

UC San Diego

UC San Diego Electronic Theses and Dissertations

Title

Essays in Computational Studies of Political Behavior on Social Media

Permalink

<https://escholarship.org/uc/item/2pb879zz>

Author

Chang, Keng-Chi

Publication Date

2024

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA SAN DIEGO

Essays in Computational Studies of Political Behavior on Social Media

A dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy

in

Political Science with a Specialization in Computational Social Science

by

Keng-Chi Chang

Committee in charge:

Professor Margaret Roberts, Chair
Professor James Fowler, Co-Chair
Professor John Ahlquist
Professor Seth Hill
Professor Will Styler

2024

Copyright
Keng-Chi Chang, 2024
All rights reserved.

The dissertation of Keng-Chi Chang is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2024

DEDICATION

Dedicated to my family for their love and support.

EPIGRAPH

這時日落的方向是西
越過眼前的柏樹。潮水
此岸。但知每一片波浪
都從花蓮開始——那時
也曾驚問過遠方
不知有沒有一個海岸？
如今那彼岸此岸，惟有
飄零的星光

— 楊牧 〈瓶中稿〉

The west is where the sun sets
over the cypresses before my eyes, waves
on this shore, but I know every breaker
begins at Hualien. Once, a confused boy
asked the distant land:
Is there a shore on the other side?
Now I'm on this shore, that's the other shore, and I see only
wandering stars

— Yang Mu, "Manuscript in a Bottle"

translated by Michelle Yeh

TABLE OF CONTENTS

Dissertation Approval Page	iii
Dedication	iv
Epigraph	v
Table of Contents	vi
List of Figures	ix
List of Tables	xi
Acknowledgements	xii
Vita	xiv
Abstract of the Dissertation	xv
Chapter 1 Introduction	1
Chapter 2 COVID-19 Increased Censorship Circumvention And Access To Sensitive Topics In China	5
2.1 Crisis is a Gateway to Censored Information	5
2.2 The COVID-19 Crisis in China	7
2.3 The Effect of Crisis on Information Seeking and Censorship Circum- vention	8
2.3.1 Crisis Increased Censorship Circumvention	9
2.3.2 Increases in Circumvention Occurred Throughout China	13
2.3.3 Crisis Provided a Gateway to Censored Political Information	14
2.3.4 Comparison with Other Countries Affected by the Crisis	22
2.4 Materials and Methods	23
2.4.1 Data	23
2.4.2 Models	26
2.4.3 Data Availability	26
2.5 Discussion	26
2.6 Acknowledgments	28
Chapter 3 Mapping Visual Themes among Authentic and Coordinated Memes	32
3.1 Introduction	32
3.2 Prior Works and Limitations	33
3.3 Methodology	34
3.4 Preliminary Findings	38

	3.5	Discussions and Future Steps	42
	3.6	Appendix	43
	3.7	Acknowledgments	43
Chapter 4		Characterizing Image Sharing Behaviors in US Politically Engaged, Random, and Demographic Audience Segments	46
	4.1	Introduction	46
	4.2	Data and Methodology	48
	4.2.1	Demographic Inference from Profile Pictures	50
	4.2.2	Collecting and Characterizing Images in Tweets	50
	4.3	Results	51
	4.4	Discussion and Conclusions	53
	4.5	Acknowledgments	57
Chapter 5		Do Images Lend Credibility to News Articles?	58
	5.1	Introduction	58
	5.2	Research Questions	59
	5.3	Experimental Design	61
	5.3.1	Sample	61
	5.3.2	Experimental Procedure	61
	5.3.3	Story and Image Selection	62
	5.4	Inferring Latent Treatments from Treatment Images	63
	5.4.1	Break the Dependency between Discovery and Estimation	63
	5.4.2	Overall Workflow in Analyzing Image Treatments	65
	5.4.3	Low-dimensional Latent Treatment Discovery	67
	5.5	Results	69
	5.5.1	Average Treatment Effect	69
	5.5.2	Heterogeneous Treatment Effect	72
	5.6	Evaluate Gains from Treatment Targeting	72
	5.7	Additional Method for Treatment Discovery by Direct Dimension Reduction on Embeddings	76
	5.7.1	The Indian Buffet Process	78
	5.7.2	Extract Latent Dimensions via SVD on Embeddings	78
	5.7.3	Score Treatment Images by Latent Dimensions	84
	5.7.4	Estimated Effects Based on Alternative Method	85
	5.8	Conclusions and Discussions	85
	5.8.1	Future Extensions	88
Appendix A		Supporting Information for: COVID-19 Increased Censorship Circumvention And Access To Sensitive Topics In China	89
	A.1	Twitter Activity by Province	89
	A.2	Twitter Data	91
	A.3	Mobility and Twitter Usage	92

A.4	Effect Size	98
A.4.1	New Twitter Users	98
A.4.2	Followers	99
A.4.3	Number of unique devices accessing Wikipedia with cookies enabled	102
A.5	Robustness Checks	103
A.6	Wikipedia Country Comparisons	109
A.6.1	Page view analysis	109
A.6.2	Analysis of an expanded set of historical political pages and ‘politically sensitive’ pages using Wikipedia2vec	114
A.7	Text Analysis of Tweets	120
A.7.1	Hand labels	120
A.7.2	Topic models	124
Appendix B	Supporting Information for: Do Imageries Lend Credibility to News Articles? 136	
B.1	Stimuli	136
B.1.1	Treatment Titles, Images, and Excerpts	136
B.1.2	(T1) Chinese hackers seeking to disrupt communications between US and Asia in event of crisis, Microsoft says	136
B.1.3	(T2) Biden Administration Delayed Sanctions over Spy Balloon to ‘Limit Damage’ to China Ties	137
B.1.4	(F1) China threatens to shoot Nancy Pelosi’s plane down if she visits Taiwan	138
B.1.5	(F2) Soviet and Chinese communists have grabbed control of U.S. entertainment, movies, television, music, academia, K-12 education and the news media	138
B.2	Evaluate Image Clustering	139
B.3	Additional Tables	140
B.4	Survey Questionnaire	142

LIST OF FIGURES

Figure 2.1:	Download Rank of iPhone Application in China: Facebook, Twitter, and Wikipedia. Data from AppAnnie.	10
Figure 2.2:	(Top) Number of Unique Geo-Locating Users in China Posting in Chinese. (Bottom) The Fraction of Active Users Who Joined Twitter in the Last 30 Days.	12
Figure 2.3:	Views of Wikipedia Pages in Chinese	13
Figure 2.4:	Increases in GeoLocated Twitter Activity by Province (modeled)	15
Figure 2.5:	Increases in Twitter Followers from China vs Hong Kong By Category	18
Figure 2.6:	Increases in Twitter Followers China vs Hong Kong By Category (Regression Estimate)	19
Figure 2.7:	Views of Blocked, Current Leader, and Historical Leader Wikipedia Pages in Chinese, German, and Italian.	30
Figure 3.1:	Distribution of predicted probability of memes for IRA images	35
Figure 3.2:	DeepCluster Pipeline (from Chaudhary (2020))	36
Figure 3.3:	Example memes from Clusters 5 (top) and 46 (bottom).	37
Figure 3.4:	t-SNE projection of IRA and Reddit memes.	39
Figure 3.5:	Cluster labels and shares of IRA/Reddit memes	40
Figure 3.6:	Confusion matrix for logistic regression predicting IRA memes using only visual representations	42
Figure 3.7:	Representative memes from selected clusters.	44
Figure 3.8:	Predicted memes (top) and non-memes (bottom) of IRA images.	45
Figure 4.1:	Distributions of Predicted Account-Level Demographics—Race, Gender, and Age—Across Random and Politically Engaged Audiences.	49
Figure 4.2:	Cluster Quality Metrics.	51
Figure 4.3:	Cluster Distribution by Audience.	52
Figure 4.4:	Regression Coefficients of Account-Level Cluster Distributions on Demographics.	54
Figure 4.5:	Random images from predictive clusters (cluster ids on top left of each panel)	55
Figure 5.1:	Screenshot of Example Treatment Stimuli	62
Figure 5.2:	Flow of Survey Experiment	63
Figure 5.3:	BLIP-2 Model Architecture (from Li et al. (2023)).	65
Figure 5.4:	UMAP Projection of the Image-Block Embeddings.	67
Figure 5.5:	Top Clusters of Image Blocks by Latent Treatments Z_1 – Z_{10} Inferred by SIBP.	68
Figure 5.6:	Average Treatment Effects of Latent Treatments.	70
Figure 5.7:	Individual-level and Mean Outcome by Treatment Image.	71
Figure 5.8:	Heterogeneous Treatment Effects.	73
Figure 5.9:	Qini Curve.	74

Figure 5.10: Fong and Grimmer (2016): Top words for latent treatment and estimated effects	76
Figure 5.11: Variance Explained and Cumulative Variance Explained by Top Latent Dimensions.	80
Figure 5.12: Treatment Images on the 1st & 2nd Latent Dimensions.	81
Figure 5.13: Treatment Images on the 3rd & 4th Latent Dimensions.	82
Figure 5.14: Treatment Images on the 5th & 6th Latent Dimensions.	83
Figure 5.15: Top 8 activated images on the latent dimensions for each latent treatment	86
Figure 5.16: Average Treatment Effects of Alternative Latent Treatments.	87
Figure A.1: Increases in Geolocated Twitter Activity by Province (modeled)	90
Figure A.2: Within city movement index by Province (black: 2020, dotted: same period in 2019).	93
Figure A.3: Reduction in within city movement and increase in geolocated Twitter users during the month of Wuhan lockdown (left); degree of moving out and increase in geolocated Twitter users on the day of Wuhan lockdown (right).	94
Figure A.4: Weekly changes in within city movement and geolocated Twitter users relative to pre-lockdown period, after adjusting for the same period in 2019.	97
Figure A.5: Excess Followers, absolute (top) and ratio normalized by category growth rate (bottom). Growth rate is calculated based on the December 2019 average number of new followers by category.	101
Figure A.6: Increases in Twitter Followers from mainland China versus Hong Kong by Week	103
Figure A.7: Increases in Twitter Followers from China versus Taiwan	105
Figure A.8: Increases in Twitter Followers from China versus US	106
Figure A.9: Increases in Twitter Followers from China versus Others (Regression Estimate)	107
Figure A.10: New Users Stay on Twitter at the Same Rates across Locations	108
Figure A.11: Views of Blocked, Current Leader, and Historical Leader Wikipedia Pages in Other Countries	113
Figure A.12: Changes in views of historical leader Wikipedia pages (expanded set of pages)	117
Figure A.13: Changes in views of ‘politically sensitive’ Wikipedia pages (expanded set of pages)	118
Figure A.14: Changes in views of Alexei Navalny related Wikipedia pages	119
Figure A.15: Changes in views of current leader Wikipedia pages (expanded set of pages)	121
Figure A.16: Tone of tweets by popular Twitter accounts across main countries mentioned	123
Figure A.17: Tone of tweets by popular Twitter accounts across topics of tweets	123
Figure B.1: Treatment Images for Treatment T1.	136
Figure B.2: Treatment Images for Treatment T2.	137
Figure B.3: Treatment Images for Treatment F1.	138
Figure B.4: Treatment Images for Treatment F2.	139
Figure B.5: Elbow and Silhouette Methods for Evaluating K-Means.	139
Figure B.6: Distribution of Clusters from K-Means.	140

LIST OF TABLES

Table 2.1: Empirical Tests	9
Table 2.2: During the lockdown period, Wikipedia views in Chinese increased relative to overall views for politically sensitive Wikipedia pages and political leader pages, as well as for historical political leaders.	31
Table 4.1: Summary statistics of Twitter sample	50
Table 5.1: Summary Statistics and Balance Table	64
Table A.1: Persistence of Followers by Account Type and Period Following Starts	100
Table A.2: Top relative increases for Wikipedia pages January 24 through March 13 compared to December 2019.	110
Table A.3: Top absolute daily increases for Wikipedia pages January 24 through March 13 compared to December 2019.	111
Table A.4: Lockdown dates	112
Table A.5: List of opposition-related pages in Russian that were checked for significant increases during lockdown.	120
Table A.6: Topic model on tweets that mention China.	126
Table A.7: Topic model on tweets that mention any country.	131
Table B.1: Average Treatment Effects	141

ACKNOWLEDGEMENTS

I am forever indebted to my advisors, Molly Roberts and James Fowler. At every step of my PhD journey, Molly provided me with continuous guidance, unwavering support, and endless encouragement. James offered a broad perspective and always provided helpful feedback whenever I needed it. They have given me opportunities I did not deserve but strive to live up to.

I am also immensely grateful to my committee members: John Ahlquist, Seth Hill, and Will Styler. Their extensive knowledge and generous assistance have been crucial in my PhD journey. Learning from their high-quality work and feedback has motivated me to continue improving my own work.

Many of my collaborators and coauthors also served as mentors during my PhD journey, especially Cody Buntain, Kevin Munger, Katie McCabe, Tiago Ventura, Zack Steinert-Threlkeld, Will Hobbs, Jake Shapiro, and Tuomas Oikarinen. It was collaborations that made my PhD journey enjoyable.

I would like to thank the American Political Science Association and the Rapoport Family Foundation for their generous funding of my research. I also thank the UCSD SHORE program for providing guaranteed housing throughout my PhD, without which I would not have survived the housing crisis.

Professors in the CSE department and the Halicioğlu Data Science Institute reshaped my understanding of machine learning and artificial intelligence, especially Julian McAuley, Sanjoy Dasgupta, Gal Mishne, and Berk Ustun. I am grateful for the open-mindedness of CSE in accepting internal Master's degree students. HDSI also hired me as a teaching assistant for the capstone project for three years, where I learned many skills in mentorship and management.

To my parents, who always gave me the freedom to pursue my dreams and always accepted me for who I am. Mom, you shoulder unimaginable responsibilities at home and work, and I admire you eternally. Dad, although you cannot see the colorful world with your own eyes, your musical talent makes our lives more colorful. Navigating interests in both social science

and math as a child was tricky, but somehow, we made it work. To my brother and sister, thank you for taking care of our parents over these years, especially during the pandemic.

Finally, I would like to thank my partner, husband, and lifelong friend, Norton Cheng, for his love, patience, and understanding. Getting through our PhD journeys together in South Mesa 9192B under the San Diego sunshine has been the most beautiful memory of my life. It's a memory I will continue to dream of in the years to come.

Chapter 2, in full, is a reprint of the material as it appears in Keng-Chi Chang, William R. Hobbs, Margaret E. Roberts, and Zachary C. Steinert-Threlkeld, "COVID-19 Increased Censorship Circumvention and Access to Sensitive Topics in China", *Proceedings of the National Academy of Sciences*. 2022, 119 (4). The dissertation author was the primary researcher and author of this paper.

Chapter 3, in full, is a reprint of the material as it appears in Keng-Chi Chang, "Mapping Visual Themes among Authentic and Coordinated Memes", *Workshop on Images in Online Political Communication of the 16th International AAI Conference on Web and Social Media (ICWSM)*. 2022. The dissertation author was the primary investigator and author of this paper.

Chapter 4, in full, is a reprint of the material as it appears in Keng-Chi Chang and Cody Buntain, "Characterizing Image Sharing Behaviors in US Politically Engaged, Random, and Demographic Audience Segments", *Workshop on Images in Online Political Communication of the 17th International AAI Conference on Web and Social Media (ICWSM)*. 2023. The dissertation author was the primary investigator and author of this paper.

VITA

- 2015 Bachelor of Arts in Economics, National Taiwan University
- 2016–2017 Visiting Graduate Student, Department of Economics,
University of Wisconsin–Madison
- 2017–2018 Research Assistant, Behavioral and Data Science Research Center,
National Taiwan University
- 2018–2024 Graduate Student Researcher, University of California San Diego
- 2020–2024 Teaching Assistant, Halicioğlu Data Science Institute,
University of California San Diego
- 2024 Master of Science in Computer Science,
University of California San Diego
- 2024 Doctor of Philosophy in Political Science with Specialization in Computational
Social Science, University of California San Diego

PUBLICATIONS

Justin Lee, Tuomas Oikarinen, Arjun Chatha, **Keng-Chi Chang**, Yilan Chen, and Tsui-Wei Weng. “The Importance of Prompt Tuning for Automated Neuron Explanations.” *NeurIPS 2023 Workshop on Attributing Model Behavior at Scale (ATTRIB)*.

Keng-Chi Chang, and Cody Buntain. “Characterizing Image Sharing Behaviors in US Politically Engaged, Random, and Demographic Audience Segments.” *PhoMemes Workshop of 2023 International AAI Conference on Web and Social Media (ICWSM-2023)*.

Keng-Chi Chang, Will Hobbs, Margaret Roberts, and Zachary Steinert-Threlkeld, “COVID-19 Increased Censorship Circumvention and Access to Sensitive Topics in China.” *Proceedings of the National Academy of Sciences*, 2022, 119 (4).

Keng-Chi Chang. “Mapping Visual Themes among Authentic and Coordinated Memes.” *PhoMemes Workshop of 2022 International AAI Conference on Web and Social Media (ICWSM-2022)*.

Victoria Asbury, **Keng-Chi Chang**, Katherine McCabe, Kevin Munger, and Tiago Ventura, “The Effect of Streaming Chat on Perceptions of Debates.” *Journal of Communication*, 2021, 71(6): 947–974.

Keng-Chi Chang, Chun-Fang Chiang, and Ming-Jen Lin “Using Facebook Data to Predict the 2016 U.S. Presidential Election.” *PLoS One*, 2021, 16(12).

Tiago Ventura, Kevin Munger, Katherine McCabe, and **Keng-Chi Chang** “Connective Efferescence and Streaming Chat During Political Debates.” *Journal of Quantitative Description: Digital Media*, 2021, 1: 1–49.

ABSTRACT OF THE DISSERTATION

Essays in Computational Studies of Political Behavior on Social Media

by

Keng-Chi Chang

Doctor of Philosophy in Political Science with a Specialization in Computational Social Science

University of California San Diego, 2024

Professor Margaret Roberts, Chair
Professor James Fowler, Co-Chair

Although social media is often optimized for broadcasting visual information, studies of political social media have primarily focused on textual content. Concerns about synthetic images and influence operations have only widened our knowledge gap. This dissertation, comprising four papers, examines the supply of vision-based misinformation generated by autocratic actors, the demand for visual content in democracies across demographic groups, and the extent to which images can shape political behaviors, such as credibility perceptions toward news articles. Additionally, I investigate how information control in autocracies loses its effectiveness during crises. To answer these questions, I combine large-scale digital trace data, open-source

pretrained models for computer vision and natural language processing, and causal inference methods. Overall, my work offers data-driven insights into the intersection of visual media, political communication, and information control.

Chapter 1

Introduction

I study political behaviors on social media using computational methods: how authoritarian regimes utilize new technologies for influence, and how new technologies are shaping political behaviors such as trust toward news media. My research combines large-scale digital trace data with methods from computer vision and natural language processing that facilitate use of novel imagery and text for causal inference around these questions. So far my research has been focused on three main themes: (1) How do societies with information control interact with ones without information control? (2) How do authoritarian regimes utilize visual media on foreign digital platforms? How do people in these societies consume visual media? (3) Does visual information shape political behavior differently from other formats? I also develop methodological tools to answer these questions. I show how we can make use of learned representations from machine learning algorithms to understand observational and causal processes of online political behavior.

In the coauthored paper in my dissertation Chapter 2 titled “COVID-19 Increased Censorship Circumvention And Access To Sensitive Topics In China” we investigated how Covid lockdowns in China has incentivized people jump the Great Firewall to seek out information on Twitter, subsequently followed Twitter accounts that are sharing topics deemed sensitive by

the Chinese government. We first show that the app download rankings such as VPN, Twitter, Facebook, and Wikipedia in China have boosted significantly since the lockdown in Wuhan. We then use geolocated Tweets in Chinese to show prolonged increase in Twitter usage in China across provinces. We then analyze the followers of popular Chinese Twitter accounts. Using a difference-in-differences design, we show that, by comparing followers before and after the lockdown in Wuhan, international news agencies and citizen journalists receive disproportionately more new followers from China than followers from Hong Kong. In contrast, entertainment or pornography accounts receive the same proportion of new followers from China, relative to Hong Kong.

Since our study, Chinese diplomats and state-link propaganda accounts have also become active on Twitter, often engaging the so-called “wolf warrior diplomacy”. Even so, the White Paper Protests, one of the largest series of demonstrations in China in response to the zero-covid policy, have also been largely mobilized by Chinese accounts on Twitter. One Twitter account, @whyoutouzhele, has gained 1 million followers in a week by sharing real-time protest images and videos. I am in the process of analyzing how the largest protest in China since 1989 was mobilized on a censored platform overseas.

The second paper in my dissertation Chapter 3 is titled “Mapping Visual Themes among Authentic and Coordinated Memes” I try to understand the **supply** of political visual misinformation initiated by foreign actors. Specifically, I study how political memes are utilized by the Russian Internet Research Agency (IRA) on Twitter. Conceptually, I hypothesize that visual memes are especially useful for repeating group-based contentious narratives. Political actors can utilize visual memes to reinforce group distinctions. Although social media is often optimized for broadcasting visual information, studies of political social media have primarily focused on textual content. With the development of large-scale pretrained open-source models, social scientists are now equipped with computational tools to analyze these visual content. I use information operations data released by Twitter and memes on Reddit to develop a computational

pipeline to understand what kinds of visual frames are commonly promoted by state actors, compared to authentic actors. First, I train a deep learning classifier classify images shared by the Russian IRA accounts into memes and non-memes. I find that around 40% of the images shared by the Russian IRA accounts can be classified as memes. Second, I use a pretrained model to extract visual embeddings from Russian IRA memes and authentic memes from Reddit. I then use K-means clustering to identify common visual themes in two sets of memes, and compare the prevalence of these themes between the two sets. I find that, compared with authentic memes from regular users, coordinated memes from Russian IRA accounts promote more themes around scenes of military strength, close-ups of faces emphasizing gender, screenshots, and slogans. I also find that a simple logistic regression on pretrained visual embeddings can discern between Russian IRA memes and authentic memes with an F1 score of 0.84.

In a third paper in my dissertation Chapter 4 titled “Characterizing Image Sharing Behaviors in US Politically Engaged, Random, and Demographic Audience Segments” we sought to understand the **demand** for visual content on social media. Specifically, we conjecture that preferences for visual content are heterogeneous across different audience segments, especially by gender, age, race, and political engagement. To study the relationship between audience segments and visual content, we collected timelines from two sets of Twitter users: a random sample of users geolocated to the US and a sample of politically engaged users (follow at least 5 political Twitter accounts) in the US. We use the profile pictures of these users and use model adjusted for minority representation to proxy the gender, ethnicity, and age information among these users. We then download 10 million images shared by these users and use a pretrained model to extract visual embeddings and K-means to identify common visual themes. Lastly, we regress the prevalence of visual themes on the learned demographic segments to understand what kinds of visual themes are predictive of user’s gender, race, age, and political engagement. We find that while most visual themes are shared among politically engaged and non-political Twitter users, around half of the visual themes are predictive of the user’s gender, race, and age segments. This

suggests it is possible to target demographic groups with specific visual content on social media.

The last paper in my dissertation Chapter 5 is titled “Do Images Lend Credibility to News Articles?” I try to understand, experimentally, whether there are additional advantages to visual media while consuming online news. I design and implement an online visual survey experiment to quantify the effects of visual media frames on credibility of news articles. Respondents are exposed to news containing only textual stimuli or news with both text and visual stimuli, and measure their perceptions of news credibility. I use online image search to collect 30 distinct images as treatments for each new story. The design would allow me to compare the treatment effects of the visual-based treatment to textual treatment, as well as the treatment effects of each types of visual frames. I utilize pretrained models to extract visual embeddings from the treatment images, and use causal discovery models to find latent treatments among the treatment images. I find that, compared to text-only news, news articles with added visual images, overall, are not perceived as more or less credible. However, using causal discovery models, I find that certain latent treatments, such as symbols of communism, images with female, and comics would decrease credibility perception. On the other hand, photos of press conferences and images with males would increase credibility perception. I also find effect heterogeneity among respondents: images with female only has negative effects on credibility perception among male respondents, and black respondents don’t trust images with white males as much as white respondents do. Lastly, I also develop additional methods to learn latent treatments directly from pretrained embeddings, without relying on clustering or sacrificing interpretability.

In summary, the work in this dissertation advances our understanding the role of visual media—how autocratic actors utilize it, how regular users consume it, and how researchers can analyze and experiment on it at scale. I hope that by taking an interdisciplinary approach, informed by new advances in other fields such as computer science, I can provide more descriptive evidence and experimental tests on important behaviors not well documented by social scientists.

Chapter 2

COVID-19 Increased Censorship

Circumvention And Access To Sensitive Topics In China

2.1 Crisis is a Gateway to Censored Information

In many authoritarian countries, traditional and online media limit access to information (Morozov, 2011; MacKinnon, 2012; Deibert et al., 2011; Sanovich, Stukal and Tucker, 2018). While this control is imperfect, studies have shown that media control in autocracies has large effects on the opinions of the general public and the resilience of authoritarian regimes (Stockmann and Gallagher, 2011; Enikolopov, Petrova and Zhuravskaya, 2011; Adena et al., 2015; Yanagizawa-Drott, 2014; Stockmann, 2012; Huang, 2015; Roberts, 2018), even though there are moments when it can backfire (Pan and Siegel, 2020; Jansen and Martin, 2003; Nabi, 2014; Hassanpour, 2014; Hobbs and Roberts, 2018; Gläbel and Paula, 2019; Boxell and Steinert-Threlkeld, 2019). Evidence from China suggests that media control may be effective in part because individuals generally do not expend significant energy to find censored or alternative sources of

information.¹

While many have studied the impact of information control in normal times in authoritarian regimes, less is known about information seeking during crisis. In democracies, information seeking intensifies during crisis, increasing consumption of mass media. Ball-Rokeach and Defleur (Ball-Rokeach and DeFleur, 1976) describe a model of dependency on the media where audiences are more reliant on mass media during certain time periods, especially when there are high levels of conflict and change in society. These findings are largely consistent with research on emotion in politics, which concludes that political situations that produce anxiety motivate people to seek out information (Marcus, Neuman and MacKuen, 2000). While in normal times information seeking is strongly influenced by pre-existing beliefs, several studies have suggested that crisis can cause people to seek out information that might contradict their partisanship or worldview (Marcus and MacKuen, 1993; MacKuen et al., 2010), although they may pay disproportionate attention to threatening information (Albertson and Gadarian, 2015).

Similar patterns may exist in authoritarian environments. Because the government controls mass media, citizens aware of censorship may not only consume more mass media that is readily available during crises, but also seek to circumvent censorship or seek out alternative sources of information that they may normally not access. For example, during the SARS crisis in China in 2003, Tai and Sun (Tai and Sun, 2007) find that people in China turned to SMS and the Internet to gather and corroborate information they received from mass media. Cao (Cao, 2020) shows an increase in censorship evasion and use of Twitter from China during “regime-worsening” events, such as worsening of trade relations between the U.S. and China and the removal of Presidential term limits in the constitution in 2018.

Outside of facilitating access to information about the crisis, evasion of censorship during

¹Stockmann (2012) provides evidence that consumers of newspapers in China are unlikely to go out of their way to seek out alternative information sources. Chen and Yang (2019) provided censorship circumvention software to college students in China, but found that students chose not to evade the Firewall unless they were incentivized monetarily. Roberts (2018) provides survey evidence that very few people choose to circumvent the Great Firewall because they are unaware that the Firewall exists or find evading it difficult and bothersome.

crisis could also provide information that has long been censored. In particular, a crisis could create spillovers of information, where evasion to find one piece of information facilitates access to a broad range of content. This phenomenon is related to the entertainment-driven “gateway effect” documented in (Hobbs and Roberts, 2018), where sudden censorship of an entertainment website (Instagram) motivated censorship evasion and thus facilitated access to unrelated political information. At the same time, crisis is a very different context than sudden censorship of an entertainment website. Anxiety about the epidemic, perhaps especially when accompanied by boredom during quarantine and lockdown, could lead consumers of information to be more likely to seek out information that has long been censored after they have evaded censorship to better understand the trustworthiness of their government. On the other hand, the crisis itself may be sufficiently distracting to make them less likely to seek out unrelated and long censored information. Further, crisis-induced spillover effects are more difficult for autocrats to avoid than gateways created through censorship of entertainment websites, which could be reduced by avoiding the initial censorship altogether or implementing less visible censorship. While the overall impact on the autocrat is unknown and could be outweighed by a successful, rapid government response to the crisis, such a gateway would strengthen the ability of consumers to read sources outside of China.

2.2 The COVID-19 Crisis in China

On December 31, 2019, officials in Wuhan, China confirmed that a pneumonia-like illness had infected dozens of people. By January 7, 2020, Chinese health officials had identified the disease – a new type of coronavirus called novel coronavirus, later renamed COVID-19. By January 10, the first death from COVID-19 was reported in China, and soon the first case of COVID-19 was reported outside of China, in Thailand. As of December 2020, COVID-19 has infected over 91,000 people in China with over 4,500 deaths, and at least 73.5 million people

worldwide with over 1.6 million deaths.²

While initial reports of COVID-19 were delayed by officials in Wuhan (Buckley and Myers, 2020), Chinese officials took quick steps to contain the virus after it was officially identified and the first deaths reported. On January 23, 2020, the entire city was placed under quarantine – the government disallowed transportation to and from the city and placed residents of the city on lockdown (Qin and Wang, 2020). The next day, similar restrictions were placed on 9 other cities in Hubei province (Griffiths, John and George, 2020). While Hubei province and Wuhan were most affected by the outbreak, cities all over China were subject to similar lockdowns. By mid-February, about half of China – 780 million people – were living under some sort of travel restrictions (Griffiths and Woodyatt, 2020). Between January 10 and February 29, 2020, 2,169 people in Wuhan died of the virus (Belluck, 2020).

2.3 The Effect of Crisis on Information Seeking and Censorship Circumvention

We use digital trace data to understand the effect of the COVID-19 crisis on information seeking. Table 2.1 summarizes the empirical tests conducted in this paper. First, we show that the crisis increased the popularity of virtual private network (VPN) applications, which are necessary to jump the Great Firewall, downloaded on iPhones in China. We also show that the crisis expanded the number of Twitter users in China, which has been blocked by the Great Firewall since 2009. The crisis further increased the number of page views of Chinese language Wikipedia, which has been blocked by the Great Firewall since 2015. We also show that the areas more affected by the crisis – such as Wuhan and Hubei Province – were more likely to see increases in circumvention.

²Source: New York Times, December 15, 2020. <https://www.nytimes.com/interactive/2020/world/coronavirus-maps.html>

Next, we show that the increase in circumvention caused by the crisis not only expanded access to information about the crisis, but also expanded access to information that the Chinese government censors. On Twitter, blocked Chinese language news organizations and exiled dissidents disproportionately increased their followings from mainland China users. On Wikipedia, sensitive pages such as those pertaining to Chinese officials, sensitive historical events, and dissidents showed large increases in page views due to the crisis. Last, the fourth subsection shows that these dynamics do not occur on Italian, German, Persian, or Russian Wikipedia – languages of countries with similar crises but where Wikipedia is uncensored.

Table 2.1: Empirical Tests

Question	Test
1. Do individuals circumvent censorship more during crisis?	VPN ranking; increased use of blocked services; new Twitter users.
2. Do individuals access crisis information?	Wikipedia traffic about current leaders; new mainland China followers for certain account types.
3. Do individuals access non-crisis sensitive information?	Wikipedia traffic to blocked pages; new mainland China followers for activists and foreign political figures.
4. Do these same dynamics occur in democracies and less censored environments?	Wikipedia page views in German, Italian, Persian, and Russian.

2.3.1 Crisis Increased Censorship Circumvention

We show that censorship circumvention increased in China as a result of the crisis using data from application analytics firm AppAnnie, which tracks the ranking of iPhone applications in China. While most VPN applications are blocked from the iPhone Apple Store, we identified one still available on it. Around the time of the Hubei lockdown, its rank popularity increased significantly and maintained that ranking (top panel of Figure 2.1).³

³To protect the application and its users, we are not disclosing its name or the exact ranking.

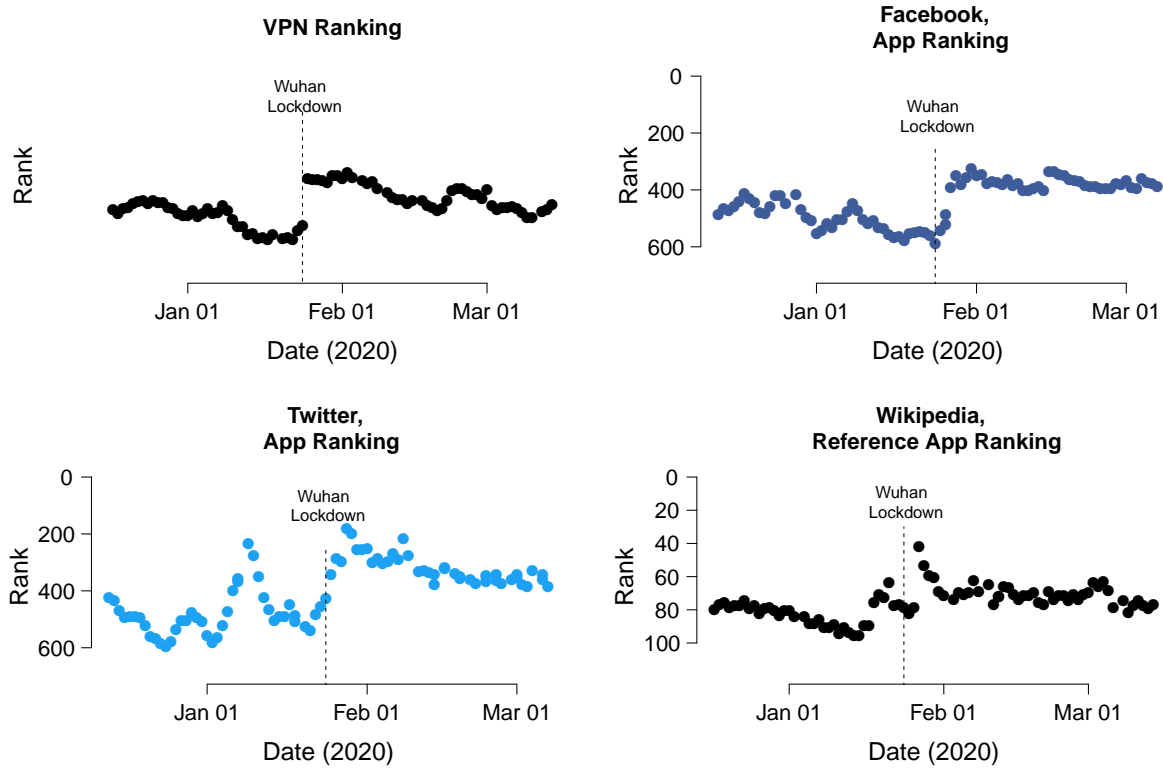


Figure 2.1: Download Rank of iPhone Application in China: Facebook, Twitter, and Wikipedia. Data from AppAnnie.

