The Gender Dimensions of Foreign Influence Operations

Summary of Revisions

April 2021

Thank you for the opportunity to revise and resubmit our manuscript for publication with *The International Journal of Communication (IJoC).* We are grateful for the thoughtful comments provided by the reviewers. We believe they have helped significantly strengthen the theoretical contributions of the paper, as well as improved the clarify of our overall findings and results. Our response to each of the suggested changes are detailed below. We are delighted to have our work be considered for publication with *IJoC* and believe that its readers will appreciate the contribution.

Improving the Introduction

The reviewers provided some valuable feedback about the introduction and the framing of our research questions, as well as emphasizing the importance of why this study matters. We have re-written the introduction to reflect the feedback of the reviewers, which involved the following changes. First, to avoid the purely gloomy picture of social media and disinformation, we also emphasized the work from social movement studies and women and (digital) politics to describe the empowering nature of technology. We hope this provides more nuance as the reviewer suggested. Finally, we removed the findings from the introduction and used this space to expand upon the "why" of the study and the contributions we hope to make with this work. We also made some minor changes to the introduction which involved diversifying our sources about the field of disinformation and foreign influence operations and softening our language around the growth of these kinds of campaigns overtime. The new introduction we have written has been copied into this document below.

Introduction

Following the outcome of the 2016 US Presidential Election, foreign influence operations have been a growing topic of academic inquiry (Author 2018; Lin & Kerr, 2019; Martin & Shapiro, 2019; McGregor et al., 2021; Moore, 2018; Rid, 2020; Singer & Brooking, 2019; Starbird et al., 2019). Since then, a wave of research has examined how disinformation—the purposeful spread of false or misleading information—was used by foreign states, among others, to manipulate public opinion via social media (Allcott & Gentzkow, 2017; Benkler et al., 2018; Bennett & Livingston, 2018; Grinberg et al., 2019). Researchers have documented how these campaigns have homed in on national issues and local communities, spreading disinformation to inflame debates about race and religion and deepen

polarization across the political spectrum (DiResta et al., 2018; François et al., 2019; Freelon et al., 2020; Friedberg & Donovan, 2019; Howard et al., 2018; Woolley & Joseff, 2019). One aspect of these operations that has received less attention is the gender dimension of these campaigns.

Social media is used by millions of women around the world to express their freedom of speech, contribute to the global digital economy, and participate in public life. Platforms like Facebook and Twitter provide space for women and girls to connect with others, develop a sense of collective identity, and participate in online activism (Crossley, 2015; Locke et al., 2018). But for many women – especially women of colour or with diverse gender identities – social media is a place where they experience sexism, harassment, and threats (Brown et al., 2017; Eckert, 2018; Jankowicz et al., 2021; Mantilla, 2015; Mendes et al., 2018; Sobieraj, 2020; Philips, 2015). The vocabulary of gender-based attacks has also been adopted by state actors to suppress women's rights and civic participation through accusations of collusion, sexualized tropes, and threats of rape and violence (Jankowicz, 2017; Judson et al., 2020; Monaco et al., 2018). Qualitative accounts have highlighted how these kinds of targeted attacks against women can have a measurable impact on the online behaviours of victims (Amnesty International, 2017; Zeiter et al., 2020).

Although women experience disinformation and harassment differently, there have been no systematic studies that examine the gender dimensions of foreign influence operations. Some theories about contemporary influence operations describe how state actors exploit ideological differences between groups of users to inflame racial division or increase polarization (Freelon et al., 2020; Freelon & Lokot, 2020). Rather than galvanizing support for a particular person or idea, contemporary influence operations exploit divisions between communities and groups. Indeed, academic, journalistic, and independent investigations into the US 2016 elections highlighted concentrated efforts targeting Conservative and Black American voters with polarizing disinformation (DiResta et al., 2018; Howard et al., 2018; Mueller, 2019; Nimmo, 2016). However, no studies have systematically examined how foreign state actors instrumentalize narratives about gender—and in particular, narratives about women—in contemporary influence operations.

Our research fills this gap by exploring how foreign state actors engage in covert influence operations targeting feminist activists and politicians. Drawing on a qualitative analysis of 7506 tweets from Twitter's *Election Integrity Initiative (EII)*, we examine the narratives and strategies foreign state actors use to target feminist movements and their advocates. Our research provides a unique opportunity to not only comparatively assess how different states approach foreign influence operations but provides insight into how these campaigns interact with gender and feminism. While critical research about foreign influence operations have focused largely on racial and political asymmetries that emerge between Conservatives and Liberals or between black and white American, we explore how gender identity is instrumentalized.

