Research Plan
Gender and Purported Audiences in Pro-CCP Tweets Targeting the Uyghur Diaspora

Allison Koh[1], Jonathan Nagler[2], and Joshua A. Tucker[2]

[1]Hertie School

[2]New York University

November 1, 2022

Abstract

Chinese authorities and supporters of the Chinese Communist Party have leveraged Western social media platforms to lodge smear campaigns against Uyghur, Kazakh, and other Turkic women living abroad. The women targeted have spoken out against human rights violations targeting Muslim minorities in the Uyghur region of northwestern China. These state-sponsored social media attacks reflect a global trend in the use of digital tools for engaging in transnational repression. Against this background, this study investigates the gender dimensions and contours of pro-CCP smear campaigns on Twitter. We use large-scale quantitative data from Twitter and information on Chinese authorities' attacks targeting Uyghur individuals abroad to analyze whether women in the Uyghur diaspora are disproportionately attacked with vitriol on pro-CCP Twitter. Controlling for gender, we also empirically test whether content produced by overtly state-affiliated accounts differs from that of other pro-CCP users, and whether pro-CCP tweets in English differ from those published in Chinese. Finally, we explore the topics discussed in targeted mentions of Uyghurs in the diaspora and possible differences based on the gender of a targeted individual. Our findings have important implications for studying gender and state-sponsored disinformation from authoritarian contexts on Western media platforms.

Keywords: disinformation; digital transnational repression; authoritarianism; gender; Twitter

1 Introduction

Repressive governments are increasingly leveraging the affordances of social media to harass online opposition beyond their borders. Women and individuals with other marginalized identities are especially vulnerable to these attacks. While relevant studies have provided much insight on how states leverage social media to influence politics at home and abroad, and how targeted harassment has been deployed in a handful of cases, little is known about the overall contours and possible gendered aspects of state-aligned content that targets the online opposition—especially in the transnational dimension. This study addresses this gap by identifying of how pro-CCP Twitter accounts communicate with diaspora activists, whether women are targeted with more negative rhetoric, and the extent to which official state accounts participate in targeted smear campaigns.

State-aligned campaigns can be initiated via official accounts, state-backed media, or pro-regime influencers. This content is amplified by genuine regime loyalists and inauthentic accounts. High-profile dissidents in the diaspora are often on the receiving end of states' online attacks because of their influential role in transnational advocacy networks (Moss 2018; Schenkkan and Linzer 2021). As the nature of online attacks reflects power dynamics of the offline world, state-aligned online harassment also stems from patriarchal contexts in which women are disproportionately attacked with violent rhetoric online (Thakur and Madrigal 2022).

In Section 2, we outline hypotheses and research questions. In Section 3, we present the research design of this project. We elaborate on variable measurement in Section 4. Our plan for analysis is detailed in Section 5, and we discuss ethical considerations for this study in Section 6.

2 Hypotheses and Research Questions

The overarching question guiding this study is, *Are Uyghur women diaspora activists more likely to be targeted in pro-CCP smear campaigns on Twitter?* Our primary hypothesis tests the extent to which women activists in the Uyghur diaspora targeted with toxic language on pro-CCP Twitter:

> **Hypothesis 1** *In pro-CCP content on Twitter, we expect that women diaspora activists are more likely to be targeted with toxicity than men diaspora activists.*

We also test the differences between tweets authored by officially state-affiliated accounts and those from other pro-CCP users. In the offline world, violent repression is often outsourced to "thugs-for-hire" to provide the state with some plausible deniability regarding their involvement in specific attacks (Ong 2020). In the digital age, repressive states outsource the production of pro-regime content to curate the illusion of having their positions "endorsed" by seemingly independent voices (Nyst and Monaco 2018). In line with the provision of these repressive tactics that take place online and offline, we generally expect that official state-affiliated accounts use less violent rhetoric when targeting the online opposition. Accordingly, we expect the following differences in content produced by state-affiliated accounts:

> **Hypothesis 2** *Controlling for gender, tweets by state-affiliated accounts are less likely to be toxic towards activists compared to other pro-CCP users.*