Note: The top panel of this figure intentionally omits the name of the VPN app and its precise ranking.

Concurrent with the increase in popularity of the VPN application is a sudden increase in popularity of Facebook, Twitter, and Wikipedia applications, as Figure 2.1 shows.⁴ These increases indicate that those jumping the Firewall as a result of the crisis were engaging in part with long blocked websites in China – Twitter and Facebook have been blocked since 2009 and Chinese language Wikipedia since 2015.

This finding is consistent with data we collected directly from Twitter and Wikipedia. The top panel of Figure 2.2 shows the number of geolocating users in China posting to Twitter in Chinese in the time period of interest. Immediately following the lockdown, Chinese language

⁴Note that increase in popularity is not comparable across applications because popularity is measured in terms of ranks. More highly ranked applications (like Facebook and Twitter) may need many more downloads to achieve a more popular ranking.

accounts geolocating to China increased 1.4 fold, and post-lockdown, 10% more accounts were active from China than before. The bottom panel of Figure 2.2 shows that the crisis also coincided with increases of new users, indicating that increases are due to new users and not dormant ones reactivating.⁵ We provide a rough, back-of-the-envelope calculation for the absolute size of these effects. If there were 3.2 million Twitter users in China (Mozur, 2019) prior to the COVID-19 pandemic and the 10% increase in usage applies generally to Twitter users (i.e. not just those geotagging), then 320,000 new users joined Twitter because of the crisis, including users who do not post or post publicly. We assess this estimate in SI Appendix, section 4 using the estimated fraction of posts in Chinese that are geotagged (1.95%) and the total number of unique Twitter users in our sample (47,389 users posting in Chinese and in China).

Data from Wikipedia on the number of views of Wikipedia pages by language matches the App Annie and Twitter patterns.⁶ We measure the total number of views for Chinese language Wikipedia by day from before the coronavirus crisis to the time of writing. Figure 2.3 reveals large and sustained increases in views of Chinese language Wikipedia, beginning at the Wuhan lockdown and continuing above pre-COVID levels through May 2020. Views of all Wikipedia pages in Chinese increased by around 10% during lockdown and by around 15% after the first month of lockdown. This increase persisted long after the crisis subsided. In absolute terms, the total number of page views increases from around 12.8 million views per day in December 2019, to 13.9 million during the lockdown period (January 24 through March 13), and up to 14.7 million views per day from mid-February through the end of April.

⁵SI Appendix, section 2 provides more detail, and SI Appendix, Figure A1 shows trends per province.

⁶Wikipedia page view data is publicly available: <https://dumps.wikimedia.org/other/pagecounts-ez/merged/>. Note that this data does not track where users are from geographically; we use language as an imperfect proxy for geography.

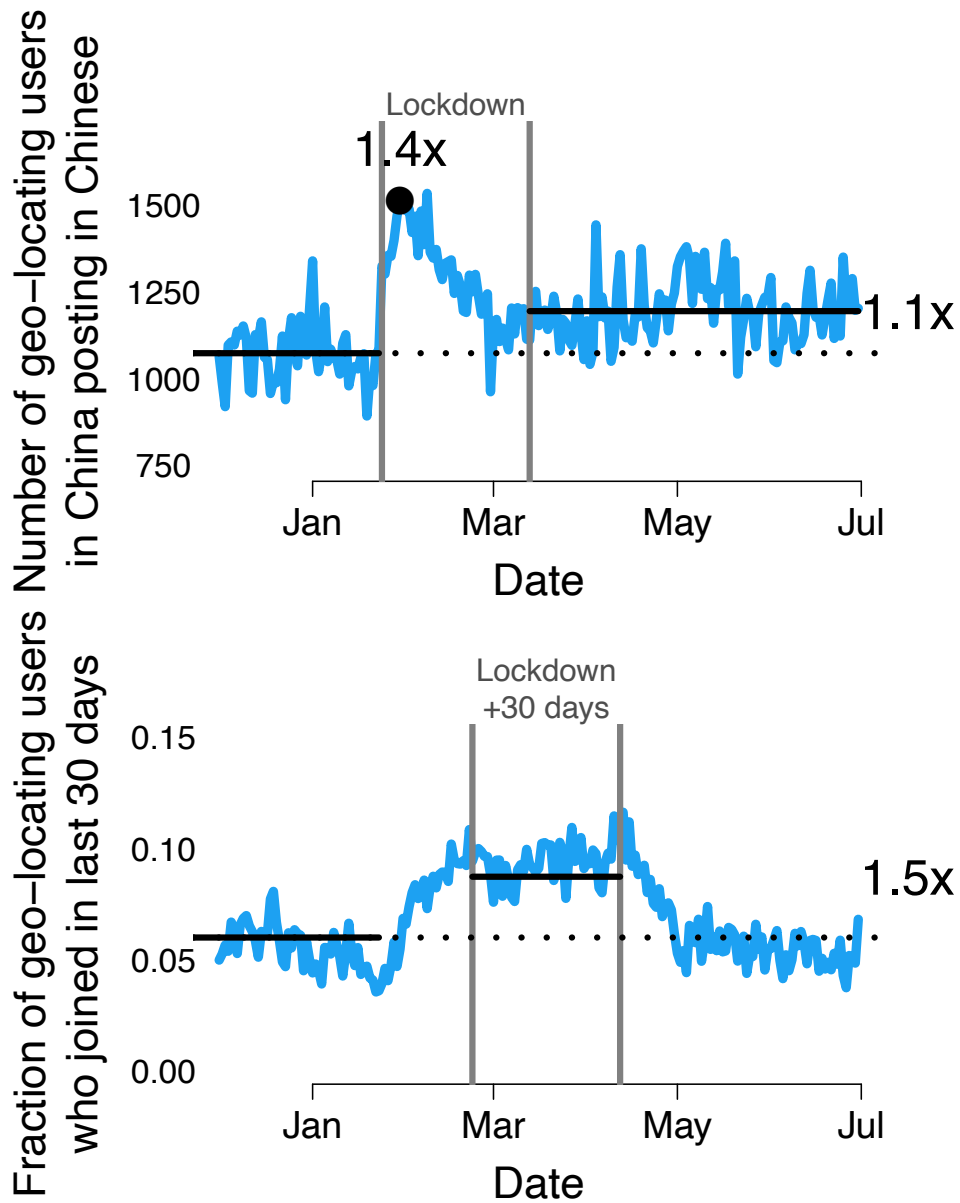


Figure 2.2: (Top) Number of Unique Geo-Locating Users in China Posting in Chinese. (Bottom) The Fraction of Active Users Who Joined Twitter in the Last 30 Days.

Note: These figures display unique users and unique users who signed up within the last 30 days. The decline in ‘new’ users after the end of lockdown in the right panel is driven by a decline in new sign-ups after lockdown easing, rather than lockdown users leaving the site (they are no longer considered new after 30 days).

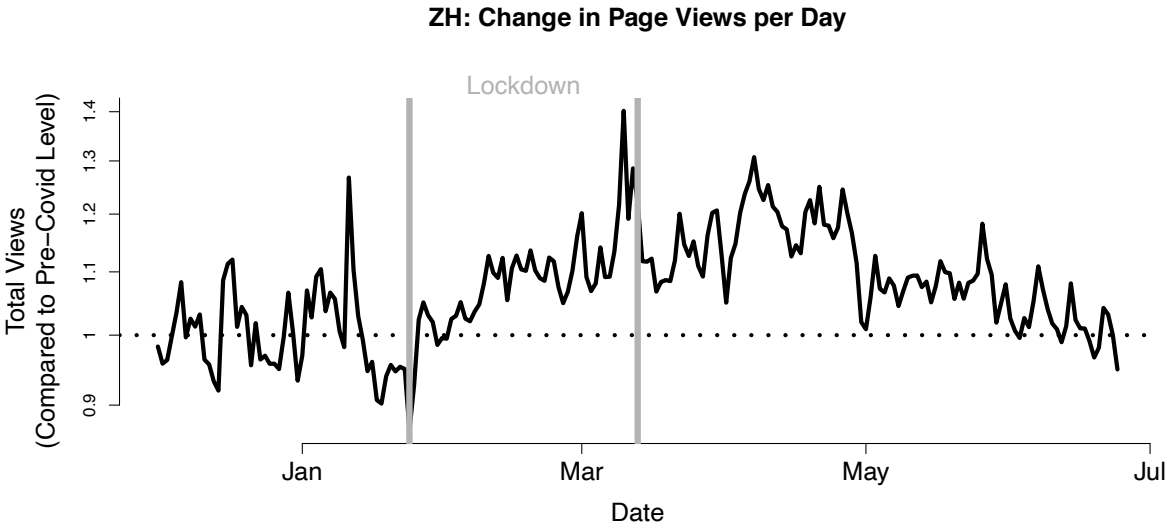


Figure 2.3: Views of Wikipedia Pages in Chinese

Note: This figure shows the ratio of total daily views of Wikipedia pages in Chinese compared to December 2019 views (12.7 million views per day in December 2019). The beginning of the Hubei lockdown and the first relaxation of lockdown in Hubei are indicated in gray.

2.3.2 Increases in Circumvention Occurred Throughout China

Whereas the data from AppAnnie and Wikipedia cannot distinguish between circumvention patterns within China, the geo-location in the Twitter data enables the examination of subnational variation. Circumvention occurred in provinces throughout China as a result of the Wuhan lockdown; Hubei, the most impacted province, experienced the most sustained increase in geolocated users.

Figure 2.4 measures the initial increase of Twitter volume on January 24, 2020, the day after Wuhan’s lockdown and the start of lockdown in twelve other cities in Hubei, in comparison to the average from December 1, 2020 to January 22, 2020 in each province in China (the x-axis). The y-axis measures how sustained the increase was – the ratio of Twitter volume 30 days after the quarantine to the baseline before the outbreak. Hubei is in the top-right corner of the plot: Twitter volume there doubled in comparison to the previous baseline, and the doubling persisted

30 days after the crisis.⁷ These estimates are drawn from polynomial models fit to the daily number of users per province – SI Appendix, Figure A1 displays the modeled lines over the raw data for each province.

To further validate that this increase in Twitter usage in China is related to the Wuhan lockdown, we collected real-time human mobility data from Baidu, one of the most popular map service providers in China. The decrease in mobility in 2020 is correlated with the increase in Twitter users across provinces in China, net of a New Year’s effect (SI Appendix, Figure A3). However, as the crisis spreads, the demobilization effect disappears, while Twitter usage remains elevated. The overall increase in Twitter users across China two weeks after the lockdown and beyond cannot be explained by further decreases in mobility or New Year seasonality (SI Appendix, Figure A4). SI Appendix, section 3 presents more detail.

2.3.3 Crisis Provided a Gateway to Censored Political Information

This subsection examines how the crisis impacted what content Twitter users from mainland China and users of Chinese language Wikipedia were consuming. Both Twitter and Wikipedia facilitate access to a wide range of content, not just information sensitive to the Chinese government. New users of Twitter from China might follow Twitter accounts producing entertainment or even Twitter accounts of Chinese state media and officials, who have become increasingly vocal on the banned platform (Zhou, 2020). New users of Wikipedia might only seek out information about the virus and not about politics. If the crisis produced a gateway effect, we should see increases in consumption of sensitive political information unrelated to the crisis.

⁷While almost all provinces experience a sustained increase in Twitter volume, Beijing and Shanghai have an overall decrease in Twitter volume after the outbreak. We suspect many Twitter users in Beijing and Shanghai left those cities during the outbreak, which is corroborated by the Baidu mobility data we detail in SI Appendix, section 3.

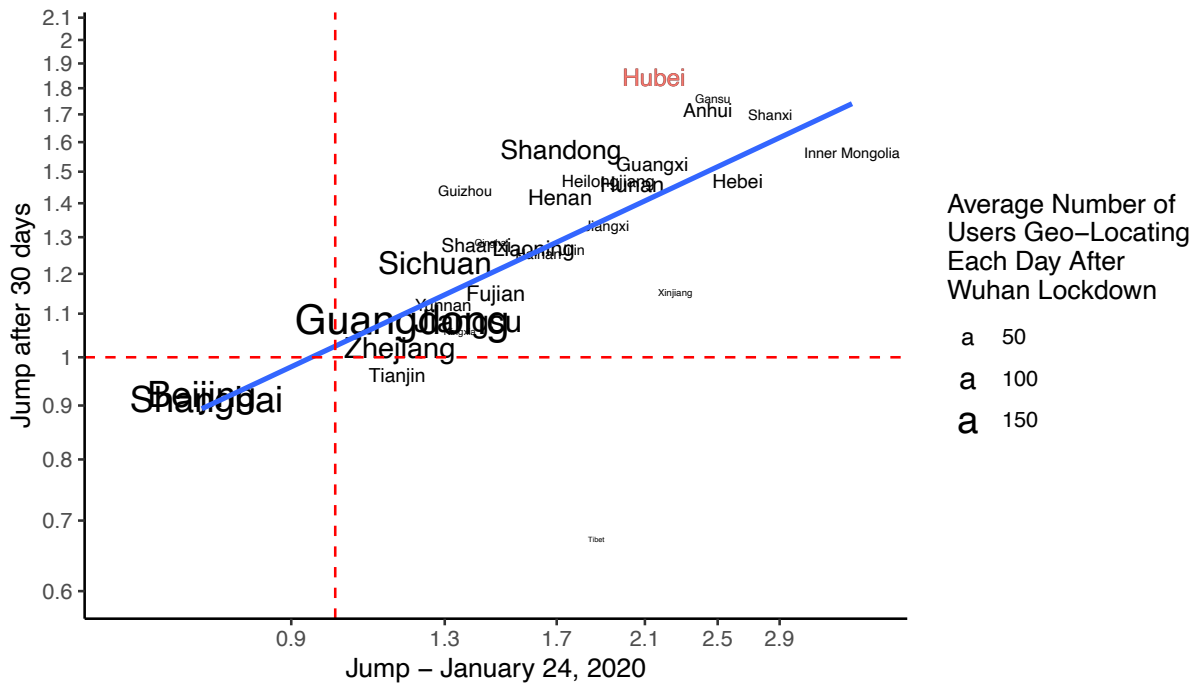


Figure 2.4: Increases in GeoLocated Twitter Activity by Province (modeled)

Note: This figure shows the increase in geo-located Twitter users compared to the average number of geo-located Twitter users in a province before the Hubei lockdown. Estimates for 30 days after and day of lockdown are drawn from a five term polynomial regression on the number of unique geo-located Twitter users per day after the lockdown. These province-by-province polynomials are displayed over the raw data in SI Appendix, Figure A1.

Types of Twitter accounts mainland China Twitter users started to follow as a result of the crisis

We use data from Twitter to examine what types of accounts received the largest increases in followers from China due to the crisis. For this purpose, we identify 5,000 accounts that are commonly followed by Twitter users located in China.⁸ The Materials and Methods and SI Appendix, section 2 detail how we identified these accounts.

We assigned each of the 5,000 popular accounts into one of six categories: 1) international sources of political information, including international news agencies; 2) Chinese citizen journalists or political commentators, which include non-state media discussions of politics within China; 3) activists, or accounts disseminating information about politics in the U.S., Taiwan, or Hong Kong; 4) accounts disseminating pornography; 5) state media and political figures; and 6) entertainment or commercial influencers. Categories 1, 2, and 3 are accounts that might distribute information sensitive to the Chinese government, such as international media blocked by the Great Firewall (e.g. New York Times Chinese and Wall Street Journal Chinese); Chinese citizen journalists and political commentators such as exiled political cartoonist Badiucao and currently detained blogger Yang Hengjun; and political activists such as free speech advocate Wen Yunchao and Wu'er Kaixi, former student leader of the 1989 Tiananmen Square Protests. Accounts in Category 4 are pornography, which we consider sensitive because it is generally censored by the Chinese government, but not politically sensitive like Categories 1-3. Accounts in Category 5 include accounts linked to the Chinese government, including the government's news mouthpieces Xinhua and People's Daily, as well as the Twitter accounts of Chinese embassies in Pakistan and Japan. Category 6 is also not sensitive, as these accounts mostly do not tweet

⁸We note that follower behavior is a useful window into user behavior, and has advantages over other metrics in this context like the content of the new users' tweets. First, merely following accounts is likely a less risky behavior than publicly posting content about politics, especially that related to China. That is, we expect users to self-censor their posts but not (to the same extent) who they only follow. Second, tweet activity is right skewed in our data, which is common in social media data. The median account in the stream tweets twice, and the top 1% of active users author 40.3% of tweets. Analyzing tweets would therefore create a less complete analysis of user behavior than analyzing following relationships.

about politics, but instead are entertainment or commercial accounts or accounts of non-political individuals.

We want to understand how the coronavirus crisis affected trends in follower counts of each of the six categories, and in particular, compare how the crisis affected the followings of categories 1-3 to those in categories 5 and 6. We therefore downloaded the profile information of all accounts that began following popular accounts in categories 1-3 and 5-6 and a random sample of popular accounts from category 4 after November 1, 2019. We then use the location field to identify which of the 38,050,454 followers are from mainland China or Hong Kong (see SI Appendix, section 2 for more details).

Because Twitter returns follower lists in reverse chronological order, we can infer when an account started following another account (Steinert-Threlkeld, 2017). For the accounts in the six categories, we compare the increase in followers from mainland China to the increase in followers from Hong Kong accounts relative to their December 2019 baselines; we chose Hong Kong because it is part of the PRC but is not affected by the Firewall. The ultimate quantity of interest is the ratio of these two increases. If the ratio is greater than one, then the increase in following relationships is more pronounced among mainland Twitter users as compared to those from Hong Kong.

Figure 2.5 shows this ratio by category-day. Relative to Hong Kong, the crisis in mainland China inspired disproportionate increases in the number of followers of international news agencies, Chinese citizen journalists, and activists (some of whom might otherwise, without exposure on Twitter, be obscure within China, especially ones who have been banned from public discourse for a long time) – users who are considered sensitive and often have long been censored. In comparison, there is only a small increase in mainland followers of Chinese state media and political figures during the lockdown period and a slight decrease for non-political bloggers and entertainers. Figure 2.6 reports the regression estimate for the relative ratio of number of new followers (akin to a difference-in-differences design with Hong Kong as control group and De-

New Followers Compared to Baseline, China / Hong Kong

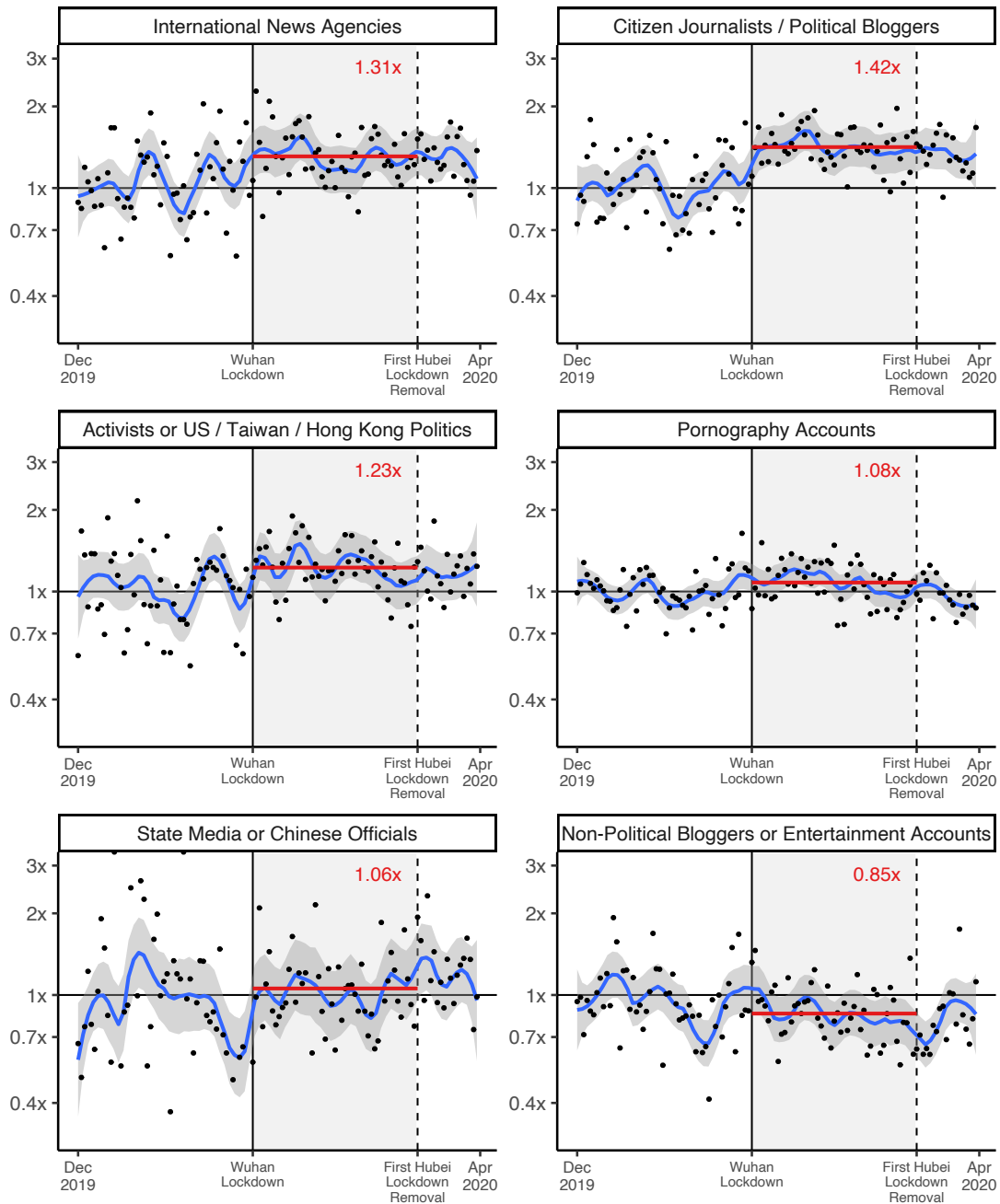


Figure 2.5: Increases in Twitter Followers from China vs Hong Kong By Category

Note: Gain in followers from mainland China compared to Hong Kong across six types of popular accounts, relative to December 2019 trends. Ratios here approximate the incidence rate ratios estimated in the models for Figure 2.6. A value greater than 1 means more followers than expected from mainland China than from Hong Kong. Accounts creating sensitive, censored information receive more followers than expected once the Wuhan lockdown starts. Accounts that are not sensitive or censored, such as state media or entertainment, do not see greater than expected increases.

Relative Size of New Followers, China / Hong Kong

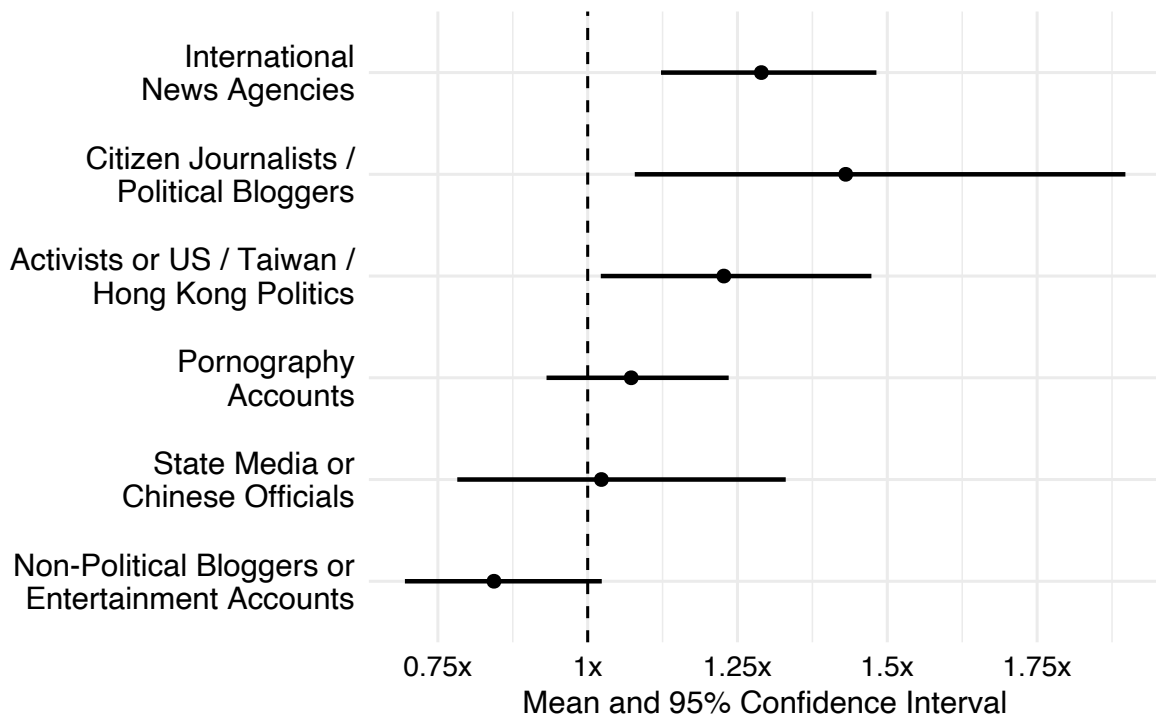


Figure 2.6: Increases in Twitter Followers China vs Hong Kong By Category (Regression Estimate)

Note: Incidence rate ratios shown above are from negative binomial regressions of number of new followers on the interaction between indicator variables for ‘in lockdown period’ and ‘in mainland China’, with December 2019 as control period and Hong Kong as control group.

cember 2019 as pre-treatment period). The result is the same.

We then demonstrate that the result does not depend on the choice of comparison group, the relative increase starts no earlier than the Wuhan lockdown. SI Appendix, Figure A6 conducts a placebo test by running weekly regressions, showing that the relative increase in followers in China starts precisely during the week of lockdown. SI Appendix, Figures A7, A8, and A9 show that the same pattern holds with alternative comparison groups such as overseas Chinese in Taiwan and the United States.

Chinese government information operations on Twitter do not explain the results. Of the 28,991 accounts Twitter identified as belong to a Chinese government information operation⁹, none author a tweet in the 1,448,850 streamed geo-located corpus. To confirm this paucity, we then analyze the 14,189,518 tweets Twitter provided from the information operation accounts. Only .03% of those tweets are geotagged. Twelve of the 1.45 million tweets mention five information operation accounts. We then download tweets from 1,000 users from China and find zero mentions or retweets of the information operation accounts. We also find that none of these information operation accounts follow any of the popular accounts for which we collected followers.

SI Appendix, section 4 provides effect size estimates. There, we roughly estimate that around 320,000 new users came from China. Further, based on December 2019 follower growth rates, 53,860 excess accounts follow citizen journalists and political bloggers; 52,144 for international news agencies. By the end of the lockdown, citizen journalists and political bloggers benefit from 3.63 times the number of followers they otherwise would have had; activists, 2.97. Importantly, 88-90% of the followers from China follow accounts in these categories one year later, and these rates are higher than for accounts which start following in the weeks after the end of the Hubei lockdown. In addition, SI Appendix, Figure A10 shows that new users from China

⁹In June 2020 and September 2019, Twitter released datasets containing 28,991 accounts it identified as being part of pro-China information operation campaigns (<https://transparency.twitter.com/en/reports/information-operations.html>). Twitter granted us access to the unhashed version of the data they do not publicly release, meaning we had the information operation campaigns' accounts' actual screen names and user identification numbers.

persist in tweeting at the same rates as those from Hong Kong and Taiwan.

Types of Chinese language Wikipedia pages that received the most attention

To better understand patterns of political views in the Wikipedia data, we leverage existing lists (see Materials and Methods for additional details) to categorize the Chinese language Wikipedia views into three different categories: 1) Wikipedia pages that were selectively blocked by the Great Firewall¹⁰ prior to Wikipedia’s move to https (after which all of Chinese language Wikipedia was blocked), 2) pages that describe high level Chinese officials¹¹, and 3) historical leaders of China since Mao Zedong. Whereas we would expect that a crisis in any country should inspire more information seeking about current leaders in Category 2, only if crisis created a gateway to historically sensitive information would we expect proportional increases in information seeking about historical leaders in Category 3 or information about sensitive events that were selectively blocked by the Great Firewall on Wikipedia prior to 2015 in Category 1.

Figure 2.7 shows the increase in page views for each of these categories on Chinese Wikipedia relative to the rest of Chinese language Wikipedia. We find that the lockdown not only increased views of current leaders (purple), but also views of historical leaders (yellow) and views of pages selectively blocked by the Great Firewall (red). SI Appendix, Tables A2 and A3 show specific pages disproportionately affected by the increase in views of Wikipedia. While pages related to coronavirus experienced a jump in popularity, other unrelated sensitive pages including the “June 4 Incident,” “Ai Weiwei,” and “New Tang Dynasty Television” (a television broadcaster affiliated with Falun Gong) also experienced an increase in page views.¹²

For more detail on this analysis as well as the Wikipedia pages that received the largest

¹⁰Using data from <https://www.greatfire.org/>.

¹¹These lists are based on offices in the CIA World Facebook. We use this list for ease of comparisons with other countries and remove the Ambassador to the United States from each list. China’s list is available here (and there are links to leaders of other countries on the same page): <https://www.cia.gov/library/publications/resources/world-leaders-1/CH.html>, excluding Hong Kong and Macau.

¹²The June 2020 increase in China is due to the anniversary Tiananmen Square protests. Our claim is not that only the COVID-19 crisis causes increases in views of sensitive content. That the same behavior is observed around another crisis event supports this paper’s argument.

absolute and relative increases in traffic, see SI Appendix, section 6.

2.3.4 Comparison with Other Countries Affected by the Crisis

Since information seeking during crisis is common (Ball-Rokeach and DeFleur, 1976), we investigate Wikipedia data in other languages to explore how other countries were affected by the crisis. We show that the gateway effect of crisis on historically sensitive information is unique to the currently censored webpages in China. For comparison, we focus on Iran, another authoritarian country affected by COVID-19 that previously censored Wikipedia (but does not any longer), and Russia, an authoritarian country that does not censor Wikipedia – for Iran, like China, we know which Wikipedia pages were previously censored (Nazeri and Anderson, 2013). We also show data from democracies without censorship affected early on by the COVID-19 crisis, Italy and Germany.¹³

To make the comparison, we use lists of current leaders from these countries (based on office lists in the CIA World Factbook, see Materials and Methods), and create lists of historical leaders using de facto country leaders since World War II (see SI Appendix, Table A4 for a list of these titles and offices). All of these countries were affected by the crisis in late February or early March and Italy imposed relatively stringent lockdowns. Therefore, we expect increases in information seeking for current leaders, as citizens begin to pay more attention to current politics as the crisis hits. However, none of these countries block Wikipedia. Information seeking about the current crisis therefore should not act as a gateway to information about historical events or controversies, as these pages are always available to the public.

Table 2.2 shows these results. While overall Wikipedia views and page views of current leaders increase in three out of four comparison languages, only for Chinese language Wikipedia do historical leaders increase disproportionately and consistently throughout the whole time pe-

¹³Like China, citizens in each of these countries speak languages relatively specific to their country, and therefore we expect most of the page views of Italian, German, Persian, and Russia Wikipedia to originate in Italy, Germany, Iran, and Russia respectively.

riod. That is, we see an overall effect on information-seeking throughout the world, including for historical leaders; for Chinese language Wikipedia, we see larger increases for historical leaders compared to Wikipedia page views in general. The small increases in historical political leader page views in German and Italian did not correspond with the start of the COVID-19 crisis or their respective lockdowns (Figure 2.7).

Further, we do not see increased attention to pages previously blocked in Iran (Nazeri and Anderson, 2013) during the crisis – Wikipedia pages that can now be accessed without restriction in Iran.

In SI Appendix, section 6.2, we replicate these results for much larger sets of 1) historical leaders and 2) ‘politically sensitive’ pages (pages related to the pre-https blocked pages in Iran and China, and political opposition pages in Russia). We expand these sets of pages using Wikipedia2vec (Yamada et al., 2020), and find that very broad information-seeking about historical leaders and politically sensitive topics occurred only for Chinese language Wikipedia.

2.4 Materials and Methods

2.4.1 Data

Application download rank data. Download rank data for Facebook, Twitter, Wikipedia, and the VPN app come from application analytics firm AppAnnie (<https://www.appannie.com>), which tracks the popularity of iPhone application downloads in China. While most VPN applications are blocked from the iPhone Apple Store (and there are other means of obtaining VPNs), we identified one still available on it. VPN download rank shown in the text is for that VPN application. This data contains the ranking of an application – for Wikipedia, its rank within the Reference App category – rather than the number of downloads. To protect the VPN application and its users, we do not disclose its name or the exact ranking.

Twitter data. For the Twitter analyses, we collected 1,448,850 tweets (101,553 ac-

counts) from mainland China from December 1, 2019 until June 30, 2020. These tweets were identified using Twitter’s POST statuses/filter endpoint. Our analyses are limited to the 367,875 that were posted in Chinese (47,389 accounts that posted in Chinese, 43,114 that had names or descriptions in Chinese).

The Twitter follower analysis examines accounts that Twitter users from China commonly follow. To find those accounts, we randomly sampled 5,000 users geolocated to China. For each of these users, we gathered the entire list of whom they follow, their Twitter “friends.” From these 1,818,159 friends, we extracted the 5,000 most common accounts. We also selected only accounts that were Chinese language accounts or had Chinese characters in their name or description field to ensure that we were studying relevant accounts: those disseminating information easily accessible to most Chinese users. SI Appendix, section 2 provides more detail.

We downloaded the profile information of all accounts that began following these popular accounts after November 1, 2019. Because Twitter returns follower lists in reverse chronological order, we can infer when an account started following another account (Steinert-Threlkeld, 2017). We then use the location field to identify which of these 38,050,454 followers are from mainland China or Hong Kong (see SI Appendix, section 2 for more details). We downloaded all new followers of non-pornography accounts and all new followers of a random selection of 200 pornography accounts (the majority of the accounts were pornography). This sampling allows us to estimate the impact of the coronavirus on pornography while decreasing our requests to the Twitter API.