Situating the Research Questions and Findings within a Theoretical Framework

Both reviewers provided some very constructive feedback on the theoretical framework for the paper. First, we decided to situate our findings by bringing together two important strands of literature (1) theories about contemporary influence operations that describe how foreign agents instrumentalize racial and ideological asymmetries to polarize and divide citizenry, and (2) theories from social movement studies about collective identity formation. What is missing from the research on foreign influence operations is the role of gender. Additionally, while most of the research on foreign influence operations highlights how covert influence operations will affirm group identities while provoking anger or outrage from oppositional groups, our studies show how these covert strategies also co-opt inter-group critiques of feminism. We believe that by focusing on these two bodies of literature, we can better emphasize our key findings and contributions to scholarly work on gender and foreign influence operations.

Literature Review

Contemporary Foreign Influence Operations

Social media platforms are becoming vectors for foreign influence operations – clandestine operations by foreign state actors that seek to undermine information systems and manipulate civic discourse (Martin & Shapiro, 2019; Waltzman, 2017). They are considered tool a of "asymmetric," "non-linear" or "hybrid" warfare and are used as an alternative for or compliment to traditional kinetic warfare (Lin & Kerr, 2019; Rid, 2020; Starbird et al., 2019). Although foreign influence operations are not new, the affordances of social media platforms—including algorithms, automation, and data—can enhance the scope, scale, and precision of these campaigns (Author 2019; Author 2018; Hwang & Rosen, 2017). And the ubiquity of social networking technologies, combined with the low cost to produce and disseminate content online, presents a qualitatively new landscape for persuasion, manipulation, and hybrid forms of warfare (Lin & Kerr, 2019; Moore, 2018; Rid, 2020).

The study of foreign influence operations regained prominence following the 2016 US Presidential Election when major social media companies confirmed their platforms had been co-opted by both foreign and domestic actors (Sanders, 2016). Since then, a wave of research has examined how disinformation was used to influence public opinion via social media (Allcott & Gentzkow, 2017; Benkler et al., 2018; Bennett & Livingston, 2018; Grinberg et al., 2019). Academic studies and journalistic investigations have looked at the variety of actors, narratives, and technological tools used to influence publics during critical political moments and around contentious political topics such as immigration, climate change, or the novel coronavirus (Benkler et al., 2018; Freelon et al., 2020; Marwick & Lewis, 2017; Nguyen & Catalan, 2020; Starbird et al., 2019; Woolley & Howard, 2018). One aspect of these operations that has received less attention is the gender dimension of these campaigns.

State actors have used social media to exclude women from politics and public life through intimidation tactics, harassment, and gendered disinformation (MacKinnon, 2012; Monaco et al., 2018). Anecdotal evidence about these state-sponsored gender attacks highlights how fake government-run accounts target women with threats of rape and violence, accusations of collusion, and disinformation to undermine their legitimacy in public spaces. By relying on gender stereotypes and sexualized tropes, state-sponsored gender attacks suppress the participation of women in public life while advancing other political objectives (Fichman & McClelland, 2021; Jankowicz et al., 2021; Judson et al., 2020). Qualitative accounts have highlighted how targeted attacks against women have a measurable impact on the behaviours, professions, and economic security of victims (Amnesty International, 2017; Zeiter et al., 2020).

Although women experience disinformation differently, most academic research on foreign influence operations has focused on political, and to a lesser extent, racial identities. Contemporary theories about foreign influence operations highlight how campaigns exploit ideological differences between groups of users to inflame racial division or increase polarization (Freelon et al., 2020; Freelon & Lokot, 2020; Friedberg & Donovan, 2019; Tucker et al., 2018). When it comes to political identity, empirical studies have found that foreign state actors generate more content targeting conservative users, and conservative users tend to share disinformation more than their liberal counterparts (Fichman & McClelland, 2021; Guess et al., 2019; Osmundsen et al., 2020). There have also been extensive influence operations targeting Black American communities, where fake accounts masquerading as Black Americans spread disinformation to demobilize activist communities and suppress voter turnout (DiResta et al., 2018; Freelon et al., 2020; Howard et al., 2018). Although research shows that certain groups of users experience disinformation disproportionately, no studies have systematically examined how foreign state actors target high-profile women or feminist movements as part of their influence operations.