Our final hypothesis tests whether we observe variation in the toxicity of posts by the language that they are posted in. In general, English-language tweets represent content that is more easily accessible to international audiences. Tweets in other languages are generally less accessible to a global audience, and might broadcast different messages accordingly. In the Uyghur context, Mandarin Chinese represents the language of the oppressor. The predominant language in the Uyghur region is Uyghur, which is a Turkic language similar to Uzbek but written in Arabic script. However, since the early 2000s, Mandarin Chinese has been imposed on the Uyghur population as part of their strategy for escalating repression in the region. With this variation in purported audiences and the role of Mandarin Chinese as a repressive language in the context of this study, we expect to observe the following:

> **Hypothesis 3** *Controlling for gender, English-language tweets on pro-CCP Twitter are less likely to be toxic towards activists compared to tweets published in Chinese.*

We also aim to identify the diversity of topics discussed in pro-CCP tweets that target Uyghur diaspora activists. To better understand how pro-CCP Twitter attacks Uyghur diaspora activists, and possible variation by the gender of a targeted activist, we address the following research questions:

> **Research Question 1** *Which topics appear in pro-CCP smear campaigns targeting Uyghur diaspora activists on Twitter?*

> **Research Question 2** *Do pro-CCP accounts discuss different topics when targeting women Uyghur diaspora activists, compared to when they target men?*

3 Data Collection and Preparation

3.1 Data Sources

In this study, we draw from multiple sources of data:

- Incidents of China's transnational repression targeting the Uyghur diaspora, 2002-2021[1] (Hall and Jardine 2021; Jardine, Lemon and Hall 2021)
- Twitter mentions of Uyghur diaspora activists, 2017-2022
  - Historical Data from the Twitter Academic API's full-archive search endpoint
  - Deleted tweets from the Twitter Information Operations Archive
- List of accounts officially affiliated with the Chinese government, monitored by the Hamilton 2.0 dashboard (Alliance for Securing Democracy 2022)

We will evaluate the extent to which women Uyghur activists residing outside of China are targeted on pro-CCP Twitter by analyzing tweets that mention (by name and Twitter handle) Uyghur activists abroad. Information on targeted individuals in the Uyghur diaspora will be drawn from a gender-balanced sample of individuals included in the data on incidents of China's transnational repression targeting the Uyghur diaspora[2]. Gender is inferred from the pronouns used to describe the targeted individual in the description of an incident in the database on transnational repression, or in cited sources (=1 if she/her pronouns are used, 0 otherwise). Our plan for obtaining gender-balanced samples from this dataset is elaborated in Section 3.2.

Our primary source of social media data is the Twitter Academic API's search endpoint for collecting historical data from all publicly available tweets. To overcome the limitations of exclusively collecting publicly available data, we will also collect deleted tweets from Twitter's transparency reporting on suspended accounts that amplified pro-CCP content related to the treatment of Muslim minorities in the Uyghur region (Twitter Safety 2021; DiResta et al. 2021; Zhang, Jacob and Meers 2021). From these data sources, we will query all tweets with the keyword "Xinjiang" mentioning

[1] `https://oxussociety.org/wp-content/uploads/2021/06/chinas-transnational-repression-of-the-uyghursv1.xlsx`; `https://oxussociety.org/wp-content/uploads/2021/11/Stage-1-Cases.xlsx`

[2] Information on these incidents were collected from searches in topically relevant reports from human rights organizations (e.g. Amnesty International, Human Rights Watch, World Uyghur Congress, and the Uyghur Human Rights Project), keyword searches from Radio Free Asia and newswires, and existing datasets on global transnational repression (Jardine, Lemon and Hall 2021).

the names and Twitter handles in our input list between January 1, 2017 and April 30, 2022[3]. The resulting dataset will comprise of all potentially relevant tweets.

3.2 Sampling Strategy

The population of interest in this study comprises of any activist in the Uyghur diaspora who advocates for the rights of Muslim minorities in the Uyghur region and is active on Twitter. In this study, we primarily focus on Uyghur activists who are in exile, but plan to expand our analysis to any Uyghur advocate who shares a common heritage with Muslim minorities in the Uyghur region. In the context of this study, we establish that an individual is part of the Uyghur diaspora if they indicate it in their profile description.