Mobility data. Human mobility data is publicly available from Baidu Qianxi (<https://qianxi.baidu.com/2020/>), which tracks real-time movement of mobile devices and is used in studies of human mobility and COVID-19 containment measures (Kraemer et al., 2020). Our robustness checks use data across China during the Lunar New Year period in both 2020 and 2019. We scraped the daily within city movement index (an indexed measure of commute population relative to the population of the city), as well as daily moving out index (an indexed measure

based on the volume of population moving out of the province relative to the total volume of migrating population on that day across all provinces in China). See SI Appendix, section 3 for more details.

Wikipedia data. Data on the number of Wikipedia page views is publicly available here: <https://dumps.wikimedia.org/other/pagecounts-ez/merged/>. To better understand patterns of political views in the Wikipedia data, we use existing lists to categorize the Chinese language Wikipedia views into three different categories: 1) Wikipedia pages that were selectively blocked by the Great Firewall (<https://www.greatfire.org/> maintains a list of websites censored by the Great Firewall) prior to Wikipedia’s move to https, after which all of Wikipedia was blocked, 2) pages about high level Chinese officials (using offices listed in the CIA World Factbook <https://www.cia.gov/library/publications/resources/world-leaders-1/CH.html>, excluding Hong Kong and Macau as well as the Ambassador to the United States), and 3) historical ‘paramount’ leaders of China since Mao Zedong.

In comparing multiple languages and countries, we use the same offices listed in the CIA World Factbook to create lists of current leaders from Iran, Russia, Italy, and Germany (for office holders as of February 2020), and create lists of historical leaders using de facto country leaders since World War II. See SI Appendix, Table A4 for a list of these titles and offices, as well as the lockdown start and end dates used for the language by language Wikipedia page view models displayed in Table 2.2. The list of pages of Wikipedia pages blocked in Iran was published by (Nazeri and Anderson, 2013).

In SI Appendix, section 6.2, we replicate the Wikipedia page view results for much larger sets of 1) historical leaders and 2) ‘politically sensitive’ pages (pages related to the pre-https blocked pages in Iran and China, and political opposition pages in Russia). We expand these sets of pages using Wikipedia2vec (Yamada et al., 2020).

2.4.2 Models

Incidence rate ratios for the follower analyses and the Wikipedia page view analyses are from negative binomial regressions. In the follower analysis, this models the number of new followers per day, with a separate model for each account category. Independent variables are ‘in lockdown period’ and ‘in mainland China’, and the effect of interest is the interaction between these indicator variables (i.e. a difference-in-difference), with December 2019 as control period and Hong Kong as control group. The Wikipedia page view analyses use the same specification, reporting the coefficient for ‘in lockdown period’ and ‘in page set’ (current leader, historical leader, previously blocked) relative to December 2019 and relative to page views for the rest of Wikipedia. Observations are the total views per category by day. Figures displaying (log scale) ratios of followers/Wikipedia page views approximate coefficients from these negative binomial regressions. Negative binomial regressions were estimated using the MASS library in R.

Increases in geolocated Twitter activity (unique users) by day and by province were modeled using a five-term polynomial regression (by day) for time trends after the Hubei lockdown and a mean without any time trend prior to lockdown (see SI Appendix, Figure A1 for a province by province visualization of this model). The points in Figure 2.2 are predicted values by province for the first day of lockdown and day 30 of lockdown.

2.4.3 Data Availability

Replication materials are posted on Dataverse (<https://doi.org/10.7910/DVN/W2NSLS>).

2.5 Discussion

Crisis in highly censored environments creates widespread spillovers in exposures to sensitive, censored information, including information not directly related to the crisis. Like in

democracies, consumers of information in autocracies seek out information and depend on the media during crisis. However, in highly censored environments, increased information seeking also incentivizes censorship circumvention. This new ability to evade censorship allows users to discover a wider variety of information than they may have initially sought, and users could also be particularly motivated to seek out accumulated, hidden information during a crisis. Our results suggest that informational spillovers produced by censorship evasion are a result of the structure of censorship, and that they occur beyond government-induced backfire from sudden censorship of popular entertainment websites (Roberts, 2018).

Public exposure to censored information during crisis is almost certainly not the intention of any regime with widespread censorship. However, the effect of this crisis-induced gateway to censored information on public opinion is unknown. In the case studied in this paper, surveys in China show increased support for the CCP over the course of the pandemic (and over the same time as large declines in favorability toward the U.S.) (Guang et al., 2020), even though we show that this increase in support occurs in conjunction with increased access to censored information. These findings could reflect favorable reactions to the government's pandemic policy response that may have overwhelmed negative impacts of access to censored information (Stasavage, 2020). Or, the increase in support at a time of greater evasion of censorship could lend support to previous findings that access to Western news sources can counter-intuitively increase support for the regime (Whyte, 2010; Huang and Yeh, 2019). Studying the impact of evasion during the crisis on public opinion is left to future research. However, we include in SI Appendix, section 7 an exploratory analysis of the content posted by the popular accounts followed by our sample. While we see quite negative coverage of China on these accounts and coverage of sensitive topics such as human rights, the 1989 Tiananmen Square protests, and protests in Hong Kong, we also find that coverage of the United States by international news agencies was much more negative or neutral than positive, and the U.S. could have served as a favorable comparison for China and the Chinese government's handling of the pandemic.

While evaluations of responses to an ongoing crisis and comparisons to other governments' responses to the same crisis may have benefited government officials in China in this particular circumstance (Stasavage, 2020), beyond these evaluations, increased access to historical and long-censored information, as documented here, has the potential to dampen positive or compound negative changes in trust, and may also contribute to easier access to uncensored information about a government in the future. Natural disasters, including epidemics, tend to alter trust in government officials. When a policy response is perceived as efficacious, support for the level of government perceived to have directed the response increases (Lazarev et al., 2014; You, Huang and Zhuang, 2020). On the other hand, neglectful responses can induce subsequent protest participation (Flores and Smith, 2013). In China, the average effect of natural disasters from 2007-2011 was to decrease political trust, and internet users have decreased baseline levels of political trust (You, Huang and Zhuang, 2020; Lee, 2021). At the same time, political surveys in China suffer from preference falsification (Jiang and Yang, 2016; Robinson and Tannenber, 2019; Shen and Truex, 2020), complicating our efforts to understand the political consequences of these events.

While the results here do not link the COVID-19 crisis gateway effect to the political fortunes of the Chinese government, they do suggest that a country with a highly censored environment sees distinctive and wide-ranging increases in information access during crisis. While in normal times censorship can be highly effective and widely tolerated, crisis heightens incentives to circumvent censorship, and regimes cannot rely on the same limits on information access during crisis, even for topics long controlled.

2.6 Acknowledgments

Chapter 2, in full, is a reprint of the material as it appears in Keng-Chi Chang, William R. Hobbs, Margaret E. Roberts, and Zachary C. Steinert-Threlkeld, "COVID-19 Increased Cen-

sorship Circumvention and Access to Sensitive Topics in China”, *Proceedings of the National Academy of Sciences*. 2022, 119 (4). The dissertation author was the primary researcher and author of this paper.

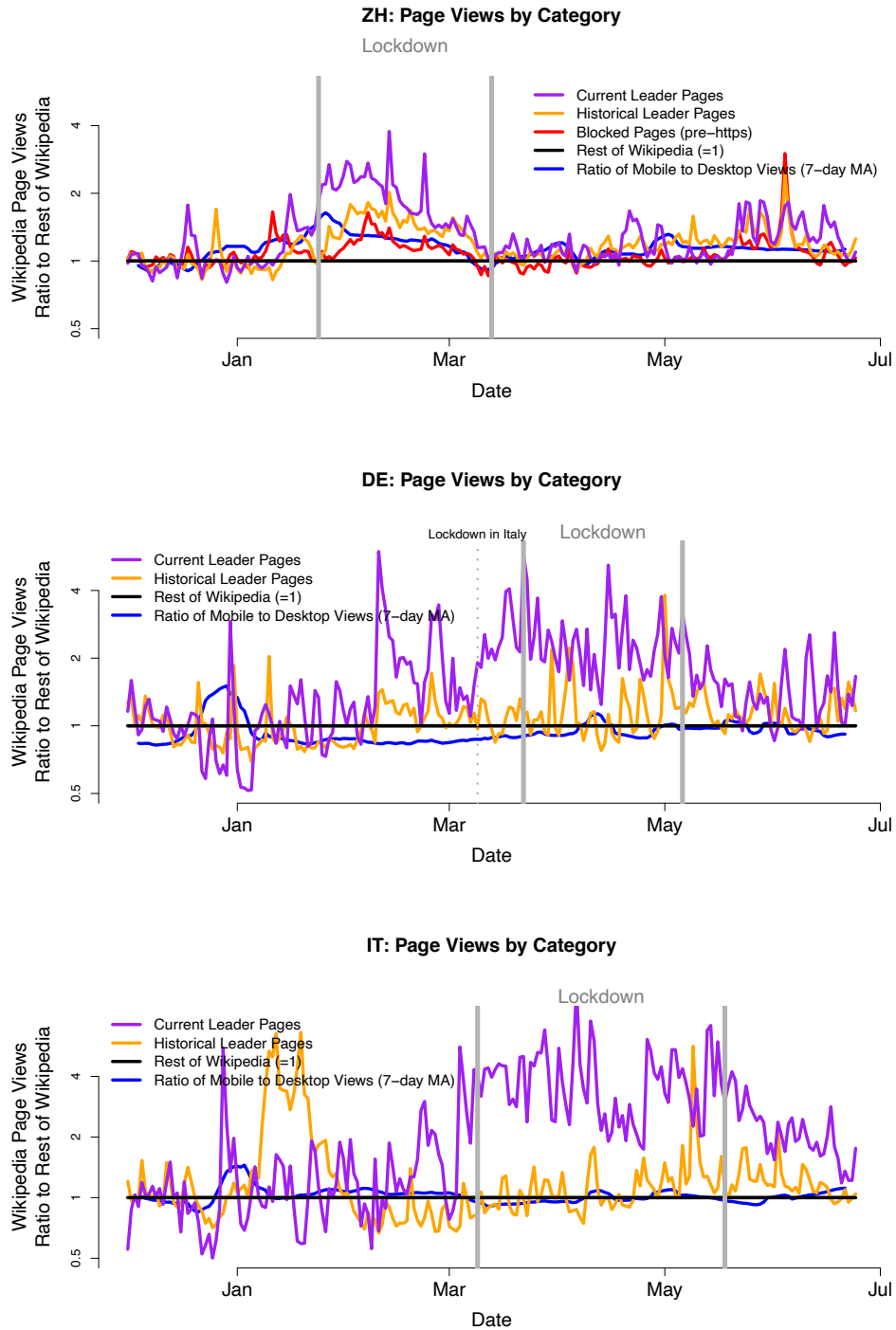


Figure 2.7: Views of Blocked, Current Leader, and Historical Leader Wikipedia Pages in Chinese, German, and Italian.

Note: Vertical lines indicate the starts and ends of lockdown periods—see SI Appendix, Table A4 for specific dates.

Table 2.2: During the lockdown period, Wikipedia views in Chinese increased relative to overall views for politically sensitive Wikipedia pages and political leader pages, as well as for historical political leaders.

Change: Language	Overall	Blocked, pre-https relative to overall:	Leaders	Historical Leaders
Chinese	1.09 (1.05 - 1.12) <0.001	1.15 (1.09 - 1.22) <0.001	1.86 (1.67 - 2.07) <0.001	1.42 (1.32 - 1.52) <0.001
Persian	1.42 (1.37 - 1.46) <0.001	0.84 (0.79 - 0.89) <0.001	0.91 (0.80 - 1.05) 0.20	0.82 (0.75 - 0.90) <0.001
Russian	1.23 (1.18 - 1.28) <0.001		1.73 (1.48 - 2.02) <0.001	0.90 (0.82 - 0.99) 0.03
German	1.16 (1.12 - 1.20) <0.001		2.36 (2.02 - 2.76) <0.001	1.21 (1.05 - 1.40) 0.01
Italian	1.47 (1.40 - 1.53) <0.001		3.29 (2.72 - 4.00) <0.001	1.17 (1.02 - 1.34) 0.03

Note: Incidence rate ratios shown above are from a negative binomial regression estimating the daily number of views within a category in the lockdown period compared to December 2019 relative to the number of views across the rest of Wikipedia compared to December 2019 (using the same difference-in-difference specification as the Twitter follower analysis). Observations are the total views per category by day. 95% confidence intervals are shown in parentheses, and p-values are shown in the third row for each language. See the SI Appendix for over-time ratios by day for all comparison languages (SI Appendix, Figure A11), and for the dates of the lockdowns used (SI Appendix, Table A4). German and Italian pages of historical leaders (shown in orange in the figures above) saw several large and short-lived spikes in views not clearly related to those countries' lockdowns. SI Appendix, Figures A12, A13, A14, and A15 replicate these results for much larger sets of Wikipedia pages, including Russian language pages related to opposition leaders and movements (which did not see broad increases in views).

Chapter 3

Mapping Visual Themes among Authentic and Coordinated Memes

3.1 Introduction

Visual memes (broadly defined as images-with-text) are everywhere on social media; a large fraction is political. According to a panel of 490K Twitter users with voter registration, 19% of their tweets are classified as memes, and 30% of the memes are politically relevant (Du, Masood and Joseph, 2020). Another study on political misinformation in Indian WhatsApp groups finds that 30% of the visual misinformation are memes (Garimella and Eckles, 2020). There are legitimate concerns around state-linked online information operations affecting political behavior, but most studies to date do not leverage the wealth of data in images.

This project aims to document what kinds of visual frames are commonly promoted by state actors compared to generic, authentic memes promoted by regular non-state users. I use a large sample of data from Russian IRA accounts released by Twitter and collect a large sample of authentic memes from `r/memes` on Reddit. I feed a balanced sample of both coordinated state-linked memes and authentic memes (memes promoted by regular users) into a self-supervised

vision model (Caron et al., 2019) to learn the lower-dimensional representations for each meme. I then apply the standard K-means clustering algorithm to the representations to find the clusters of memes. I find that coordinated and authentic memes differ in the visual themes and that a simple logistic regression on the lower-dimensional representations can achieve reasonable accuracy in predicting coordinated vs. authentic memes (on the test set, AUC 0.91, accuracy 0.84, F_1 -score 0.84).

Compared with similar methods relying on multimodal neural networks (Beskow, Kumar and Carley, 2020; Du, Masood and Joseph, 2020), Bag-of-Visual-Words (BOVW), or Perceptual hashing (pHash) (Zannettou et al., 2018, 2019), this transfer learning framework does not rely on extensive tagging (cf. multimodal models), does not only learn on local visual features (cf. BOVW), and does not require memes to be nearly identical (cf. pHash).

3.2 Prior Works and Limitations

Twitter released a ground truth dataset of state-linked operations, including the 1.8M images posted by the accounts controlled by the Russian Internet Research Agency (IRA) during the 2016 US Presidential election. Qualitative studies pointed out that IRA employees are assessed for meme-making capabilities (DiResta, Grossman and Siegel, 2021). Other studies used textual data from IRA found asymmetric flooding of entertainment, not necessarily politics, as a strategy (Cirone and Hobbs, 2022), and that textual content is a reasonable predictor for state-linked campaigns (Alizadeh et al., 2020). Previous work, in a similar effort, also documented the spread of IRA memes online using Perceptual hashing (pHash) of images (Zannettou et al., 2019). However, there is a lack of systematic understanding of the amplification of visual themes by state-linked actors compared to organic ones.

3.3 Methodology

This project has the following steps, which will be explained in subsections.

1. Collect state-linked images, organic memes, and non-meme image-with-text data (as negative samples).
2. Classify state-linked images into memes vs. non-memes.
3. Extracting embeddings of visual feature jointly for both coordinated and authentic memes.
4. Cluster memes based on the vectors of embeddings.
5. Label the clusters and compare the difference in proportions between coordinated and authentic memes.
6. Train a simple baseline model based on visual embeddings to distinguish coordinated and authentic memes.

Data collection

The primary coordinated data are the images shared by IRA on Twitter. Other than the dataset from Twitter, I collected 26K generic memes collected from the `r/meme` subreddit and 15K non-meme image-with-text data (COCO-text, Veit et al. (2016)) as negative samples for training meme vs. non-meme classifier. The reason for choosing non-meme image-with-text data as negative samples for training is that we would not want models simply picking up textual features in images and classifying images into with vs. without text. We contend there there can still be coordinated memes in the `r/meme` subreddit, but the percentage should be low and is outside of the scope of this project.

Classify images into memes

I trained a state-of-the-art deep learning classifier based on ResNet-50 to classify the images shared by IRA into memes vs. non-memes. The accuracy is greater than 0.97 on the test set. Most predicted probabilities are either 0 or 1. Based on this, I find that around 40% of the images shared by the IRA accounts can be classified as memes (see Figure 3.1 for a histogram of predicted probability).

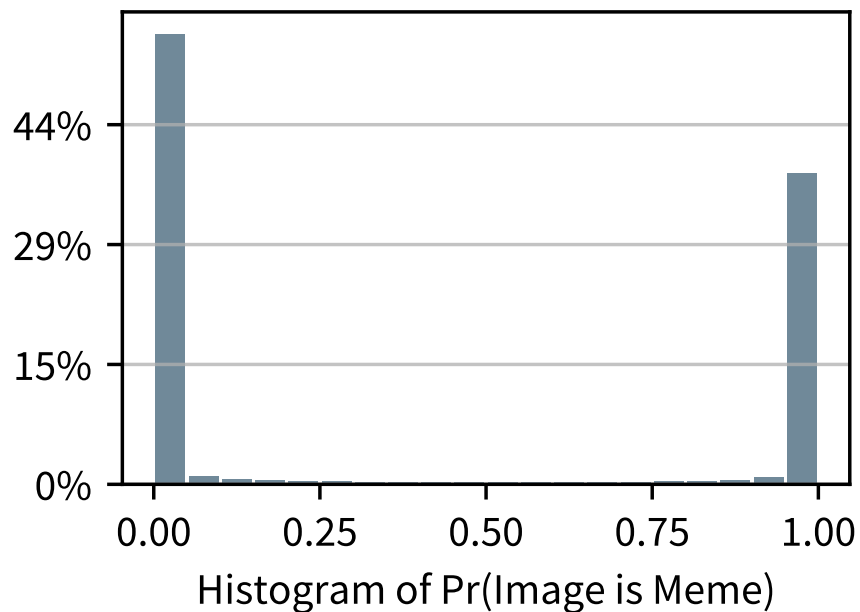


Figure 3.1: Distribution of predicted probability of memes for IRA images

Figure 3.8 in Appendix shows a sample of predicted meme vs. non-meme IRA images. There is room for improvements in accuracy. But since there is no universally accepted definition to serve as ground truth for labeling memes, this might be a simpler procedure without human labeling. For better comparisons, the later analysis will only use the IRA subsample that I classified as memes (predicted probability >0.9) to compare with authentic memes on Reddit.

Extracting representations of visual features

Most traditional image classification tasks are based on supervised learning, but it is hard to scale, especially for memes. Another approach, recently picked-up by social scientists (eg, Torres (2018)), is to extract keypoints via scale-invariant feature transform (SIFT) and find Bag-of-Visual-Words (BOVW) feature representation (Sivic and Zisserman, 2003) by building patches around the neighbor of keypoints and finding clusters of patches. However, this method tends to only focus on local features around the patches and can be not meaningful enough for interpretation.

This paper leverages DeepCluster (Caron et al., 2019, also introduced to the social scientists by Zhang and Peng (2021) that found successes in social science applications), a recent self-supervised method for clustering images. See Figure 3.2 for the pipeline for DeepCluster. Specifically, DeepCluster learns pseudo-labels iteratively by grouping features into clusters and uses the subsequent assignments as supervision to update the weights of the convolutional neural network (ConvNet).

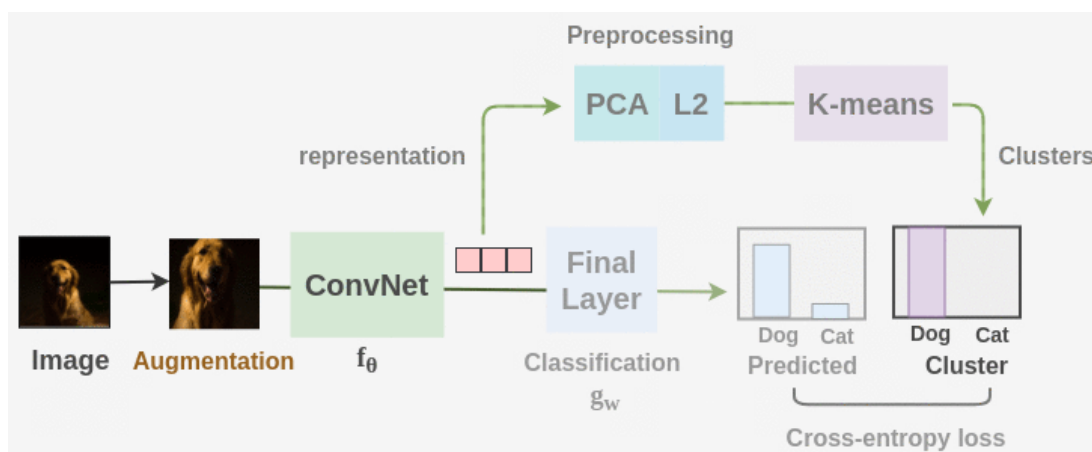


Figure 3.2: DeepCluster Pipeline (from Chaudhary (2020))

I feed a balanced sample of coordinated IRA memes and authentic Reddit memes (each of size 26K) into DeepCluster jointly. Notice that the IRA/Reddit labels are not inputs of the model since this is an unsupervised algorithm, and we also don't want the model to memorize the labels

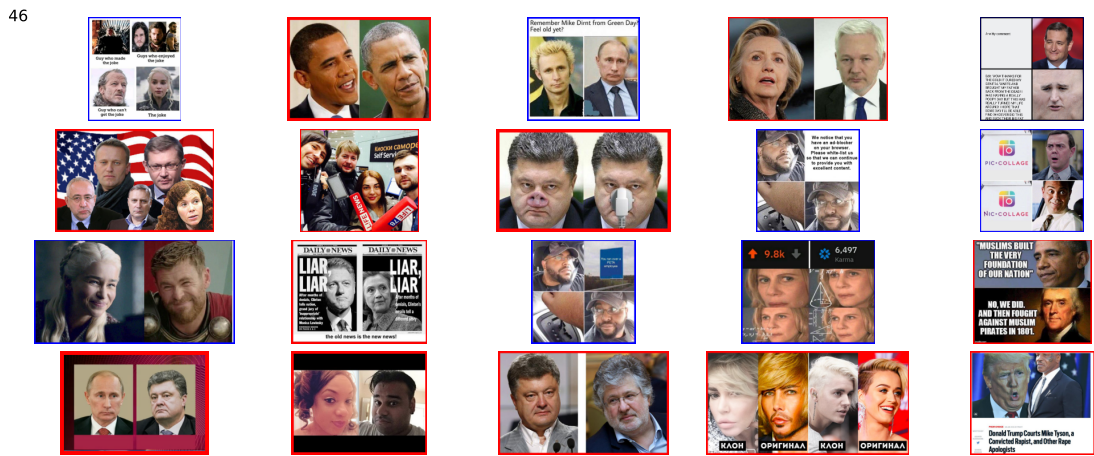


Figure 3.3: Example memes from Clusters 5 (top) and 46 (bottom).

Note: For each, the first row contains the 5 memes nearest to the center of the cluster; the 2-4 rows contain 15 random memes from that cluster. Red border indicates the meme is from IRA; blue border indicates the meme is from Reddit.

at this stage. For faster training, I use pre-trained weights released by the authors (based on the VGG-16 model trained on the ImageNet dataset). ImageNet contains 1.28M images of 1000 categories such as scenes, places, and objects. Thus, the learned embeddings should be helpful in finding general themes in images, not just localized features (e.g., textures) like those in BOVW. The code and pre-trained weights are publicly available on the GitHub repository of Facebook Research. I extract the final layer of the ConvNet before classification (a 4096-dimensional vector) for each meme as a representation of the visual features.

Cluster the memes and label the clusters

After getting representations for each meme, I train the standard K-means algorithm with $K=100$ and Euclidean distance on the 4096-dimensional embedding space. The choice of K is still arbitrary at this stage. The idea is to choose a large enough K and combine similar clusters at the later stage.

After clustering, I see the images within each cluster to label the clusters. Specifically, I sample 5 representative memes (memes that are nearest to the center of the clusters) and 15 random memes within that clusters (to ensure the robustness of the distance measure). See Figure 3.3 for examples from Clusters 5 and 46. Some more examples are in Figure 3.7 in the Appendix.

3.4 Preliminary Findings

Figure 3.4 plots the t-SNE projection of the learned visual embeddings for each meme. Each point represents a meme from IRA or Reddit; each color indicates a clustering result from K-means. Each number indicates the index of the cluster, labeled at the centroid of the cluster. For each cluster, we also calculate the percentage of memes in that cluster (for IRA and Reddit memes separately). Red indicates that the cluster has the highest percentage of memes from IRA; blue indicates that the cluster has the highest percentage of memes from Reddit.

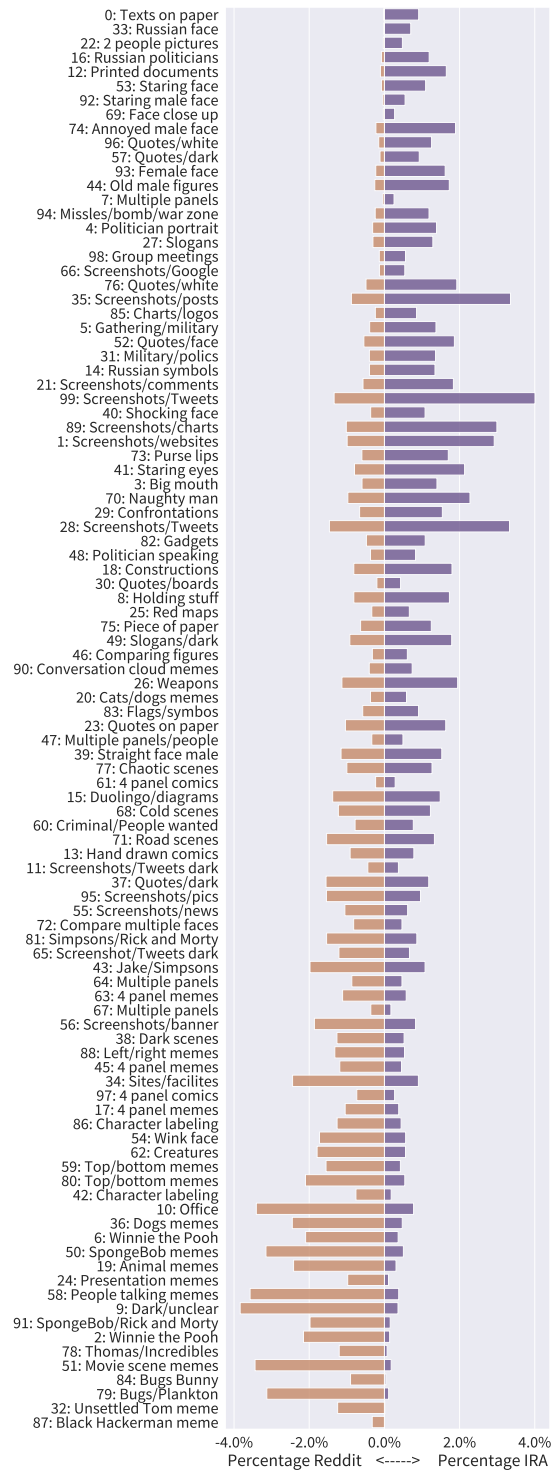


Figure 3.5: Cluster labels and shares of IRA/Reddit memes

Note: Percentages are calculated by the number of memes from IRA (Reddit) in a cluster out of total number of memes from IRA (Reddit), respectively. The clusters are ordered by the percentages.

One can see that clusters located near the top of Figure 3.4 involve mostly pictures of public figures (politicians, celebrities). Clusters located near the bottom involve mostly screenshots (Twitter/Facebook posts, quotes/slogans, news websites/headlines, messages, etc.). Clusters located near the top left consist of the common “memes”: pictures surrounded by text. Clusters located near the bottom left consist of comics, maps, charts, etc. Most clusters are complex mixing of images, text, pictures, screenshots, and comics.

Figure 3.5 presents a complete list of clusters, and labels, along with the percentage count within IRA/Reddit memes. For example, cluster 99 (one of the clusters involving Screenshots/Tweets) accounts for 4% of the IRA memes. The top row indicates that the cluster has the highest relative percentage of IRA memes (around 1% of IRA memes and no Reddit memes); the bottom row indicates that the cluster has the lowest relative percentage of IRA memes (0.5% of Reddit memes and no IRA memes). One can see that, towards the top of the list (more common in IRA memes), there are more themes around pictures of public figures, quotes, slogans, screenshots, and scenes related to military or gender. In comparison, towards the bottom of the list (more common in Reddit memes), there are more comics, cats/dogs, superheroes, and movie scenes. Noticeably, Reddit memes usually evolve around fixated “frames” where free online meme-creating tools can help you create memes with the same frame without editing the whole meme yourself. Although these tools are wildly available in the West, it seems like the IRA accounts are not utilizing these tools to generate memes with popular frames.

Can machine learning discern IRA memes from Reddit memes simply by using the 4096-dimensional visual representations? I train a simple logistic regression using only the visual representations learned by DeepCluster with a 70/30 train/test split. This simple baseline using visual representations alone achieves training accuracy 0.90, AUC 0.91, testing accuracy 0.84, precision 0.84, recall 0.84, F_1 -score 0.84. See Figure 3.6 for the confusion matrix for this logistic regression.

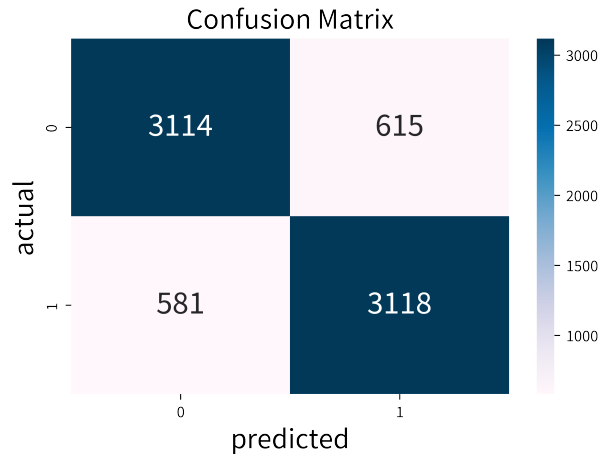


Figure 3.6: Confusion matrix for logistic regression predicting IRA memes using only visual representations

3.5 Discussions and Future Steps

In these preliminary experiments, I find that coordinated IRA memes and authentic Reddit memes share a large set of visual themes but with varying degrees. IRA memes promote more pictures of celebrities, quotes, screenshots, and images related to military and gender. Reddit memes involve more comics and movie characters. I also find that using a simple logistic regression on the learned visual representations can reasonably discern coordinated memes from authentic ones.

The proposed method, based on DeepCluster (Caron et al., 2019), does not rely on labels and can learn broader themes of images. In contrast, BOVW only learns about local visual features within patches, and pHash requires that images be nearly identical. They can be less useful in identifying the visual themes of memes.

I plan to extend this framework to find better representations of visual themes:

- With the successes of multimodal transformer models (such as VisualBERT, ViLBERT, and VL-BERT) in Facebook’s Hateful Memes Challenge, we can extract texts and entity/race tags and learn a more flexible model to get richer embeddings not only based on vision but also interacts with texts and other augmented information.

- It is possible to better preprocess memes to strip off structures that are less related to themes (such as a number of panels within a meme) so that meme structures would not dominate during clustering.
- It is also possible to utilize more flexible clustering models so that each meme does not only belong to one cluster but a distribution of clusters (similar to the Latent Dirichlet Allocation, Blei, Ng and Jordan (2003)) or even to include covariates such as source, time, or other metadata for building clusters (similar to the Structural Topic Model, Roberts et al. (2013)).

3.6 Appendix

- Figure 3.7: examples of representative memes from selected clusters.
- Figure 3.8: examples of IRA images predicted as memes and non-memes.

3.7 Acknowledgments

Chapter 3, in full, is a reprint of the material as it appears in Keng-Chi Chang, “Mapping Visual Themes among Authentic and Coordinated Memes”, *Workshop on Images in Online Political Communication of the 16th International AAAI Conference on Web and Social Media (ICWSM)*. 2022. The dissertation author was the primary investigator and author of this paper.



Figure 3.7: Representative memes from selected clusters.

Note: For each panel, the index of cluster is on the top left. The first row contains the 5 memes nearest to the center of the cluster; the 2-4 rows contain 15 random memes from that cluster. Red border indicates the meme is from IRA; blue border indicates the meme is from Reddit.