While some traditional theories of propaganda focused on fostering support for a particular person or idea, contemporary theories suggest that influence and persuasion are a function of political and group identity where covert operations exploit digital affordances to share, amplify, and target content that provokes resentment against oppositional groups (Freelon et al., 2020; Freelon & Lokot, 2020; Friedberg & Donovan, 2019; Tucker et al., 2018). Scholars have enriched our understanding of these processes by examining the role of race in these "ideological asymmetries", which can be a highly divisive and engaging narrative (Freelon et al., 2020). We build on this critical work by asking whether gender leads to a qualitatively different experience or understanding of contemporary influence operations. In addition to exploring how harassment is perpetuated through foreign influence operations, we also explore how state-actors co-opt narratives around feminism as part of their campaigns.

Feminism, Online Social Movements & Collective Identity Formation

To properly conceptualize how feminist movements are co-opted for foreign influence operations, it is important to discuss contemporary feminist movements and the relationship between collective identity formation and mobilization. Understanding how and why people come together and mobilize clarifies how these processes can be instrumentalized for demobilization. Theories about collective action mobilization suggest that the development of collective identity is a key factor for mobilizing action (Buchan et al., 2011; Jenkins, 1983; Van Zomeren et al., 2004). Before mobilization occurs, individual members must identify and share common concerns, experiences, and feelings (Klandermans, 1997; Polletta & Jasper, 2001; Simon et al., 1998). This sense of collective identity is not static, rather, it is constantly defined, re-defined and re-negotiated through every-day interactions between group members (Barassi, 2018; Hopkins & Blackwood, 2011; Khazraee & Novak, 2018). Empirical studies suggest that sharing a sense of collective identity can facilitate feelings of togetherness and commonality, which can increase trust and support and improve a groups' ability to collectively mobilize. When individual group members attach their sense of self to their group membership, the pursuit of group interests becomes interchangeable with pursuing one's own interests (Buchan et al., 2011; Van Zomeren et al., 2004).

Feminist movements are no exception to these findings, and social media provides a new environment for women and girls to learn about feminist activism, connect with others around the globe, and develop a sense of collective identity (Crossley, 2015; Locke et al., 2018). Scholars have documented the rise of various digital feminist campaigns that speak to broad issues of violence, rape, injustice, and inequality facing women in both online and offline spaces (Fileborn & Loney-Howes, 2019; Loney-Howes, 2018). By leveraging the unique affordances of digital platforms, feminist movements have used platforms like Twitter to organize, structure, and make accessible the movement to an ever-growing audience, facilitating a sense of collective identity, support, and empathy (Turley & Fisher, 2018). But the high visibility around feminist movements opens participants to a number of vulnerabilities, such as the further proliferation of misogyny and harassment (Boynton, 2012). Global, visible, and performative movements also have the potential to exclude people based on specific elements of their identity and many scholars and activists have critiqued feminist movements for maintaining white and Western-centric perspectives (Daniels, 2015; Liska, 2015; Phipps, 2019). These intersectional critiques and counter-movement narratives are spaces where collective identity and shared norms are challenged.

Like other movements, there has been some evidence about foreign influence operations co-opting the language of feminism, its counter-movements, and critiques. For example, pro-Kremlin outlets have spread content around notions of "totalitarian feminism" where women supporting feminist movements want to bully men and discourage "men's interest in women" by "turning them homosexual" (EU vs. Disinfo, 2019). Similarly, other narratives have described feminism as being incompatible with Islam, where women are "forced to be prostitutes", "wear hijabs", or "undergo genital mutilation" (EU vs. Disinfo, 2019). Despite the use of these frames, there have been no systematic analyses on how

feminism is co-opted as part of foreign influence operations, and our study contributes to filling this gap.

In sum, our data allows us to look both at how high-profile women are targeted and harassed, and to explore how foreign agents engaged in rhetoric and narratives on women's rights and empowerment as part of contemporary influence operations. Thus, our core research questions ask:

1. What are the key narratives foreign state actors use in covert influence operations that discuss women's rights?
2. What tactics do foreign state actors use when targeting high-profile feminist politicians, journalists, and activists?