For our primary analyses, we will study pro-CCP tweets, published in English and Chinese, which mention a gender-balanced sample of 22 individuals who are active on Twitter and have been targeted in more than one publicly recorded incident of Stage 1 Chinese transnational repression. This sampling strategy allows us to focus on attacks targeting more "prominent" activists who are targeted with tactics that align with what we observe on social media. The findings from tweets mentioning this smaller sample of individuals will be important for exploring the broader implications of our findings. To ensure that results are consistent beyond tweets mentioning the individuals in this sample, we will run robustness checks on larger gender-balanced samples drawn from different subsets of the data on publicly recorded incidents of Chinese transnational repression.

Our data collection strategy for primary analyses focuses on prominent individuals who were targeted by Chinese authorities with tactics that do not involve social media. This strategy gives rise to potential issues of selection bias for drawing inferences on the broader implications of pro-CCP social media attacks targeting Uyghur diaspora activists. To address these issues, we will expand on our data collection to include tweets mentioning individuals who have "Uyghur" included in their profile description or username. We use this keyword to identify Uyghur activists due to the partisan divide in how pro-CCP actors and Uyghur advocates refer to the region.

[3] We use January 2017 as the starting point of our data collection to encompass tweets that were published around the time that the Chinese government enacted "XUAR De-Extremification Regulation" legislation in early 2017. This legislation had important implications for how foreign governments treated Uyghur and Turkic Muslims trying to flee the region. For instance, potential escape routes for Muslim minorities throughout Asia and the MENA region were dwindling (Jardine 2022). Other notable events that occurred within the data collection period include the Trump administration's declaration of genocide in the region (Pompeo 2021) and the 2022 Beijing Olympics, also referred to as the "Genocide Games" by individuals around the world who boycotted the Olympics because of human rights abuses committed by the Chinese government in the region.

3.3 Data Annotation Process

Our data annotation pipeline for labeling tweets is illustrated in Figure 1. With names and Twitter handles as inputs for collecting social media data, we will filter out irrelevant content by training crowd-sourced annotators to classify a gender-stratified random sample of tweets mentioning Uyghur diaspora activists. Annotators will label tweets based on whether they target a Uyghur diaspora activist and—if a Uyghur diaspora activist is targeted—the tweet author's stance towards the targeted individual. They will also label tweets based on whether state-sponsored human rights abuses in the Uyghur region are discussed, and whether tweets discussing this topic in any way question whether it is happening. Accordingly, the criteria for this labeling task are:

1. Stance on Uyghur activists
   (a) Whether a Uyghur activist is the target of a tweet
   (b) Stance towards a targeted Uyghur activist is "negative"[4]
2. Stance on human rights abuses in the Uyghur region
   (a) Discusses (alleged) attacks against Muslim minorities in the Uyghur region
   (b) Questions the veracity of these allegations[5]

We consider all tweets that satisfy either set of criteria on stance toward Uyghur activists (1a and 1b) or stance on human rights abuses in the Uyghur region (2a and 2b) as relevant to this study.

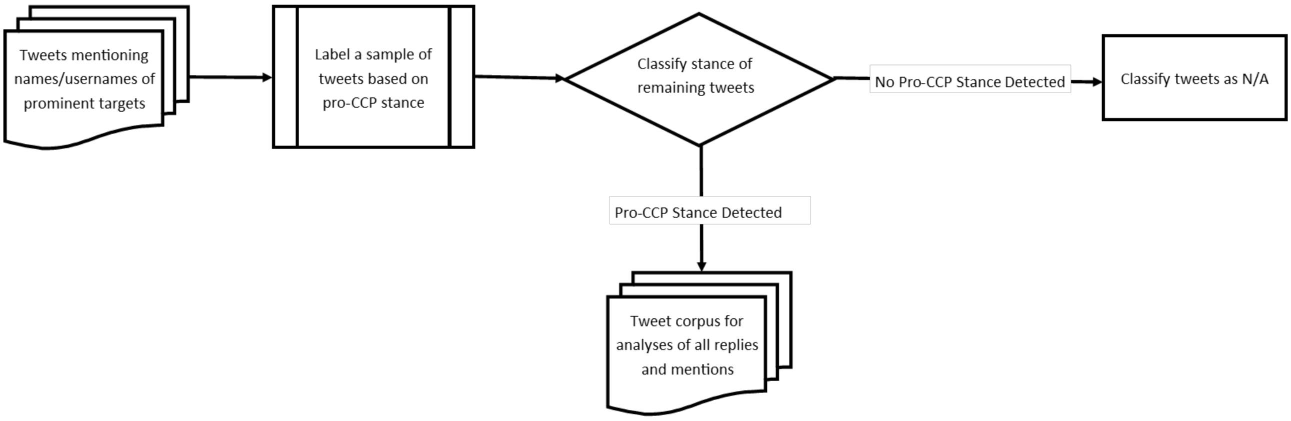
Tweets mentioning names/usernames of prominent targets
Label a sample of tweets based on pro-CCP stance
Classify stance of remaining tweets
No Pro-CCP Stance Detected
Classify tweets as N/A
Pro-CCP Stance Detected
Tweet corpus for analyses of all replies and mentions