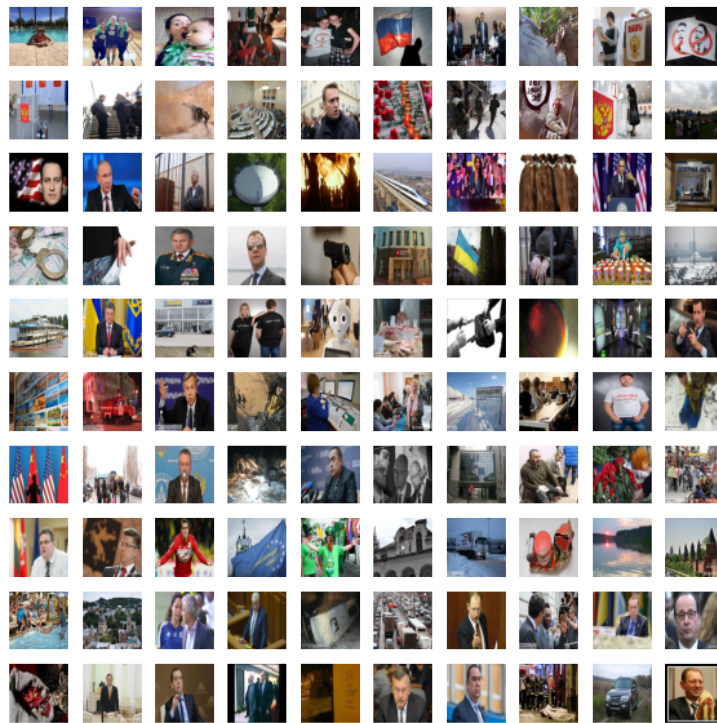
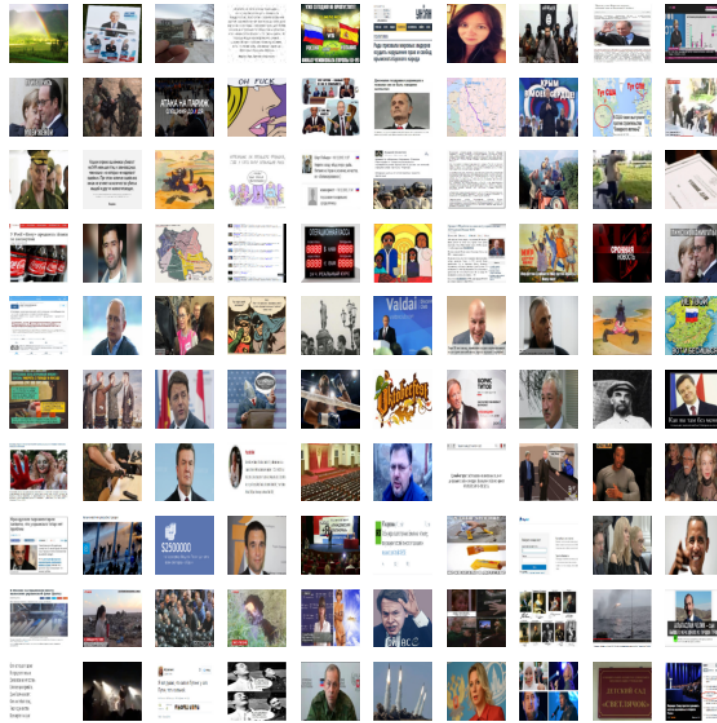


Figure 3.8: Predicted memes (top) and non-memes (bottom) of IRA images.

Note: See Methodology for details.

Chapter 4

Characterizing Image Sharing Behaviors in US Politically Engaged, Random, and Demographic Audience Segments

4.1 Introduction

As visual media—i.e., images and video—become increasingly popular in the online information space, insights into and methods for measuring how such media is used and the audiences most engaged with such media are increasingly important for understanding and improving both online and offline behaviors and information spaces. Many studies support these online/offline implications for media in online spaces, as we have good evidence that including visual media in textual posts increases engagement (Li and Xie, 2020a), mobilizes individuals to protest (Casas and Williams, 2019a), exposes individuals to anti-social QAnon content (Buntain et al., 2022), and often provides a vector for hate speech (Kiela et al., 2020). Recent advances in image generation and large multi-modal models like GPT4,¹ only amplify these needs, as they reduce cost

¹<https://openai.com/research/gpt-4>

and effort necessary to create visual media.

Visual media plays a substantial role in contemporary political discourse, and while individuals on Twitter rarely post political content (Bestvater et al., 2022), posting political imagery has the potential to increase exposure to political content. At the same time, ideologically cross-cutting exposure on social media can drive polarization (Bail et al., 2018), and studies suggest that the political right enjoys additional amplification in online spaces (Huszár et al., 2022), especially with respect to visual content (Munger and Phillips, 2022). Barberá further explores how better capturing audience demographics is needed to improve our understanding of these online dynamics (Barberá, 2016). As a community, we must understand these dynamics thoroughly to better mitigate such inequities and improve information spaces. It is in this context that this paper is situated, where we contribute to this space by characterizing how different types of images are used by politically engaged Twitter audiences versus a general Twitter audience and how different demographic segments engage with images. To this end, we answer two main research questions:

RQ1: Do politically engaged Twitter accounts share different types of imagery on Twitter than the general US Twitter audience?

RQ2: What types of image sharing behavior are predictive of the account’s demographic backgrounds?

To answer these questions, this paper combines two large-scale samples of US Twitter audiences from Alizadeh et al. (Alizadeh et al., 2020) with an automated method for demographic inference from profile pictures, called FairFace (Karkkainen and Joo, 2021). This analysis uses behavior from 10,000 US Twitter accounts, covering more than 66 million tweets, and 10 million images. Uniquely, we apply FairFace to public profile images from these accounts at scale, inferring age, race, and presented gender—a departure from prior work, that has leveraged surveys (Barberá, 2016) or matching Twitter accounts with “voter files” (Hughes et al., 2021; Barberá,

2016). Then applying a clustering scheme to these 10 million images to construct a set of image types, we assess the predictive power of these image types as they relate to demographics and political engagement.

Results show that demographics exhibit little variation between politically engaged and randomly selected US Twitter audiences. For image types, we likewise see several types of images appear common across ages, gender presentations, and political engagement. That said, using logistic regression models to predict demographics, we see about half of the image types (i.e., around ten of the image clusters) correlate significantly with race, gender, age, and political interest; which clusters correlate with these attributes vary across the attributes, however. These logistic regressions capture only a limited amount of variation in these attributes though, ranging from McFadden pseudo $R^2 = 0.04$ for race and $R^2 \in [0.10, 0.12]$ for gender, age, and political engagement.

4.2 Data and Methodology

Our sample consists of two sets of Twitter users, gathered from Alizadeh et al. (Alizadeh et al., 2020): one based on timelines from a random set of 5,000 accounts geolocated to the United States (“random”), and another based on timelines from a collection of 5,000 US accounts who are politically engaged (“political”). Politically engaged accounts are defined as those accounts following at least 5 political Twitter accounts—that is, they follow Twitter accounts of US congresspeople in the Senate, House, Governors, and in the executive branch. These samples are also restricted to users who posted at least 100 times in 2015–2017. Table 4.1 summarizes statistics for these two audience samples.

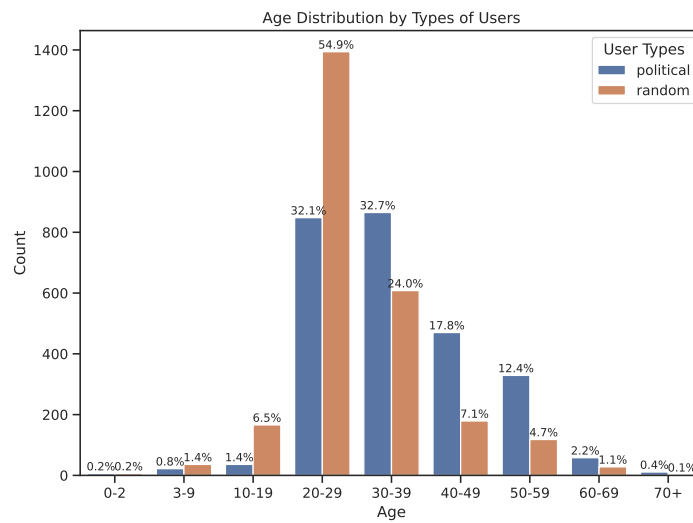
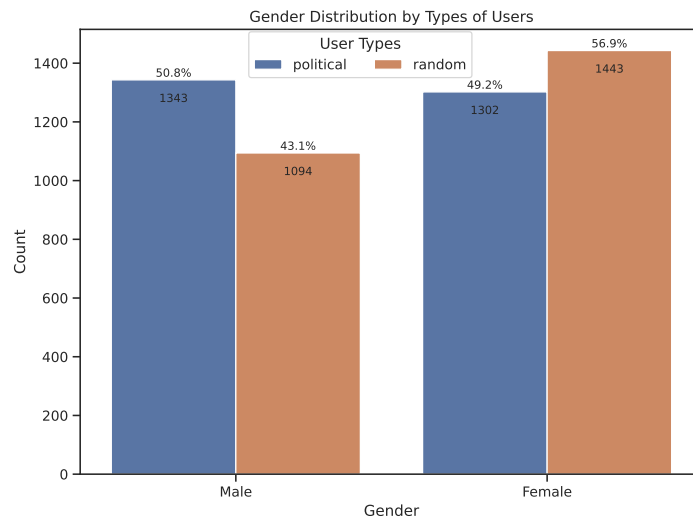
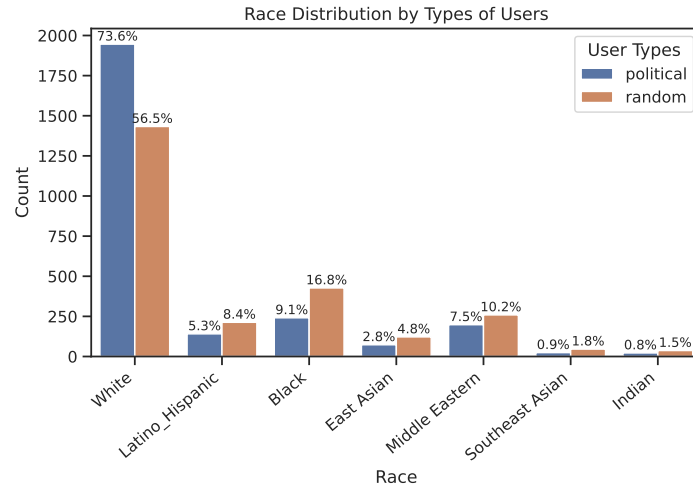


Figure 4.1: Distributions of Predicted Account-Level Demographics—Race, Gender, and Age—Across Random and Politically Engaged Audiences.

Notes: Demographic attributes predicted using FairFace (Karkkainen and Joo, 2021).

Table 4.1: Summary statistics of Twitter sample

Sample	Random	Political
# users	5,000	5,000
# tweets & retweets	31,038,705	35,932,231
% tweets	40%	37%
% retweets	60%	63%
% has image	18%	14%

4.2.1 Demographic Inference from Profile Pictures

From these audience samples, we assess differences in how accounts of various demographic attributes present themselves via the images they share. To that end, beyond dividing between politically engaged versus random accounts, we also investigate account-level demographics using FairFace (Karkkainen and Joo, 2021) applied to an account’s profile picture. FairFace uses a face attribute model that is balanced on race, gender, and age to mitigate the potential bias toward Caucasian faces. Figure 4.1 shows distributions of the predicted demographics. Compared with the Twitter samples linked to voter files (Hughes et al., 2021), the marginal distributions of race and gender are similar (e.g., 70% White, 11% Black; 52% Female), while predicted age from profile pictures appears biased toward younger age.

4.2.2 Collecting and Characterizing Images in Tweets

After assessing demographics, we then collect images in the timeline of these account samples, so that we can assess differences in visual presentation. We focus on the original tweets—i.e., we exclude retweets—on the user timelines containing images. To featurize images, we use a pre-trained ResNet50 deep learning model (He et al., 2016) to generate 2048-dimensional embedding for each image.

To characterize disparate types of visuals, as well as the general content of these image types, we apply k-means clustering on the image embeddings to group these images into clusters.

Using cluster-quality metrics of within-cluster sum of squared distances and silhouette scores to determine the number of clusters (see Figure 4.2), we set $k = 20$. See Figure 4.5 for random images for some clusters.

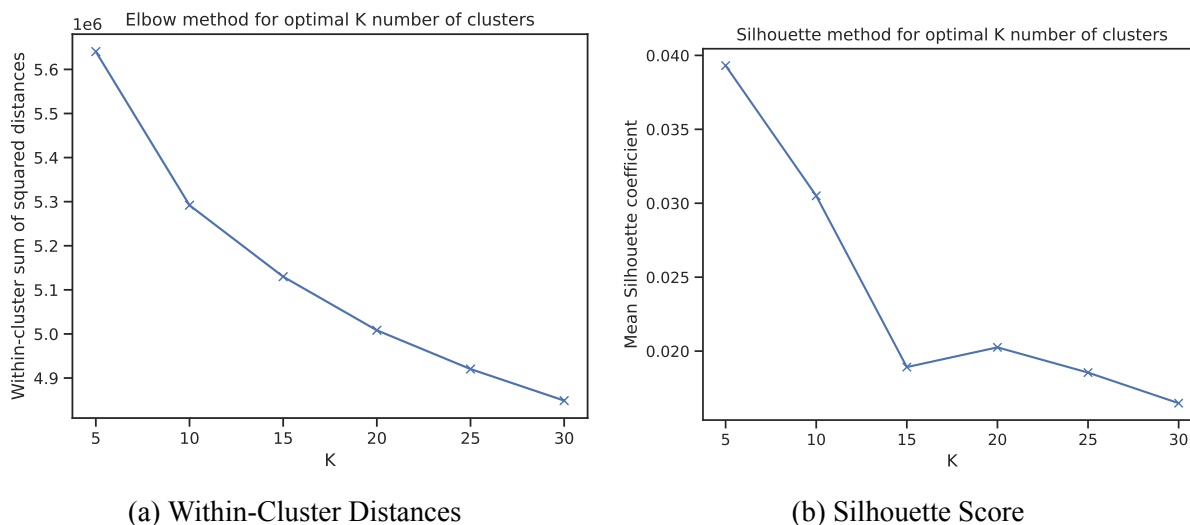


Figure 4.2: Cluster Quality Metrics.

Note: Cluster counts between 15-20 seem reasonable based on elbows in these curves.

4.3 Results

To address **RQ1**, on whether politically engaged accounts share different types of images than general US Twitter audiences, we first present the distribution of clusters by different types of audiences—politically engaged versus random accounts—as shown in Figure 4.3. Results show overlap between image-types shared by political and random users: Most clusters are not discriminant in separating political and random accounts. That said, a few clusters appear over-represented among politically engaged accounts, specifically clusters 3, 4, and 7, whereas clusters 0 and 16 are over-represented among general audiences.

To understand predictive power of image clusters for demographics (**RQ2**), we run an account-level logistic regression of demographic variables (race, gender, age, political engage-

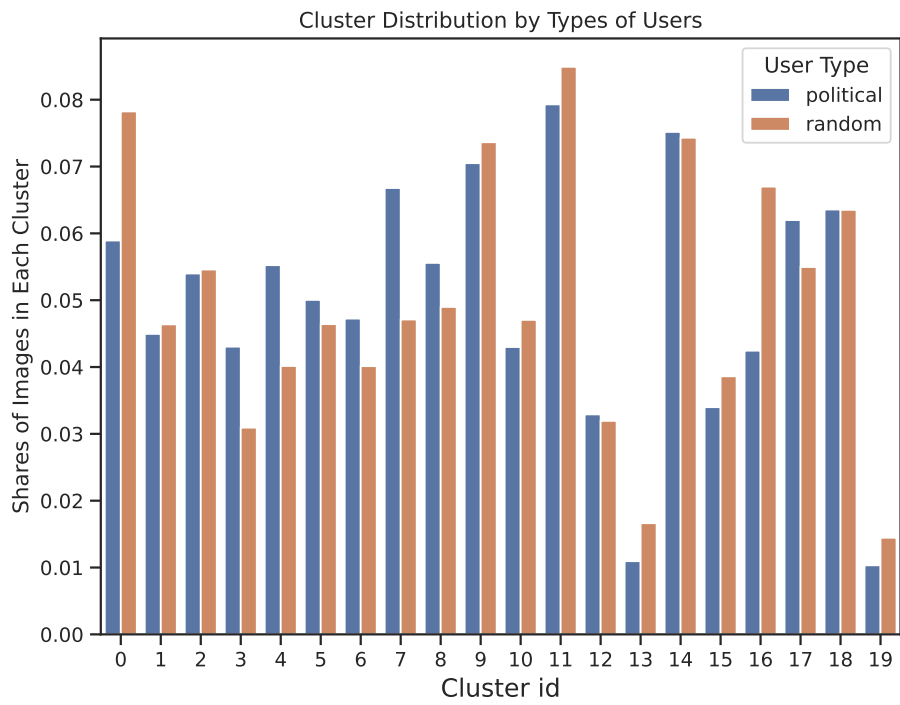


Figure 4.3: Cluster Distribution by Audience.

Note: Most clusters have similar distributions across politically engaged and random accounts, but a few clusters appear over-represented in one of the other sets (e.g., clusters 0, 3, 4, 7, and 16).

ment) using an account’s percent of images in each cluster; that is, for user i with demographics y_i , we estimate:

$$\sum_{k=1}^{20} \beta_k \frac{\# \text{ images in cluster } k \text{ shared by user } i}{\# \text{ images shared by user } i} + \varepsilon_i$$

In this equation, β_k show the correlation between sharing cluster k and demographic variables. To binarize demographic attributes returned by FairFace, we collapse race to white or non-white, gender to female or non-female, age to less than 30 years old or older, and politically engaged or not. Figure 4.4 shows the results. Our findings are in general interpretable, providing insight into how increases in sharing a particular type of image change the probability .

Each logistic regression model has an associated McFadden Pseudo- R^2 , in the range [0.04 – 0.12], with race appearing the most difficult to predict given an account’s distribution of images over image clusters. In contrast, gender appears to perform best with a pseudo- $R^2 = 0.12$, while identifying age and political engagement have the same $R^2 = 0.1$. Though these values are low, guidance on interpreting McFadden’s Pseudo- R^2 state it is never 1 (Hu, Shao and Palta, 2006) and suggest values in the range [0.2 – 0.4] “represent an excellent fit” (McFadden, 2021). As such, we interpret at least the gender, age, and politically engaged models to have some predictive power.

4.4 Discussion and Conclusions

Our finding that the distributions of clusters over politically engaged versus general audiences is relatively stable, as shown in Figure 4.3, is interesting in that it suggests, in general, information sharing behaviors on Twitter are not massively driven by political interest. This result may be a reflection of our choice to exclude retweeted images, as retweets are a significant indicator of political affiliation (Conover et al., 2012), coupled with the rarity of posting political

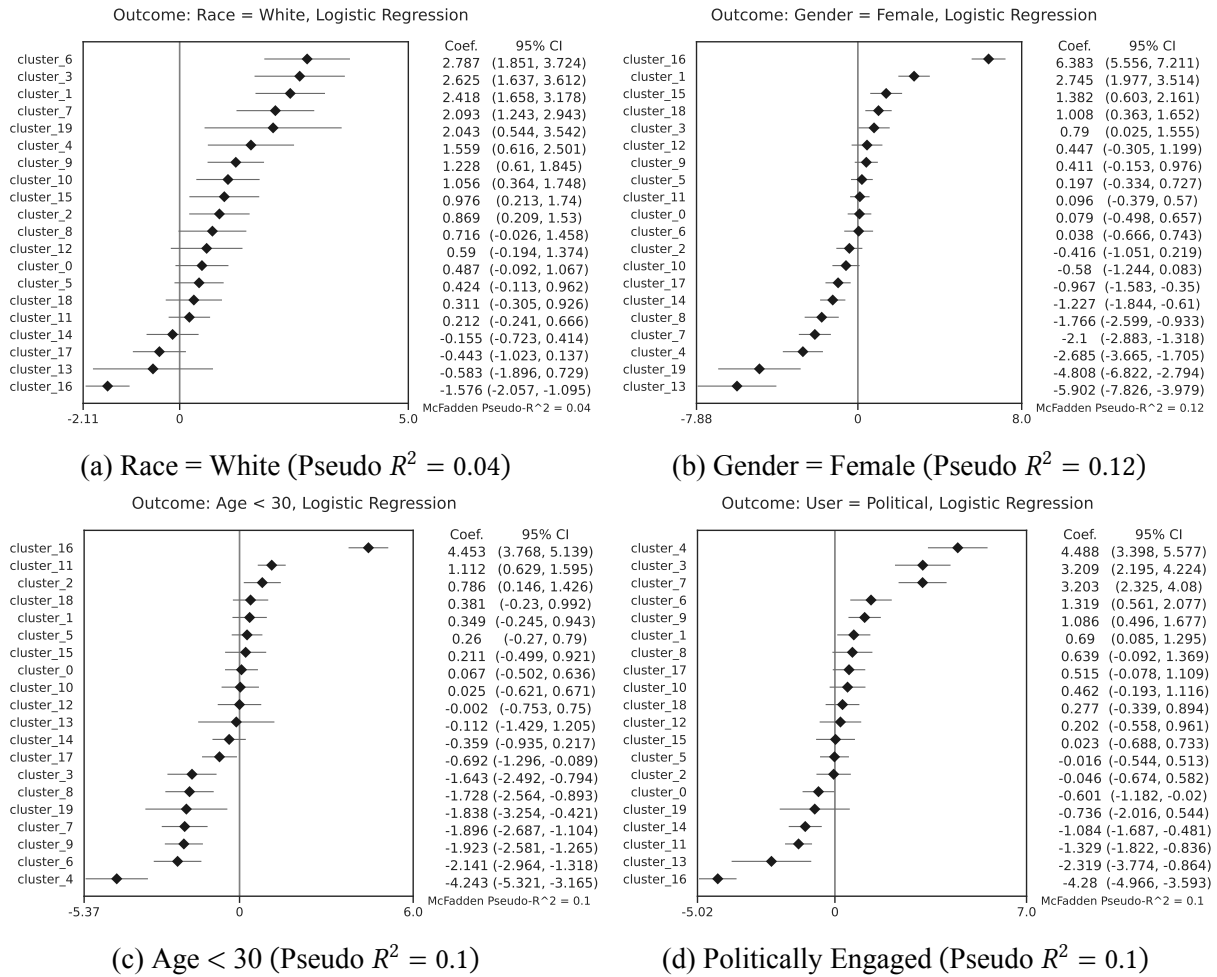


Figure 4.4: Regression Coefficients of Account-Level Cluster Distributions on Demographics.

Note: Cluster distributions appear to have some predictive power for gender, age, and political engagement, whereas white-versus-non-white appears difficult to capture with this data.

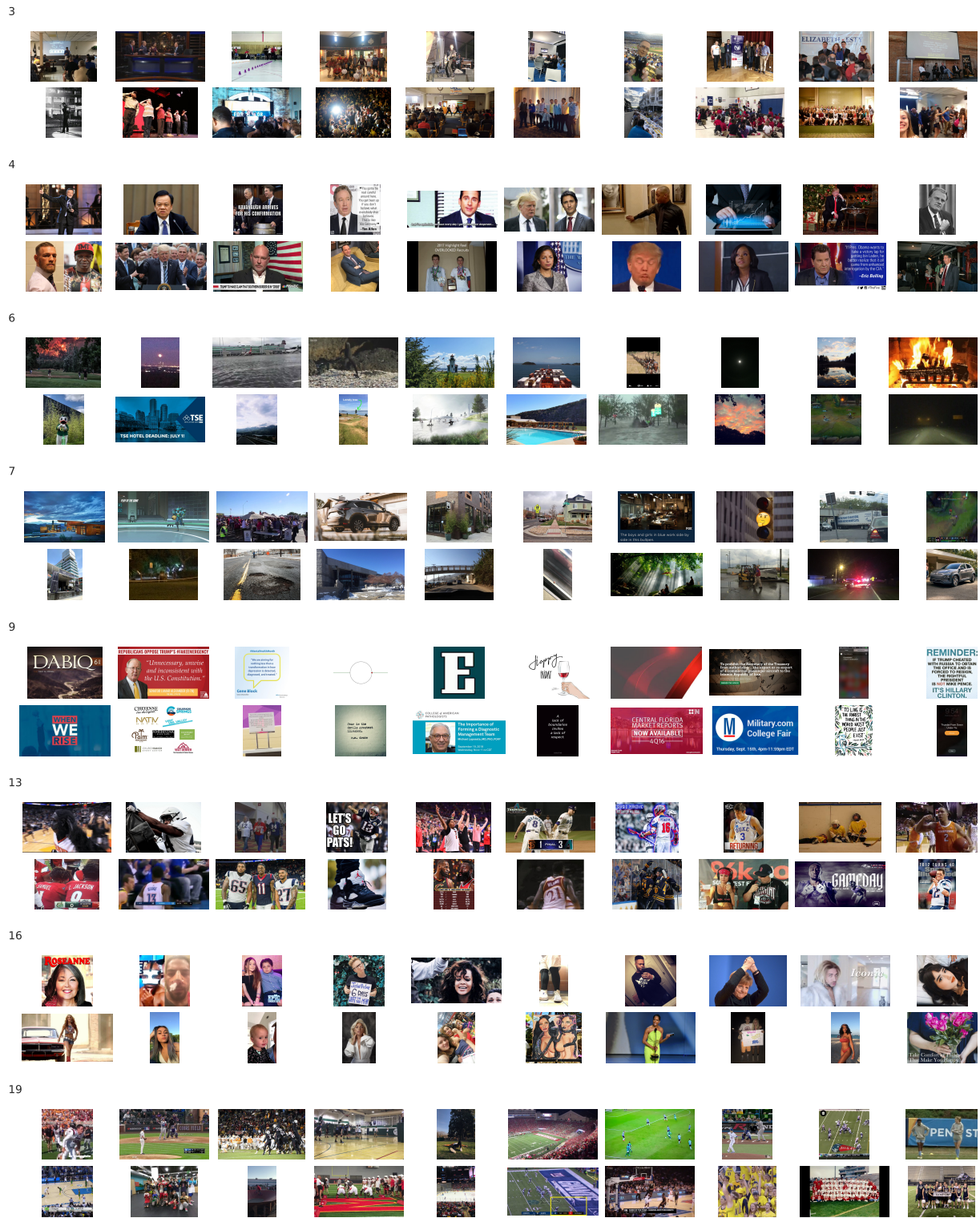


Figure 4.5: Random images from predictive clusters (cluster ids on top left of each panel)

content organically (Bestvater et al., 2022). Additional research is needed to assess whether these results on image-sharing behaviors are consistent at the account level—that is, whether politically engaged and general audiences post similar distributions of political imagery as they do political text.

If we inspect the clusters in Figure 4.5 that are over-represented in politically engaged audiences, these clusters of images do appear particularly politically relevant. For example, in cluster 3, we see many images of groups of people; this finding is consistent with Joshi and Buntain (Joshi and Buntain, 2022), where politicians share images of constituents. Likewise, cluster 4 appears comprised of images containing one or two faces, mainly politicians (e.g., Donald Trump, Justin Trudeau, etc.), with clear political relevance. Cluster 7 is less clearly political relevant though, suggesting more research is needed to assess how these images are being used in a potentially political context.

Regarding other demographic attributes, sharing infographics (cluster 9), faces of politicians (cluster 4), natural scenery (cluster 6), or street views (cluster 7) is predictive of users being older than 30; sharing images of natural scenery (cluster 6) or social gatherings of White people (cluster 3) is highly predictive of users being White; sharing sports (clusters 13, 19) is highly predictive of users being male.

Decades of social science research suggest sociodemographic traits are major drivers for behaviors online and offline. Our study illustrates a way to proxy such information from profile pictures and infers how image-sharing behavior varies with demographic segments. Using visual features extracted from deep learning, our initial finding suggests that around half of the image clusters contain predictive information about the account’s race, gender, age, and political engagement. The proposed method is interpretable and scalable, allowing for more images, more fine-tuned feature extractors, and more fine-grained demographic variables.

Our study has clear implications for studies of digital literacy and misinformation. If certain users tend to share certain types of images, information actors can utilize this information

to design visuals if they would like to target some particular populations. It is also possible that this “content-based targeting” is harder to achieve in text than images.

Limitations include the fact that profile pictures are not always representing the users themselves. However, manual inspections and the customs on Twitter as a platform (e.g., unlike Reddit, where most users do not use profile pictures) convinced us that it is a reasonable measure. Second, we have only studied a narrow aspect of online image-sharing behavior—tweeting. Retweeting or reacting to visual-based content is also of interest to the study of information space.

4.5 Acknowledgments

Chapter 4, in full, is a reprint of the material as it appears in Keng-Chi Chang and Cody Buntain, “Characterizing Image Sharing Behaviors in US Politically Engaged, Random, and Demographic Audience Segments”, *Workshop on Images in Online Political Communication of the 17th International AAAI Conference on Web and Social Media (ICWSM)*. 2023. The dissertation author was the primary investigator and author of this paper.

Chapter 5

Do Images Lend Credibility to News Articles?

5.1 Introduction

In the past decade, the general public concerns over the spread of misinformation or partisan news has elevated, due to considerations of their potentials in shaping beliefs and perceptions. Scholars have responded to these concerns by estimating the prevalence of misinformation (Allcott and Gentzkow, 2017; Berinsky, 2017; Flynn, Nyhan and Reifler, 2017; Guess, Nyhan and Reifler, 2020), finding automated or crowdsourced detection algorithms (Kumar, West and Leskovec, 2016; Wu et al., 2019), and inventing corrective interventions (Pennycook et al., 2021; Badrinathan, 2021). However, most scholarly research have focused on textual misinformation. There are substantially less attention paid to visual misinformation, especially what determines the perceived credibility of visual contents.

More recent studies have started to document some facts about the prevalence of visual misinformation (Garimella and Eckles, 2020; Yang, Davis and Hindman, 2023) and provide **observational** evidence that images might have some advantages over text in driving engagements

(Li and Xie, 2020b) or mobilizing protests (Casas and Williams, 2019b). Despite some **experimental** studies in testing credibility perception of visuals or videos (Wittenberg et al., 2021; Hameleers et al., 2020; Barari, Lucas and Munger, 2021), their experimental design only treat image as a black box—it remains unclear what it is about the images that matters for the credibility of information.

In this paper, we address the question of when and why images impact the credibility of information. Does including an image with a statement lend to its credibility? What is it about an image that makes a fact more credible, and for whom?

To address this problem, we design an experiment that randomizes **different versions** of real-world news images **within** each experimental news stories. By leveraging recent advances in open source pretrained Large Image Models and new frameworks in causal inference (Egami et al., 2022; Fong and Grimmer, 2023; Pugh and Torres, 2023), we are able to discover and estimate aspects of images that changes credibility perception.

Contrary to findings in earlier experiments where there is no variation in visual treatments (although in other contexts such as social media posts (Hameleers et al., 2020) or videos (Wittenberg et al., 2021)), we do not find that the presence of visuals can always increase credibility perception of news. The **type** of image used as treatment matters—there are types of images that can increase perception and others can decrease perception. We account for these **latent treatments**—aspects of image treatments that are indirectly manipulated by the researcher—in a principled manner.

5.2 Research Questions

Research on news credibility seeks to understand which characteristics of news articles make them more or less credible to individuals and how such perceived credibility affects communication outcomes. One understudied aspect of news articles are the images (**RQ1**).

RQ1: Do images change perceived credibility of news articles?

To our knowledge, there was no direct experimental evidence for this, but there are some suggestive evidences from related studies. Wittenberg et al. (2021) tested the effectiveness of political campaigns in videos vs. texts, showing respondents believe the event more if shown the video version. Hameleers et al. (2020) tested Tweets with and without images, showing images increased perceived credibility.

Even if we can answer **RQ1**, what exactly is about images that make them perceived more or less credible? We vary characteristics of visual images in treatments to answer this question (**RQ2**).

RQ2: What characteristics of images change perceived credibility of news articles?

This question (**RQ2**) cannot be answered with existing experimental designs (Wittenberg et al., 2021; Hameleers et al., 2020; Barari, Lucas and Munger, 2021) where researchers assign one fixed version for each modality (image/text/video) in treatments and compare directly across modalities, leaving it unclear what it is about images that matter.

We have good evidence that predictors of misinformation dissemination include age and partisanship (Guess, Nagler and Tucker, 2019). Other research also finds that the types of image individuals shared is correlated with demographic attributes (Chang and Buntain, 2023). Since there are far less unified education for reading images, we conjecture (**RQ2**) there would be divergent responses to image treatments across known societal groups.

RQ3: Is there heterogeneity across gender, race, and age in the effects of image on perceived credibility?

Image treatments are high-dimensional, and simply randomize the images are not sufficient for valid causal inference (Fong and Grimmer, 2023) since there will be confounding by unmeasured treatments. We use general-purpose vision embeddings and adapt method address-

ing this issue from Natural Language Processing (Fong and Grimmer, 2016) to analyze image treatments (**RQ3**), without relying on human labeling or pre-defined categories.

RQ4: How do we discover and estimate the various latent aspects of visual treatments?

5.3 Experimental Design

5.3.1 Sample

Our sample consists of U.S. adults recruited on Cloud Research Prime Panels and Connect in June 2023. To identify heterogeneity in demographics, quotas were set for race and region. Table 5.1 presents the summary statistics and balance tests between treatment and control groups. The demographic characteristics are balanced between treatment conditions. Samples that did not pass the attention checks (e.g. “Please select ‘Disagree’ from the options below.”) are dropped from the analysis. Given the design, the attrition rate was neither asymmetric between treatments nor attributable to observed sociodemographic characteristics.

5.3.2 Experimental Procedure

After asking participants’ informed consent and measuring some pretreatment variables such as party affiliation and social media consumption, the respondents were asked to read four news stories. Two of the stories are false news (fact-checked by PolitiFact, a nonpartisan fact-checking website), and two of the stories are real news. The respondents were not provided with the veracity or the source of the news articles, nor were they primed to think about the accuracy of the stories.

We randomly assign respondents into treatment and control groups. Control group will read the story with the title and a two-paragraph excerpt of the story, while treatment group will read the story with the title, a randomly assigned image that fits the story, and the same two-



Figure 5.1: Screenshot of Example Treatment Stimuli

paragraph excerpt of the story. In other words, the only difference between the treatment and control condition is the presence of image while reading news articles. Figure 5.1 presents an example screenshots for the treatment stimuli.

Figure 5.2 plots the flow of the survey experiment. We use block randomization so that each respondents are assigned to two treatment conditions (reading two stories with image) and two control conditions (reading two stories without image). After reading each article, the respondents were then asked to rate the credibility of the story on a 0-100 scale (“On a 0-100 scale, how likely this news is true?”). The stories and treatment conditions were also presented in randomized order so there will be no learning effect when averaging across respondents.

5.3.3 Story and Image Selection

For this first part of a series of studies, we pick news related to our substantive interest in US public opinion towards China from PolitiFact, the Google Fact Check Tools API, and Polygraph.info (a fact-checking website produced by Voice of America) with the following criteria: (1) Relevant to the general US population (not just stories about things happened in China); (2) Reported recently (up to one year). The frequency of fact-checked stories about China is quite low. Stories that satisfy both criteria is only a handful while designing the survey.