Finally, one reviewer noted that our explanation of the affordances of Twitter was too long and does not directly answer the research questions laid out in the paper. We completely agree with this and cut the section. We believe that the discussion about Twitter as a platform for foreign influence operations are covered sufficiently in our methodology section and have used the space to engage more deeply with relevant theory and literature outlined above.

Clarifying Data & Analysis

Both reviewers helpfully noted that the targets of state-sponsored operations were not always clear. We did go back to review the announcements made by Twitter to see if they had discussed the themes found in the various EII datasets. Recently, their announcements about Coordinated Inauthentic Behaviour have included small descriptions of the targets and themes discussed by accounts in the dataset. However, Twitter did not have any similar analysis for the datasets we examined (we suspect because this was early in the policy process). Thus, we decided it was best to scope our Tweets based on the language. Because we focused on content that was labelled by Twitter as "US-English-language" content, we clarified the scope (and subsequent limitation) in the body of the manuscript (in red):

> Finally, since the influence operations in this data come from diverse geographic locations, the accounts published content across a variety of languages. Twitter did not provide tweet language information for all the data in public repositories. For those where language information was available, the top-5 most used languages were Russian, US-English, French, Hungarian, and Arabic. For the purposes of our analysis, only tweets in US-English (3,940,094 tweets) were analysed. Thus, while influence operations target a variety of countries and audiences, our findings are only relevant for US-English-speaking audiences where most of the English-language Tweets were labelled.

One reviewer asked for the rationale for sampling with predetermined terms, as well as a discussion of the limitations of this approach. We added the following text (in red) to address these comments in our methodology section.

> All Twitter studies begin the sampling process by using search terms or hashtags to identify relevant data (Kim et al., 2013). To build a clean and exhaustive sample of tweets about feminism, we developed a pre-determined list of hashtags related to women's rights and activism queried the data set for tweets that contained these terms. The list included: #feminism, #feminist, #feminists, #womensrights, and #genderequality, leading to a sample of 5149 tweets. We selected these hashtags to avoid under- and overestimating the volume of discussion about feminist movements.
>
> …
>
> Overall, topic-based sampling can be accurate and representative, particularly because of the flexible nature that allow researchers to adapt and extend their list of terms (Cameletti et al., 2020). Nevertheless, this approach introduces limitations concerning comprehensiveness (Bartlett et al., 2014; Cameletti et al., 2020; Hull & Grefenstette, 1996). By purposively selecting specific terms and manually reviewing collocated hashtags our sampling strategy tended to value precision over comprehensiveness (Tulkens et al., 2016). Although this sample might not capture *all* of the gender-related conversations being held by foreign state actors, best efforts were made to ensure the data in this sample were clean, representative, and relevant for the topic of study.

One reviewer asked for a clarification around our approach to coding the discourse types. They asked how we defined our 11 discourse types (and when this happened in the coding process). We have clarified this in the manuscript, noting that we used a line-by-line review of data on a small subset (20% or 1500 tweets), and then used forced coding to label the rest of the data.

> Tweets were coded using a grounded theory of coding (Charmaz, 2006). We began with a line-by-line review of a small subset of data (20% or 1,500 tweets) to investigate discourse types used by state accounts in the data set. To refine the concepts emerging from the data, we manually grouped the discourse types into 11 categories (e.g. tweets that celebrated a specific person for promoting female empowerment or women's rights in some way, tweets that distorted feminist principles or values, polemic tweets that attacked individuals or the women's rights movement, or informational tweets that shared resources relevant to those interested in feminism). We then applied focused coding to the rest of the data set (Charmaz, 2006). Reliability scores showed that there was substantial agreement among the coders in terms of how to categorize tweets relative to our discourse types (Krippendorff's α = .84). All disagreements were reviewed and collectively resolved.

Finally, one reviewer asked if we did an automated or manual sentiment analysis. We opted for a manual sentiment analysis since many of the automated tools we tested for sentiment analysis did not accurately label our Tweet data. We have clarified this in the manuscript.

> To identify relational patterns, we also manually grouped the 11 discourse types into 3 sentiment categories: (1) positive tweets that supported, celebrated or promoted women's rights, (2) negative tweets that attacked individuals or the movement more broadly, or (3) neutral tweets that were informational in nature, without a positive or negative sentiment.

