Strategies and Vulnerabilities of Participants in Venezuelan Influence Operations

Ruben Recabarren Bogdan Carbunar Nestor Hernandez Ashfaq Ali Shafin
FIU

Abstract

Studies of online influence operations, coordinated efforts to disseminate and amplify disinformation, focus on forensic analysis of social networks or of publicly available datasets of trolls and bot accounts. However, little is known about the experiences and challenges of human participants in influence operations. We conducted semi-structured interviews with 19 influence operations participants that contribute to the online image of Venezuela, to understand their incentives, capabilities, and strategies to promote content while evading detection. To validate a subset of their answers, we performed a quantitative investigation using data collected over almost four months, from Twitter accounts they control.

We found diverse participants that include pro-government and opposition supporters, operatives and grassroots campaigners, and sockpuppet account owners and real users. While pro-government and opposition participants have similar goals and promotion strategies, they differ in their motivation, organization, adversaries and detection avoidance strategies. We report the Patria framework, a government platform for operatives to log activities and receive benefits. We systematize participant strategies to promote political content, and to evade and recover from Twitter penalties. We identify vulnerability points associated with these strategies, and suggest more nuanced defenses against influence operations.

1 Introduction

Social networks have become the central medium for *influence operations* (IOs), enabling them to disseminate and amplify disinformation, and compromise the integrity of information posted by others. We observe a parallel between disinformation (information designed to mislead [96]) and malware (e.g., self-propagating worms) where disinformation runs on human minds instead of computers. From this perspective, influence operations seek to fraudulently boost the search rank of the content they distribute, and increase the number of human hosts exposed and infected.

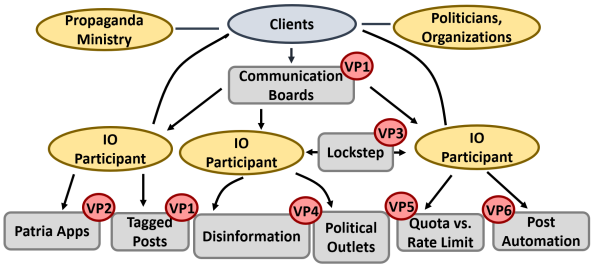

Figure 1: Map of discovered strategies (gray rectangles) of influence operations participants (yellow ovals). Red circles represent influence operations vulnerabilities that we identified, and discuss in the context of Twitter changes, in § 5.

Goals of influence operations include manipulating or corrupting public debate, undermining trust in democratic processes and scientific evidence, and even influencing elections [25, 36, 67, 82]. Such efforts are becoming increasingly prevalent: Facebook reported the discovery of 150 covert IOs on its site between 2017 - 2020, that originated from countries all over the world [36].

Influence operations were shown to be well organized [37, 54, 60, 61, 82, 94, 96], control many social network sockpuppet accounts [54, 59, 82], and employ inauthentic behaviors [36, 61, 85, 87, 110]. This knowledge was collected through journalistic efforts [28, 61, 63, 67, 82, 85, 87] and forensic analysis of social networks [24, 46, 52, 59, 109, 110] and released datasets [1, 39, 69, 79, 96, 102].

To develop information assurance solutions that can control influence operations, we need however to understand the experiences, challenges and vulnerabilities of their contributors. We lack such information due to difficulties to identify, reach, recruit and establish trust with such participants.

In this paper, we investigate the perspective of participants in influence operations. For this, we leverage unique background and insights into Venezuela, a country where influence operations have replaced verified news [44, 68, 82, 83]. Since

2013, Maduro's regime has taken over the country's institutions, electoral and justice system, and has censored standard news delivery solutions [18]. To bypass censorship, communicate and organize, the opposition uses social networks and mobile apps [42]; conversely, the government uses them to distribute hyperpartisan news and disinformation [82].

In a first contribution, we developed a protocol to identify and recruit participants in Venezuelan influence operations. Our protocol uses Telegram groups and Twitter to identify candidates, and to contact them over direct messaging. Second, we recruited 19 relevant participants, and conducted semi-structured interviews to study the following key questions:

- **RQ1**: What are concrete (a) organization and communication mechanisms, (b) resources, capabilities, and limitations, (c) motivation, and (d) promotion strategies of participants in Venezuelan influence operations?
- **RQ2**: Do they participate in influence operations that target other countries? (a) Are they willing to be hired to participate in external influence operations?
- **RQ3**: What is the participant perception on disinformation? (a) Do they contribute or do they have strategies to avoid their distribution?
- **RQ4**: Are participants aware of, and affected by social network defenses and penalties? (a) Have they developed strategies to circumvent and recover from detection? (b) Are these strategies effective?

Third, to validate participant claims, we performed a quantitative investigation with data collected over four months, from 34 Twitter accounts they control. Our findings include:

(1) Interview participants are diverse, e.g., pro-government vs. opposition supporters, paid operatives vs. grassroots campaigners, and sockpuppet owners vs. real users (§ 4.6). We found consistency with the "communication constitutes organization" perspective of organizational theory [77] (§ 4.5);

(2) Both pro-government and opposition participants revealed a history of contribution to foreign campaigns. Many on both sides are willing to be hired to participate in influence operations, including targeting US politics (§ 4.4);

(3) Participants claimed strategies to verify information they post. However, we report concrete instances of pro-government participant distribution of disinformation that received significant community engagement (§ 4.2);

(4) Adversarial environments: Pro-government participants reported efforts by social nets to thwart their activities; the opposition revealed pro-government operative attacks against their accounts (§ 4.8). Both sides disclosed strategies to avoid detection and recover from account suspensions (§ 4.9);

(5) Pro-government and opposition participants differ in their motivation, organization, adversaries, and detection avoidance strategies. Based on our findings, we present the IO strategy map of Figure 1 (§ 4.7).

In a fourth contribution, we identify vulnerability points (VPs) associated with promotion strategies revealed by our participants (Figure 1), and suggest changes to social networks' handling of influence operations (§ 5).

2 Background, Model and Goals

2.1 The Venezuelan Crisis

Venezuela is experiencing the worst economic crisis in its history [72]. The government controls every aspect of daily life ranging from food to gas supply, while the country struggles with hyperinflation, unemployment and poverty. In recent years, Maduro's government has implemented a takeover of the Venezuelan government and institutions, the electoral and justice systems, and the army. The government has used lethal force against protesters, exiled critics, and held political prisoners. Six million people have migrated to neighboring countries [88].

A large majority of Venezuelans oppose the current regime [19]. The government has however invested heavily in media censorship efforts [18]. This has led to a migration of anti-government movements to social media and mobile apps [44, 68]. In turn, this was followed by the creation of a government-sponsored online army [83], to disrupt the opposition activities and promote the government propaganda.

Political allies of Venezuela provide support, by distributing hyperpartisan news and disinformation through their Spanish language news organizations [21, 81], e.g., Russia Today (RT) Español [2], Sputnik Mundo [11], the Iranian Hispan TV [7] and Cuban [5, 6] news outlets.

2.2 Adversary

Influence Operations. Influence operations (IOs) also known as information campaigns [101] or strategic information operations [96], are coordinated efforts to manipulate or corrupt public debate for a strategic goal [36]. Influence operations were shown to have at least short-term effects, that include political beliefs and behavior changes [26], increased xenophobia [104], and increased uncertainty about vaccines [76].

In this work we distinguish between centralized influence operations and grassroots movements. While grassroots movements are bottom-up, often spontaneous decision making efforts [107], centralized influence operations have a command and control (C&C) center (e.g., government, institution, or interest group) that designs the operation's goals and message.

We now define several types of participants in influence operations. Not all are adversarial. We discuss them here because they are used or manipulated by adversarial C&Cs.

IO Participants. To avoid detection, centralized operations were shown to emulate online grassroots movements [36, 70]. They achieve this by recruiting real people, that include *operatives* and *grassroots campaigners*. Operatives receive incentives to promote the operation's message online [36, 70]; grassroots campaigners believe the information they distribute, do not get paid and are unwilling to be hired for activities that

contradict their beliefs. Our study includes both types of participants. Such participants were shown to create and amplify posts that promote the operation's message and to communicate and coordinate activities [96].

In contrast, *unwitting agents* [29,96] are human participants that receive and occasionally engage with influence operations content, but do not receive external incentives and do not coordinate activities.

Influence operations contributors also include *trolls*, that use anonymous accounts and post inflammatory and digressive messages, designed to trigger conflict and disrupt online discussions [62, 82]. Datasets of Russia's Internet Research Agency (IRA) trolls in Twitter [40,102] revealed several types of troll accounts, each performing a specialized function [59]. While IRA tweets reached many users [109], they had minor impact in making content viral [109]. In contrast, we found that many of our participants accumulated significant community engagement for their posts, including disinformation.

Influence operations also use *bots*, automated accounts that require little human supervision [23, 38, 47]. Both trolls and bots use sockpuppet accounts to hide their identities. Previous work however has found that many IO participants are not bots, and manage their online identities in complex ways [46]. Most of our interview participants use their own identity to establish a personal brand and a follower base.

Coordination Apps. IO participants use apps to communicate and coordinate activities [96]. Previous work studied misinformation and fear speech in WhatsApp [48, 51]. In particular, Javed et al. [48] analyzed the spread of information through WhatsApp, and documented information flows between WhatsApp and Twitter. Our participant recruitment process builds on a similar finding, that Venezuelan operators use communication apps to organize, coordinate and disseminate messages to be promoted in Twitter (§ 3). We also found similar information flows in Venezuelan operations, between Telegram groups and Twitter.

2.3 Research Goals

Social networks implement various techniques to address influence operations. They include mechanisms to detect [106], verify [53] and penalize accounts and activities that violate their terms of service. Twitter penalties include suspending accounts detected to post spam, suspected to be compromised, or reported to violate rules surrounding abuse [15]. Further, social networks were reported to *shadowban*, i.e., remove or limit the distribution or visibility of certain content [17, 90].

Such mechanisms are often unable to address influence operations in real time [36], and some consider them to be censorship [90]. Our study confirms this.

Instead, in this work we seek to provide insights into influence operations, by studying the perspective of IO participants. We document experiences, motivation, organization and communication mechanisms, capabilities, goals and strategies of participants in Venezuelan influence operations, in order to understand their strengths and vulnerabilities, and help inform future efforts to design more appropriate, inclusive and effective solutions to address influence operations.