After selecting the stories, we feed the title of the stories into Google and Bing Image

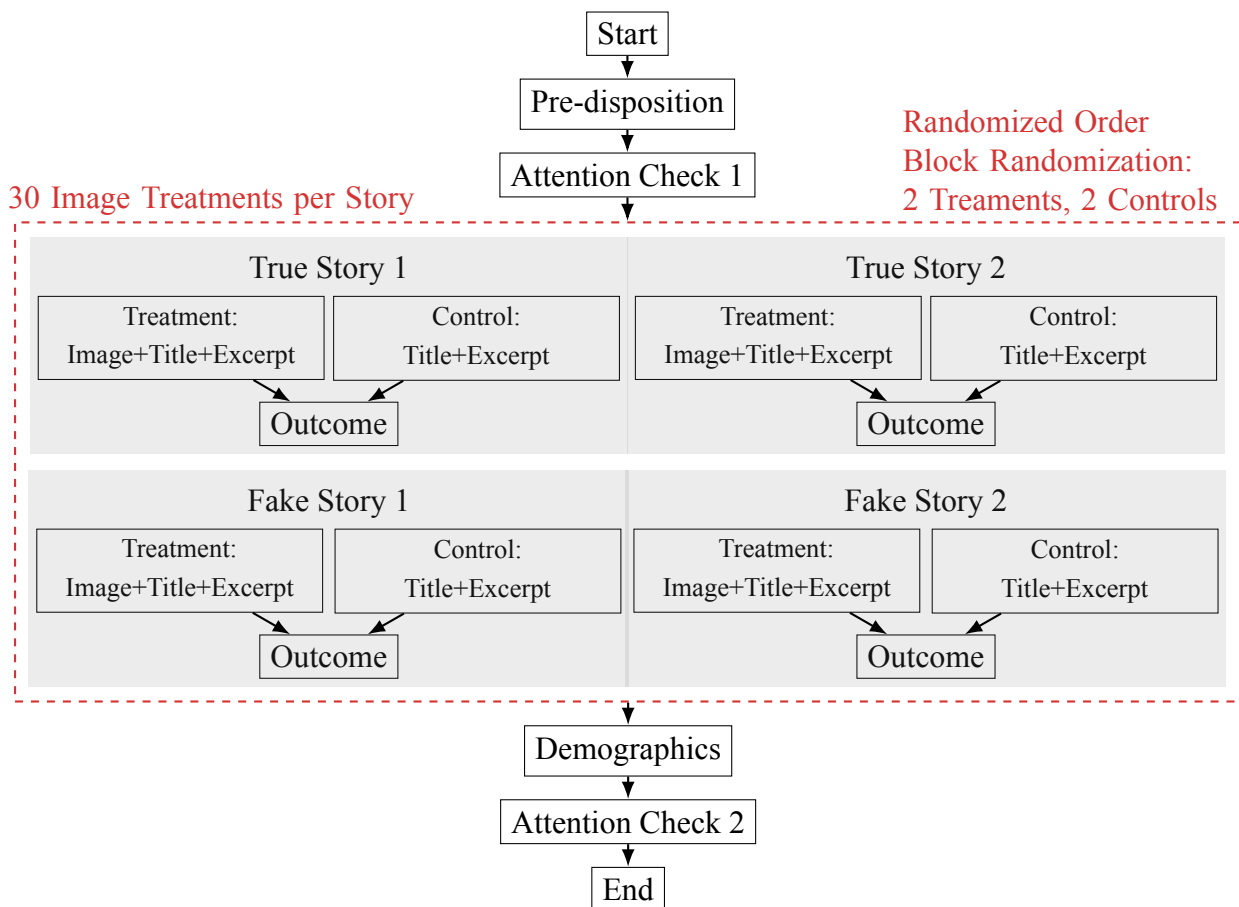


Figure 5.2: Flow of Survey Experiment

Search. We then pick 30 unique images for each stories that satisfy the following criteria as image treatments: (1) Images that are consistent with the textual content of news articles; (2) Covers distinct types of imagery among the universe of images that we sampled from.

5.4 Inferring Latent Treatments from Treatment Images

5.4.1 Break the Dependency between Discovery and Estimation

Social scientists are generally interested in effects from low-dimensional interventions, while the interventions of text—as well as images—are almost always high-dimensional (Fong and Grimmer, 2023). Additional care has to be taken for valid causal inference, since SUTVA

Table 5.1: Summary Statistics and Balance Table

Variable		Total	Control	Treatment	χ^2 p-val
Party Id	Democrat	523 (59.7%)	263 (50.3%)	260 (49.7%)	0.4906
	Independent	96 (11.0%)	50 (52.1%)	46 (47.9%)	
	Republican	257 (29.3%)	119 (46.3%)	138 (53.7%)	
Gender	Female	459 (52.4%)	217 (47.3%)	242 (52.7%)	0.2055
	Male	417 (47.6%)	215 (51.6%)	202 (48.4%)	
Race	Black	137 (15.6%)	72 (52.6%)	65 (47.4%)	0.8390
	Latino	100 (11.4%)	49 (49.0%)	51 (51.0%)	
	Others	99 (11.3%)	50 (50.5%)	49 (49.5%)	
	White	540 (61.6%)	261 (48.3%)	279 (51.7%)	
Age	Older than 40	402 (45.9%)	193 (48.0%)	209 (52.0%)	0.4768
	Younger than 40	474 (54.1%)	239 (50.4%)	235 (49.6%)	
Education	College Graduate	536 (61.2%)	256 (47.8%)	280 (52.2%)	0.2481
	Under College	340 (38.8%)	176 (51.8%)	164 (48.2%)	

violation is likely if the researcher does not break the dependency between discovery and estimation (Egami et al., 2022). To solve this problem, we follow the framework of Egami et al. (2022) to split the sample and use the training set (50% of the data) for treatment discovery and test set (50% of the data) for treatment effect estimation.

To be more specific, in the treatment discovery phase, we combine a pre-trained image model (Li et al., 2023) and the Supervised Indian Buffet Process (Fong and Grimmer, 2016) to learn a model that maps high-dimensional visual features X_i into a low-dimensional binary representation of latent treatment Z_i , indicating whether an image $i \in \mathcal{T}$ conveys such a treatment, using training data \mathcal{T} .

In the treatment effect estimation phase, we use the learned model to predict latent treatment \hat{Z}_i on the test set, $i \notin \mathcal{T}$. We then use the predicted latent treatment \hat{Z}_i to estimate the treatment effects of the images. Fong and Grimmer (2023) shows that the ATE is identified using standard regression adjustments under suitable (while not generally testable) assumptions. This procedure is also consistent with the emerging design-based framework of using annotations from Large Language Models for valid downstream causal inference (Egami et al., 2023).

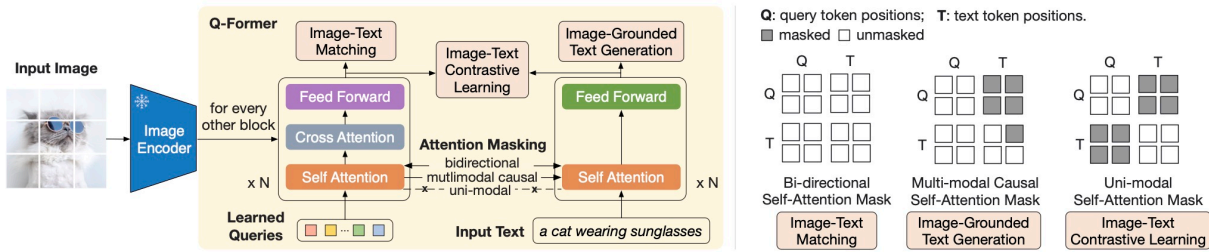


Figure 5.3: BLIP-2 Model Architecture (from Li et al. (2023)).

5.4.2 Overall Workflow in Analyzing Image Treatments

Analyzing visual images is a growing field in computational social science (Torres and Cantú, 2021; Williams, Casas and Wilkerson, 2020; Joo and Steinert-Threlkeld, 2018). However, since there is no clear pre-defined categories, we take an unsupervised approach in the early steps and rely on large-scale pretrained image models. This overall workflow is similar to Pugh and Torres (2023) except that (1) Pugh and Torres (2023) uses a fixed 32 pixel by 32 pixel size blocks to split the images, and (2) Pugh and Torres (2023) uses a ResNet50 model pretrained on ImageNet to extract embeddings. I also derived an additional method in Section 5.7 to learn latent treatments without relying on clustering.

Specifically, I take the following several steps:

1. Split each image into 1×1 , 2×2 , and 3×3 blocks.
2. Map each block into a pre-trained multimodal deep learning model to get the embeddings of each block. Here, we use the new state-of-the-art multimodal model BLIP-2 (Li et al., 2023) developed by Salesforce Research. There are at least two novelties about BLIP-2 that make it more suitable for social science tasks than traditional CNN models such as ResNet-50:
 - Unlike ResNet-50, BLIP-2 is not pretrained just for object detection (classifying cats vs. dogs), but for **large-scale image captions** such as those on Flickr.

- Unlike ResNet-50, BLIP-2 is not trained for classification loss, but for **contrastive loss** where predefined categories (e.g. cats and dogs) are not needed.
3. After getting the embeddings for each image blocks, we cluster (K-Means) on the embeddings to generate $K = 100$ image-block clusters. Figure B.5 plots the evaluation metrics for the K-Means, and Figure B.6 shows the empirical distribution of clusters. The clusters would assign visually similar blocks of images into the same cluster. Figure 5.4 plots the UMAP projection of the image embeddings of the image-blocks, where the colors and numbers indicates K-Means results. Similar methods have been used in other papers (Zhang and Peng, 2022) and our earlier works (Chang, 2022; Chang and Buntain, 2023).
 4. Build an image-by-cluster matrix where each row represents one image while each column represents one visual cluster and the elements indicates the presence of each cluster in each image:

$$[\mathbf{X}]_{i,j} = \mathbb{1}(\text{image } i \text{ has blocks from cluster } j).$$

5. Divide experimental data into training and test sets.
6. In the training set, learn SIBP model (Fong and Grimmer, 2016) to identify latent treatments that maps image-cluster matrix to latent treatments: $g : \mathbf{X}_i \mapsto \mathbf{Z}_i, i \in \mathcal{T}$.
7. In the test set, infer latent treatments using learned SIBP model: $\hat{\mathbf{Z}}_i = \hat{g}(\mathbf{X}_i), i \notin \mathcal{T}$.
8. Estimate treatment effects using inferred latent treatments in the test set.

Figure 5.5 reports the discovered 10 latent treatments Z_1 to Z_{10} (in the rows) by image clusters (in the columns, 4 image blocks per cluster). The ids of the clusters are on the top left of each cluster. First, we can see that clusters group visually similar image blocks together. Second, the discovered latent treatments tend to be not so visually similar but still semantically related and interpretable. For example, Z_2 is a latent treatment about white male wearing suits; Z_6 is

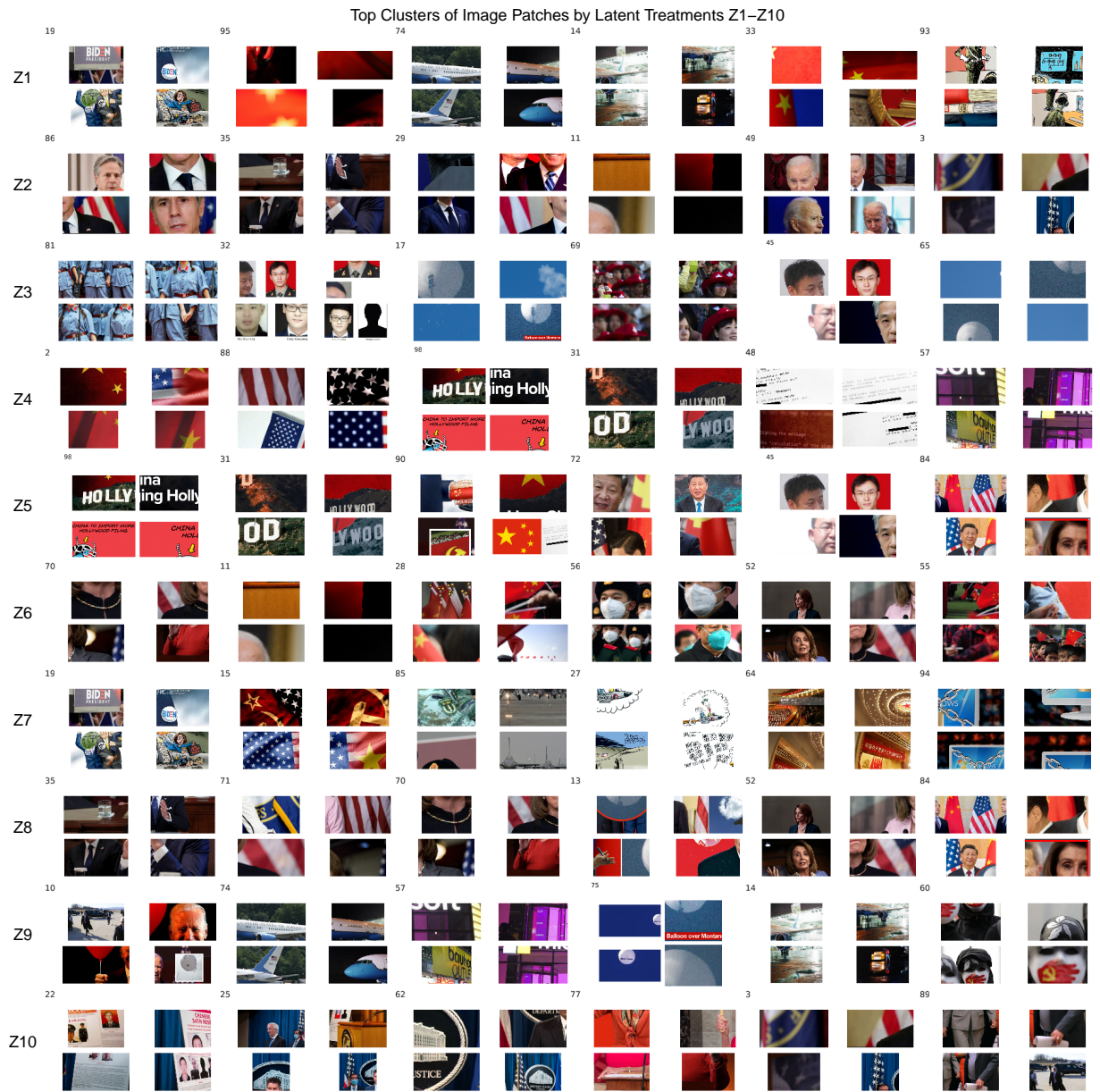


Figure 5.5: Top Clusters of Image Blocks by Latent Treatments Z_1 – Z_{10} Inferred by SIBP.

just about the relationship between texts. Second, unlike LDA or STM, but consistent with the design-based causal inference paradigm, its topic assignment Z_i is **binary**. Both LDA and STM assumes that the document topics Z_i lie on a multinomial simplex. In other words, with SIBP there is no inherent tradeoffs or interference between topic assignments—assigning document to topic A does not necessarily reduce the probability of assigning document to topic B . In this way, we are able to define potential outcomes $Y_i(Z_i = 1)$ and $Y_i(Z_i = 0)$.

5.5 Results

5.5.1 Average Treatment Effect

Since the treatment conditions are randomly assigned, we can estimate the ATE by simply regressing the outcome (perception of news being true) on treatment assignment (has image):

$$\text{NewsPerception} = \alpha + \beta \times \text{HasImage} + \epsilon \quad (5.1)$$

where β is the ATE. To quantify the treatment effect of each latent treatment, we can further interact the treatment assignment (has image) with indicators of latent treatment (Z_1 to Z_{10}):

$$\text{NewsPerception} = \alpha + \sum_{j=1}^{10} \beta_j \times \text{HasImage} \times \text{LatentTreatment } Z_j + \epsilon \quad (5.2)$$

where β_1 to β_{10} is the ATE by latent treatments Z_1 to Z_{10} .

Figure 5.6 reports the overall ATE and ATE by latent treatments. We find that, overall, reading news with images does not causally increase the truth perception of the news article. However, the **type of image treatment** in the news article actually matters.

When treated with images of white male wearing suits (Z_2) and images of photos from press conferences (Z_{10}), perceptions that news being true increases, relative to the text-only con-

Perception of News being True (0–100 Scale)

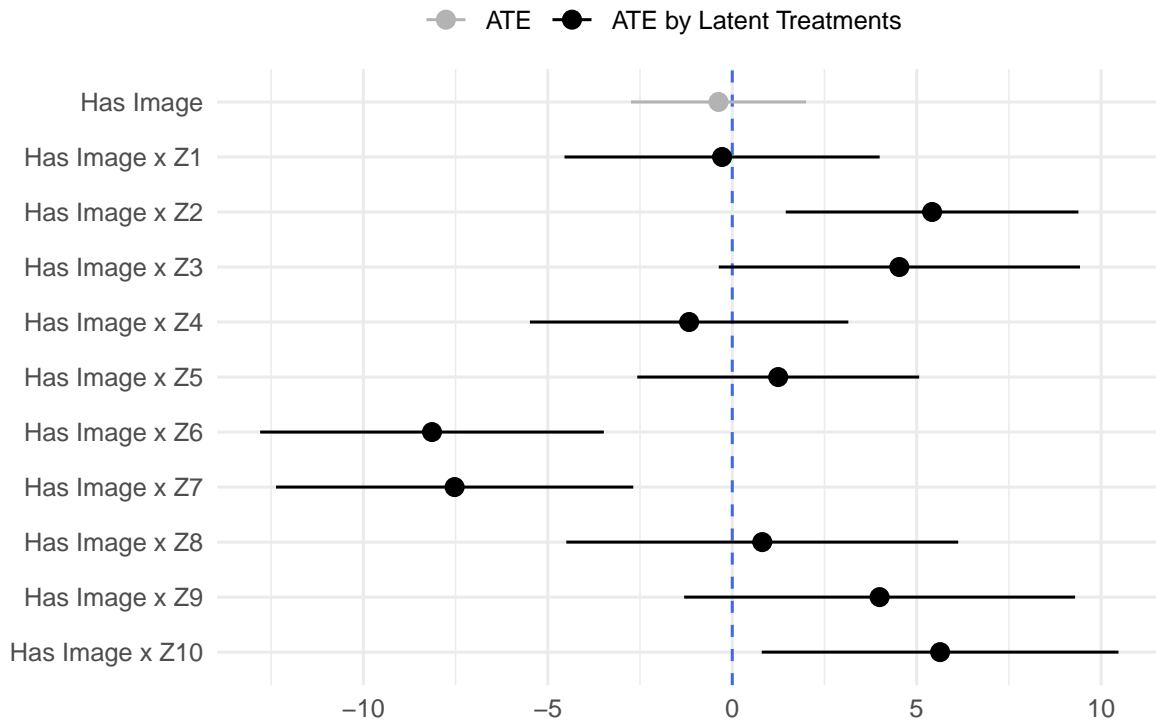


Figure 5.6: Average Treatment Effects of Latent Treatments.

trol group. On the other hand, when treated with images containing female clothes or masks (Z_6) and images with comics or symbols (Z_7), perceptions that news being true decreases.

In other words, the null effect of the overall ATE can be viewed as a weighted average of ATEs by latent treatments 1–10. The table version of Figure 5.6 is reported in Figure B.1. We also included group-level covariates such as fixed effects of partisanship, gender, race, age, education, and region. The effect size does not change much while it costs statistical power.

Figure 5.7 further plots the raw outcome by treatment condition for each story. Red point indicates treated units; blue point indicates control units. Dark red bars indicate mean outcome by treatment image; dark blue bars indicate mean outcome in text-only control groups. One can see that the outcomes are distributed quite smoothly across treatment conditions. The observed ATE difference across latent treatment is not driven by a single image or a few outliers.

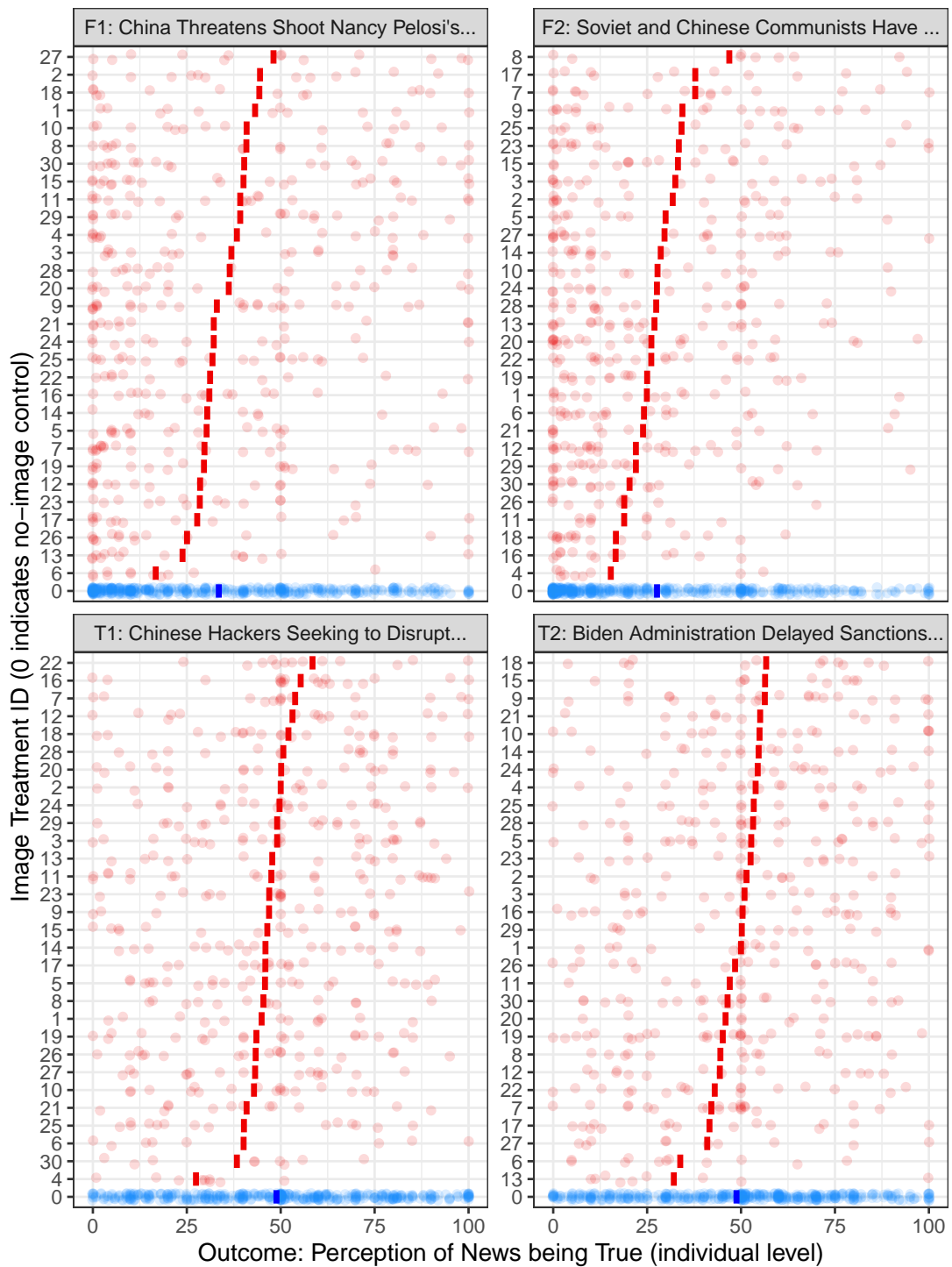


Figure 5.7: Individual-level and Mean Outcome by Treatment Image.

Note: Each panel presents the raw outcome data by story. Dark red bars indicate mean outcome by treatment image. Dark blue bars indicate mean outcome in control groups. No single image or small set of images are driving the observed difference in estimated ATE by latent treatment.

5.5.2 Heterogeneous Treatment Effect

Whether and how do the effect of latent treatments vary across demographic groups? To understand this, we estimate the Conditional Average Treatment Effect (CATE) of the latent treatments Z_1 to Z_{10} across demographic groups g :

$$\mathbb{E} [Y_i(Z_j = 1) - Y_i(Z_j = 0) \mid G_i = g], \quad \text{for } j = 1, \dots, 10.$$

The groups include party identification, gender, race, and age.

Figure 5.8 plots the regression estimated CATEs. Overall, we don't find substantial difference in group-level responses to image treatments as a whole on perception of news credibility. However, we do find differences in response to latent treatments across demographic groups. In the CATE by gender, we find that male respondents do view latent treatment with images containing female clothes or masks (Z_6) as less credible as that of female respondents. In the CATE by race, we find that white and latino respondents view latent image treatment with white male wearing suits (Z_2) as more credible than that of black respondents. In the CATE by age, we also find that respondents older than 40 do view latent treatment with image containing comics or symbols (Z_7) as less credible than that of younger respondents.

5.6 Evaluate Gains from Treatment Targeting

With the presence of some treatment effect heterogeneity, it is natural to ask whether there are benefits in targeting respondents with different treatments by known group-level variables such as demographics. As an example, using data from randomized field experiments, Imai and Strauss (2011) evaluated the benefits from targeting respondents in sending GOTV campaign messages. There is also societal concerns about targeting in political ads, in which scholarly research about targeting based on visual media is still limited.

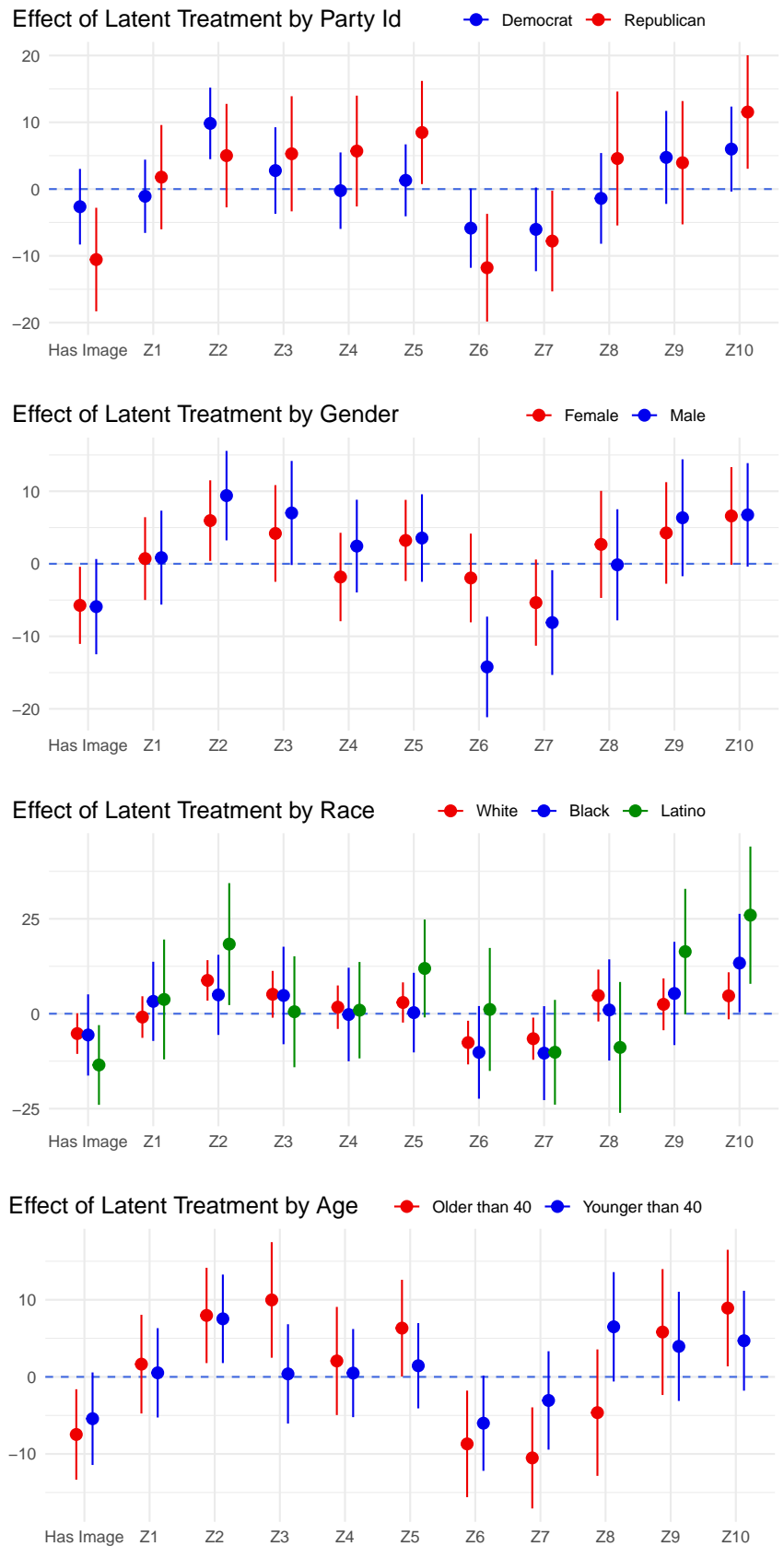


Figure 5.8: Heterogeneous Treatment Effects.

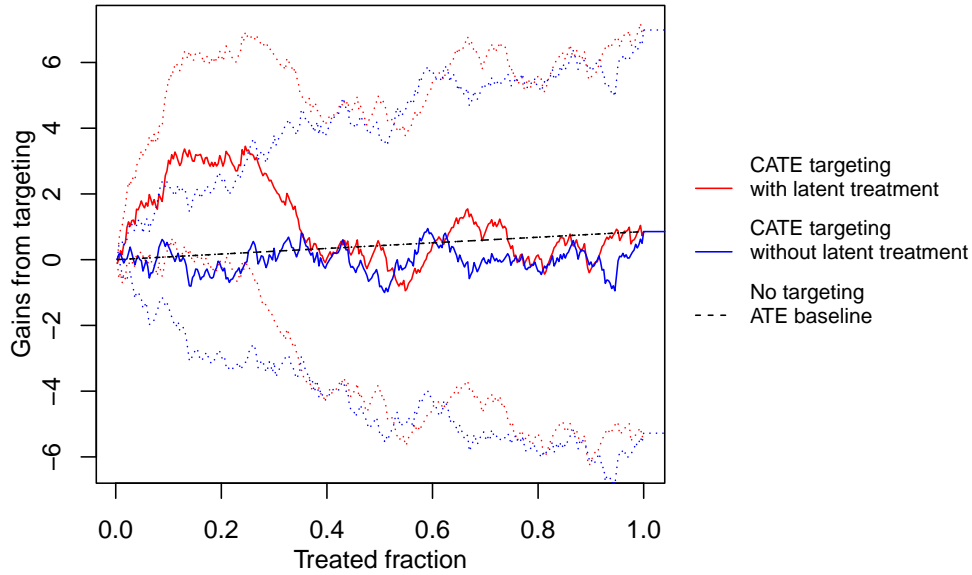


Figure 5.9: Qini Curve.

Note: This plot shows the average policy effect of treating the units most responsive to the treatment as we increase the scale of the treated population.

To account for the multi-treatment nature of our experiment, we adopt the framework of Sverdrup et al. (2023) to evaluate the benefits of targeting using the Qini curve, which plots the average policy effect of treating the units most responsive to the treatment as we increase the scale of the treated population. This involves the following steps:

1. We first train a Causal Forest (Wager and Athey, 2018) on training set to learn the CATEs across partitions of demographic variables and latent treatments:

$$\tau(G_i = g, Z_j = z) = \mathbb{E} [Y_i(1) - Y_i(0) \mid G_i = g, Z_j = z]$$

The benefit of tree-based Causal Forest is that it can flexibly model the relationship between covariates and outcome without strict model specification.

2. We then use the learned Causal Forest Model to infer CATE on test set, i.e., estimate $\hat{\tau}(\cdot)$. To compare the efficacy of latent treatments, we do this both with and without including

information about latent treatments Z_i 's.

3. Form targeting policy $\hat{\pi}(G_i)$ that maps covariates to an optimal treatment decision:

$$\pi : G_i \mapsto \{\text{text-only, image 1, image 2, } \dots\}$$

by comparing across the estimated CATEs $\hat{\tau}(\cdot)$.

4. Calculate the Qini coefficients across treated fractions: In the test data, for each treated fraction, use inverse-propensity weighting (IPW) to estimate the gains from treatment targeting:

$$\frac{1}{n} \sum_i^n \hat{\pi}(G_i) \left(\frac{D_i Y_i}{P[D_i = 1 | G_i]} - \frac{(1 - D_i) Y_i}{P[D_i = 0 | G_i]} \right).$$

Figure 5.9 shows the Qini curve for our visual survey experiment. Solid red curve indicates treatment targeting with CATE including information about latent treatments. Solid blue curve indicates treatment targeting with CATE but without including latent treatments. Black dashed line indicates no targeting—just randomly assign respondents to the overall best performing treatment.

First, we can observe from Figure 5.9 that when treated fraction is not large (smaller than 40%), there is a benefit in treatment targeting. When treated fraction increases, the overall gains converges to the level of non-targeting ATE baseline. This is expected, since with lower treated fraction we can start with targeting respondents most responsive to the treatments. Second, and perhaps more importantly, the benefit of targeting with CATE over ATE **only** occurs when latent treatment information is also included in the CATE. In other words, to reap the full benefits of treatment targeting, it is not sufficient to only include demographic covariates. The presence of heterogeneous responses to latent treatment suggests it is the **interaction** between **demographics** and the **types of imagery** that is driving the benefits of targeting in visual treatments.

5.7 Additional Method for Treatment Discovery by Direct Dimension Reduction on Embeddings

Fong and Grimmer (2016) proposes the Supervised Indian Buffet Process to discover latent treatment from the document-term matrix in text. However, we cannot directly port this method for images due to different featurization mechanism across modality. Words (tokens) are directly interpretable, while pixels are not. Specifically, for Fong and Grimmer (2016), let \mathbf{X} be the document-term matrix where the features are simply the words, the function $g : \mathbf{X} \mapsto Z$ can gain tractions for interpretability by leveraging the words. Figure 5.10 presents the example in Fong and Grimmer (2016) where the authors use the top activated words to understand the latent treatments.