Strengthening the Findings & Analysis

Both reviewers provided helpful feedback around strengthening the findings and analysis section of the paper. One reviewer suggested that we re-structure our findings around the characteristics of state sponsored attacks. Thus, we decided to re-structure our results section around the four major themes we found in our data (polarization and gender asymmetries, undermining women's collective identity, targeted attacks against high-profile women, and amplification through automation and hashtag hijacking). We believe these themes speak more directly to the research questions and theory of the paper (i.e. what are the narratives and what are the strategies/tools of contemporary foreign influence operations). We also included more examples and better situated some of the unique state-by-state findings within research about these states as suggested by the peer reviewers.

Findings and Analysis

The Narratives about Feminism & Women's Rights

Polarization & Gender Asymmetries

Research about foreign influence operations describes how foreign state actors instrumentalize political and group identity to share, amplify, and target content that promotes ingroup solidarity and provokes resentment against oppositional groups (Freelon et al., 2020; Freelon & Lokot, 2020; Guess et al., 2019; Osmundsen et al., 2020; Tucker et al., 2018). This is particularly relevant for racial identity, where empirical research has demonstrated a concentrated strategy to amplify racial divisions in the United States (DiResta et al., 2018; Freelon & Lokot, 2020; Howard et al., 2018). We found similar patterns in the narratives foreign state actors employed around feminist movements, but rather than promoting in-group solidarity, conversations about feminism were skewed towards amplifying negative sentiment about the movement and its activists more broadly.

First, foreign influence operations disproportionately shared content that attacked feminist movements over content that promoted ingroup solidarity. The IRA was the most antagonistic, where over 64% of the IRA tweets focused on spreading negative sentiment about feminism compared to less than 20% of tweets that were coded as expressing a positive sentiment. IRA accounts focused largely on misogynistic polemics that attacked the feminist movement and its activists. For example, one IRA tweet stated: "Feminists are

more likely to rely on petty insults in spite of giving valid arguments #FeminismIsAwful". Like other gender-based attacks, state-sponsored accounts discussing feminism on Twitter focused on the appearance of women and activists: "RT Handful of self-entitled fat ugly feminists trying to get arrested in desperate attempt to impress any man" or "RT #FeministsAreUgly is trending by feminists themselves just post pics amp get attention don't y'all have women to save?". Overall, tweets by state actors, and to a greater extent, the IRA, about feminism and feminists focused on amplifying negative sentiment rather than promoting solidarity and would use the appearance of women in their state-sponsored gender-based attacks.

Additionally, research has highlighted how racial and political identities are leveraged to provoke oppositional groups and exacerbate polarization. We wanted to explore the asymmetries between men and women in the context of foreign influence operations. IRA and GRU accounts were the only state actors to share content that discussed asymmetries about gender identity juxtaposing male versus female rights. Here, tweets focused on distorting principles of equality – for example, by making arguments that feminism is about suppressing male rights opposed to supporting equality for men and women. Narratives often presented feminists as "man-hating" oppressors, claiming that modern-day feminism leads to the suppression of male rights. For example: "#IHaveADreamThat feminism will stop being a man hating ideology and girls stop accusing us of rape if sex wasn't good". Content describing these divisions—between men and women—made up only a small proportion of the overall tweets (6% of the total tweets shared by both the IRA and GRU accounts). Thus, theories about contemporary influence operations that focus on amplifying divisions *between* identity groups do not fully fit with an analysis on gender.

Undermining Women's Collective Identity

Rather than focusing on divisions between men and women, foreign influence operations made use of narratives that undermined women's sense of collective identity. This was a strategy primarily adopted by the Russian-backed accounts, where Russian state actors co-opted intersectional critiques that have arisen from and exist within the broader feminist movement. Their strategies were not based on driving divisions between men and women, or between feminism and society more broadly, but rather were about fracturing the feminist movement and undermining women's ability to formulate a shared identity.