Algorithm 1: Pseudocode for study protocol. The recruitment takes place on a social network SN (Twitter in our case). IOGroups lists influence operations communication groups used to seed the candidate search (from Telegram in our case).

```
1.StudyProtocol(SN: SocialNet, IOGroups: list)
2.    while (true) do{
3.      IOMembers = getSNAccounts(IOGroups);
4.      IOActive = followBack(IOMembers);
5.      IOFollowers = getFollowers(IOActive);
6.      candidates = getOpenToDM(IOFollowers);
7.      respondents = sendDM(candidates);
8.      groups = {};
9.      for each R in respondents
10.       (answers, accounts, g) = interview(R);
11.       accountData = SN.collectData(accounts);
12.       validate(answers, g, accountData);
13.       groups = groups ∪ g;
14.     if (groups ∈ IOGroups) then break;
15.     IOGroups = IOGroups ∪ groups;}
```

3 Methodology

Our study consists of a qualitative exploration and a quantitative investigation into various aspects of participation in influence operations. We first detail the recruitment procedure and ethical considerations, then describe the studies.

3.1 Participant Recruitment

We focused recruitment efforts on identifying Twitter accounts with a verifiable history of participation in influence operations. Our recruitment protocol identifies active operatives, by starting with a seed set of communication groups. To identify this set, we leveraged observations that Venezuelan operatives use Telegram to communicate about their goals and objectives. We have used Telegram's search (keyword "Twiteros activos") to identify three Telegram groups and channels dedicated to Venezuelan influence operations.

Members of these groups often disclose their Twitter handles to follow one another. We have selected a set of Twitter accounts that were revealed by members of these groups. We did not contact these accounts directly: To minimize the chance of interference with our study, we wanted to delay news of our efforts from reaching the influence operations command and control center (§ 2.2). Members of these Tele-

gram groups may have communication channels with the command and control, thus may quickly alert many participants and influence their perception about our study.

Instead, we followed these accounts from our lab's Twitter account. For the accounts that followed us back, we collected, via breadth-first search, and using the Twitter API, their Twitter followers. From these, we identified the accounts that were open to direct messaging from our account.

We sent an interview invitation to these accounts, over direct messaging (DM). We then sent personalized messages, including a consent form, to the accounts that replied. We inspected the accounts that accepted the consent form, and interviewed those that were active, were posting tweets with political topics and had at least 500 followers.

During the interview, we also collected other accounts claimed to be controlled by participants, and groups they claimed to use for communications. We then iterated our recruitment activities over these groups.

Algorithm 1 summarizes our study procedure, including recruitment, interviews, and data validation steps.

In total, we followed 1,543 Twitter accounts. From the 109 accounts that followed us back, we collected their 256,770 followers. We sent DMs to the subset of 2,843 accounts that were open for DM, then sent personalized messages to the 99 accounts that replied. From the 35 Twitter accounts that accepted the consent form, we selected 19 for interview.

3.2 Ethical Considerations

The study procedure was scrutinized and the full study was approved by the institutional review board of our university (IRB-20-0550). We followed ethical practices for conducting sensitive research with vulnerable populations [30]. For instance, we tailored the consent process to the participant, and re-confirmed consent. We sent the consent form link and obtained consent both during recruitment, and at the beginning of the interview. We obtained consent both electronically and verbally. We accommodated participant requests for private payments: cryptocurrencies, intermediaries in Venezuela, and sending money over snail mail.

During recruitment and the interview, we clearly declared the identity of the researchers, the research objective, the data that we collect (including Twitter account data) and how we process it, and potential impact on the participant. More specifically, our invitation and consent form made clear that our intention is to study political content promotion capabilities, resources and behaviors on social platforms. During recruitment, we followed candidate accounts from our lab's Twitter account, where we made clear its association to our lab, and its use strictly for research purposes.

We also explained any risks that their work may have through our research. We asked several times during the interview if they are comfortable discussing potentially sensitive topics and told them that they could skip any question.

We were careful to hide participant identity. Following the account data collection and participant payment steps, we removed all participant PIIs from our data (e.g., names, IDs, handles, locations). From the participants' Twitter accounts, we kept only account statistics, their posts, and their followers. We used multiple solutions to securely store de-identified research data. The data was stored on a physically secure Linux server in our university, and accessed through encrypted channels only from the password-protected authors' laptops. Further, all data was processed only on the server.

In § 6 we further revisit ethical considerations from the perspective of the impact of our findings.

Team Positionality Statement. The research team consists of Venezuelan and international investigators, all located outside Venezuela. The investigators support neither the Venezuelan government or the opposition. The interviews were conducted by a politically neutral team member, who shared context with study participants, including the language and some knowledge of the country's political and economical circumstances.

3.3 Qualitative Study

Interviews. We conducted semi-structured interviews with the recruited participants: one pilot interview to test our interview guide and method, then in-depth interviews with 19 participants. The interview focused on participant (1) incentives to contribute to influence operations, (2) organization structure, (3) resources, capabilities and limitations, (4) strategies employed to promote IO goals, (5) operations in which they participated and in which they are interested to participate, (6) perception of disinformation in influence operations, (7) perception on the impact of Twitter's defenses on IO activities, and (8) strategies to evade and recover from detection.

The interviews were conducted over the phone, in Spanish, by one author who is a native speaker. All audio interviews were recorded with participant permission. The interviews lasted between 17 and 98 minutes (M = 52, SD = 19.43). We paid 2.5 USD for every 15 minutes spent in the interview.

Analysis Process. We analyzed responses using a grounded-theory open-coding process [99], performed by two co-authors: the one who conducted the interviews and a non-Spanish speaker. We conducted the interviews over 6 weeks. During this time we also transcribed and anonymized the recorded interviews, then translated them into English. Given time constraints, we started the analysis after data collection was complete. Following each interview, the interviewer discussed impressions, observations and findings with the rest of the team. This enabled us to detect reaching data saturation [41], where the interviewer reported no new insights from the last two participants. We confirmed this during analysis.

In the preliminary analysis stage, we independently read five transcripts to establish a thematic framework of the interview data. We coded participant responses to each interview question including relevant information provided later in the

interview. We organized the themes into an initial codebook. We then independently coded and met to revise the codebook. We used these themes to organize codes emerging from the remaining 14 transcripts. Two co-authors met to discuss the themes and codes after processing each set of two to three interviews. In total, we created 177 codes from 410 pages of transcripts. Since we reviewed the coded transcripts jointly, we do not include the inter-rater reliability score [64].

3.4 Quantitative Investigation

To validate participant claims, we performed a quantitative analysis with data from several sources:

Participant Twitter Accounts. We have collected information from 19 Twitter accounts that we know are controlled by the participants, i.e., they replied to DMs we sent to these accounts during recruitment. We call these *recruitment validated* accounts. The accounts were between 11 months and 11.5 years old (M=93.14 months, SD=48.82). We have monitored the Twitter timelines of these accounts over four months in 2021. We have collected their tweets and retweets, the engagement received by each tweet, the number of followers and accounts that they follow. We have collected the trending hashtags for all nine Venezuelan regions available in Twitter during that interval. In total, we have collected 264,043 timeline posts and 5,499,700 trending hashtag reports.

In addition, 11 participants revealed during interviews, 15 other Twitter accounts they claimed to control.

Telegram Groups. During participant recruitment, we have identified and joined six Telegram groups (TuiterosDeChavez, Tuiteros Patriotas, TuiterosActivos, Twiteros Patriotas, Twiteros Activos, Bonos de la Patria) used by participants to communicate and coordinate activities. The groups had a total of 3,352 members, and were active at the time of submission. The groups provide members with instructions regarding the work they are expected to perform.

4 Results

In this section we first classify the participants, then explore perception and participation in the distribution of disinformation (§ 4.2), motivation (§ 4.3), and willingness to participate in paid campaigns (§ 4.4). We then describe participant reported organization and communication channels (§ 4.5), and capabilities (§ 4.6). We discuss reported strategies to promote content (§ 4.7), perceptions of Twitter defenses (§ 4.8), and strategies to evade and recover from detection (§ 4.9).

4.1 Participant Classification

Demographics. Our participants have diverse backgrounds. Thirteen male, seven female; age range between 18 and 67 (M=50.8, SD=11.5); job types include self-employed (2), teacher (7), engineer (2), lawyer (1), public accountant (2), communications expert (2), manager (1), TV actor (1) and assistant (2). The highest education level was high-school (4), bachelors (10), masters (5) and PhD (1). 18 participants lived in Venezuela, one in Nicaragua.

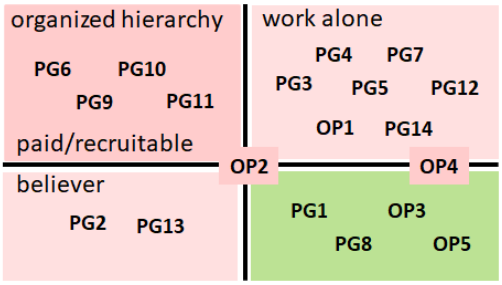

Figure 2: Participant classification across two dimensions: (1) member of organized hierarchy (left column) vs. working alone (right column), and (2) paid or willing to be hired (top row) vs. believer (bottom row). Our participants include both influence operators and grassroots members.

Pro-government vs. Opposition. Fourteen participants were pro-government and five supported the opposition. We verified this using their Twitter account data. In the following, for simplicity, we use PG1, .., PG14 for the pro-government participants and OP1, .., OP5 for the opposition participants.

Operatives vs. Grassroots Campaigners. Figure 2 shows the classification of our participants on two dimensions: (1) members of an organization vs. working alone, and (2) having received benefits or being willing to be hired vs. being a believer. Six pro-government and one opposition participant operated in hierarchical operations, and received or issued instructions. Overall, twelve pro-government and one opposition participants were either part of an organized hierarchy, had received rewards, or were willing to receive rewards for their activities.

Two pro-government and two opposition participants have strong political convictions, and may be considered grassroots campaigners (§ 2). Two other opposition participants are hybrid (OP2, OP4, shown on borderlines in Figure 2). For instance, OP2 used to be pro-government, and had leadership roles in influence operations. OP2 later became disillusioned, started supporting the opposition, and was even a political prisoner. OP2's activities are driven by political beliefs, but is also recruited to participate in campaigns during special events, e.g., before elections.

4.2 RQ3: Disinformation

The Twitter Truth. All participants have created or amplified hyperpartisan news in Twitter. We observed consistency among the views of pro-government participants, who often re-tweeted the same posts. This includes images from staged events, showing efforts by various institutions and politicians

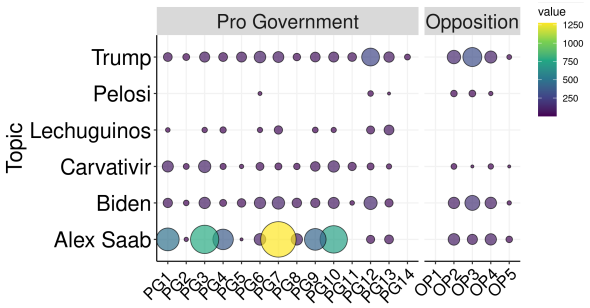

Figure 3: Per-participant number of posts over four months, on select controversial and US politics-related topics. Both pro-government and opposition participants have interests in US politics and controversial topics, but with opposing views.

to improve the lives of Venezuelans. Given the government's obliteration of independent reporting, such events are impossible to verify, and highly suspicious. We observed more diverse interests among opposition participants. However, they also post and promote anti-government messages and accusations, often without providing trustworthy proofs.