Treatment 1	Treatment 2	Treatment 3	Treatment 4	Treatment 5
appointed	fraternity	director	received	elected
school_graduated	distinguished	university	washington_university	house
governor	war_ii	received	years	democratic
worked	chapter	president	death	seat
older	air_force	master_arts	company	republican
law_firm	phi	phd	training	served
elected	reserve	policy	military	committee
grandfather	delta	public	including	appointed
office	air	master	george_washington	defeated
legal	states_air	affairs	earned_bachelors	office
Treatment 6	Treatment 7	Treatment 8	Treatment 9	Treatment 10
united_states	republican	star	law	war
military	democratic	bronze	school_law	enlisted
combat	elected	germany	law_school	united_states
rank	appointed	master_arts	juris_doctor	assigned
marine_corps	member	awarded	student	army
medal	incumbent	played	earned_juris	air
distinguished	political	yale	earned_law	states_army
air_force	father	football	law_firm	year
states_air	served	maternal	university_school	service
air	state	division	body_president	officer

Table 2: Top Words for 10 Treatments sIBP Discovered

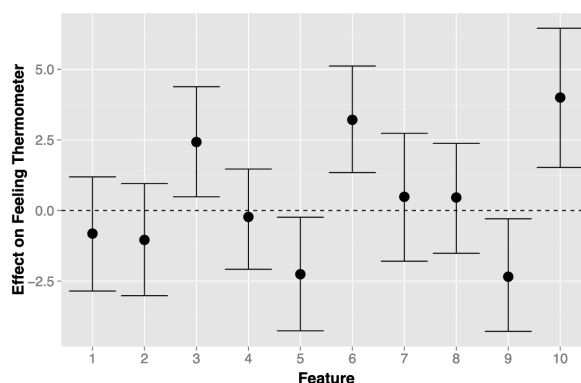


Figure 5.10: Fong and Grimmer (2016): Top words for latent treatment and estimated effects

To identify the latent treatments in an image-based survey experiment, Pugh and Torres (2023) adopts a strategy analogous to Fong and Grimmer (2016), which utilizes K-means clustering on pre-trained image embeddings to create “visual words”—the “tokens” for the images—to form a (image, visual word) matrix \mathbf{X} , and then map these clusters to the low-dimensional treatments Z . One can interpret the latent treatments by investigating the correspondence between latent treatments and the visual words. This is also the procedure in the previous sections.

There are three main drawbacks to this approach:

1. Conceptually, since K-means relies on a single distance metric on the embedding space, an image closer to the centroid of one particular cluster is further from all other clusters, presenting inherent trade-offs between cluster assignments.
2. Embeddings encode some potentially useful information for inferring the latent treatments, while clusters collapse all information from the embeddings to disjoint classes. By discarding information from the embeddings and inferring latent treatments only from the clusters, we risk discarding large shares of useful information.
3. Practically, some visual clusters can be hard to interpret. For example, see Clusters 11, 57 in Figure 5.5. Latent treatment discovery might also map seemingly unrelated clusters to the same treatment. For example, see Z_6 , Z_8 , and Z_9 in Figure 5.5.

This section proposes an alternative method to Pugh and Torres (2023) to alleviate these drawbacks. First, since the inference of Supervised Indian Buffet Process (SIBP) requires approximating the posterior distribution via variational inference, having a large number of features in the \mathbf{X} matrix is not feasible (hence the image-cluster featurization proposed). To solve this, I utilize the Singular Value Decomposition to project the 24576-dimensional image embedding (from the pre-trained BLIP-2 model) to the top latent dimensions, and then use the projection of images on these top latent dimensions, the principal components, to discover the latent treatments via SIBP. This is analogous to the Latent Semantic Analysis (but on embeddings matrix instead of document-term matrix) in Natural Language Processing or the Matrix Factorization (but on embeddings matrix instead of user-item matrix) used in recommender systems.

Second, the learned latent treatment would be a vector of weights given to each of the top latent dimensions. However, both the weights and the top latent dimensions are not readily interpretable. To solve this, I use the principal components of the treatment images and the learned weights for the latent dimensions to score the images. The most highly scored images can be used to interpret the latent treatments. This is akin to a nearest neighbor search of image

in the dual space of the top latent dimensions.

5.7.1 The Indian Buffet Process

The Indian Buffet Process (IBP) is a general stochastic process defining a probability distribution that can be used to define probabilistic models that represent objects with “infinitely” many “binary” features (Griffiths and Ghahramani, 2011). It is still irreplaceable in this alternative framework since it is (to our knowledge) the only stochastic process that satisfies all the following properties:

1. It can represent objects using a **potentially infinite** number of features. In the case of modeling latent treatments, it avoids pre-selecting the number of features.
2. Unlike clustering, the Dirichlet process, or the Chinese Restaurant Process, each data point can be assigned to **multiple** latent classes, so the row sums are not a fixed number and there is no inherent trade-offs between associating a data point to class A versus class B.
3. It associates a data point to a latent class via a **binary** indicator, making treatment effects well-defined.

The Supervised Indian Buffet Process (SIBP) builds on IBP to provide a framework of using outcome variables Y as distant supervision to discover latent treatments Z by assuming that Z follows the IBP (Fong and Grimmer, 2016).

5.7.2 Extract Latent Dimensions via SVD on Embeddings

Similar to the earlier sections, I use pretrained BLIP-2 (Li et al., 2023) to get the embeddings (the last fully connected layer) from each treatment image. This will map each treatment image to a 24576-dimensional embedding.

The SIBP is built on variational inference, which is not scalable for high-dimensional data.¹ Pugh and Torres (2023) proposes to cluster the embeddings to reduce the dimensionality of the image-embedding matrix, but this approach has the three main drawbacks discussed above. To solve this, I propose to take the Singular Value Decomposition on this image-embedding matrix to get a low-rank approximation of the image-embedding matrix and to reduce the number of feature dimensions needed for SIBP inference. Let \mathbf{X} be a $n \times m$ image-embedding matrix, where each row is an image and each column is a dimension in the embedding space. The SVD decomposes \mathbf{X} into three matrices:

$$\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$$

where \mathbf{U} and \mathbf{V} are orthogonal matrices of $n \times n$ and $m \times m$, respectively. The first d columns of each are the right and left singular vectors respectively. $\mathbf{\Sigma}$ is a $n \times m$ diagonal matrix with the singular values in the diagonal. The top d singular values and corresponding singular vectors recovers the best rank- d approximation of \mathbf{X} in terms of the Frobenius norm (Eckart and Young, 1936). In other words,

$$\hat{\mathbf{X}}_d^* = \mathbf{U}_d \mathbf{\Sigma}_d \mathbf{V}_d^\top = \arg \min_{\text{rank}(\hat{\mathbf{X}}) \leq d} \|\mathbf{X} - \hat{\mathbf{X}}\|_F,$$

where $\|\mathbf{A}\|_F = \sqrt{\sum_{i,j} |a_{ij}|^2}$.

Since $\mathbf{\Sigma}$ is diagonal, we can further derive the following more interpretable expression:

$$\hat{\mathbf{X}}_d^* = (\mathbf{U}_d \mathbf{\Sigma}_d^{1/2})(\mathbf{\Sigma}_d^{1/2} \mathbf{V}_d^\top) := \mathbf{A}\mathbf{B}^\top,$$

where \mathbf{A} encodes the representation of the images on the top d latent dimensions (the “semantics” for Latent Semantic Analysis (Deerwester et al., 1990)), and \mathbf{B} encodes the representation of the embedding space on the top d latent dimensions. The i -th row of $\mathbf{A} = \mathbf{U}_d \mathbf{\Sigma}_d^{1/2}$ is a vector of

¹This is less of a problem for document-term matrix, since the researcher can choose to trim the size of the tokens. For example, the analysis of Trump’s campaign messages in Fong and Grimmer (2016) is based on the document-term matrix of 303 tokens.

coordinates of the i -th image on the top d latent dimensions, which is the projection of the image on the space spanned by the top d latent dimensions.

The singular values also quantify the variations of the images in the embedding space. The left panel in Figure 5.11 plots how the variance explained by top dimensions decay as the order of the dimension increases. In our dataset, the first latent dimension explains 46% of the total variance, and the second latent dimension explains 8% of the total variance.² The right panel in Figure 5.11 plots the cumulative variance explained by the top dimensions. One can see that the top 20 dimensions explain around 80% of the variance in the embedding space among the treatment images. The analysis below will be based on the top 20 latent dimensions.

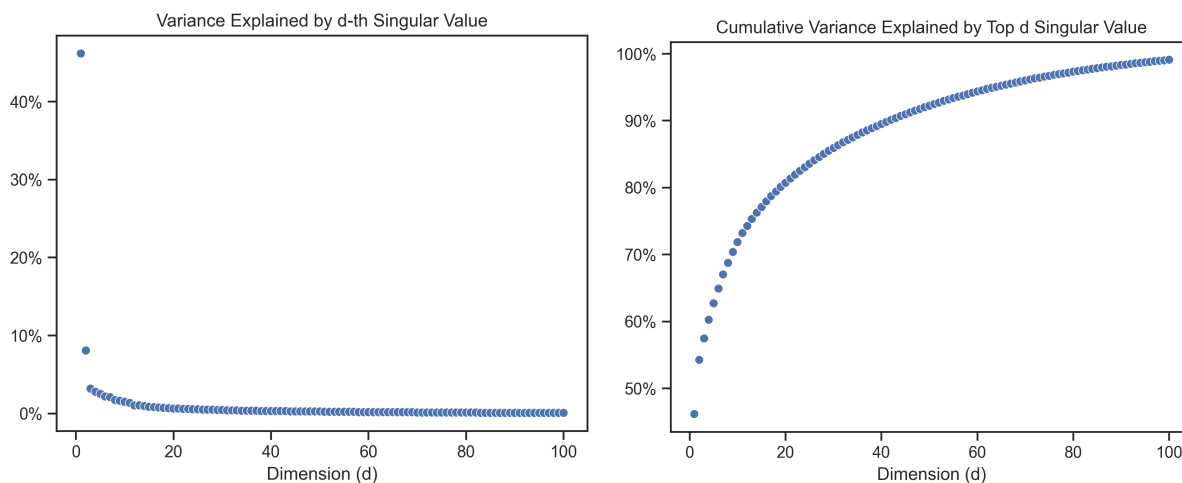


Figure 5.11: Variance Explained and Cumulative Variance Explained by Top Latent Dimensions.

To verify that the top latent dimensions captures some underlying concepts, we can plot the treatment images on the components of the top latent dimensions. Figures 5.12–5.14 show the treatment images on the top 5 columns of \mathbf{A} . One can see that visually similar images would be placed into nearby coordinates on the latent dimensions. Having successfully reducing the dimensions without losing information due to clustering, we can then use this image-latent-dimension matrix for inference in SIBP.

²Since there are only 120 treatment images, despite the embedding is 24576-dimensional, the maximum possible latent dimension is still 120.

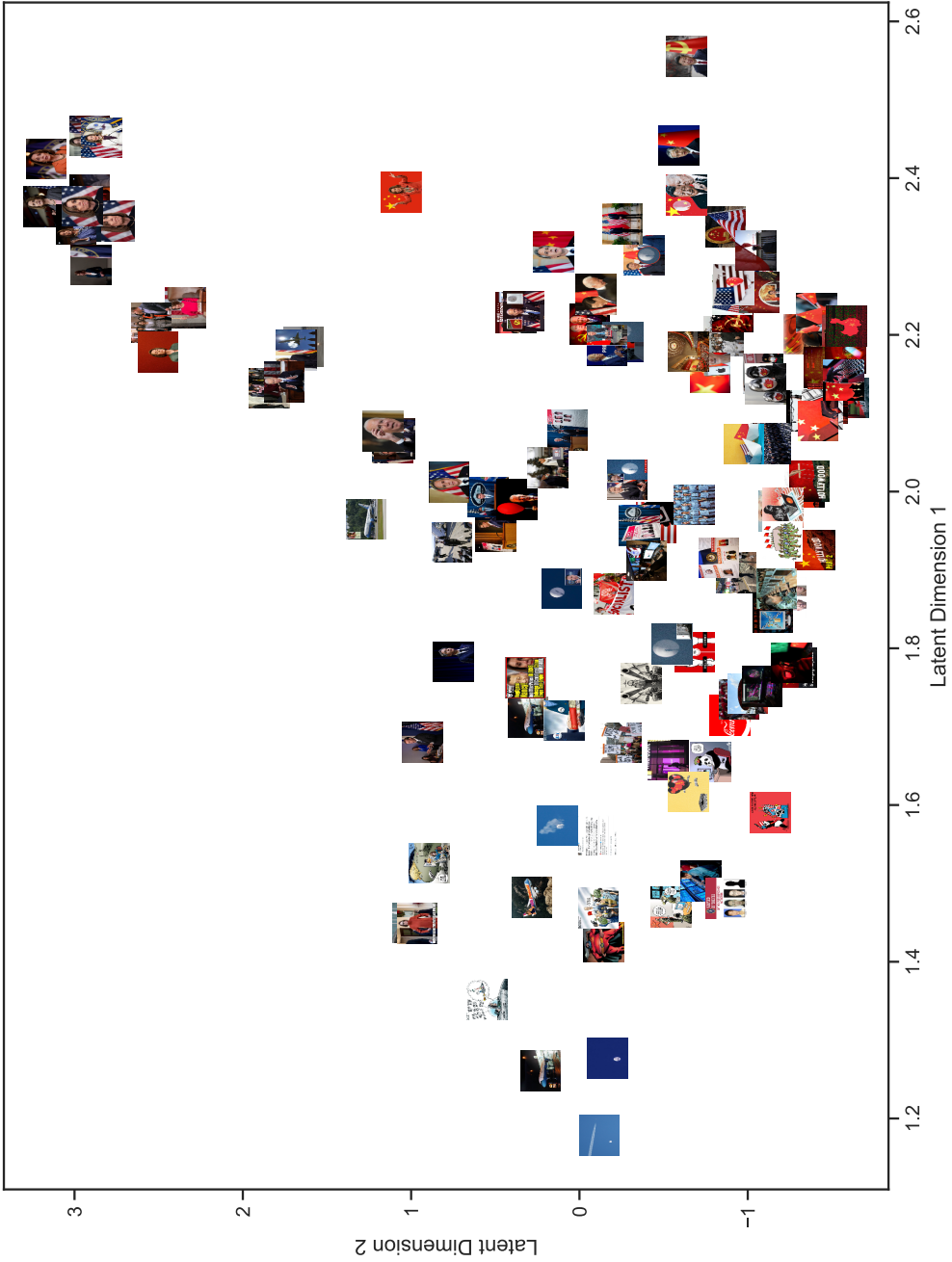


Figure 5.12: Treatment Images on the 1st & 2nd Latent Dimensions.

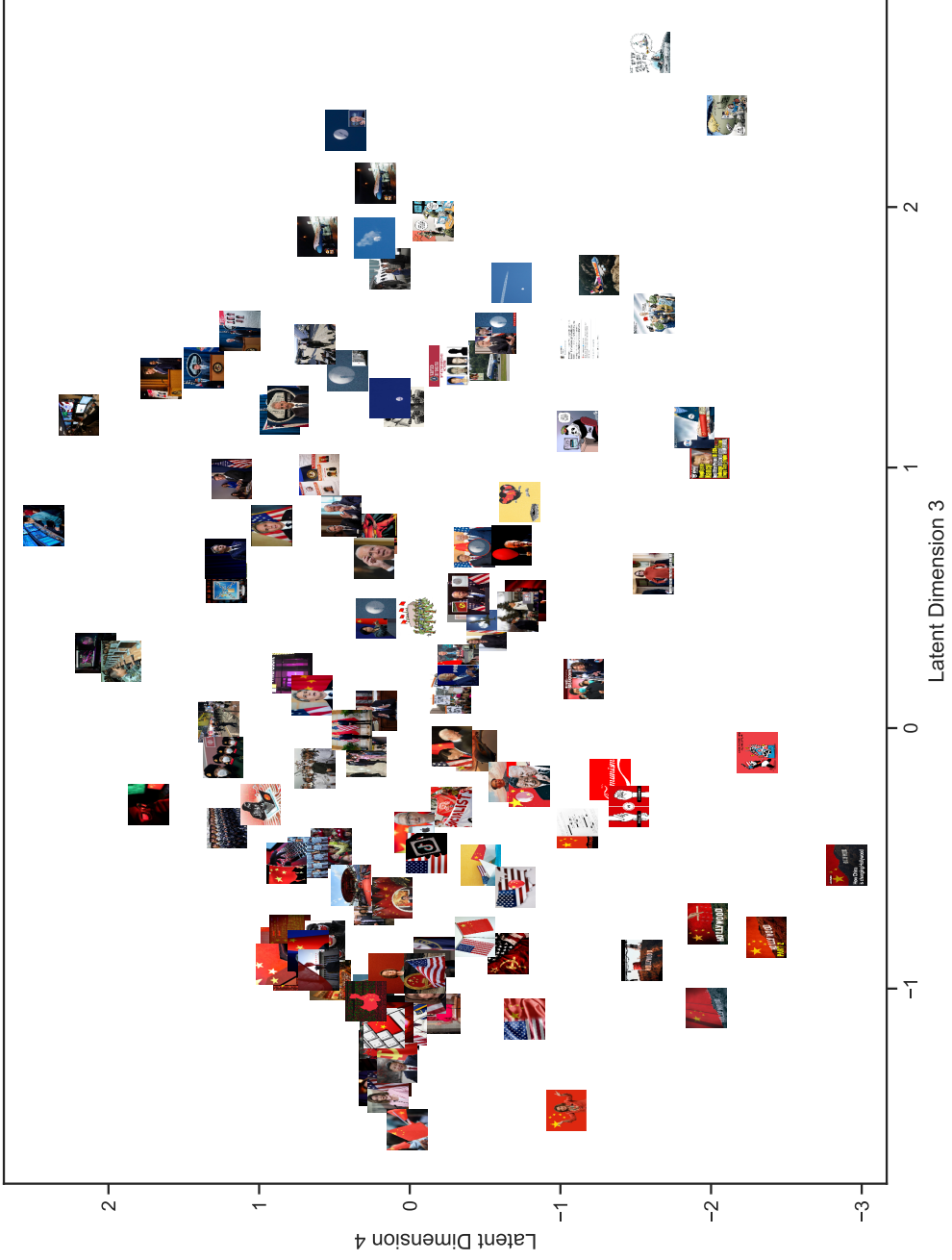


Figure 5.13: Treatment Images on the 3rd & 4th Latent Dimensions.

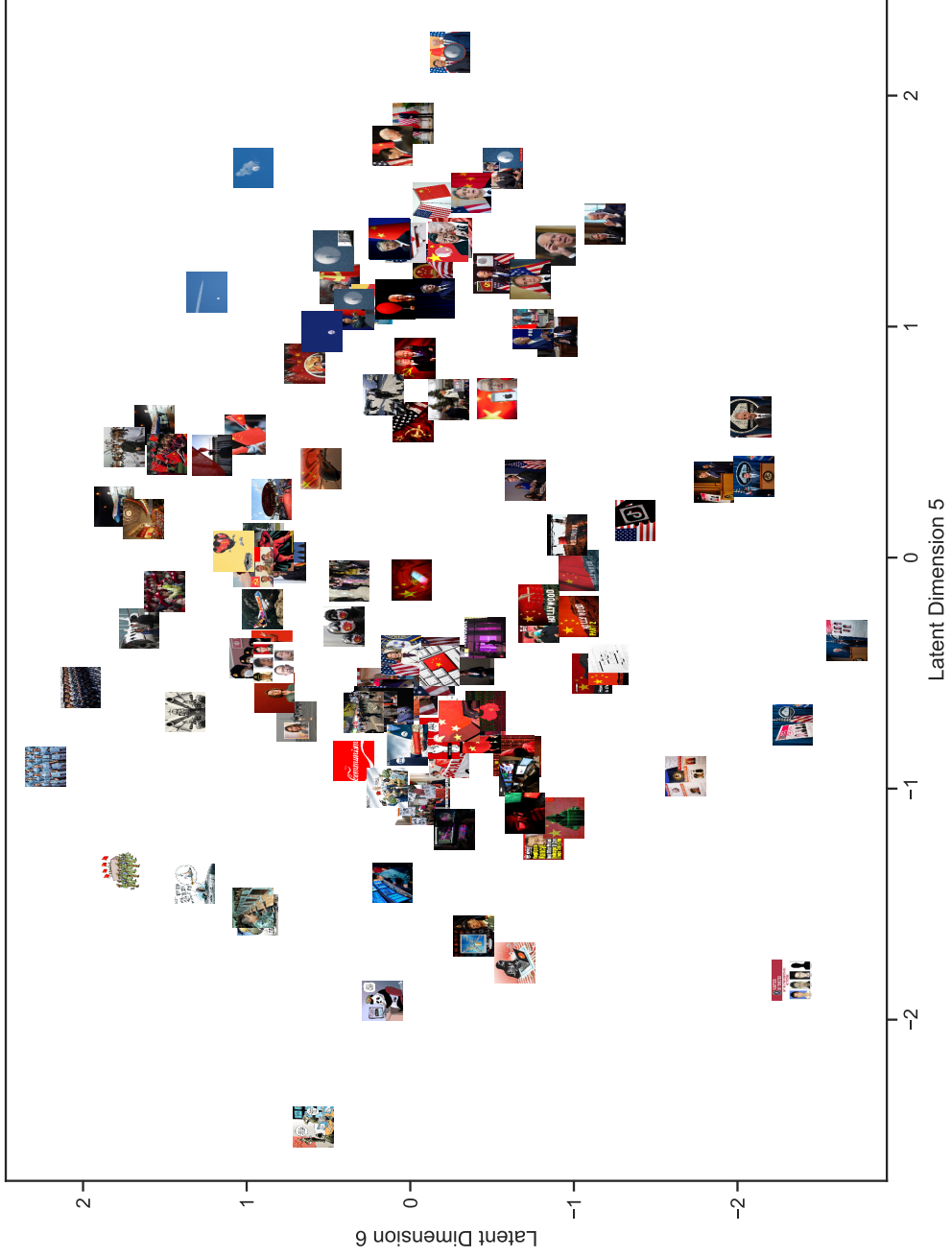


Figure 5.14: Treatment Images on the 5th & 6th Latent Dimensions.

5.7.3 Score Treatment Images by Latent Dimensions

After constructing the image-latent-dimension matrix \mathbf{A} , we follow the same procedure to split the matrix into training set (for treatment discovery) and test set (for effect estimation). We can then take the training set in \mathbf{A} , along with their responses, to learn the SIBP. The SIBP would output the posterior distributions of how each latent dimensions (columns in \mathbf{A}) contribute to the latent treatment Z_i 's. Specifically, the posterior means of these distributions would be the importance weights of the latent dimensions among the latent treatments. Following the notation in Fong and Grimmer (2016), we use $\boldsymbol{\phi}_k = (\phi_{k1}, \dots, \phi_{kd})$ to denote the d -dimensional importance weights of the latent dimensions on the k -th latent treatments.

However, these dimensions in themselves are not interpretable. This creates a challenge in interpreting the latent treatments. To interpret the latent treatments, I derive a score of how each treatment images are activated by the latent treatments. Given that the i -th row of \mathbf{A} represents the i -th treatment image on the d latent dimensions, and that $\boldsymbol{\phi}_k$ represents the importance weights of the latent dimensions on the k -th latent treatments, we can calculate an activation score for image i by latent treatment k as

$$\text{Score}(i, k) = \mathbf{A}_{i*} \cdot \boldsymbol{\phi}_k = \sum_{j=1}^d \mathbf{A}_{ij} \phi_{kj}.$$

This procedure differs from Fong and Grimmer (2016) since, in this case, the weights in $\boldsymbol{\phi}_k$'s correspond the words to the latent treatments, where we can infer the latent treatments by looking at the top words. This procedure also differs from the procedure in Pugh and Torres (2023) since the weights in this case correspond the clusters to the latent treatments, where we can infer the latent treatments by looking at the top clusters. In our proposed method, we transform the space of latent dimensions back to the images to gain interpretability.

Figure 5.15 provides the top 8 activated images for each latent treatments Z_1 – Z_{10} . Contrasting this to the previous method, as shown in Figure 5.5, this new method based purely on

factorization on the embeddings provides more interpretable latent treatments. For example, Z_6 is about white in Figure 5.5 would group clusters about female clothes, necklaces, masks, and Chinese flags into the same latent treatment. In contrast, the closest latent treatment in the alternative method, Z_3 in Figure 5.15, would group all images of female into the same latent treatment. Take Z_7 in Figure 5.5 as another example, which groups clusters about comics, stars on the Chinese flag, and the hammer and sickle Communism symbol into the same latent treatment. In contrast, in the alternative method, Z_6 , Z_5 , and Z_8 in Figure 5.15 would capture them separately into different latent treatments.

5.7.4 Estimated Effects Based on Alternative Method

Based on these alternatively discovered latent treatments, we can then estimate the same set of Average Treatment Effects. Figure 5.16 reports the regression results based on these alternative latent treatments. We get both some similar and some distinct result as the latent treatments discovered by clustering (Figure 5.6). For example, in a similar light, we find that respondents treated with images of female features (Z_3) decreases credibility perception while treated with male features (Z_1) increases credibility perception, which is consistent with the clustering method. On the other hand, we find that respondents treated with images of comics (Z_6) does not have a detectable effect on credibility perception, while treated with Chinese flags (Z_5) or the hammer and sickle Communism symbol (Z_8) has a detectable effect to decrease credibility perception. These three treatments are bundled in Z_7 in the latent treatments discovered by clustering (See Figures 5.5 and 5.6).

5.8 Conclusions and Discussions

Do images lend credibility to news articles? The answer from this work is that—it depends. We show this by first designing a novel visual survey experiment where respondents are

Top Images by Latent Treatments

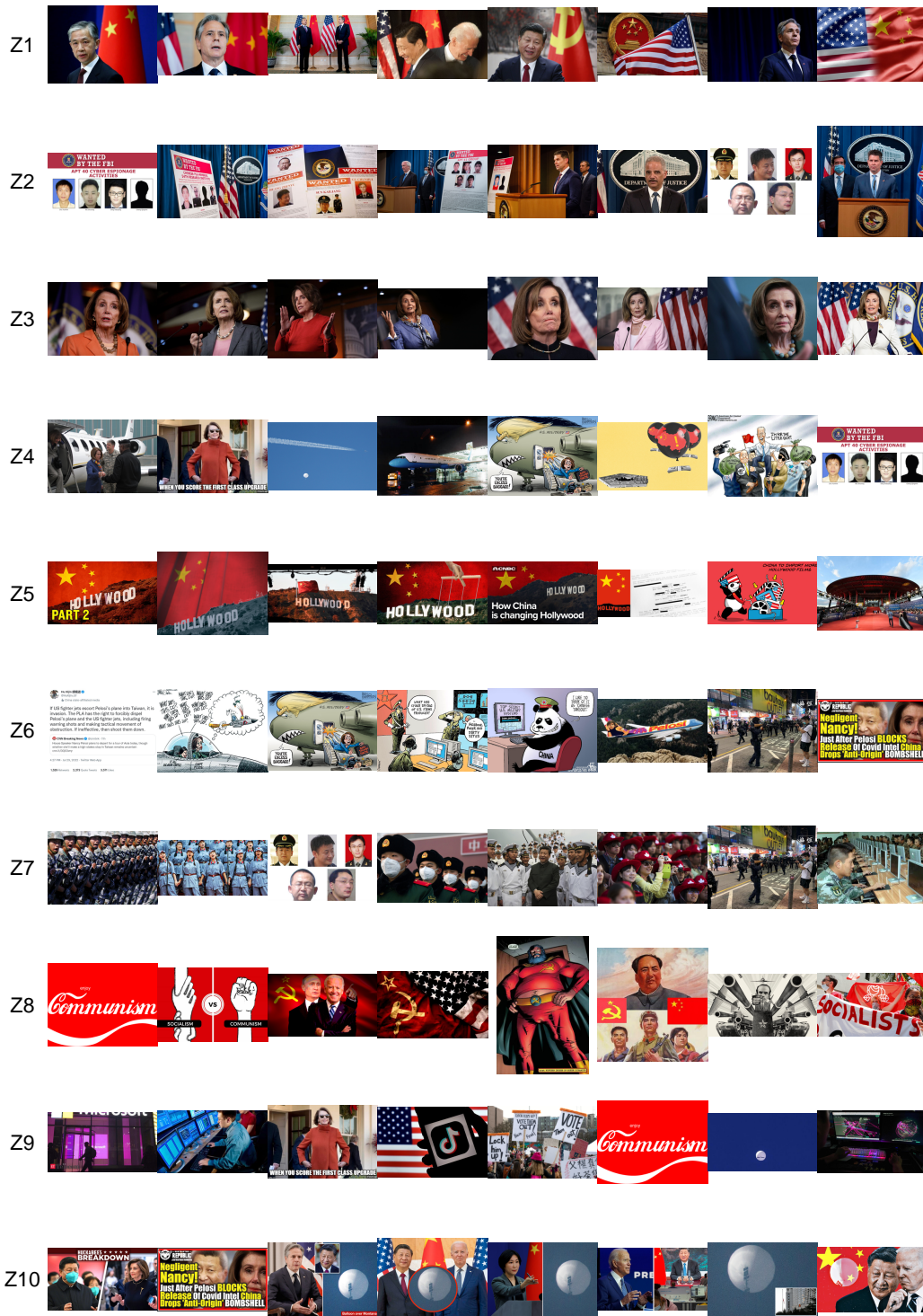


Figure 5.15: Top 8 activated images on the latent dimensions for each latent treatment

Perception of News being True (0–100 Scale)

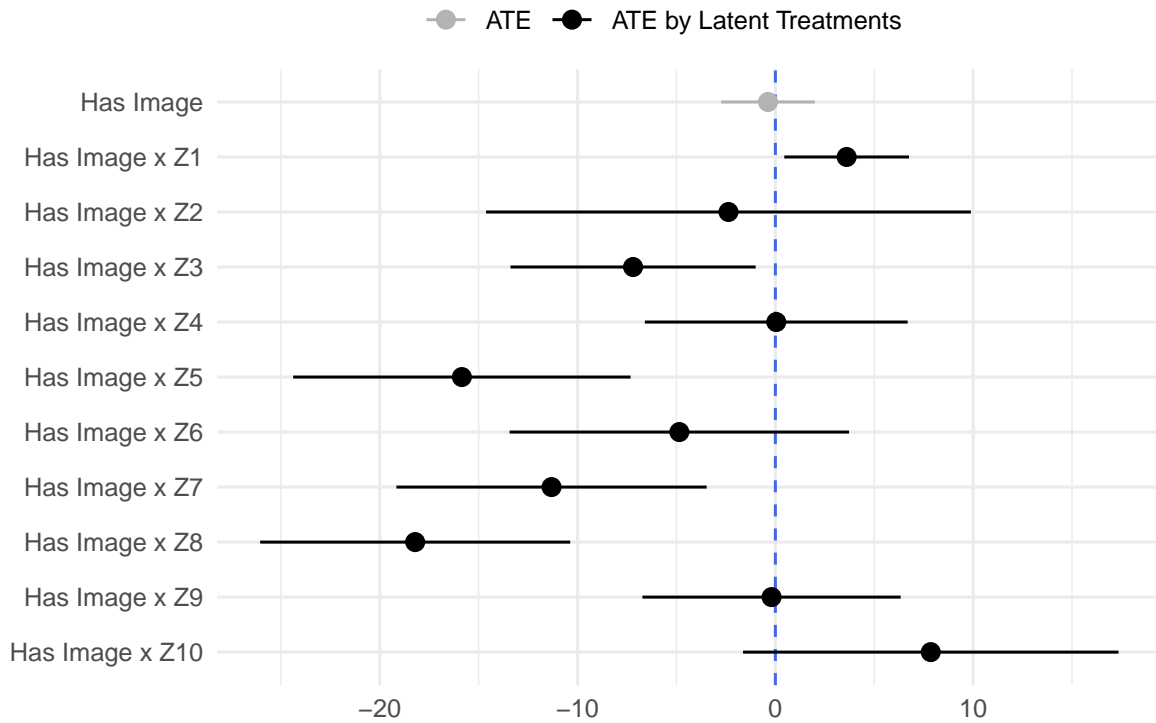


Figure 5.16: Average Treatment Effects of Alternative Latent Treatments.

treated with different types of images across four news stories. We then utilize Large Vision Models to extract the visual information from these images to group visually similar image blocks into clusters, without human labeling or supervision. Thirdly, we leverage established method in NLP to identify and adjust for latent treatments—by learning a topic model in training set and use the predicted topic in test set for downstream inference—to separate treatment discovery from treatment effect estimation.

Results suggest that adding images to news articles in itself has no clear overall advantage over text-only controls. However, this overall null effect is masked by blending together latent visual treatments, some of which can increase respondent’s perception of news credibility while others decrease credibility perception. Heterogeneous Treatment Effect analysis also shows there are differences in group-level responses to certain latent treatments by gender, race, and age. We

also provide evidence that targeting respondents by demographics into different visual treatments can be moderately valuable.

5.8.1 Future Extensions

There are several extensions to continue this research agenda. First, I plan to expand this experiment to a much larger set of news stories—not just those related to US-China relations—to understand the broader relationship between visual media, news credibility, and demographics. Second, it is possible to derive a more scalable method to learn SIBP than the current Bayesian approach to infer latent treatments. However, the proposed alternative method in Section 5.7 to reduce the dimension of the embeddings can enhance the applicability of SIBP for both text and images, without sacrificing the desired properties of SIBP (binary, no trade-offs between classes, as discussed in Section 5.7.1) and interpretability. Thirdly, researchers might also leverage large generative models to generate synthetic images to directly manipulate certain aspects of images, although most open source models to date are trained for generating human portraits or movie scenes instead of images on news websites. Lastly, hosting image-based survey experiment and making inferences on images have a pretty high barrier to entry due to the close to non-existent of researcher-friendly infrastructures. Building these tools are also important next steps.

Appendix A

Supporting Information for: COVID-19 Increased Censorship Circumvention And Access To Sensitive Topics In China

A.1 Twitter Activity by Province

Figure A.1 shows the number of unique, geolocating users who are tweeting in Chinese by province. The x-axis is the number of months before (negative) or after (positive) the initial coronavirus lockdown in Hubei province. The blue line is a pre-lockdown average for x less than 0 and a five term polynomial regression for x greater than or equal to 0 (where 0 is the first day of Hubei's lockdown). The points in Figure 2 are the values of the blue line by province for x equals 1/30 (first day of lockdown) and x equals 1 (day 30 of lockdown).