Tweets in our data discussed three main intersectional critiques: that feminism was: (1) too white to represent black women; (2) too liberal to represent conservative women; and (3) too wealthy to represent poor women. In all three cases, the IRA co-opted the narratives that feminist activists have themselves raised about contemporary feminist movements in both online and offline spaces (Daniels, 2015; Liska, 2015; Phipps, 2019). Intersectional critiques about racial or political identities were predominantly discussed by Russian accounts, where more than 86% of the tweets came from IRA (66%) or GRU (20%) accounts. These accounts used discourses about feminism being rooted in "white" and "liberal" values, and that black women or conservative women were not represented by the

movement. For example, one IRA tweets stated: “RT A quick search of white feminist & #BlackLivesMatter yielded ZERO results. Gee I wonder why?”. Another example includes: “RT Funny how feminists are all for the advancement of women unless that woman is a Conservative #TCOT #MAGA @FoxNews”. In addition to race and political identity, IRA accounts also discussed class identity and the representation of poor or marginalized groups of women. For example: “RT Liberal feminists ignore poor women entirely in their activism That’s why their activism is garbage”. The co-option of these narratives about feminism is about undermining women’s ability to form collective identities that are necessary for political mobilization.

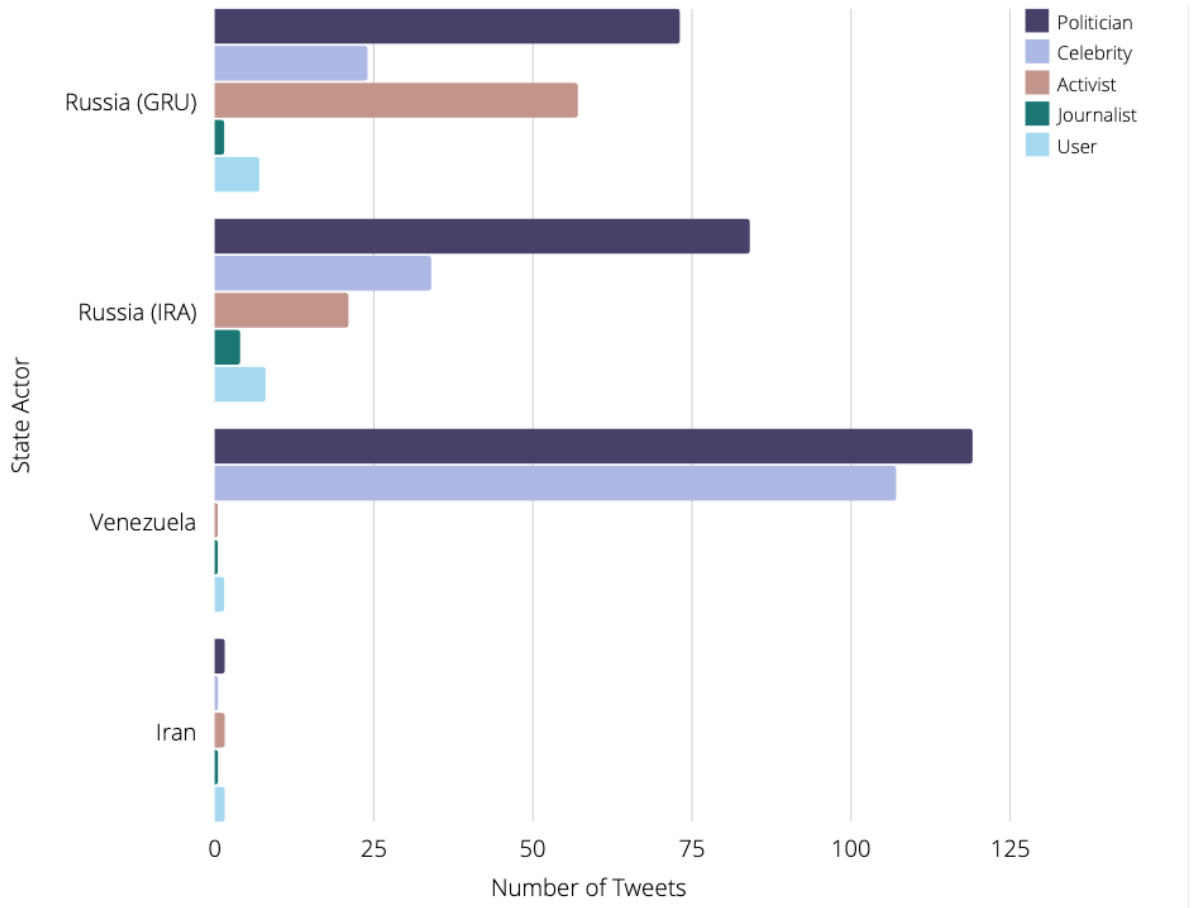
Strategies for Targeting Feminists & Their Movements