Figure 3 shows the number of posts from participants, on controversial subjects "Carvativir" [95] (540 posts), "Alex Saab" [3] (4,577 posts) and articles from disinformation site [32] `lechuguinos.com` (72 tweets, 178 retweets from 9 participants). Carvativir is a thyme derivate that was promoted by the Venezuelan president to neutralize COVID-19 with no side effects, a claim not substantiated by data [95]. Alex Saab is a Colombian businessman, alleged financier for Venezuela's president, who was arrested and extradited to the US. He was accused by the US Department of Treasury to be part of the corruption network that stole from Venezuela's food distribution program [20], see also § 4.3.

We observed however opposing views between these groups. For instance, on Carvativir [95], pro-government participants distributed claims that FDA considers it to be safe, that it is optimal for the treatment of COVID-19, and has antiviral capacity to block SARS-CoV-2 and positive effects in COVID-19 patients. The opposition participants claimed that the government deceives people and uses Carvativir as a source of revenue. Further, while pro-government and opposition participants converge in their enmity toward the US president Biden, their reasons differ: Pro-government operations use him as a scapegoat to blame for the country's situation; opposition participants believe that his government will convert the US into Venezuela.

We further found 559 posts with links to Venezuelan government sites; also, 219 posts with links to Russian [2, 11], Iranian [7] and Cuban [5, 6] news outlets, known to distribute disinformation [22, 71]. This is consistent with strategies of integration of government and externally-funded media as source content for narratives in countries like Syria [96, 97].

These findings confirm a "firehose of propaganda and falsehood" model [73] employed by pro-government participants, where propaganda and disinformation is used to drown out the opposition [70] and reduce the ability of readers to make sense of information [73, 75].

Perception of Disinformation. Both pro-government and opposition participants explained that they have witnessed disinformation in Twitter, e.g., "There is a lot of fake news" (OP3), "*Many people tend to post fake news*" (PG5). To avoid distributing such posts, some explained that they research the content they receive, e.g., "*I research [my publications] otherwise I could become an amplifier of what is known as fake news.*" (PG7), "*many times, we learn about something then we research the truth*" (PG5). We emphasize however the lack of trustworthy news sources, the remote nature of many reported events, and restricted communications.

Some participants validate the sources of tweets, e.g., *"I retweet posts from journalists and politicians that publish truthful information, and not accounts with pseudonymous and unknown names"* (OP3). This is consistent with findings that people in the US rate mainstream sources more trustworthy than hyperpartisan or disinformation sources [74]. However, others found that the source has little impact on how people judge headlines (accurate vs. inaccurate) [33].

Several participants claimed that when posting original tweets, they add links to a credible source that confirms the information. 11,678 of the 237,978 posts we collected from the accounts of our participants contained links to other sites.

4.3 RQ1: Motivation

Twelve participants claimed to have received some form of rewards for their online activities. Of these, nine were not required to do this work, while three reported mechanisms suggesting coercion. Of the latter, one received medical help from the government, and two explained they are on government payroll, where posting political content is part of their work. Indeed, working for the government, which for many who cannot leave the country is the only option, entails being subject to implicit forms of both blackmail and bribery. For instance, state employees who do not tweet in favor of the government or who do not go to government-sponsored protests do not get paid or do not receive food stamps. We note that 60% of the active population is employed, of which almost 30% are working for the government [92].

One theme among pro-government participants was the central role played by the government-commissioned Patria platform [12], in recording online activities and distributing rewards. Patria was inspired by the Chinese social credit system, was developed by ZTE [27], and uses the Homeland ID Card to identify and link users across plans. The platform includes the Android vePatria app [9], the veMonedero app to connect the user wallet and receive bonuses, and the veQR app to keep track of social plans offered by the government.

Participants revealed that the Patria system provides (1) *activity awards*, for accounts that post around 50 tweets and at least 300 retweets a day, (2) a bonus if their posts receive significant engagement from other participants, and (3) monthly bonuses through the Carnet de la Patria system. Admins in the Tuiteros Activos Telegram group (§ 3.4) confirmed that the activity rewards are given on a weekly basis. Figure 11 (§ 4.7) further confirms that several interview participants had significant posting activity levels.

Several participants confirmed reports of the government's use of food distribution as a form of social control [65, 80]. Some claimed to receive monthly rewards (payments and food packages) also from individual politicians and organizations, e.g, *"They ask for publicity and offer one bag of food monthly with groceries, vegetables, proteins"* (PG9).

Seven participants claimed to receive no benefits for their activities. Six, both opposition and pro-government, explained that their motivation stems from strong political views and the desire to reveal the real situation in Venezuela, e.g., "*I only do this when I am at mad at the government, as a way to criticize*" (OP1), *"it is my mission to highlight the advantages of this political, social and economic system for us, the majority, who have been traditionally excluded"* (PG7).

These findings confirm the classification in § 4.1: participants in Venezuelan campaigns include paid and coerced operatives, and (at least part-time) grassroots campaigners. Our findings are also consistent with recent reports that campaigns are recruiting real people into their operations [36].

4.4 RQ2: International Influence Operations

Past Involvement: Spanish-Speaking Countries. Eleven participants, on both sides, claimed to contribute to campaigns for other Spanish-speaking countries. They explained that the contributions included (1) promoting certain hashtags, e.g., "*I worked the coup in Bolivia. We normally have a hashtag, something like #EvoEsPueblo [Evo Morales]*" (PG4), (2) tweeting, e.g., "*I publish political tweets for other countries when I see the risk, in this case, I see an extreme risk in Spain with Podemos*" (OP3), and (3) retweeting, e.g., "*I retweet the Nicaraguans, the Ecuadorians, the Cubans*" (PG8).

Most of these participants receive requests for help on their communication groups, from operatives in other countries. One participant finds and contributes to campaigns based on interest. None of the participants mentioned receiving explicit benefits for contributions to foreign campaigns. However, for Patria systems users, these activities may count toward their quota for, e.g., activity awards.

Willingness to be Hired. Fourteen participants said they would be willing to be hired to participate in influence operations on Twitter, including for the US. Their motivation included the impact of Trump's politics and the presence of many Latin Americans in the US. Some agreed conditionally, based on (1) payment, e.g., "*if payment is adequate and I can sustain myself, if I can buy a device able to withstand the work*" (PG5), (2) the campaign's political orientation e.g., "*If it does not go against my opinion*" (OP4), and (3) the correctness of the information to be promoted (OP3).

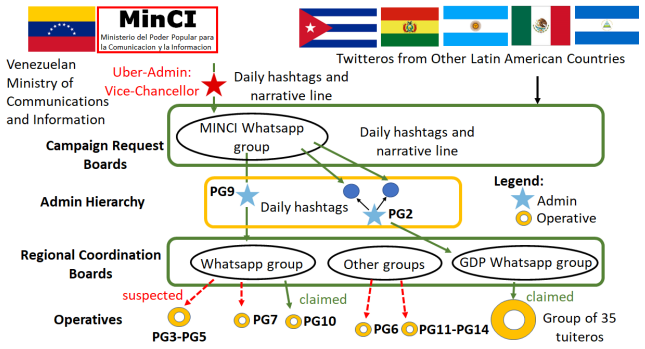

Figure 4: Organization structure inferred from pro-government participants. Campaign requests originate from MIPPCI and other Latin American groups, and are communicated and distributed to operatives through online groups organized by a hierarchy of admins.

Several participants claimed a keen interest in US politics, and a history of posting content on US politics. Figure 3 confirms these claims, showing the number of posts tweeted from the accounts of our participants that mention Trump (1,529 posts), Biden (1,141 posts), or Pelosi (49 posts).

4.5 RQ1: Organization and Communications

Several participants revealed their organization structure and communication mechanisms. Figure 4 shows information revealed by pro-government participants. Some report and receive instructions from the MIPPCI (§ 3.4) through a selective Whatsapp group: *"I am a member of the MIPPCI WhatsApp group.* **We use it to plan the hashtags for the next day**. *We are a group of around 200 people. Membership is selective, admission decisions are made by the vice-chancellor"* (PG9).

Several participants reported that requests also come from other countries (see Figure 4), e.g., *"I am a member of five international WhatsApp and Telegram groups where we share information that comes from different countries. Sometimes the 'tuiteros' from Nicaragua, Cuba, or Bolivia ask us for help and we support them."* (PG7). This confirms recent reports from Facebook about the emergence of influence operations that target both domestic and foreign audiences [36].

Two participants claimed to be *admins*, who organize other members through groups that promote each other's political content in Twitter (see regional boards in Figure 4). Consistent with previous findings [70, 82], they revealed hierarchical organizations: *"I am the admin of [anonymized group]. I have 3 sub-admins, and a total of 35 people under my command. [..]* **We are organized by regions at the national level.**" (PG2).

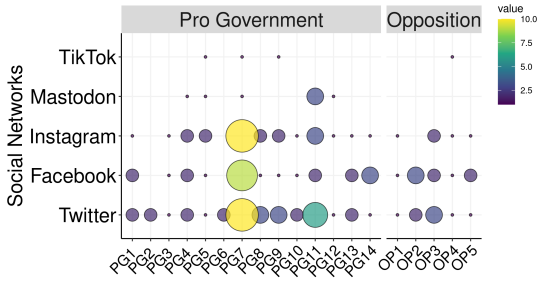

Figure 5: Per-participant number of accounts controlled on social networks. Dot sizes are proportional to the number of accounts. A few participants revealed sockpuppet accounts, but most claimed to own only backup accounts.

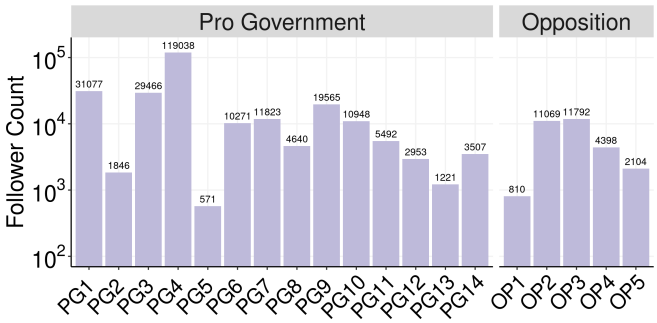

Figure 6: Number of followers for each participant at the beginning of the data collection. The y axis is in log scale. Nine participants have an audience of more than 10,000 followers.