Figure 1: Data generation and annotation workflow for primary analyses

Once we subset for relevant content, we will train crowd-sourced annotators to label a gender-stratified sample of tweets, from which we will derive information on the probability that each tweet

[4] The main objective of this task is to filter out tweets demonstrating support for the target of a pro-CCP campaign and journalistic reporting that supports claims made by exiled activists.

[5] e.g.: "Have you ever been to Xinjiang?" or "I've been, and I didn't see that"

is toxic. We focus on measuring the *toxicity* of a tweet, or the probability that a tweet will lead to a target limiting their activity on the platform. Variable measurement with these tweet labels are elaborated in Section 4.1. Once relevant data are procured and labeling tasks on samples of tweets are completed, we will use a supervised text classifier to label the remaining tweets. We will supplement this final set of tweets with information on which accounts are state-affiliated, based on a list of government/diplomatic and state-backed media accounts tracked by the Alliance for Securing Democracy (2022). We elaborate on how these labels will be used for measurement and analysis in Sections 4 and 5.

4 Variable Measurement

4.1 Dependent Variable

Our outcome measure of interest is the probability that a tweet is toxic. A tweet is *toxic* if it has the propensity to lead to a target limit their social media activity, which in the this study includes how Uyghur diaspora activists share their perspectives on human rights abuses in the Uyghur region. The probability that a tweet is toxic is measured by the Perspective API.

4.2 Independent Variables

Gender is the primary independent variable of focus for testing all hypotheses. It is measured as a binary variable, coded as 1 if she/her pronouns are used to refer to a Uyghur diaspora activist in detailed notes or cited sources from the dataset compiled by the Oxus Society for Central Asian Affairs. We presume that the reference group for this variable is predominantly male, but will account for the broader implications of our analysis if this is not the case upon further inspection of the data. The other independent variable of interest is whether a tweet author is a *state-affiliated account*. A tweet is authored by a state-affiliated account if the user appears on a comprehensive list—maintained by the Alliance for Securing Democracy's Hamilton 2.0 project[6]—of government, diplomatic, or state-backed media accounts. Our final independent variable of interest is *language*, which include English and Chinese.

[6] https://securingdemocracy.gmfus.org/hamilton-dashboard/

5 Empirical Strategy

5.1 Hypothesis Testing

All analyses will be performed at the tweet level. We plan to use multiple linear regression analysis to assess the difference in average toxicity scores based on our independent variables of interest[7]. As specified in Section 4.1, the outcome variable of interest for hypothesis testing is the probability that a tweet is *toxic*. The primary independent variable of interest for all hypotheses is the *gender* of a targeted Uyghur diaspora activist, with the aim of comparing differences in average toxicity scores between women and men. We will initially run all specifications without control variables, but will also run specifications with controls. Controls on the individual level will be selected from newswires and datasets on global transnational repression, while country-level variables (e.g. bilateral extradition treaty with China) will be derived from human rights reports on the repression of Muslim minorities in the Uyghur region.

To test our first hypothesis, our primary independent variable of interest is gender. For our second hypothesis, our independent variables of interest are gender and whether a tweet is authored by a state-affiliated account. For our third hypothesis, our independent variables of interest are gender and the language that a tweet is published in. We conduct primary analyses with a corpus of tweets that mention 22 prominent Uyghur diaspora activists who are active on Twitter.

5.2 Exploratory Analysis

In addition to the aforementioned hypothesis testing, we will engage in exploratory analysis to investigate possible sources of variation in the topics observed in pro-CCP smear campaigns that target Uyghur diaspora activists. We will also investigate whether we observe differences in the topics used in tweets that target women diaspora activists compared to men diaspora activists.