Targeted Attacks Against High-Profile Women

Research about women and disinformation has consistently demonstrated that high-profile women face distinct patterns of hate, harassment, and gender-based attacks online (Jankowicz, 2017; Judson et al., 2020; Monaco et al., 2018). When these kinds of attacks are perpetrated by state actors, oppressive regimes can combine online harassment with the threat of real-world violence or imprisonment to stifle fundamental freedoms (Deibert, 2013; MacKinnon, 2012). However, no systematic studies have examined how foreign state actors use gender-based attacks as part of their foreign influence operations.

Our data provides some insight into these questions. We found concentrated efforts to delegitimize and discredit female activists, celebrities, politicians and, to a lesser extent, journalists and users, who discussed or participated in feminist movements on Twitter. The IRA, GRU and Venezuelan accounts launched the most character attacks against high-profile feminists (see Figure 1), reflecting tactics of suppression and discrediting often seen in these countries’ domestic context (MacKinnon, 2012; Monaco et al., 2018). However, across all three datasets in the sample, the accounts rarely used the @mention feature when launching character attacks. Instead, attacks against women often used their name in plaintext along with hashtags that were popular or trending. This could suggest that foreign influence operations make less use of *publicly* targeted attacks on prominent voices as part of their campaigns abroad. We have two explanations for this less visible approach. The first is that that attacking or harassing individuals domestically comes with better technical or legal measures for enforcing actions that would silence speech – such as physical harm or arrests that are so often combined with state-sponsored attacks. Another explanation could be that targeted harassment happens more through direct-messages opposed to public facing Twitter where tweets are more likely to be flagged by users or by the automatic detection measures employed by Twitter.

Figure 1: Targeted Attacks Against High-Profile Women



Source: Authors (2020). Data Collected from Twitter EII

Nevertheless, we still identified a number of character attacks made about specific women. Across all three datasets, politicians were discussed the most by state sponsored accounts with attacks against Hilary Clinton accounting for more than 90% of all attacks made against politicians. Users (everyday Twitter users who are not considered "high-profile" i.e. not recognizable or famous) were rarely attacked, with less than 5% of the total character attack tweets targeting users. The GRU launched the most character attacks against women on Twitter (30% of all attack on character tweets). In particular, they focused on creating and amplifying ad hominem attacks about Hilary Clinton and Linda Sarsour, the American political activist who organized the 2017 Women's March. GRU accounts tweeted content suggesting she wanted to implement Sharia Law in the United States or was secretly a Jihadi terrorist infiltrating America. One explanation for the use of Islamophobic narratives could be that traditional propaganda campaigns have used narratives about Islam and orientalism more broadly to spread fear in order to justify military intervention (Khalid, 2011). While the dissemination of Islamophobic disinformation is modern in its distribution methods, the topics and content are highly similar to Soviet operations of the past (Lukito, 2020; Tucker et al., 2018). Thus, the use

of these narratives could simply be an extension of traditional military propaganda strategies being brought to the digital realm.

While GRU accounts had more character attacks targeting high-profile activists with traditional propaganda frames, the IRA focused their attacks on high-profile feminist celebrities or women supporting feminist causes. Madonna was the biggest celebrity target of the IRA accounts with over a third (35%) of all character attacks coded in the dataset mentioning her. Venezuelan accounts also made character attacks against many female and feminist celebrities, but these accounts relied on clickbait and sensation to discredit female figures. For example, one tweet stated": "CAT FIGHT Legendary Lesbian Feminist Declares Hillary Exploits Feminism". Overall, character attacks against female politicians, activists, and celebrities by accounts from the IRA, the GRU and Venezuela may have differing tactics but all work to discredit and delegitimize women involved in feminist discourse.