Admins use these groups to (1) distribute the narrative line from the MIPPCI group, e.g., *"I tell my admins what they are going to do, and they are in charge of bringing that message down to the regions through regional groups"* (PG2), (2) coordinate activities, *"We coordinate there and then we publish in Twitter.* **We have shifts in which a certain group promotes content***"* (PG2), (3) identify and nudge members that are inactive, e.g., *"I evaluate daily to verify who is working and who is not. If someone is not active, I call them up"* (PG9), and (4) find new clients, *"we talk about content and we even run into clients there"* (PG11).

Grassroots participants, on both sides, also revealed less structured coordination, where they contribute to the efforts of multiple groups, e.g., *"We started with direct messaging groups from Twitter, we exchanged numbers, and we started creating Whatsapp and Telegram groups, we even have groups in Facebook and Instagram"* (PG5), "*I have three little WhatsApp groups that we created ourselves, where we sometimes share a tweet and we give retweets among all of us*" (PG4). In particular, opposition participants claimed to post content on their own, and work without an admin: "**we work alone but together***, we do not have an organization, we do not know each other but we have the same interests"* (OP3).

For both pro-government and opposition reports, we observe consistency with the "communication constitutes organization" perspective of organizational theory, that communication and organization co-produce and co-adapt [77]. Similar to volunteer organizations in disaster response [98], the loose coordination of the opposition enabled them to evolve into an effective organization that distributes information, garners engagement, and promotes hashtags to trending status (§ 4.7).

These findings confirm that influence operations are collaborative work [36, 70, 82, 96], whether through hierarchical structures consistent with previous reports [70, 83], or through flexible, decentralized structures. The decentralized infrastructure claimed by opposition participants is also consistent with their claims of persecution by the government (§ 4.8 and §4.9) and the documented news-accessing reliance of the general population on social networks and mobile apps [44, 68].

4.6 RQ1: Capabilities

We discuss participants' insights on accounts and followers.

Social Network Accounts. Figure 5 shows the social networks where our participants claimed to be active, and the number of accounts they claimed to control on each social network. Two participants revealed control of sockpuppet accounts, e.g., "*I have three accounts plus three institutional accounts for which I am the community manager. I also have a personal account.*" (PG11). PG7 claimed to have 9-10 accounts on each of Twitter, Facebook and Instagram. Three other participants each have three Twitter accounts.

Reasons for having multiple accounts include (1) separating personal from institutional accounts, e.g., "*One is institutional and the other is not so much political but I publish different things*" (PG13), and (2) separating political from personal accounts, e.g., *"I have family members with different political beliefs [..]* **I created separate accounts so to not impose my political messages onto them***"* (PG4).

In contrast, other participants control only one or two accounts in each social networks. Most explained that at most they have a backup account, in case of account suspensions, e.g., "*I have only two, my "hard" account, and another account that I have, just in case, because Twitter treats us badly.*" (PG1). Some explained that this was due to the difficulty of managing multiple accounts: *"I can barely manage two accounts, I cannot imagine how it would be like with many accounts, I wouldn't be able to do it"* (PG1).

Eleven participants revealed control of additional 15 Twitter accounts during the interview. We have manually compared these accounts against their 19 recruitment-validated accounts (§ 3.4). We confirmed that with the exception of the two accounts revealed by PG11, all participants use their online identity on the account profile and/or Twitter handle.

This suggests diverse strategies among our participants. While a few rely on sockpuppet accounts, confirming previ-

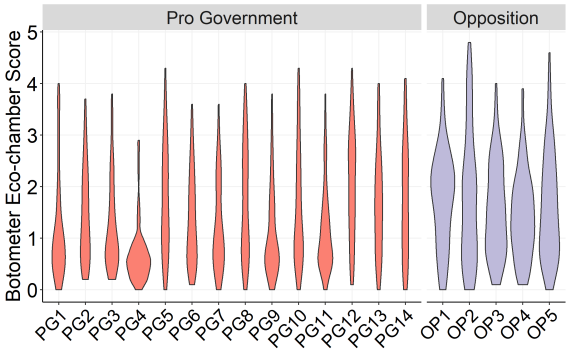

Figure 7: Per-participant distribution of Botometer echo chamber scores for a random sample of their followers.

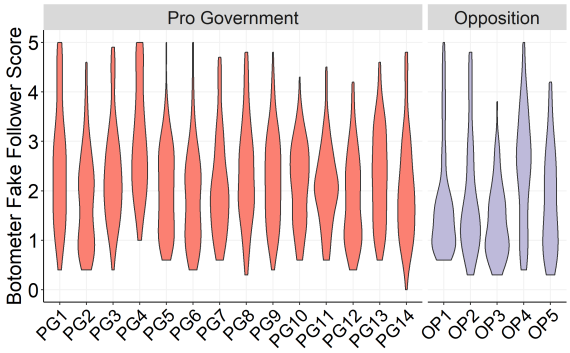

Figure 8: Per-participant distribution of Botometer fake follower scores for a random sample of their followers.

ous findings [54, 70, 82], we found many participants who only control a few personal accounts. This is consistent with participant reports of their use of the Patria system, where they need to register their Twitter account with the platform in order to receive rewards for their activities. This further supports recent Facebook reports that campaigns are starting to recruit real people into their amplification operations [36].

Followers. The followers of an account are vital for its impact. The numbers of followers revealed by participants are consistent with the ones we collected from their accounts. During several interviews, participants logged into their accounts in order to quote accurate numbers. Figure 6 shows the number of followers that we collected on February 2nd, 2021 from each participant. Our participants can reach a large audience: Nine participants had more than 10,000 followers, with the maximum being 119,038 followers (PG4).

To evaluate the ability of our participants to reach a wide audience, we used the Botometer tool [86] on a random sample of 100 followers from each participant. Botometer provides scores on a 0 - 5 scale, where high scores denote more likely bots, and scores in the middle denote uncertainty [4].

Figure 7 shows the per-participant distribution of their followers' Botometer echo-chamber scores (0 - 5 scale). This score signals accounts that engage in follow-back groups and share and delete political content in high volume [4]. Between 0 and 22.93% of the participants' followers had scores of at least 3 (M = 9.47, SD = 6.01). Figure 8 shows the per-participant distribution of Botometer fake follower scores for the same random sample of their followers. These scores identify bots purchased to increase follower counts [4]. Between 2% and 43.62% of the participants' followers had scores of at least 3 (M = 19.19, SD = 9.49).

Participants had 56.38% to 92% (M = 72.32, SD = 9.65) followers with both scores under 3. This suggests that while some participants had many suspicious followers, most also have significant numbers of genuine followers, and may reach a wide audience.

Several participants explained that their follower communities are smaller than what they should be, due to Twitter (1) removing subsets of them, (2) suspending their accounts, or (3) due to periods of inactivity, e.g., *"I was outside of Twitter for a year, because I was a political prisoner. During that time, many people stopped following me."* (OP2)

4.7 RQ1: Political Promotion Strategies

We discuss strategies to create and promote political content.

Daily Hashtags: Creation and Promotion. An admin participant provided insights into the creation of the daily hashtags, by the MIPPCI board where he is a member: *"We make hashtag proposals daily based on the political movement of the day. Once every hashtag has been proposed we start studying them [..] and cast our votes until we reach a consensus.* **The vice-chancellor has the last word***"* (PG9).

Admins distribute these hashtags through their groups, see Figure 4. Several participants explained that they monitor the posting of these daily hashtags, to include in their tweets in order to simultaneously (1) promote the hashtags, e.g., *"Everything I publish I accompany with the hashtag from the MIPPCI. We use the hashtag so that it gets more interaction"* (PG10), and (2) garner engagement for their own posts, *"If you take advantage of those first 5-10 mins after the hashtag is announced, the post will receive support [engagement] throughout the day. You can get up to 700 retweets"* (PG10).

We believe that the goal of these efforts is not only to make hashtags reach trending status, but also to maximize the time they stay in the top trending list: an account that uses a hashtag after it reaches a high rank, helps the hashtag stay trending.

We plotted the percentages and absolute counts of hashtags that (1) appeared in participant tweets or retweets in a three month interval, and (2) have become top-3 trending hashtags anywhere in Venezuela. Figure 9 shows that 16 of the 19 participant-controlled accounts, posted either a tweet or a retweet containing a hashtag that reached top-3 trending. Eleven participants posted at least one original tweet with a top-3 hashtag; up to 55% of hashtags included by one

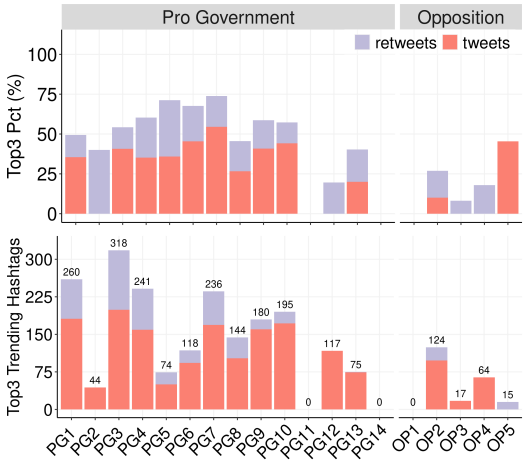

Figure 9: Trending hashtags. Top: Per-participant percentage of hashtags that appeared in a post, and were top-3 trending anywhere in Venezuela. Bottom: Absolute hashtag counts.

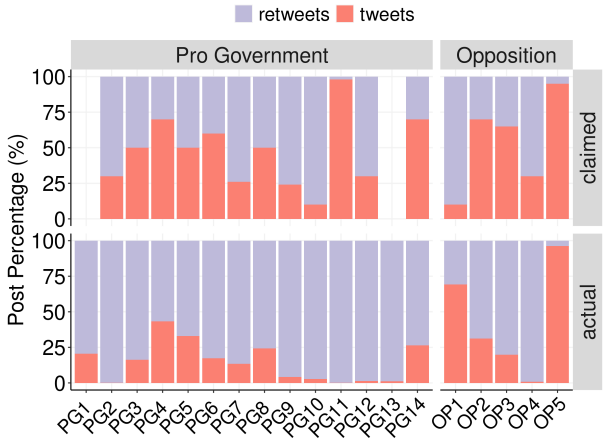

Figure 10: Participant tweets (red, bottom) vs. retweets (purple, top). Participant-claimed percentages top, real percentages bottom. PG1 and PG13 did not provide an answer.

participant (PG7) in original tweets had a top-3 trending rank.

These hashtag-promoting attacks compromise information integrity because they fraudulently promote the search rank of desired hashtags, and simultaneously demote other hashtags promoted (perhaps organically) by opposing campaigns. We note that this attack differs from the ephemeral astroturfing attacks of Elmas et al. [35] in that it (1) involves humans instead of astrobots, and (2) does not seek to remain invisible, e.g., by erasing hashtag-promoting tweets.

