 93–116.

Kranton, R., and D. McAdams (2020): "Social Networks and the Market for News," .

Little, A. T. (2017): "Propaganda and credulity," *Games and Economic Behavior*, 102, 224–232.

Matyskova, L. (2018): "Bayesian Persuasion with Costly Information Acquisition," *CERGE-EI Working Paper Series*, 614.

Ottaviani, M., and F. Squintani (2006): "Naive audience and communication bias," *International Journal of Game Theory*, 35(1), 129–150.

Papanastasiou, Y. (2020): "Fake news propagation and detection: A sequential model," *Management Science*, 66(5), 1826–1846.

Reuters (2020): "Facebook, Twitter dismantle global array of disinformation networks," Accessed on 19 April 2021.

Stone, D. F. (2011): "A signal-jamming model of persuasion: interest group funded policy research," *Social Choice and Welfare*, 37(3), 397–424.

The Guardian (2011): "Revealed: US spy operation that manipulates social media," Accessed on 19 April 2021.

The New Republic (2017): "Rodrigo Duterte's army of online trolls," Accessed on 19 April 2021.

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

Walter, N., J. Cohen, R. L. Holbert, and Y. Morag (2020): "Fact-Checking: A Meta-Analysis of What Works and for Whom," *Political Communication*, 37(3), 350–375.