Finally, is important to highlight one outlier: in contrast to the IRA, GRU and Venezuela, Iranian accounts were positive in their discussion about women's rights and empowerment, and celebrated feminist figureheads on Twitter. This was a surprising finding considering Iran is one of the worst countries in the world for women's rights (Freedom House, 2019; Human Rights Watch., 2019). Iranian accounts focused their engagement with English-speaking audience on timely and high-profile events, generating positive tweets about various feminist movements. More than 20% of all the Iranian tweets in the sample positively tweeted out solidarity statements about the 2017 Women's March or the #MeToo movement. Theories about Iranian influence operations describe how Iran promotes its geopolitical interests aboard by employing a "distorted truth" strategy that exaggerates "Iran's moral authority while minimizing the repression of its citizens" (Brooking & Kianpour, 2020). Supporting feminist movements could be an extension of this "distorted truth" strategy, especially considering both the #MeToo movement and 2017 Women's March were highly critical of Donald Trump's presidency, which is in-line with Iran's anti-Trump stance and geopolitical interests.

Amplification through Automation & Hashtag Hijacking

The use of political bots to spread disinformation has become a growing area of academic research as investigations have uncovered networks of "bot" accounts being part of foreign influence operations (Author 2018, 2019; Bessi & Ferrara, 2016; Borak, 2019; Elswah, 2019; Murthy et al., 2016; Owen Jones, 2016). The detection and identification of bots is technically complex, with multiple definitions and measures for capturing this kind of activity (Gorwa & Guilbeault, 2020). One useful way for defining automated activity is what McKelvy & Dubois (2017) call "amplifier accounts", which are accounts that deliberately seek to increase the number of voices speaking about or the attention being paid to certain messages. Unlike traditional "bot" definitions, amplifier accounts include automated, semiautomated, and highly active human-curated accounts on social media (McKelvey & Dubois, 2017).

State-sponsored amplifier accounts were present in the dataset. In total, ten accounts tweeted over a third of the content in the sample. Nine out of the ten accounts were Russian backed, with four accounts being operated by the GRU and five operated by the IRA. All nine Russian-operated accounts amplified negative content about feminism and women's rights. The most active GRU account amplified character attacks in-line to supress and delegitimize high-profile feminists. Similarly, the most active IRA account amplified demobilization messages that attacked the feminist movement more broadly. These amplifier accounts did not only focus on issues to do with feminism and women's rights but tweeted about many different issues in data provided by Twitter. For issues covered in the sample, the GRU amplifier accounts mainly retweeted content. This fits in line with other analyses that suggest the GRU creates content for fake media websites, think tanks, or political commentators and amplifies this content with other fake accounts to generate a false sense of legitimacy and credibility (DiResta & Grossman, 2019). In contrast, the IRA amplifier accounts created original content, with the most active amplifier account tweeting 502 original tweets out of the IRA sample.

It is important to note that amplifier accounts might be over-represented in the dataset because automated account activity is a common detection method Twitter uses for identifying coordinated inauthentic behaviour. Additionally, amplified content does not need to be automated by highly active amplifier accounts. For example, many tweets in the Venezuelan dataset appeared to be copy and pasted. For example, a tweet stating: "Donald Trump Responds To #WomensMarch Liberal Snowflakes Explode On Twitter" was posted by the Venezuelan accounts over 70 times.

In addition to amplifier accounts, Iranian operations used conversations about feminism to amplify other state-sponsored propaganda unrelated to women's rights. Almost a third (29.2% of Iranian tweets) of the Iranian tweets in the sample "hijacked" hashtags, where accounts would simply use prominent or trending hashtags to spread unrelated messages by linking them to popular conversations. One example would be the use of the #WomensMarch hashtag to share news about Saudi air strikes in Yemen. Iran also had the largest proportion of tweets coded as NA – where tweets were simply spam or were not comprehendible to the coders. Hashtag hijacking and spam are reflective of trends identified by other investigations into Iranian influence operations, where accounts are much less sophisticated and make use of crude automation techniques to hijack conversations, rather than engaging with issues that were important to the audiences discussing them (Leprince-Ringuet, 2018; Nimmo et al., 2020).

One reviewer critically noted that bots might cover a large portion of the data-set because it is one of the easier ways platforms attribute accounts, thus there might be an over-representation of bots because of the inherent bias in the way the data was compiled. This is an excellent point and we have added a sentence in our findings & analysis section to reflect this (highlighted in red above).

Further Minor Revisions

The reviewers brought to our attention that certain wording choices, particularly the use of "trolling" could be confusing or misleading to the reader. In accordance with their suggestion, we decided to avoid the term altogether and went with "gender attacks" instead.

Thank you for the opportunity to revise and resubmit our manuscript. We are looking forward to engaging with the peer reviewers on the changes we have made to the document in light of their comments.