Content Creation vs. Engagement: Perception. Sources of inspiration for the content created for original tweets include the mission and vision of the client (for participants who claimed to work for various clients) and also news portals: *"We access news portals like RT Actualidad [Russian news outlet], HispanTV [Iranian news outlet in Spanish], and CNN"* (PG10). We confirm that 219 posts of our participants included links to Russian [2, 11], Iranian [7] and Cuban [5, 6] news outlets, and 559 posts had links to government sites. These behaviors confirm strategies of integration of government and externally-funded media as source content for narratives, previously reported in countries like Syria [96, 97].

Pro-government participants confirmed their use of information distributed through MIPPCI, e.g., *"We receive daily the information from the MIPPCI. They say, look, today's line is this .. and so we read the news release and we compose the final content with our own words."* (PG2).

Participants explained the importance of the audience reached and engagement received (number of times people interacted with their tweets, i.e., retweets, quotes, replies and likes) by their tweets. They are used by admins for evaluation purposes, e.g., "*admins look at the amount of retweets that I received, the amount of followers that the account has gained, the projection of the publications"* (PG11), and also by the Patria system [12] (§ 4.3) to assign bonuses, e.g., "*My account is mentioned quite a lot, and the more engagement I have, this bonus arrives without me needing to publicize"* (PG10).

A widespread theme is a peer-based strategy to acquire engagement, where participants share their tweets in communication groups, and retweet the tweets posted there by other members, e.g., "*When I see a tweet from someone in my group, I retweet immediately. This is the work each and every one of us does.*" (PG2), "*I am a member of six groups, I go into each group and I retweet whatever they publish during the day, no matter the content.*" (PG10). Participants explained that they expect to receive engagement for their tweets, once they share them in their groups. This likely results in lockstep behaviors, which can be exploited to detect influence operations (§ 5).

Participants further reported unexpected strategies, i.e., (1) receiving requests for retweets through direct messages, (2) retweeting their own tweets, and (3) mentioning select accounts in their tweets to encourage reciprocation: *"On every tweeted news I mention at least six accounts of people that follow me and consistently retweet the information that I post. I look up the number of their followers so that I know that it is worth mentioning their account "* (OP2).

We confirmed that most participants have posted or retweeted such content. Some explained that they preferentially retweet posts from certain sources, e.g., *"We retweet the information from the government work and political figures, and show the work done by state institutions"* (PG11). This explains our reports of hyperpartisan news with images from staged events (§ 4.2). The "Tuiteros Activos" Telegram group (§ 3.4) had claims that the Patria system gives bonuses only for retweets of accounts controlled by government officials.

We further investigated the participant perception of the distribution of their original posts versus retweets. Figure 10 (top) shows claimed percentages. A majority of participants claimed to post a mix of original content and retweets, thus

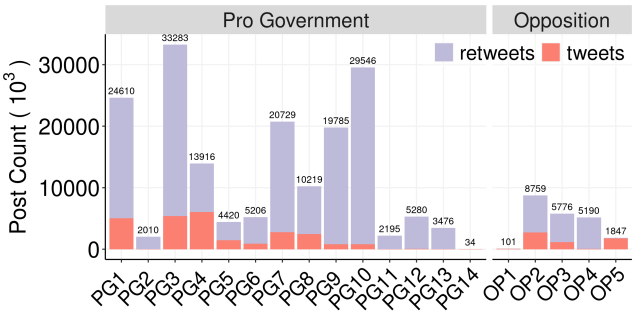

Figure 11: Number of tweets vs. retweets collected from participant accounts over three months. We observe substantial efforts to create engagement for content posted by others.

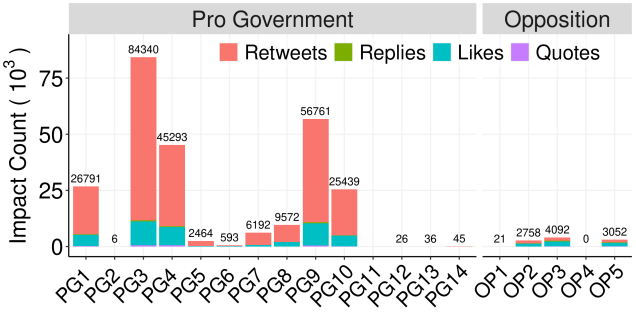

Figure 12: Engagement received over one month, by original posts of our participants. We observe influencer potential or a well-oiled propaganda machine, for several participants.

to be both content and engagement creators.

One participant motivated this strategy by the need to appear influential to clients, *"It does not look good if the people that hire us see that all we do is retweet. So, we try to have more tweets than retweets in our main accounts"* (PG9). Two opposition participants said they post more retweets due to self-censorship, e.g., *"I sometimes express an opinion, but very little, because they punish people, politically. If you go directly against the government, then they look for you"* (OP4).

Content Creation vs. Engagement: Twitter Truth. To take steps toward verifying several of these claims, we collected all the Twitter posts of the interview participants over three months. Figure 10 (bottom) shows the real percentages during this interval. Figure 11 shows the absolute values. Only two participants (OP1 and OP5) posted more original tweets than retweets. In fact, seven participants have posted less than 65 original tweets each, over three months.

To analyze engagement claims, we collected the per-participant engagement, i.e., the total number of retweets, replies, likes and quotes received by the original tweets posted from their Twitter accounts during a one month interval. We collected this engagement almost one month after the end of the posting interval. Figure 12 plots this data. PG11's account

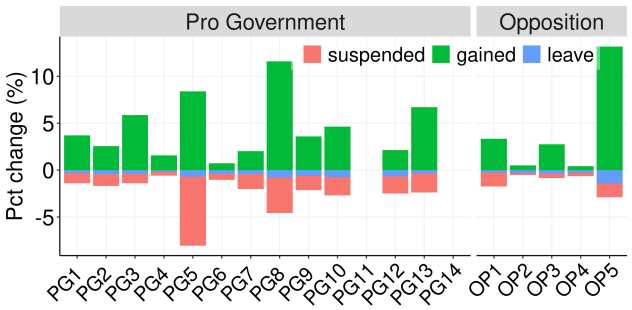

Figure 13: Follower percentage change from the initial count, after two months. Overall, most participants gained followers.

was suspended on the day we collected this snapshot and OP4 had not received any engagement during the period examined. We observe that the posts of five participants received towering engagement, each with a total over 25,000. PG3 had a one-month engagement of 84,340. The average per-tweet engagement of PG9 and PG10 was 291 and 189 respectively.

4.8 RQ4: Exploration of Twitter Defenses

Interview participants reported a suite of penalties they experienced in Twitter. We discuss these in the following.

Account Closure, Restriction, Suspension. Most participants confirmed to have had at least one account suspended by Twitter. Several reported frequent suspensions, that prevent them from accessing their accounts for days, e.g., *"That account got blocked, suspended, restricted, it used to be between 5 days to a week when I could not use it"* (PG4). During our monitoring interval, we recorded five suspension events for the accounts of our participants.

Several participants revealed that Twitter directly closed their accounts, e.g., *"I had another account that got canceled, not even suspended. The SEBIN [Venezuela's political police] objected [to Twitter] and asked information about the account. Twitter didn't give any information to SEBIN. Thank God, otherwise I wouldn't be talking with you today"* (OP3).

Pruning of Followers. Two participants claimed that Twitter removed followers from their accounts on multiple occasions, e.g., *"My account used to have 8,500 followers and after the first suspension it had 1,100 followers. After that, it reached 12,000 and after the second suspension they returned it with 9,800."* (PG10). He surmised that such events could also occur because some of his follower accounts were suspended by Twitter. Figure 13 shows the per participant number of follower accounts that were suspended by Twitter during a two months interval. It provides evidence toward confirming these claims.

Shadowbans and Content Flags. Two participants reported that Twitter shadowbanned their accounts. They mentioned Shadowban [10], a webservice popular in their community, to detect if an account has been shadowbanned. One participant described an experiment performed by her Twitter group to

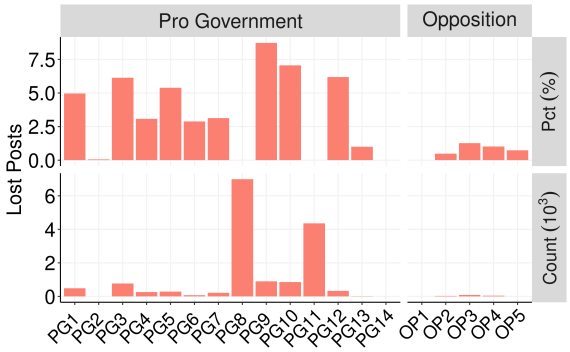

Figure 14: Number of posts over two months, that were suspended or deleted by Twitter one month later: (bottom) actual values in thousands, (top) percentage from their total posts.

discover inconsistent counting of retweets by Twitter: *"We have a group of 50 people. We each posted exactly the same tweet. We then all retweeted that same tweet for everyone. So, every tweet should have 50 retweets. One didn't reach 10, some reached 20, some other reached 30-40ish"* (OP3).

Figure 14 shows the number of tweets and retweets that were posted by the participants over two months, and were suspended or deleted by Twitter one month later. Only retweets were removed, and most because the posting account was suspended by Twitter. The accounts of PG8 and PG11 were in a suspended state during this interval.

4.9 RQ4: Detection Avoidance and Recovery

Participants revealed several strategies to bypass Twitter's penalties. We discuss these in the following.

Rate Limiting Efforts vs. Activity Quotas. Several participants explained that the above penalties are due to high posting rates, e.g., "*if I do too many retweets and replies Twitter interprets this as if I was a robot*" (PG5). Even short bursts of tweets can lead to account suspensions, "*after the 6th or 7th tweet, bam, my account would get restricted.*" (PG4).

To avoid detection, participants report limiting their posting rates, e.g., *"When I get to around 20 retweets I stop, I logout of the account because the limitation may be about to pop up"* (PG5). They also claimed to space out their posts, e.g., *"In the morning I posted 20 tweets. Then I wait 2 hours, and I post 20 more"* (PG11). This is consistent with strategies shared on Telegram groups, of waiting at least 5s between posts, or posting five tweets slowly every 10 mins.

We find a conflict for participants who reported quotas on their daily original tweets. Quotas include (1) upper bounds, e.g., "*We do not go beyond 10 posts a day when we are hired"* (PG11), (2) lower bounds, e.g., *"Our job is to post 10 tweets a day, but if you want more, it's fine"* (PG2), and (3) contract-based, e.g., *"Depends on the contract I have with my customer, the payment they offer"* (PG9). We exploit this conflict in § 5.

Figure 11 shows that 17 participants posted more than 10 tweets and retweets on average per day over three months, with nine posting more than 60; PG3 posted 341 daily posts.