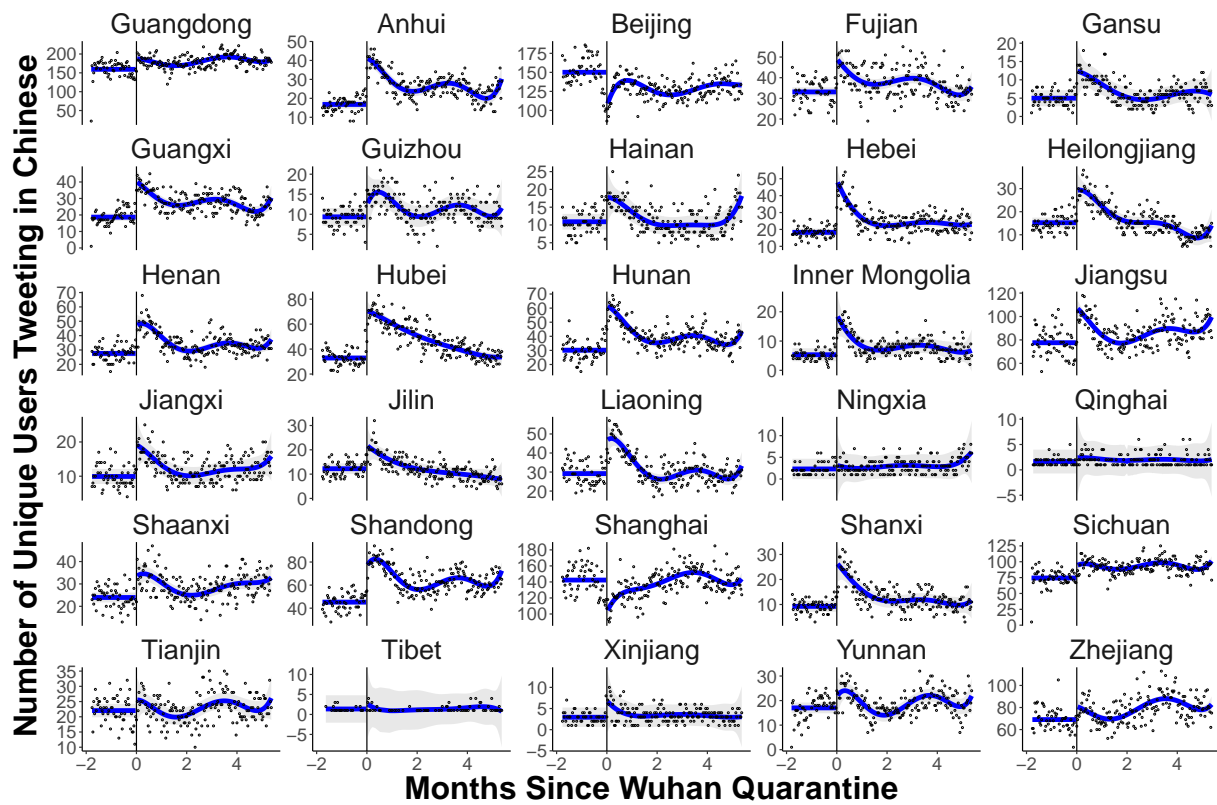


Figure A.1: Increases in Geolocated Twitter Activity by Province (modeled)

A.2 Twitter Data

From a global sample of tweets with GPS coordinates, we found the 1,448,850 tweets from China from December 1, 2019 through June 30, 2020, 367,875 of which are in Chinese. This corpus contains 101,553 unique users, 43,114 of whom had names or descriptions in Chinese. These dates were chosen to encompass a baseline period and the height of COVID-19 in China. This corpus is used for Figures 2 and 4, evaluating the impact of the lockdown on tweeting behavior.

For the follower analysis (Figures 5 and 6), we sample 5,000 of these 43,114 accounts. For these 5,000 random users in China, we download who they follow, their “friends” in Twitter parlance. From these friends, we identify the 5,000 most commonly followed accounts that are either a Chinese language account or have Chinese characters in their name or description field. Of these 5,000 most common friends, the vast majority were pornography accounts. We therefore hand-categorized the accounts into pornography or not pornography. We keep the 354 non-porn accounts and sample 200 from the remaining 4,646 porn accounts.

We then download the followers of these 554 accounts. We identify 38,050,454 total followers. For each, we identify the location of the users. Because very few of these followers have geolocated information, we rely on the language of their Twitter status and their self-reported location to distinguish between mainland and overseas followers. We only include users whose status language is Chinese in order to study only Chinese language followers of these accounts. Followers are classified as Mainland Chinese if the location field contains the name of a Chinese city, town, or province. Followers are classified as from Hong Kong if the location field contains the name of a district in Hong Kong. Followers are classified as Taiwanese if the location field contains the name of a Taiwanese city, county, or district. Followers are classified as US if the location field contains the name of states or state abbreviations (in capital letters).

A.3 Mobility and Twitter Usage

To better understand the relationship between lockdown and Twitter usage, we use the publicly-available human mobility data from Baidu Qianxi (<https://qianxi.baidu.com/2020/>), which tracks real-time migration (including moves in & out of provinces and within city movements) across China during the Lunar New Year period in both 2020 and 2019. The move out data is downloaded from Harvard Dataverse (China Data Lab, 2020; Hu et al., 2020), and we scrape the within-city movement data from Baidu Qianxi.

Figure A.2 plots the average within city movement index in both 2020 (real black line) and 2019 during the same period in the Chinese Lunar New Year (dotted line). Specifically, since the New Years day is on February 5 in 2019 and January 25 in 2020, we shifted the dates in 2019 backwards for 12 days to match the dates in 2020. Red vertical line indicates the day of Wuhan lockdown. One can see that almost all provinces experienced a huge decrease in human mobility after January 23 in 2020, compared to the same period in 2019. In 2019, we only see significant decreases in mobility in Beijing, Shanghai, and Tianjin.

We also validate that the increase in geolocated Twitter users is correlated with the decrease in human mobility. The left panel of Figure A.3 plots this correlation. Let $M_{i,t}$ denote the mobility index for province i on date t . The x-axis plots the decrease in within city movement index from January 22 (the day before the Wuhan lockdown) to February 22, $M_{i,\text{Jan22}} - M_{i,\text{Feb22}}$. The y-axis plots the increase in geolocated Twitter users for 30 days after the Wuhan lockdown compared to the average number of geolocated Twitter users in a province before the Wuhan lockdown. This shows that the more reduction in human mobility, the more increase in geolocated Twitter users, comparing to the levels before the lockdown. Hubei province experience the most reduction in mobility and most increase in the number of geolocated Twitter users.

In Figure A.1 we see increases in geolocated Twitter users in most provinces except Beijing and Shanghai. One explanation for this is that Twitter users in Beijing and Shanghai left

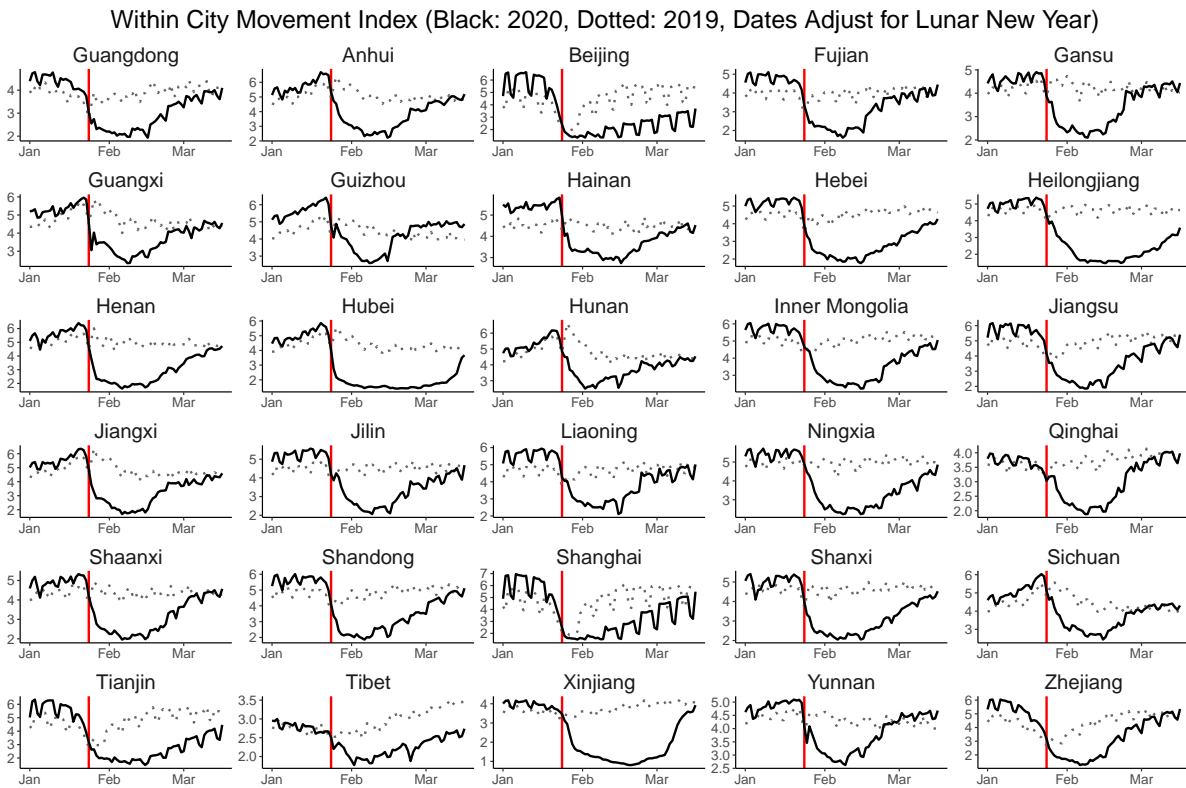


Figure A.2: Within city movement index by Province (black: 2020, dotted: same period in 2019).

Note: Real black line indicates the time series for the average within city movement index by province in 2020. Dotted line indicates the average within city movement index by province during the same Chinese Lunar New Year period in 2019. Red vertical line indicates the day of Wuhan lockdown. Chongqing is excluded since it is not counted in Twitter’s geolocation map.

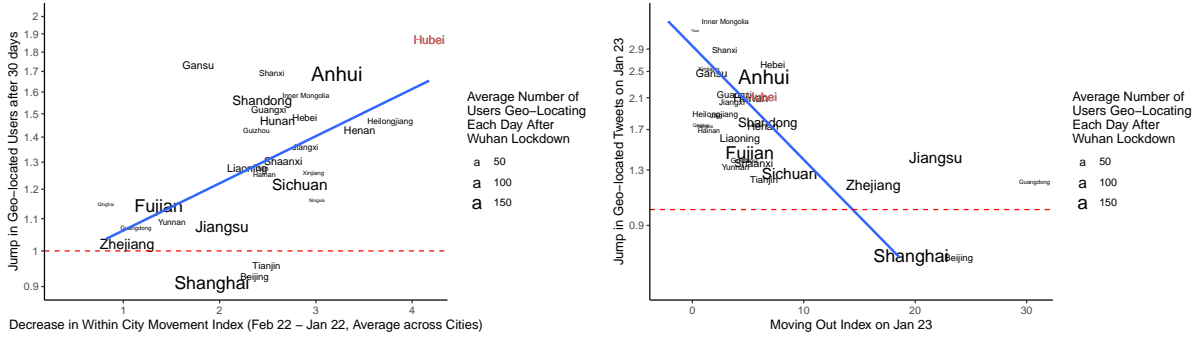


Figure A.3: Reduction in within city movement and increase in geolocated Twitter users during the month of Wuhan lockdown (left); degree of moving out and increase in geolocated Twitter users on the day of Wuhan lockdown (right).

Note: The left panel plots the correlation between decreased mobility and increased geolocated Twitter users during the first 30 days of Wuhan lockdown. The x-axis plots the decrease in within city movement index from January 22 (the day before Wuhan lockdown) to Feb 22. The y-axis plots the increase in geolocated Twitter users for 30 days after the Wuhan lockdown compared to the average number of geolocated Twitter users in a province before the Wuhan lockdown. The right panel plots the relationship between moving out of province and the increase of geolocated Twitter users on January 23, the day of Wuhan lockdown. Estimates and day of lockdown are drawn from a five term polynomial regression on the number of unique geolocated Twitter users per day after the lockdown. These province-by-province polynomials are displayed over the raw data in Figure A.1.

the cities during the outbreak. Mobility data supports this explanation. The right panel of Figure A.3 plots the relationship between moving out of a province on January 23, the day of Wuhan lockdown, and the increase of number of geolocated Twitter users on the same day, compared to the average number of geolocated Twitter users in a province before the lockdown. One can see that the more people moving out, the less jump in Twitter user on the day of lockdown. Beijing, Shanghai, and Guangdong all experienced large outflows of individuals on the day of Wuhan lockdown.

Since the period of Wuhan lockdown overlaps with the Chinese Lunar New Year, increased Twitter usage could partly be due to general boredom during the New Year. To explore New Year versus pandemic effects, we normalize both mobility and number of Twitter users in 2020 by those in the same period in 2019. To do so, we first adjust the dates in 2019 backwards for 6 days to match the dates of 2020 Lunar New Year. Then, we create normalized mobility and Twitter usage. Specifically, denote $M_{i,y,t}$ the mobility index and $T_{i,y,t}$ the Twitter usage for

province i in year y on date t . The normalized mobility index would be

$$M_{i,2020,t}/M_{i,2019,t}$$

and the normalized Twitter usage would be

$$T_{i,2020,t}/T_{i,2019,t}$$

We then plot the weekly change in mobility and Twitter usage after Wuhan lockdown, comparing to the period before Wuhan lockdown. Figure A.4 shows the plots. In mathematical notations, for the first week of Wuhan lockdown, we plot

$$\frac{M_{i,2020,Week\ 1}/M_{i,2019,Week\ 1}}{M_{i,2020,Week\ 0}/M_{i,2019,Week\ 0}}$$

on the x-axis and

$$\frac{T_{i,2020,Week\ 1}/T_{i,2019,Week\ 1}}{T_{i,2020,Week\ 0}/T_{i,2019,Week\ 0}}$$

on the y-axis. This is shown in the top left panel in Figure A.4. The other panels shows the corresponding ratios for the 2nd, 3rd, and 4th week, respectively (all relative to the week before lockdown).

Figure A.4 shows that we still find effects of reduced mobility on increased Twitter usage, after adjusting for the decrease in movement driven purely from New Year, in the early periods of lockdown (at least for the first week of lockdown, the correlation for the second week is not statistically significant). In the 3rd and 4th weeks, we find a general increase in Twitter usage in most provinces, regardless of the relative decrease in mobility in these weeks. In other words, the mobility-induced effect specific to Wuhan lockdown fades out in around 2 weeks, and there's a general increase in Twitter usage across China that is not related to reduced human mobility. This pattern suggests that the increase in Twitter usage is not driven only by people's staying at home

because, if that is the case, we expect to see a continued relationship between relative reduction in mobility and increase in Twitter usage, as other Provinces started to announce stay-at-home orders. This pattern is also not driven only by New Year because we should not expect to see an overall increase in Twitter usage after normalizing with the same New Year period in 2019.

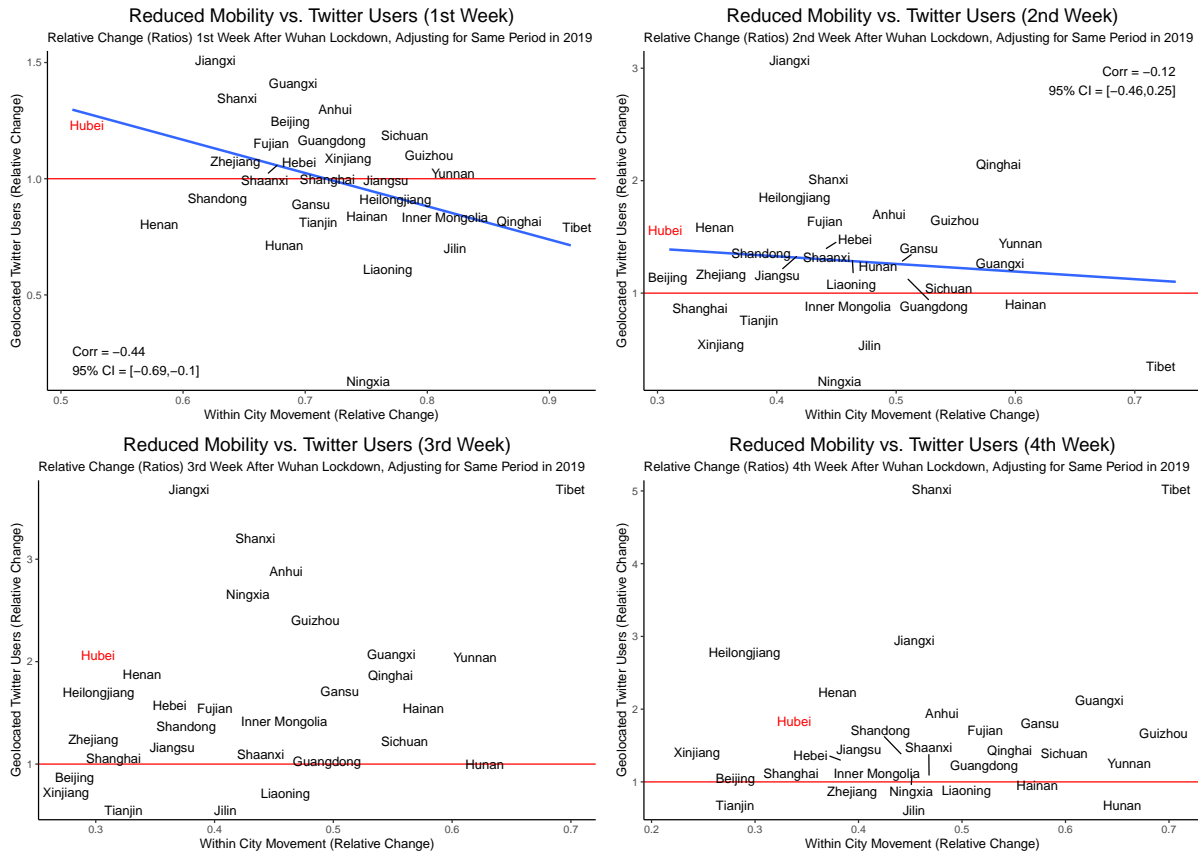


Figure A.4: Weekly changes in within city movement and geolocated Twitter users relative to pre-lockdown period, after adjusting for the same period in 2019.

Note: We plot the weekly relative change in mobility and Twitter usage after Wuhan lockdown, adjusting for the same period in 2019 and comparing to the period before Wuhan lockdown. Specifically, denote $M_{i,y,t}$ the mobility index and $T_{i,y,t}$ the Twitter usage for province i in year y on date t . For the first week of Wuhan lockdown, we plot (in the top left panel) $\frac{M_{i,2020,Week\ 1}/M_{i,2019,Week\ 1}}{M_{i,2020,Week\ 0}/M_{i,2019,Week\ 0}}$ on the x-axis and $\frac{T_{i,2020,Week\ 1}/T_{i,2019,Week\ 1}}{T_{i,2020,Week\ 0}/T_{i,2019,Week\ 0}}$ on the y-axis. The other panels shows the same for the 2nd, 3rd, and 4th weeks, respectively.

Note also that the first week of Twitter use in 2020 was not much higher than in 2019 because 2019 saw a very large number of posts on Chinese Lunar New Year. This increase was presumably related to New Year related posts, and these celebratory posts did not increase to the same extent during the start of the COVID-19 pandemic.

A.4 Effect Size

A.4.1 New Twitter Users

This section provides rough estimates of absolute increases in Twitter use in China, and sections below expand it to consider increased Twitter followings and increased Wikipedia use. Note that these are estimates for increased usage on only these sites, which require that new users from Mainland China (where these sites are blocked) 1) create an account to view Twitter content and 2) use cookies (to be recorded in the Wikipedia unique device data). Other sites that do not require accounts could have seen larger increases, and the Wikipedia unique device counts are underestimates.

The top panel of Figure 2 shows a 10% long-term increase in the number of geolocating users from China. In 2019, as reported in (Mozur, 2019), Professor Daniela Stockman of the Hertie School of Governance surveyed 1,627 internet users in China and found .4% of them use Twitter; the article reports that number as 3,200,000. Roughly, if the same 10% increase applies to all users from China and the long-term increase reflects a new pool of users (the number of unique geotagging users in our sample in May 2020 was around 10% higher than in December 2019), then 320,000 new users joined Twitter because of the crisis.

We can assess this estimate by considering 1) the fraction of (posting) users who geotag and 2) the number of unique geotagging users in our data. For 1), using a sample of 100 hours of non-geolocated tweets from 2019.01.01-2020.12.31, we found 37,957 in Chinese. Assigning location to these tweets using the same code that was used to assign location to followers of the most commonly followed accounts, we then found that 1.79% of tweets and 1.95% of users from China geotag. For 2), we find that 47,389 unique Twitter users geotagged (in Chinese and in China) in our sample (note, however, that our 1% sample captures approximately 56% of tweets that are geotagged). Dividing this number by 0.0195 gives us 2.4 million Twitter users, suggesting that somewhere around 70% ($\frac{2.4}{3.2 \times 1.1}$) of geotagging Twitter users in China publicly

geotagged posts and were in our sample.

Though this number is small in the context of China's 1.4 billion inhabitants, it is nonetheless important for three reasons. First, the effects in this paper are a minimum effect size for Twitter since accounts do not have to use geolocation or provide an accurate self-reported location in their profile. Second, the effects documented herein focus only on one banned platform (Twitter) and website (Wikipedia), and there is no reason to think the same behavior did not occur on other banned platforms like Facebook, Telegram, and Instagram as well as banned websites such as Reddit or The New York Times. Third, the Chinese government behaves as though these relatively small numbers threaten it. Since 2018, it has become increasingly repressive in response to comments its citizens make on platforms unavailable in China. Recently, several individuals from China have been arrested for comments made on platforms such as WhatsApp (owned by Facebook, unavailable inside the Great Firewall) and Twitter (Mozur, 2018). The government has also started large influence campaigns on social media platforms that are unavailable domestically, including Twitter (Kinetz, 2021). If the behaviors documented in this paper were immaterial, then we believe the government would not put such a priority on attempting to control speech on these platforms.

A.4.2 Followers

In addition to causing new people to join Twitter, the crisis caused more people to follow accounts posting sensitive content. Here, we estimate the number of surplus followers from China and show that they persist after the crisis, perhaps at greater rates than users who follow after.

Figure A.5 shows the absolute number of excess followers (top) and its ratio (bottom). The absolute number is the total number of new followers minus the total number of predicted new followers based on the December daily average growth rate per category; the bottom panel divides the new follower count by the predicted number of new followers. Several interesting

patterns emerge. First, the crisis clearly causes all account types to gain followers; some, such as pornography and international news agencies, may even have served as early warning indicators since they receive excess followers before the Wuhan lockdown. Second, the categories with the most excess followers, citizen journalists/political bloggers and international news agencies, are exactly those people would seek out in a crisis. By the end of March, 53,860 more accounts follow citizen journalists/political bloggers than would have happened without the crisis; for international news agencies, 52,144. Third, normalizing for the expected number of new followers reinforces that attention was paid to sensitive categories. Extra, early attention is paid to the citizen journalists and activist categories (which received almost 4 times as many new followers during the lockdown as we would expect based on December’s following rate), while international news agencies’ importance decreases to third place. Normalizing emphasizes the increased attention activists receive since they have relatively fewer followers than the other categories. Fourth, Chinese accounts increase their following of state media or Chinese officials once Hubei’s lockdown lifts, though from a low base.

Importantly, these excess followers persist a year after the lockdown. To make this claim, we crawled the follower list of the same popular accounts starting on May 31, 2021, more than one year after the first crawl, and assigned location using the same procedure as before. Comparing the 2021 follower lists to 2020 shows which followers stopped following the popular accounts. We then calculate the percentage of the 2020 followers that persist in 2021 by account type, follower location, and date. Table A.1 shows these results.

Table A.1: Persistence of Followers by Account Type and Period Following Starts

	Pre-Lockdown			Lockdown			Post-Lockdown		
	China	Hong Kong	Taiwan	China	Hong Kong	Taiwan	China	Hong Kong	Taiwan
International News Agencies	87.31	87.71	85.51	90.80	90.19	83.20	89.09	87.35	88.70
Citizen Journalists / Political Bloggers	72.84	79.58	78.73	87.49	86.35	85.00	81.69	83.01	79.64
Activists or US / Taiwan / Hong Kong Politics	78.27	76.74	76.45	88.02	86.40	83.83	85.82	85.46	83.99
Pornography Accounts	85.56	84.28	83.00	88.84	87.51	89.52	86.32	87.19	86.88
State Media or Chinese Officials	82.72	81.38	84.62	87.99	86.36	84.94	85.90	86.61	82.45
Non-Political Bloggers or Entertainment Accounts	73.75	72.94	65.66	87.80	87.01	87.19	81.90	85.13	83.61

Note: Each cell is the percent of followers from April 2020 that still follow the six account types (row) in May 2021, by follower location and period the follower started following the account. The lockdown period is January 23, 2020 - March 13, 2020. Post-lockdown refers to March 14-April 1.

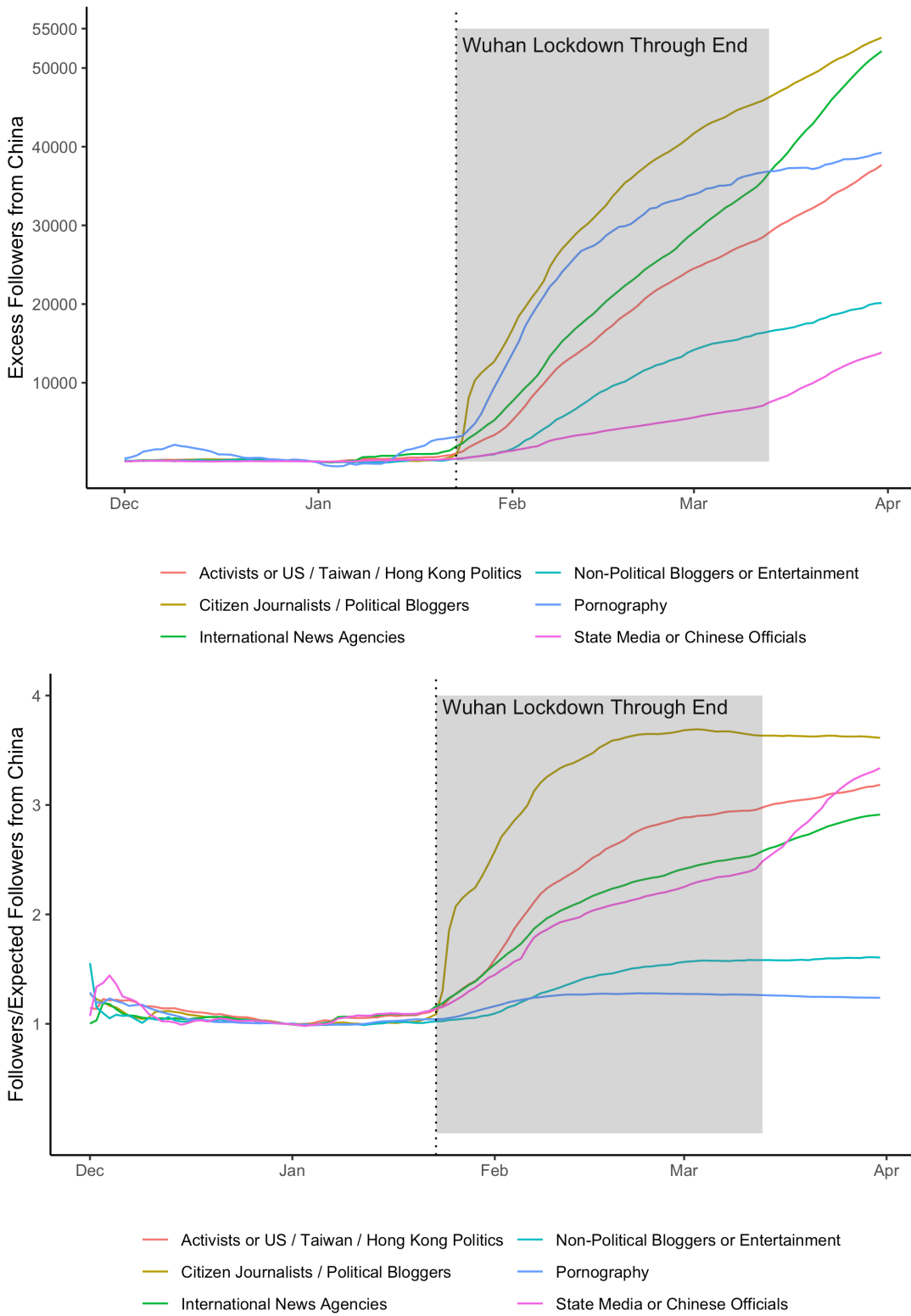


Figure A.5: Excess Followers, absolute (top) and ratio normalized by category growth rate (bottom). Growth rate is calculated based on the December 2019 average number of new followers by category.

Accounts from China that start following the popular accounts during the lockdown period persist at the same to slightly higher rates than those that start following before or after then. 87.31% of accounts from China that start following international news agencies before the lockdown persist versus 89.09% that start following after the lockdown. The difference is especially stark for citizen journalists/political bloggers. Finally, since older followers should have a lower persistence rate since more time has passed, it is striking that accounts that start following during lockdown have higher persistence rates than newer accounts, those that start following during the seventeen days after the lockdown ends. The increased exposure to sensitive content persists after the crisis passes at rates equal to or greater than for non-crisis periods.

A.4.3 Number of unique devices accessing Wikipedia with cookies enabled

Wikipedia tracks the number of unique devices that have accessed its site each day and month using a ‘privacy-sensitive access cookie’ (<https://dumps.wikimedia.org/other/analytics/>). By design, this number does not count devices not accepting cookies through private browsing (as we might expect from users accessing Wikipedia from within Mainland China) and so underestimates access (see <https://diff.wikimedia.org/2016/03/30/unique-devices-dataset/>). However, this estimate still provides some perspective on the number of individuals who might be accessing the Chinese language version of Wikipedia over time. For the Chinese language version of Wikipedia, 40.8 million devices accessed the site during December 2019 and 42.8 million per month during January, February, and March 2020, an increase of approximately 2 million devices. 3.34 million devices accessed the Chinese language Wikipedia per day in December 2019 and 3.66 million accessed the site per day during lockdown, an increase of approximately 300 thousand devices. These differences are somewhat smaller when comparing to the last half of 2019 (during ongoing protests in Hong Kong) – an increase of 1 million unique devices monthly during lockdown compared to July through December 2019, and an increase of 200 thousand devices daily.

A.5 Robustness Checks

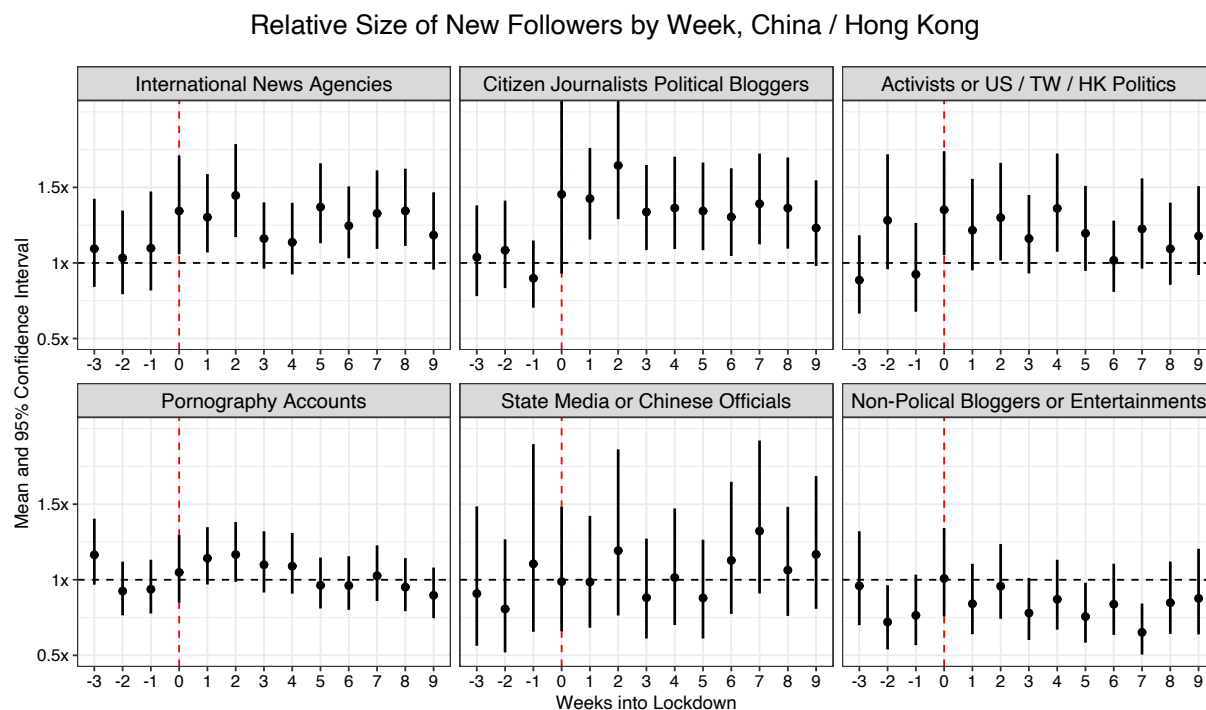


Figure A.6: Increases in Twitter Followers from mainland China versus Hong Kong by Week

Note: Incidence rate ratios shown above are from Negative Binomial regressions of number of daily new followers on the interaction between dummy for each week and China, with December 2019 as control period and Hong Kong as control group.

In this section, we assess whether the result is driven by (1) a misspecified treatment period, (2) the choice of comparison group, or (3) an increase of followers due to only a few accounts.

Figure A.6 plots the estimates based on regressions for each week before and after the lockdown. We do not see pre-treatment increases in number of followers in China, and the increase starts precisely on the week of lockdown.

Figures A.7 and A.8 verify that the results in Figure 5 are not due to choosing Hong Kong for the denominator. Figure A.7 uses accounts from Taiwan for the denominator, and Figure A.8 uses accounts in the United States. These accounts are from any user using Chinese and their self-

reported location is in Taiwan or the United States. Figure A.9 reports the regression estimate for the relative ratio of number of new followers (akin to a Difference-in-differences design with December 2019 as control period and Hong Kong/Taiwan/China as control group). The result is not driven by Hong Kong-specific trend of news cycles.

One might also be curious about whether new users stayed on Twitter at different rates. Figure A.10 plots the daily unique active users since their sign up dates in 2020. We don't find that users from one location stayed on Twitter longer than others.