5.3 Addressing Limitations

Inferences drawn from focusing on 22 prominent Uyghur diaspora activists who are active on Twitter in our primary analyses may limit the broader applicability of this research to other Uyghur

[7] We may reassess this model specification upon further inspection of the distribution of toxicity scores—our main outcome variable of interest.

diaspora activists who are targeted in pro-CCP smear campaigns on Twitter. To address this potential limitation, we will run analyses on tweets that mention a wider set of Uyghur diaspora activists advocating against Chinese repression of Muslim minorities. The data collection strategy for this corpus is specified in Section 3.2.

A potential limitation with solely using historical data from the Twitter Academic API's full-archive search is that we are missing tweets that have been removed from the platform. As a robustness check, we will compare the results of our analyses with deleted tweets from Twitter's Information Operations archive. We will compare results from this data source with a subset of the corpus used in primary analyses—all tweets that include the word "Xinjiang"—to match the original keyword used in data collection of the relevant data releases from Twitter.

6 Ethical Considerations

We have secured approval from New York University's IRB. Given the potentially hostile nature of the social media posts that this study focuses on, we will give clear instructions for annotators to ensure that they have limited exposure to distressing content. One of the main precautions we will include is that annotators should not open any URLs, as it is possible that the linked content can evoke negative reactions. We rely on publicly available data documenting relevant incidents of transnational repression, to ensure that we are not contributing any additional risk to the individuals we focus on in this study. We will also publish results of analysis with social media data at the aggregate level, with the possible exception for tweets authored by verified accounts.

References

Alliance for Securing Democracy. 2022. "Hamilton 2.0 Dashboard." https://securingdemocracy.gmfus.org/hamilton-dashboard/.

DiResta, Renée, Josh A Goldstein, Carly Miller and Harvey Wang. 2021. "One Topic, Two Networks: Evaluating Two Chinese Influence Operations on Twitter Related to Xinjiang." *Stanford Internet Observatory* p. 44.

Hall, Natalie and Bradley Jardine. 2021. "'Your Family Will Suffer': How China Is Hacking, Surveilling and Intimidating Uyghurs in Liberal Democracies." *Uyghur Human Rights Project and Oxus Society for Central Asian Affairs* .

Jardine, Bradley. 2022. "Great Wall of Steel: China's Global Campaign to Suppress the Uyghurs." *Wilson Center* .

Jardine, Bradley, Edward Lemon and Natalie Hall. 2021. "No Space Left to Run: China's Transnational Repression of Uyghurs." *Uyghur Human Rights Project and Oxus Society for Central Asian Affairs* .

Moss, Dana M. 2018. "The Ties That Bind: Internet Communication Technologies, Networked Authoritarianism, and 'Voice'in the Syrian Diaspora." *Globalizations* 15(2):265–282.

Nyst, Carly and Nick Monaco. 2018. State-Sponsored Trolling: How Governments Are Deploying Disinformation as Part of Broader Digital Harassment Campaigns. Technical report Institute for the Future.

Ong, Lynette H. 2020. *Outsourcing Repression: Everyday State Power in Contemporary China*. Oxford University Press.

Pompeo, Michael R. 2021. "Press Release: Determination of the Secretary of State on Atrocities in Xinjiang." \url{https://2017-2021.state.gov/determination-of-the-secretary-of-state-on-atrocities-in-xinjiang/}.

Schenkkan, Nate and Isabel Linzer. 2021. "Out of Sight, Not Out of Reach: The Global Scale and Scope of Transnational Repression." *Freedom House (8 February 2021), available at:{https://freedomhouse.org/report/transnational-repression} accessed* 7.

Thakur, Dhanaraj and DeVan Hankerson Madrigal. 2022. Facts and Their Discontents: A Research Agenda for Online Disinformation, Race, and Gender. Preprint Open Science Framework.

Twitter Safety. 2021. "Disclosing State-Linked Information Operations We've Removed." https://blog.twitter.com/en_us/topics/company/2021/disclosing-state-linked-information-operations-we-ve-removed.

Zhang, Albert, Wallis Jacob and Zoe Meers. 2021. "Strange Bedfellows on Xinjiang: The CCP, Fringe Media and US Social Media Platforms." *Australian Strategic Policy Institute* p. 28.