Backup Accounts Several participants claimed that to recover from longer or permanent suspensions, they create *backup accounts* whose names are a variation of their main account name. We analyzed the 15 additional Twitter account handles revealed by 11 participants during interview. These participants revealed each between 1 and 3 additional Twitter accounts. For each of the nine participants whose revealed accounts were accessible, we confirmed that these accounts have either similar Twitter handles or the same screen name. We conjecture that this occurs because of the participant need to confirm identity for the government Patria app and receive bonuses for their online activities (see § 4.3).

Avoiding Suspension Wars. Several participants claimed that Twitter penalties (§ 2.3) are due to reports from other accounts, resulting in a *suspension war*: *"People can create up to 10 accounts to report someone else.* **The government collectives use large networks to annihilate opposition accounts***"* (OP3). Twitter does provide mechanisms for users to report tweets that they consider abusive or harmful [16]. Participants avoid their tweets being reported by others [16] by being nice to their social entourage, e.g., *"I have never blocked or reported anyone. This is my policy. I accept anything, anyone can comment whatever they want. If I do not like what you publish, I do not follow you"* (PG9).

Avoiding Post Automation. While a few participants know others who use Hootsuite [8] and Tweetdeck [14] to schedule tweets, most use only the official Twitter app or Web UI to post content. This is due to fear of detection: *"Twitter closes your account because you robotized. You can detect bots because they are programmed to post at certain times"* (OP2). Our qualitative analysis mostly confirms these claims. Only a few of our participants have used tools to post their tweets: dlvr.it [34] (PG6, 181 tweets), Instagram (PG5, 14 tweets) and Tweepsmap [13] (PG9, 1 tweet; PG7, 2 tweets). In § 5 we discuss the potential to identify suspected influence operations participants, among accounts that use automation.

Further detection avoidance and recovery strategies include (1) careful management of accounts, IP addresses and browsers, (2) using self-censorship to avoid toxic language and offensive images, and (3) appealing account suspensions.

5 Information Assurance

We now discuss discovered vulnerability points (VP) of IO participants, summarized in Figure 1. We then suggest changes to social networks' handling of influence operations.

5.1 IO Vulnerability Points (VPs)

VP1: Identify and Penalize Daily Hashtags. Hashtags promoted by operatives can be found by monitoring communications groups used by operatives (§ 3.3). Such groups often have open membership, e.g., "*Maybe I put up a good*

word about you [the interviewer] with the administrator of the group. Everyone is looking for people that do retweets" (PG5). Hashtags posted in these groups are available to all members, and can be validated against posts from hyperpartisan Twitter accounts (e.g., @mippcivzla).

VP2: Patria System. The Patria system [12] reported by pro-government participants (§ 4.3) uses the Twitter API to link Twitter accounts and to keep track of tweet counts to rank and pay participants. While Twitter could block Patria's access to the Twitter API, a smarter strategy is to determine if apps like the vePatria, veMonedero and/or veQR, are co-installed on the user's device. Twitter can then use other device information (e.g., IP address, model) to identify the Twitter accounts registered on the device. Platforms like Twitter can also generalize this approach to identify other apps that use their APIs and are co-installed with their client. This would enable them to detect operatives active through other frameworks and countries.

VP3: Lockstep Behaviors. Participants revealed lockstep behaviors that arise from their rush to include newly released daily hashtags into their posts, and their peer-based strategy to acquire engagement (§ 4.7). We observe the opportunity to leverage existing lockstep behavior detection solutions [31,43, 56, 93, 100, 108], to detect groups of social network accounts with synchronized behaviors.

VP4: Disinformation and Political Outlets. Our finding that participants contribute to the distribution of disinformation and articles from hyperpartisan outlets (§ 4.2) suggests the opportunity to leverage existing efforts to detect misinformation and disinformation [55, 89, 91], and identify accounts responsible for posting or promoting such content.

VP5: Activity Quotas vs. Rate Limits. In § 4.9 we found that IO participants need to post and amplify many messages daily, while simultaneously avoiding detection and account suspensions. Even though participants reported limiting their posting rates, our analysis of their accounts revealed that in reality, many continue to post significant daily content over long periods of time. This suggests the ability to detect accounts with suspicious levels of activity, including substantial activity bursts [45,49,56–58,66,78,105], e.g., associated with the release of the daily hashtags.

VP6: Post Automation. To sustain high levels of activity, operatives may use post automation tools, see § 4.7. Participants observed that accounts that use post automation tend to post at predictable times (§ 4.9). This suggests further opportunities to identify automated accounts.

5.2 Proposed Next Steps

Our study reveals that influence operations continue to thrive despite social network defenses that include suspending accounts or shadowbanning posts. Based on our findings, we suggest that social networks could instead implement the following, more nuanced approach toward influence operations, that leverages knowledge of the different IO participant types.

Classify IO Participants. Different IO participant types (human operatives, grassroots campaigners, unwitting targets, trolls, bots) should be treated differently. We conjecture that features emerging from the above VPs (e.g., number of daily hashtags promoted, number of Patria apps installed, number and amplitude of activity spikes, counts of disinformation posted/promoted), could be used to train models to detect and even classify influence operations participants.

Previous work suggests promise for such an approach. Saeed et al. [84] found that Reddit trolls differ from regular users, e.g., in their loose coordination activities. They leveraged these differences to develop relevant features and train a Reddit troll-detection model. Volkova and Bell [103] extracted features and trained a model to detect accounts involved in the 2014 Ukraine-Russia conflict, that were deleted by Twitter. Luceri et al.'s [60] inverse reinforcement learning-based discovery that Russian Twitter trolls differ in their behavior when engaged by others or when their content is reshared, further suggests potential to generalize previous work to similarly identify other types of IO participants.

Monitor, Don't Censor. Study participants revealed strategies to groom backup accounts to address account suspensions, and also techniques to study shadowbanning. Instead, monitoring the activities of detected accounts would allow social networks to identify IO strategy shifts in real time, and implement subtler mechanisms to reduce the reach and impact of influence operations, and even turn them into echo chambers. We suggest two directions:

- **Nudge Unwitting Participants**. Social networks could deploy interventions to nudge unwitting targets of influence operations (§ 2) toward safer behaviors. This includes extending the social network client to signal to detected unwitting participants that certain accounts they follow have inauthentic behaviors, and suggest unfollowing such accounts. Further, signal when viewed posts are suspected of being promoted by influence operations, and suggest avoiding engagement. Keiser et al. [50] provide evidence that disinformation warnings that interrupt the user and require interaction can inform and guide user behaviors.
- **One Device One Vote**. For Venezuelan operations, the difficulty to access mobile devices can be used to thwart attempts by operatives to game the system, e.g., the search rank of hashtags. For instance, Twitter's hashtag ranking algorithm could reduce the weight of contributions from devices associated with influence operations, e.g., to ensure that each device can contribute at most one vote per hashtag.

6 Discussion and Limitations

Pro-government vs. Opposition IOs. Our analysis reveals a complex dynamic between pro-government and opposition participants, that have different motivations, organizations, technical ecosystems, adversaries, and strategies. We also found common goals that include (1) preventing the suspen-

sion of their accounts, (2) ensuring their continued access to Twitter and their follower base after an account suspension, (3) acquiring many followers, from diverse groups, (4) receiving engagement, (5) promoting key hashtags to trending status, and (6) creating and distributing content that challenges the version of events of opposing factions.

Sockpuppets vs. Real Users. While a few participants rely on sockpuppet accounts, confirming previous findings [54, 70, 82], we found many participants who only claim to control a few personal accounts. This supports recent Facebook reports that campaigns are starting to recruit real people into their amplification operations [36].

IO = Collaborative Work. Our work confirms that influence operations are collaborative work [36, 70, 82, 96], whether through hierarchical structures consistent with previous reports [70, 83], or through flexible, decentralized structures. The communications technologies that form the basis of both structure types point at organizations of both pro-government and opposition participants [77]. We reveal a social network influence war between participants supporting opposing sides, where each side seeks to increase its influence and reach, and thwart the opponent's efforts.

Effectiveness and Loopholes of Twitter Defenses. We observe that Twitter's defenses are effective to a certain degree to degrade the efforts of influence operations (§ 4.8). However, we also found that participants are pushing their activities beyond the limit admissible by Twitter, experiment with its defenses, learn from its responses, and share their lessons (§ 4.8 and § 4.9). Nevertheless, despite finding that participants have posting strategies that successfully evade Twitter detection, we also identified behaviors that continue to render them vulnerable to detection (§ 5).

Impact of Findings. Ethical Considerations Revisited. The goal of this study is not to help social networks detect and punish operatives. Instead, we leverage our findings to propose a shift from censoring to monitoring operatives, and detecting and nudging their unwitting targets. We believe that human participants will continue to play an important role in influence operations. However, the approach we propose in § 5.2 may reduce their perception of heavy-handed social network interference: operatives would no longer lose followers due to automatic suspensions, but only when their unwitting targets make an explicit effort to unfollow them. Their posts would no longer be shadowbanned, but may organically experience a reduced engagement from unwitting targets.

Giving users more information and control may thus also improve the public's trust in social networks. It may also indirectly nudge operatives to post higher quality, more trustworthy content, to avoid alienating their followers.

Our approach may also lead to strategy changes for the IO command and control (§ 2.2). For instance, they could provide more resources to participants (e.g., devices), and reduce the number of posts required by the Patria system to provide activity awards (now at 50 tweets and 300 retweets/-day). This may reduce the detectability of operative accounts. They could also use better bots, that leverage, e.g., generative adversarial networks, to emulate human behaviors.

Limitations. Our recruitment process was biased due to only contacting active Twitter users whose accounts were not bots, had at least 500 followers, were open to DMs from our accounts, read our DMs, and consented to our terms. We also observe the unexpected high mean participant age (50). This is perhaps due to older people having more time and willingness to discuss their experiences, among the accounts that we reached. We cannot claim that this age distribution applies to all people who post political content in Venezuela. We also note that our study focused on Venezuela, but cannot provide a complete picture of influence campaigns in Venezuela. Further, our findings do not apply to operations in other countries.

However, we present results from the first qualitative study conducted from the perspective of participants in influence operations, on their experiences and challenges associated with their online activities.

The quantitative analysis-based validation of participant claims is limited since we cannot identify all the accounts they control. Thus, our analysis can only validate a subset of the claims made by participants. This limitation applies to participants that claimed ownership of multiple, sockpuppet accounts, but are less obvious for participants that claimed only backup accounts. We confirm that all the accessible accounts revealed by interview participants were linked to their owner's identity.

7 Conclusions

In this paper we have reported findings from interviews with 19 influence operation participants that targeted Venezuela, and a quantitative investigation with data collected from participant-controlled accounts. We found that pro-government and opposition participants use similar content-promotion strategies, but have marked differences in their motivation, organization, technical solutions, adversaries, and detection avoidance strategies. Our findings complement previous work, suggesting a strategy adjustment for influence operations in Venezuela. We reveal however vulnerabilities of existing influence operations strategies, and suggest detection solutions.

8 Acknowledgments

This research was supported by NSF grants CNS-2013671 and CNS-2114911, and CRDF grant G-202105-67826. This publication is based on work supported by a grant from the U.S. Civilian Research & Development Foundation (CRDF Global). Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of CRDF.

References

[1] 300,000 Russian Troll Tweets. Kaggle, https://www.kaggle.com/vikasg/russian-troll-tweets.

[2] Actualidad RT en Espanol. actualidad.rt.com.

[3] Alex Saab. The Hill, https://thehill.com/people/alex-saab/.

[4] Botometer FAQ: What are the bot type scores? https://botometer.osome.iu.edu/faq.

[5] Cuba Informacion. cubainformacion.tv.

[6] Cuba Si. cubasi.cu.

[7] Hispan TV. hispantv.com.

[8] Hootsuite. www.hootsuite.com.

[9] JMT ST C.A (Developer of vePatria App). AppBrain, https://www.appbrain.com/dev/JMT+ST+C.A/.

[10] Shadowban. https://shadowban.eu/.

[11] Sputnik Mundo. mundo.sputniknews.com.

[12] The Patria System. https://www.patria.org.ve/.

[13] Tweepsmap. https://tweepsmap.com/.

[14] TweetDeck. https://tweetdeck.twitter.com/.

[15] Twitter: About suspended accounts. https://help.twitter.com/en/managing-your-account/suspended-twitter-accounts.

[16] Twitter: Report a Tweet, List, or Direct Message. https://help.twitter.com/en/safety-and-security/report-a-tweet.

[17] Twitter Terms of Service. https://twitter.com/en/tos.

[18] En cinco años, 55 medios impresos dejaron de circular en Venezuela. LaPatilla,, https://tinyurl.com/mrxbhbjj, 2018.

[19] El 79.9% de los venezolanos quiere que Maduro negocie ya su salida (encuesta flash Hercon). La Patilla, https://tinyurl.com/2s5txrk7, 2019.

[20] Treasury Disrupts Corruption Network Stealing From Venezuela's Food Distribution Program, CLAP. US Department of Treasury, https://home.treasury.gov/news/press-releases/sm741, 2019.

[21] An Enduring Relationship – From Russia, With Love. Center for Strategic and International Studies, https://www.csis.org/blogs/post-soviet-post/enduring-relationship-russia-love, 2020.

[22] Report: RT and Sputnik's Role in Russia's Disinformation and Propaganda Ecosystem. US Department of State, https://www.state.gov/report-rt-and-sputniks-role-in-russias-disinformation-and-propaganda-ecosystem/, 2022.

[23] Norah Abokhodair, Daisy Yoo, and David W McDonald. Dissecting a social botnet: Growth, content and influence in twitter. In *Proceedings of the 18th ACM conference on computer supported cooperative work & social computing*, pages 839–851, 2015.

[24] Aseel Addawood, Adam Badawy, Kristina Lerman, and Emilio Ferrara. Linguistic cues to deception: Identifying political trolls on social media. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 13, pages 15–25, 2019.

[25] Meysam Alizadeh, Cody Buntain, Jacob N. Shapiro, and Joshua Tucker. Are influence campaigns trolling your social media feeds? The Washington Post, https://tinyurl.com/56vuv656, 2020.

[26] Christopher A. Bail, Lisa P. Argyle, Taylor W. Brown, John P. Bumpus, Haohan Chen, M. B. Fallin Hunzaker, Jaemin Lee, Marcus Mann, Friedolin Merhout, , and Alexander Volfovsky. Exposure to Opposing Views on Social Media Can Increase Political Polarization. In *Proceedings of the National Academy of Sciences*, volume 115, page 9216–9221, 2018.

[27] Angus Berwick. How ZTE helps Venezuela create China-style social control. Reuters, https://tinyurl.com/3k7ween3, 2018.

[28] Christopher Bing, Elizabeth Culliford, and Paresh Dave. Spanish-language misinformation dogged Democrats in U.S. election. Reuters, https://tinyurl.com/yck4xyu2, 2020.

[29] Ladislav Bittman. *The KGB and Soviet disinformation: an insider's view*. Pergamon-Brassey, 1985.

[30] Dearbhail Bracken-Roche, Emily Bell, Mary Ellen Macdonald, and Eric Racine. The concept of 'vulnerability' in research ethics: an in-depth analysis of policies and guidelines. *Health research policy and systems*, 15(1):8, 2017.

[31] Qiang Cao, Xiaowei Yang, Jieqi Yu, and Christopher Palow. Uncovering large groups of active malicious accounts in online social networks. In *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security*, page 477–488, 2014.

[32] Atlantic Council's DFRLab. #AlertaVenezuela. Atlantic Council, https://www.atlanticcouncil.org/content-series/alertavenezuela/alertavenezuela-april-21-2020/, 2021.

[33] Nicholas Dias, Gordon Pennycook, and David G Rand. Emphasizing publishers does not effectively reduce susceptibility to misinformation on social media. *Harvard Kennedy School Misinformation Review*, 2020.

[34] dlvr.it. https://dlvrit.com/.

[35] Tuğrulcan Elmas, Rebekah Overdorf, Ahmed Furkan Özkalay, and Karl Aberer. Ephemeral astroturfing attacks: The case of fake Twitter trends. In *Proceedings of the IEEE European Symposium on Security and Privacy (EuroS&P)*, pages 403–422, 2021.

[36] Facebook. Threat Report: The State of Influence Operations 2017-2020. https://tinyurl.com/bdcx5m4n, 2021.

[37] Henry Farell. The Chinese government fakes nearly 450 million social media comments a year. The Washington Post, tinyurl.com/2wxu7bta, 2016.

[38] Emilio Ferrara, Onur Varol, Clayton Davis, Filippo Menczer, and Alessandro Flammini. The rise of social bots. *Communications of the ACM*, 59(7), 2016.

[39] FiveThirtyEight. 3 Million russian Troll Tweets. GitHub, https://github.com/fivethirtyeight/russian-troll-tweets/.

[40] Dan Frommer. Twitter's list of 2,752 russian trolls. tinyurl.com/2jbe4vck.

[41] Greg Guest, Arwen Bunce, and Laura Johnson. How many interviews are enough? An experiment with data saturation and variability. *Field methods*, 18(1), 2006.

[42] Drew Harwell and Mariana Zuniga. Social media remains key to Venezuela's opposition, despite efforts to block it. The Washington Post, https://tinyurl.com/485nvfac, 2019.

[43] Nestor Hernandez, Mizanur Rahman, Ruben Recabarren, and Bogdan Carbunar. Fraud de-anonymization for fun and profit. In *Proceedings of the 25th ACM Conference on Computer and Communications Security*, 2018.

[44] Isayen Herrera. How Venezuela's vice grip on the internet leaves citizens in the dark during crises. NBC News, https://tinyurl.com/ysyz55ma, 2019.

[45] Bryan Hooi, Neil Shah, Alex Beutel, Stephan Günnemann, Leman Akoglu, Mohit Kumar, Disha Makhija, and Christos Faloutsos. Birdnest: Bayesian inference for ratings-fraud detection. In *Proceedings of the SIAM International Conference on Data Mining*, 2016.

[46] Jane Im, Eshwar Chandrasekharan, Jackson Sargent, Paige Lighthammer, Taylor Denby, Ankit Bhargava, Libby Hemphill, David Jurgens, and Eric Gilbert. Still out there: Modeling and identifying russian troll accounts on Twitter. In *Proceedings of the 12th ACM Conference on Web Science*, 2020.

[47] S. K. Jan, Q. Hao, T. Hu, J. Pu, S. Oswal, G. Wang, and B. Viswanath. Throwing Darts in the Dark? Detecting Bots with Limited Data Using Neural Data Augmentation. In *Proceedings of the IEEE Symposium on Security and Privacy (SP)*, pages 1729–1745, 2020.

[48] R Tallal Javed, Muhammad Usama, Waleed Iqbal, Junaid Qadir, Gareth Tyson, Ignacio Castro, and Kiran Garimella. A deep dive into covid-19-related messages on whatsapp in pakistan. volume 12, page 5, 2022.

[49] Parisa Kaghazgaran, James Caverlee, and Anna Squicciarini. Combating crowdsourced review manipulators: A neighborhood-based approach. In *Proceedings of the 11th ACM International Conference on Web Search and Data Mining*, page 306–314, 2018.

[50] Ben Kaiser, Jerry Wei, Eli Lucherini, Kevin Lee, J Nathan Matias, and Jonathan Mayer. Adapting security warnings to counter online disinformation. In *Proceedings of the 30th USENIX Security Symposium*, pages 1163–1180, 2021.

[51] Ashkan Kazemi, Kiran Garimella, Gautam Kishore Shahi, Devin Gaffney, and Scott A Hale. Tiplines to Combat Misinformation on Encrypted Platforms: A Case Study of the 2019 Indian Election on WhatsApp. In *arXiv e-prints*, 2021.

[52] Gary King, Jennifer Pan, and Margaret E Roberts. How the chinese government fabricates social media posts for strategic distraction, not engaged argument. *American political science review*, 111(3):484–501, 2017.

[53] Fedor Kozlov, Isabella Yuen, Jakub Kowalczyk, Daniel Bernhardt, David Freeman, Paul Pearce, and Ivan Ivanov. Evaluating changes to fake account verification systems. In *23rd International Symposium on Research in Attacks, Intrusions and Defenses ({RAID} 2020)*, pages 135–148, 2020.

[54] Srijan Kumar, Justin Cheng, Jure Leskovec, and V.S. Subrahmanian. An army of me: Sockpuppets in online discussion communities. In *Proceedings of the 26th International Conference on World Wide Web*, pages 857–866, 2017.

[55] David MJ Lazer, Matthew A Baum, Yochai Benkler, Adam J Berinsky, Kelly M Greenhill, Filippo Menczer, Miriam J Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, et al. The science of fake news. *Science*, 359(6380):1094–1096, 2018.

[56] Huayi Li, Geli Fei, Shuai Wang, Bing Liu, Weixiang Shao, Arjun Mukherjee, and Jidong Shao. Bimodal distribution and co-bursting in review spam detection. In *Proceedings of the 26th International Conference on World Wide Web*, page 1063–1072, 2017.

[57] Shanshan Li, James Caverlee, Wei Niu, and Parisa Kaghazgaran. Crowdsourced App Review Manipulation. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, page 1137–1140, 2017.

[58] Ee-Peng Lim, Viet-An Nguyen, Nitin Jindal, Bing Liu, and Hady Wirawan Lauw. Detecting product review spammers using rating behaviors. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, page 939–948, 2010.

[59] D. L. Linvill, , and P. L. Warren. Troll factories: The internet research agency and state-sponsored agenda building. Resource Centre on Media Freedom in Europe, 2018.

[60] Luca Luceri, Silvia Giordano, and Emilio Ferrara. Detecting troll behavior via inverse reinforcement learning: A case study of russian trolls in the 2016 us election. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, 2020.

[61] Neil MacFarquhar. Inside the Russian Troll Factory: Zombies and a Breakneck Pace. New York Times,

https://www.nytimes.com/2018/02/18/world/europe/russia-troll-factory.html, 2018.

[62] Evita March. Psychopathy, sadism, empathy, and the motivation to cause harm: New evidence confirms malevolent nature of the internet troll. *Personality and Individual Differences*, 141:133–137, 2019.

[63] Tom McCarthy. How Russia used social media to divide Americans. The Guardian, https://tinyurl.com/32dz738m, 2017.

[64] Nora McDonald, Sarita Schoenebeck, and Andrea Forte. Reliability and inter-rater reliability in qualitative research: Norms and guidelines for cscw and hci practice. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–23, 2019.

[65] Mogollon Mery. Food isn't just a dire need in venezuela — it has become a major political tool. *Los Angeles Times*, 2019.

[66] Shirin Nilizadeh, Hojjat Aghakhani, Eric Gustafson, Christopher Kruegel, and Giovanni Vigna. Think outside the dataset: Finding fraudulent reviews using cross-dataset analysis. In *The World Wide Web Conference*, page 3108–3115, 2019.

[67] Ben Nimmo and Aric Toler. The Russians Who Exposed Russia's Trolls. DFRLab, https://medium.com/dfrlab/the-russians-who-exposed-russias-trolls-72db132e3cd1, 2018.

[68] Ciara Nugent. 'Venezuelans Are Starving for Information.' The Battle to Get News in a Country in Chaos. Time Magazine, https://time.com/5571504/venezuela-internet-press-freedom/, 2019.

[69] U.S. House of Representatives. Social Media Advertisements. https://intelligence.house.gov/social-media-content/social-media-advertisements.htm.

[70] Jonathan Corpus Ong and Jason Vincent A. Cabañes. Architects of networked disinformation: Behind the scenes of troll accounts and fake news production in the philippines. Scholarworks UMassAmherst, 2018.

[71] José Ospina-Valencia. How Russia is waging a successful propaganda war in Latin America. DW, https://tinyurl.com/3rw7rj5w, 2022.

[72] Tim Padgett. The Disaster That Is Venezuela. The New York Times, https://tinyurl.com/mr36y3fa, 2022.

[73] Christopher Paul and Miriam Matthews. The Russian "firehose of falsehood" propaganda model. *Rand Corporation*, 2(7):1–10, 2016.

[74] Gordon Pennycook and David G Rand. Fighting misinformation on social media using crowdsourced judgments of news source quality. *Proceedings of the National Academy of Sciences*, 116(7):2521–2526, 2019.

[75] Peter Pomerantsev and Michael Weiss. *The menace of unreality: How the Kremlin weaponizes information, culture and money*, volume 14. Institute of Modern Russia New York, 2014.

[76] Man pui Sally Chan, Kathleen Hall Jamieson, and Dolores Albarracin. Prospective associations of regional social media messages with attitudes and actual vaccination: A big data and survey study of the influenza vaccine in the united states. *Vaccine*, 38(40), 2020.

[77] Linda L Putnam and Anne M Nicotera. *"Building Theories of Organization: The Constitutive Role of Communication*. Routledge, 2009.

[78] Mizanur Rahman, Nestor Hernandez, Ruben Recabarren, Syed Ishtiaque Ahmed, and Bogdan Carbunar. The art and craft of fraudulent app promotion in google play. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, pages 2437–2454, 2019.

[79] Reddit. Suspicious accounts investigated. https://www.reddit.com/wiki/suspiciousaccounts.

[80] Rosario De Souza Renato. Venezuela: UN report urges accountability for crimes against humanity. *United Nations Human Rights Council*, 2020.

[81] Moises Rendon. The Fabulous Five: How Foreign Actors Prop up the Maduro Regime in Venezuela. Center for Strategic and International Studies, https://tinyurl.com/bdfs6yvx, 2020.

[82] Michael Riley, Lauren Etter, and Bibhudatta Pradhan. A Global Guide to State-Sponsored Trolling. Bloomberg, https://tinyurl.com/mr2pk23e, 2018.

[83] Michael Riley, Lauren Etter, and Bibhudatta Pradhan. Proyecto de Formación del Ejercito de Trolls de la Revolucion Bolivariana. Bloomberg, https://tinyurl.com/42ez9e7k, 2018.

[84] Mohammad Hammas Saeed, Shiza Ali, Jeremy Blackburn, Emiliano De Cristofaro, Savvas Zannettou, and Gianluca Stringhini. Trollmagnifier: Detecting state-sponsored troll accounts on reddit. In *Proceedings of the IEEE Symposium on Security and Privacy*, 2022.

[85] David E. Sanger and Zolan Kanno-Youngs. The Russian Trolls Have a Simpler Job Today. Quote Trump. The New York Times, https://tinyurl.com/yy8ddbh4, 2020.

[86] Mohsen Sayyadiharikandeh, Onur Varol, Kai-Cheng Yang, Alessandro Flammini, and Filippo Menczer. Detection of novel social bots by ensembles of specialized classifiers. In *Proceedings of the 29th ACM international conference on information & knowledge management*, pages 2725–2732, 2020.

[87] Alyza Sebenius. Russian Trolls Shift Strategy to Disrupt U.S. Election in 2020. Bloomberg, https://tinyurl.com/ycuhvzm7, 2019.

[88] Clare Ribando Seelke. Venezuela: Political crisis and us policy. *Current Politics and Economics of South and Central America*, 13(1):5–11, 2020.

[89] Karishma Sharma, Xinran He, Sungyong Seo, and Yan Liu. Network inference from a mixture of diffusion models for fake news mitigation. In *Proceedings of the 15th International AAAI Conference on Web and Social Media*, 2021.

[90] Karuna Sharma. Social media platforms deny but content creators and tracking websites confirm that shadowbanning or silent censorship is real. Business Insider, https://tinyurl.com/2ts654yu, 2022.

[91] Kai Shu, Suhang Wang, and Huan Liu. Beyond news contents: The role of social context for fake news detection. In *Proceedings of the 12th ACM WSDM*, 2019.

[92] Florantonia Singer. Tres dólares al mes por trabajar para el Estado venezolano. El Pais, https://tinyurl.com/3xtnx5dw, 2021.

[93] Jonghyuk Song, Sangho Lee, and Jong Kim. Crowdtarget: Target-based detection of crowdturfing in online social networks. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, page 793–804, 2015.

[94] Spyscape. Inside Russia's Notorious 'Internet Research Agency' Troll Farm. https://spyscape.com/article/inside-the-troll-factory-russias-internet-research-agency.

[95] Reuters Staff. Doctors skeptical as Venezuela's Maduro touts coronavirus 'miracle' drug. Reuters, https://tinyurl.com/2mhw3k8r, 2021.

[96] Kate Starbird, Ahmer Arif, and Tom Wilson. Disinformation as collaborative work: Surfacing the participatory nature of strategic information operations. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–26, 2019.

[97] Kate Starbird, Ahmer Arif, Tom Wilson, Katherine Van Koevering, Katya Yefimova, and Daniel Scarnecchia. Ecosystem or echo-system? exploring content sharing across alternative media domains. In *Proceedings of AAAI ICWSM*, volume 12, 2018.

[98] Kate Starbird and Leysia Palen. Working and sustaining the virtual" disaster desk". In *Proceedings of the ACM Conference on Computer Supported Cooperative Work*, pages 491–502, 2013.

[99] Anselm Strauss and Juliet M Corbin. *Grounded theory in practice*. Sage, 1997.

[100] Gianluca Stringhini, Pierre Mourlanne, Gregoire Jacob, Manuel Egele, Christopher Kruegel, and Giovanni Vigna. EVILCOHORT: Detecting communities of malicious accounts on online services. In *Proceedings of the 24th USENIX Security Symposium*, 2015.

[101] K. Thomas, D. Akhawe, M. Bailey, D. Boneh, E. Bursztein, S. Consolvo, N. Dell, Z. Durumeric, P. Kelley, D. Kumar, D. McCoy, S. Meiklejohn, T. Ristenpart, and G. Stringhini. Sok: Hate, harassment, and the changing landscape of online abuse. In *IEEE Symposium on Security and Privacy (SP)*, 2021.

[102] Twitter. Elections Integrity. Data Archive. https://about.twitter.com/en_us/values/elections-integrity.html#data.

[103] Svitlana Volkova and Eric Bell. Account deletion prediction on runet: A case study of suspicious twitter accounts active during the russian-ukrainian crisis. In *Proceedings of the 2nd Workshop on Computational Approaches to Deception Detection*, 2016.

[104] Matthew L. Williams, Pete Burnap, Amir Javed, Han Liu, and Sefa Ozalp. Hate in the machine: Anti-black and anti-muslim social media posts as predictors of offline racially and religiously aggravated crime. *British Journal of Criminology*, 60(1):93–117, 2020.

[105] Zhen Xie, Sencun Zhu, Qing Li, and Wenjing Wang. You can promote, but you can't hide: Large-scale abused app detection in mobile app stores. In *Proceedings of the 32nd Annual Conference on Computer Security Applications*, page 374–385, 2016.

[106] Xu, Teng and Goossen, Gerard and Cevahir, Huseyin Kerem and Khodeir, Sara and Jin, Yingyezhe and Li, Frank and Shan, Shawn and Patel, Sagar and Freeman, David and Pearce, Paul. Deep Entity Classification: Abusive Account Detection for Online Social Networks. In *Proceedings of the Usenix Security Symposium*, 2021.

[107] Kevan Yenerall. Grassroots Politics. In *Encyclopedia of American Government and Civics*. 2017.

[108] Dong Yuan, Yuanli Miao, Neil Zhenqiang Gong, Zheng Yang, Qi Li, Dawn Song, Qian Wang, and Xiao Liang. Detecting fake accounts in online social networks at the time of registrations. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, pages 1423–1438, 2019.

[109] Savvas Zannettou, Tristan Caulfield, Emiliano De Cristofaro, Michael Sirivianos, Gianluca Stringhini, and Jeremy Blackburn. Disinformation Warfare: Understanding State-Sponsored Trolls on Twitter and Their Influence on the Web. In *Companion of The 2019 World Wide Web Conference*, 2019.

[110] Savvas Zannettou, Tristan Caulfield, William Setzer, Michael Sirivianos, Gianluca Stringhini, and Jeremy Blackburn. Who let the trolls out? towards understanding state-sponsored trolls. In *Proceedings of the 10th ACM Conference on Web Science*, 2019.