New Followers Compared to Baseline, China / Taiwan

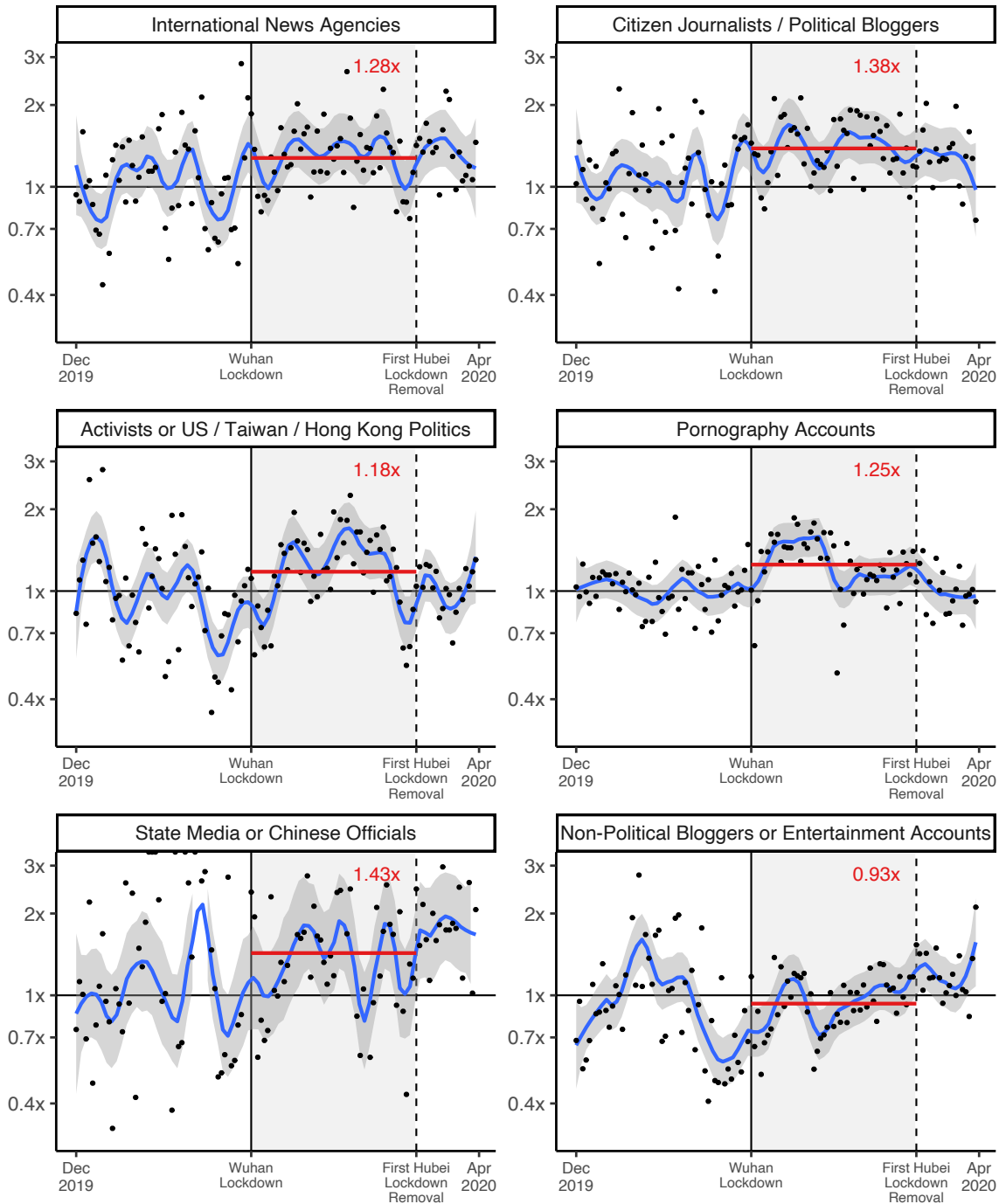


Figure A.7: Increases in Twitter Followers from China versus Taiwan

Note: Gain in followers from mainland China compared to Taiwan across six types of popular accounts, relative to December 2019 average. A value greater than 1 means more followers than expected from mainland China than from Taiwan. Accounts creating sensitive, censored information receive more followers than expected once the Wuhan lockdown starts. Fewer Taiwanese users follow Chinese state media or government officials than Hong Kong users do.

New Followers Compared to Baseline, China / US

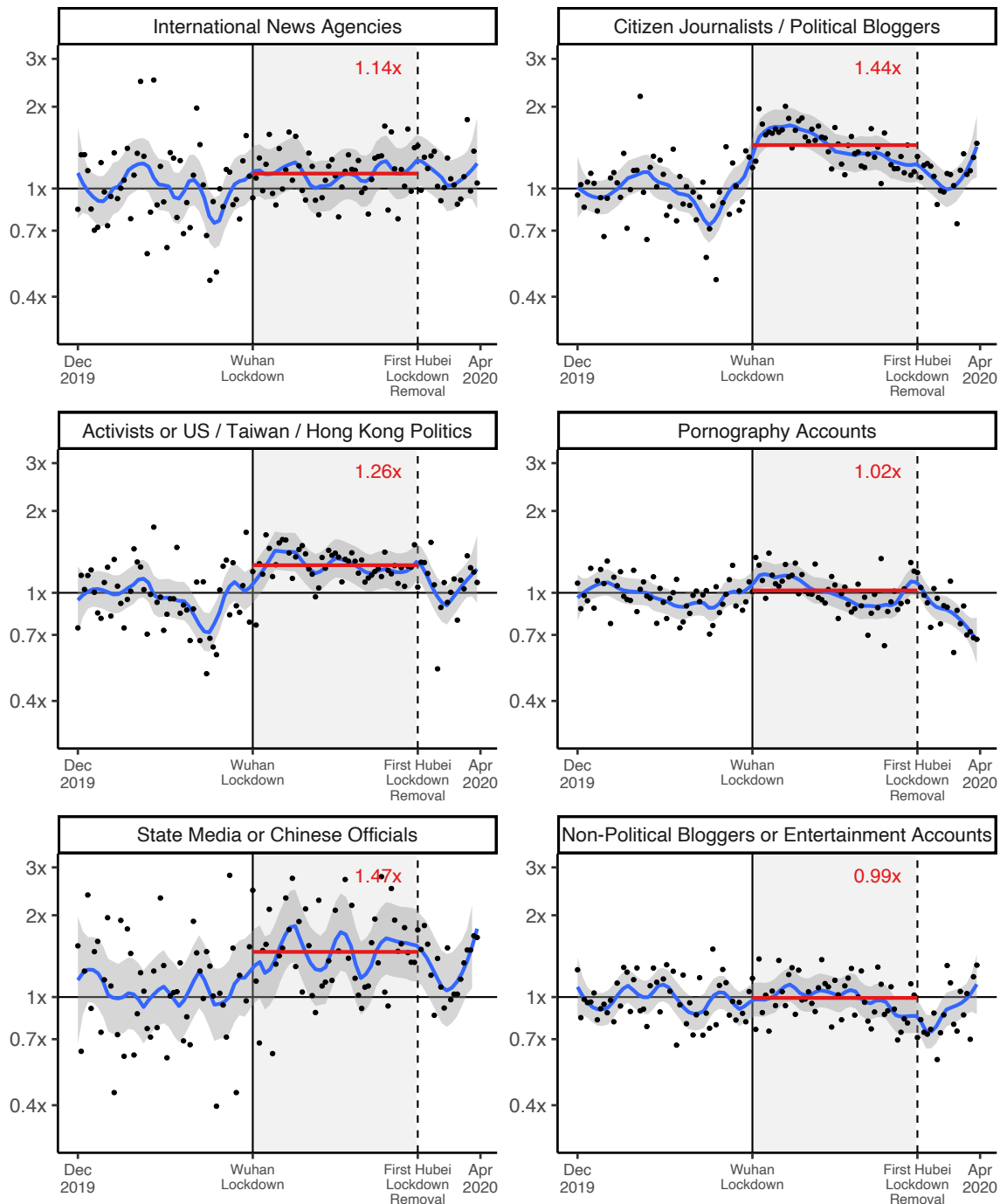


Figure A.8: Increases in Twitter Followers from China versus US

Note: Gain in followers from mainland China compared to US across six types of popular accounts, relative to December 2019 average. A value greater than 1 means more followers than expected from mainland China than from the US. Accounts creating sensitive, censored information receive more followers than expected once the Wuhan lockdown starts. Fewer US users follow Chinese state media or government officials than Hong Kong users do.

Relative Size of New Followers, China / Control Group

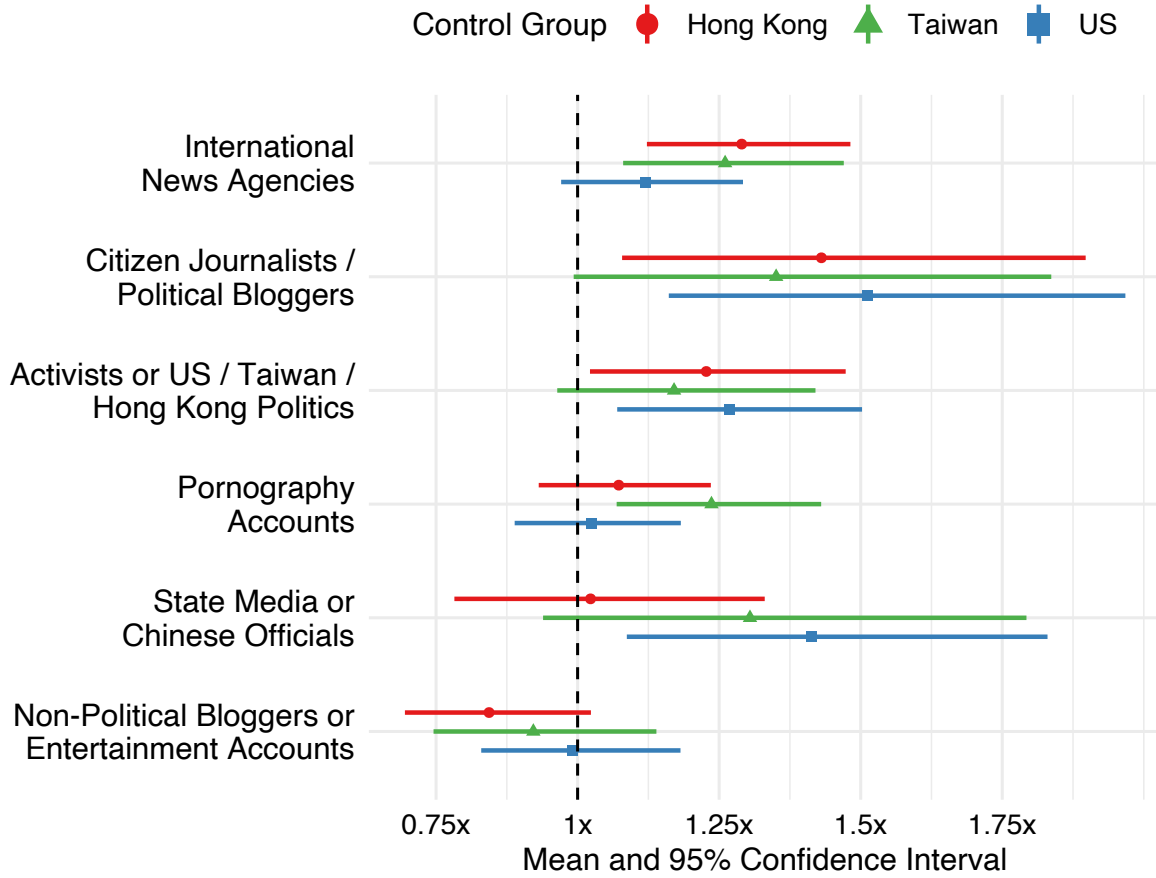


Figure A.9: Increases in Twitter Followers from China versus Others (Regression Estimate)

Note: Incidence rate ratios shown above are from negative binomial regressions of number of new followers on the interaction between indicator variables for ‘in lockdown period’ and ‘in mainland China’, with December 2019 as the control period.

Decay of Daily Unique User Activity

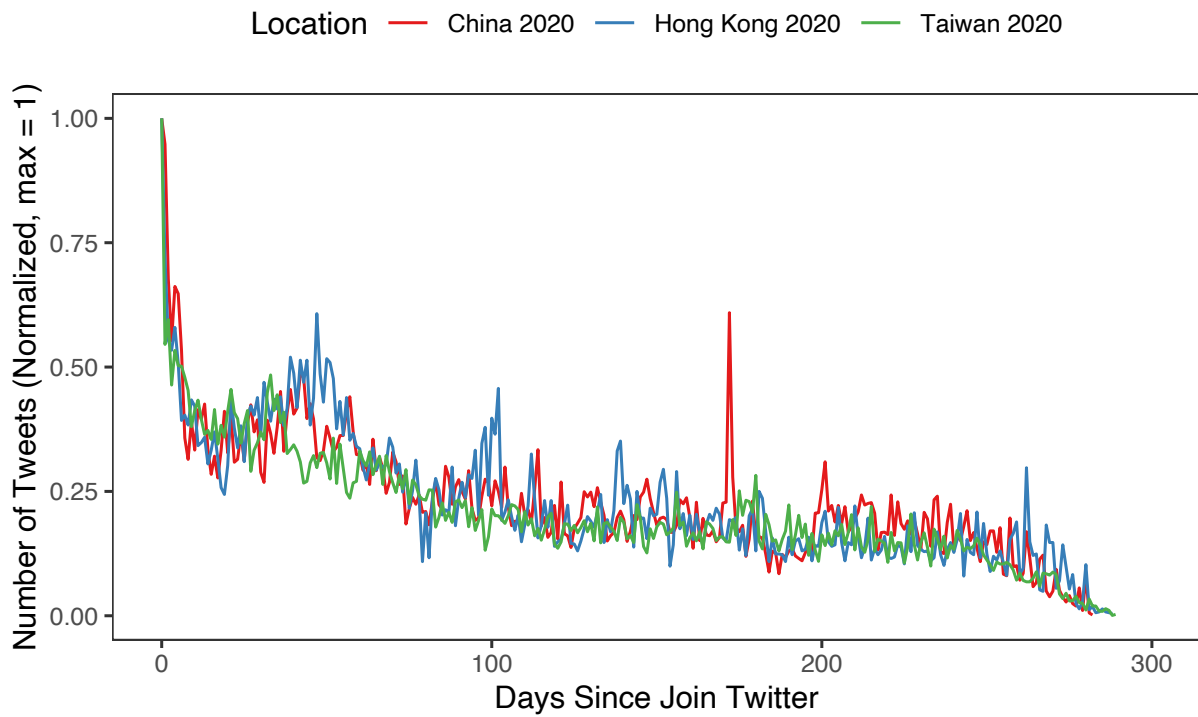


Figure A.10: New Users Stay on Twitter at the Same Rates across Locations

Note: This figure plots the daily unique active users since their sign up date using the user panel across locations. A user is considered active between their sign up date and the last day they tweet (before July 2020). We find that users stay on Twitter at the same rate across locations.

A.6 Wikipedia Country Comparisons

A.6.1 Page view analysis

Page view data analyzed in this paper is publicly available and hosted here: <https://dumps.wikimedia.org/other/pagecounts-ez/merged/>. In replication materials, we will additionally provide processed and aggregated versions of the page view data so that this paper's findings can be more quickly replicated than would be possible with the above page view files.

Below, we show the top Wikipedia pages by relative and absolute increases in page views within each of the categories we analyzed in the main text, as well as pages about the coronavirus and COVID-19 (pages considered: coronavirus, COVID-19, ventilator, flu, pneumonia, fever). The largest relative increases among these pages and for current leaders were related to coronavirus – the COVID-19 pandemic Wikipedia page and the head of China's National Health Commission. Top increases for pages that were blocked prior to the introduction of https on Wikipedia (after which China blocked all pages) were for an activist who criticized China's pandemic response.

In Figure A.11, we show the trajectories for categories matching those analyzed for China – current leaders (using offices listed in the CIA World Factbook), historical leaders, and, in Iran, pre-https blocked Wikipedia pages (Nazeri and Anderson, 2013).

Russia, Germany, and Italy (none of which block Wikipedia) saw increases in current leader views without accompanying increases in historical leader views. Germany and Italy did see spikes views of in historical leader pages in the weeks leading up to the relaxation of lockdowns in early May, but saw no change during the initial crisis.

German and Russian political pages also saw an increase in political leader page views prior to their own lockdown, and approximately at the same time as the announcement of widespread lockdown in Italy (see Figure A.11).

Table A.2: Top relative increases for Wikipedia pages January 24 through March 13 compared to December 2019.

Overall	Blocked	Current Leaders	Historical Leaders
马晓伟_(官员) (36.67) Ma Xiaowei	许志永 (16.78) Xu Zhiyong	马晓伟_(官员) (36.67) Ma Xiaowei	胡锦涛 (1.81) Hu Jintao
许志永 (16.78) Xu Zhiyong	2月17日 (9.01) February 17	孙春兰 (9.38) Sun Chunlan	邓小平 (1.75) Deng Xiaoping
孙春兰 (9.38) Sun Chunlan	西藏人民起义日 (7.04) Tibetan Uprising Day	李克强 (2.52) Li Keqiang	江泽民 (1.65) Jiang Zemin
2月17日 (9.01) February 17	台湾 (5.21) Taiwan	王岐山 (2.50) Wan Qishan	华国锋 (1.44) Hua Guofeng
西藏人民起义日 (7.04) Tibetan Uprising Day	圆周率日 (4.15) Pi Day	肖捷 (2.45) Xiao Jie	毛泽东 (1.15) Mao Zedong
肺炎 (5.38) Pneumonia	艾未未 (3.93) Ai Weiwei	韩正 (2.14) Han Zheng	
台湾 (5.21) Taiwan	李长春 (3.71) Li Changchun	胡春华 (1.99) Hu Chunhua	
流行性感冒 (5.04) Influenza	新唐人电视台 (3.51) New Tang Dynasty Television	苗圩 (1.88) Miao Wei	
圆周率日 (4.15) Pi Day	唐柏桥 (3.34) Tang Baiqiao	习近平 (1.80) Xi Jinping	
艾未未 (3.93) Ai Weiwei	长春围困战 (3.21) Siege of Changchun	杨晓渡 (1.73) Yang Xiaodu	

Note: This is where authors provide additional information about the data, including whatever notes are needed.

Table A.3: Top absolute daily increases for Wikipedia pages January 24 through March 13 compared to December 2019.

Overall	Blocked	Current Leaders	Historical Leaders
Rest of Wikipedia (1095913)	习近平 (4797) Xi Jinping	习近平 (4797) Xi Jinping	江泽民 (1197) Jiang Zemin
2019 冠状病毒病 (new page: 9236) Coronavirus disease 2019	王岐山 (2168) Wang Qishan	王岐山 (2168) Wang Qishan	邓小平 (1102) Deng Xiaoping
习近平 (4797) Xi Jinping	台湾 (2063) Taiwan	李克强 (1584) Li Keqiang	胡锦涛 (1079) Hu Jintao
肺炎 (4603) Pneumonia	六四事件 (1941) June 4 Incident (Tiananmen Square)	孙春兰 (1350) Sun Chunlan	毛泽东 (349) Mao Zedong
流行性感冒 (2463) Influenza	香港电台 (1689) Radio Television Hong Kong	韩正 (579) Han Zheng	华国锋 (255) Hua Guofeng
王岐山 (2168) Wang Qishan	中华人民共和国 (1631) People's Republic of China	胡春华 (541) Hu Chunhua	
台湾 (2063) Taiwan	李克强 (1584) Li Keqiang	马晓伟_(官员) (244) Ma Xiaowei	
六四事件 (1941) June 4 Incident (Tiananmen Square)	江泽民 (1197) Jiang Zemin	王毅 (119) Wang Yi	
香港电台 (1689) Radio Television Hong Kong	中华民国 (1128) Republic of China	傅政华 (99) Fu Zhenghua	
中华人民共和国 (1631) People's Republic of China	邓小平 (1102) Deng Xiaoping	肖捷 (63) Xiao Jie	

Note: Studying average daily increases standardizes the different lengths of time before versus after the Wuhan lockdown. Labels are limited to: blocked, leader, historical leader, COVID/coronavirus. All other pages are aggregated as “rest of Wikipedia”.

Table A.4: Lockdown dates

Country	Lockdown Start	Lockdown End	Historical Leaders
China	January 24, 2020 Hubei Lockdown	March 13, 2020	Paramount Leader
Iran	March 20, 2020 Nowruz - Tehran Easing	April 18, 2020	President, Supreme Leader
Russia	March 28, 2020 Non-Working Period	May 12, 2020	President General Secretary (Soviet Union) Chairman, Council of Ministers (1953)
Germany	March 22, 2020 National Social Distancing	May 6, 2020	Chancellor
Italy	March 9, 2020 National Quarantine	May 18, 2020	Prime Minister

Note: This table lists the time periods we use to estimate the effects of crisis lockdowns on Wikipedia page views, along with the offices considered for the historical leaders analysis. Each country's lockdown involved various levels of lockdown for different parts of the countries, and so there is no single time period for us to analyze. Figure A.11 displays Wikipedia page views with solid, vertical gray lines for the periods listed above.

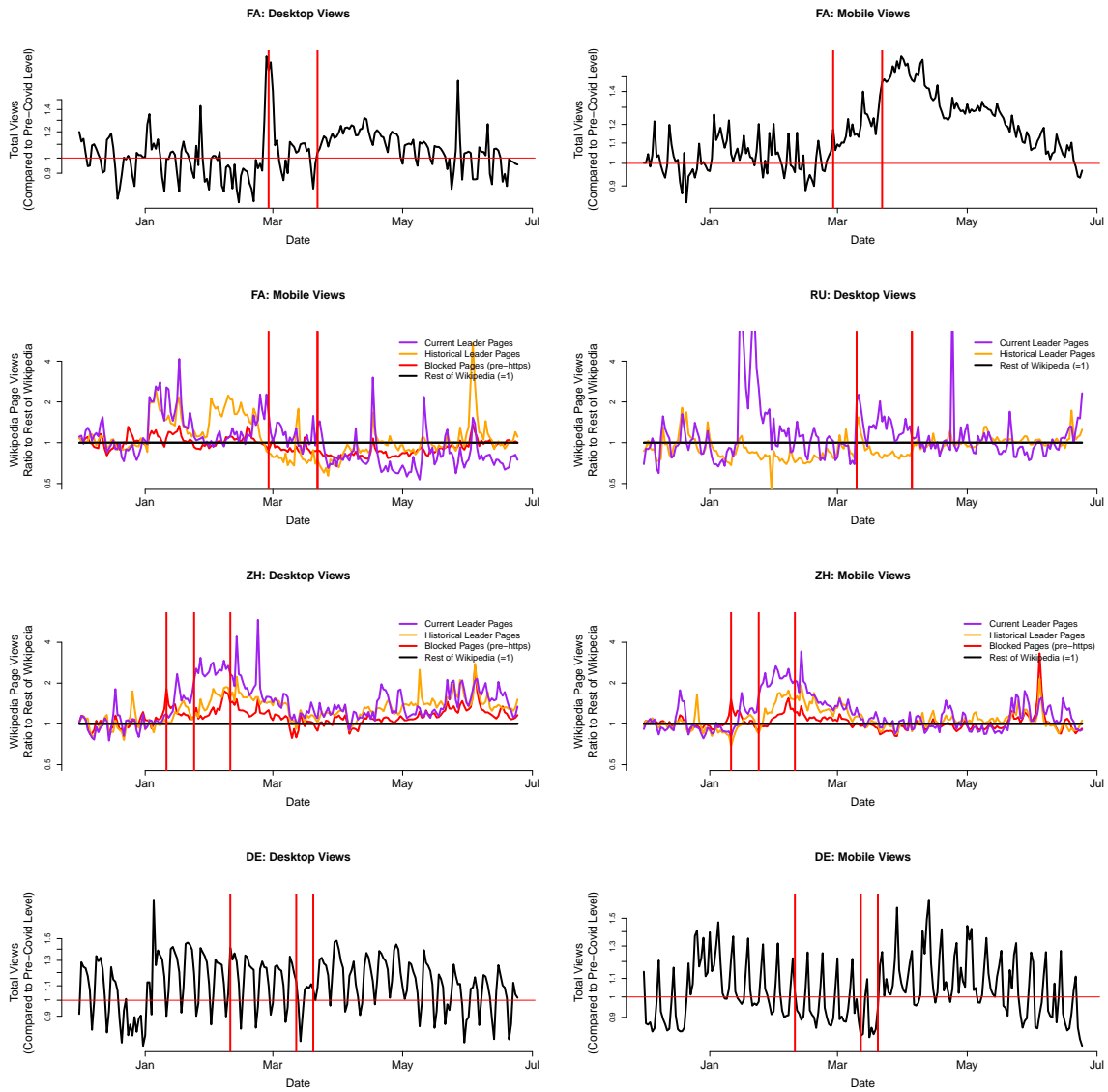


Figure A.11: Views of Blocked, Current Leader, and Historical Leader Wikipedia Pages in Other Countries

A.6.2 Analysis of an expanded set of historical political pages and ‘politically sensitive’ pages using Wikipedia2vec

We replicated our analyses of historical Wikipedia pages and “politically sensitive” (pages specifically blocked in China and Iran prior to the introduction of https) Wikipedia pages by expanding the original set of pages to a much larger set of related pages. We expanded these lists of pages using Wikipedia2vec (Yamada et al., 2020). This analysis assesses 1) whether the increase in views of Chinese historical leaders (and the lack of increase for other languages) was a relatively narrow effect or much broader one than what we see for that small set of pages and 2) whether a broader set of ‘politically sensitive’ pages are able to uncover increases in page views in Iran and Russia. Because, unlike China and Iran, Russia did not provide a list of politically sensitive pages (by blocking specific pages on Wikipedia), we assess Russian views of political opposition pages related to a) Alexei Navalny (arguably the most prominent opposition leader in Russia) and b) a list of opposition-related pages which we mine to discover increases in views – after this, we then looked to closely related pages to assess whether single page increases represented broader trends or were isolated and potentially random occurrences.

Wikipedia2vec finds similar pages (along with other entities and words) on Wikipedia by analyzing the network of page links, the co-occurrence of words, and the occurrences of specific words on pages. This analysis is accomplished using the same approach as in word2vec (Mikolov et al., 2013). At a high level, this approach involves placing words and entities into a shared n-dimensional space such that words and entities are placed close together if they frequently share contexts (e.g. page links or co-occurring words). Shared contexts must occur beyond what would be expected from the frequency of a word or page, which is accomplished through ‘negative sampling’ – predicting the co-occurrence of words and entities against frequency weighted sampling of negative cases. Once in the n-dimensional space, we can find the most similar entities (pages) for any given entity (or the mean of a set of entities’ pro-

jections) using cosine similarity – and we can incorporate dissimilar entities in this calculation by flipping the sign those entities’ locations when calculation the mean of a set of entities. Wikipedia2vec can be run with hyperparameters that affect the size of the n-dimensional space and the exact weighting scheme used in negative sampling. Our estimations for each language used the same default settings as the wikipedia2vec pre-trained embeddings provided at <https://wikipedia2vec.github.io/wikipedia2vec/pretrained/> with the number of dimensions set to 100.

For each set of pages (historical leaders, blocked pages, current leaders, Russian opposition pages), we found the top 100, 250, 500, and 1,000 pages that were most similar according to Wikipedia2vec. For historical leaders, we expanded to historical leader related pages not related to the current leader – using the current leader as a dissimilar case – and we expanded to current leader related pages not related historical leaders. With these sets, we re-estimated the changes in views during the first 30 days of lockdown. This excludes the late lockdown spikes in historical leader views visible for German and Italian (visible in Figure 7 in the main text and in Figure A.11 above). Note that the German increases in views of historical leaders (in those figures and in the results below) began well prior to the German lockdown (in February).

For Alexei Navalny specifically, we also manually collected a list of Wikipedia pages closely related to his opposition activities, and re-estimated changes in lockdown for each of these pages. The list of Russian opposition-related pages checked for increases is shown in Table A.5.

For the pages previously blocked in Iran, Nazeri and Anderson (2013) provided labels for the category of each blocked page: academic, artistic/cultural, drugs or alcohol, human rights, media and journalism, other, political, profane non-sexual, religious, and sex and sexuality. We also replicated our analyses for the Persian language set subset to page categories human rights, media and journalism, and political.

The findings from these analyses are displayed in Figures A.12, A.13, and A.14, and we

also show findings for current leaders in Figure A.15. In each cluster of estimates in the top panels, the first is the estimate for the seed pages (and is colored yellow for historical pages, red for blocked/‘politically sensitive’ pages, purple for current leaders). These exclude estimates for seeds which we mined for increases (i.e. selected them only because we saw increases during lockdowns – after a Bonferroni multiple testing correction). Given many tests when looking for increases, these pages have estimates that could very likely reflect random variation in page views, even though we are relatively certain that the increases were not zero, given the multiple testing correction.

Across the results, we see 1) that the increase in historical leader page views in Chinese also applies to a much larger set of pages and page views (bottom panel) and 2) we do not see comparable increases in historical leader pages or politically sensitive page views in other languages, despite increased interest in current leaders across almost all languages analyzed.

In the manual Alexei Navalny analysis, we see that views for his page specifically did rise and that this rise was comparable to what we see for historical leaders in Chinese. However, unlike the broad increase in views in China, we did not see similar increases for any other Navalny-related pages – and only one of the 9 considered showed a statistically significant increase without a multiple-testing correction (falling just short of significance at a 0.05 level after a Bonferroni correction for 9 tests).

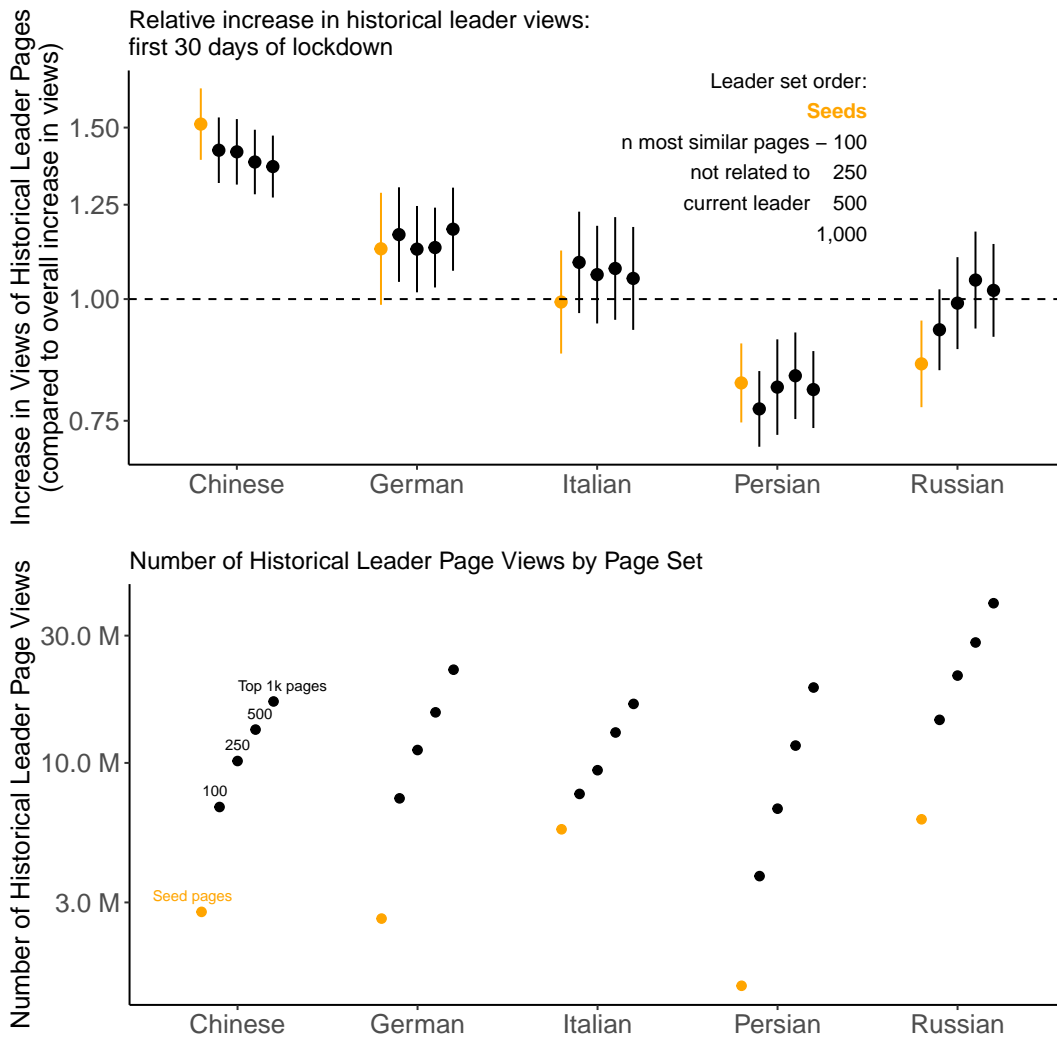


Figure A.12: Changes in views of historical leader Wikipedia pages (expanded set of pages)

Note: German increases in views of historical leaders began in February (see Figure A.11 above)

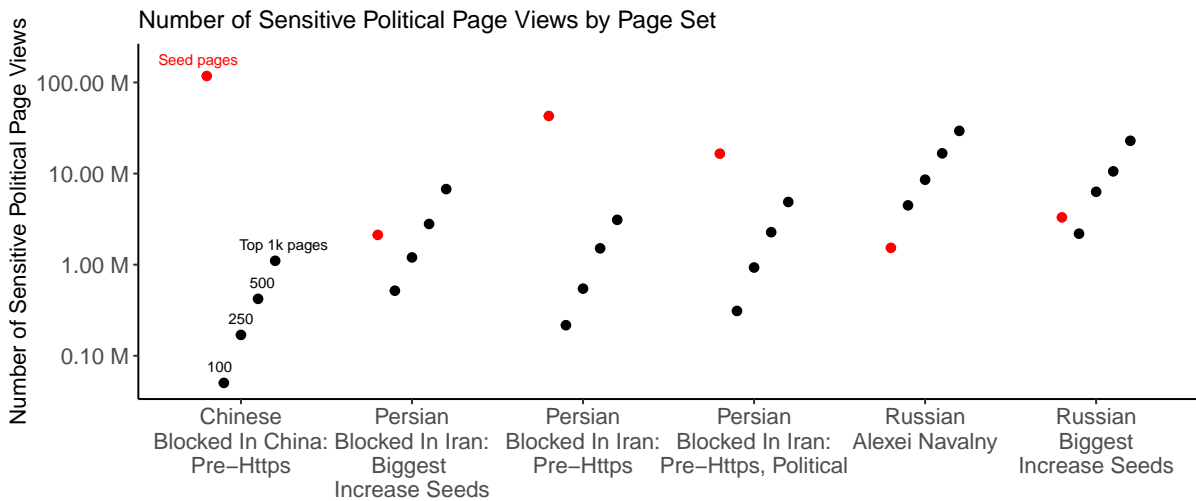
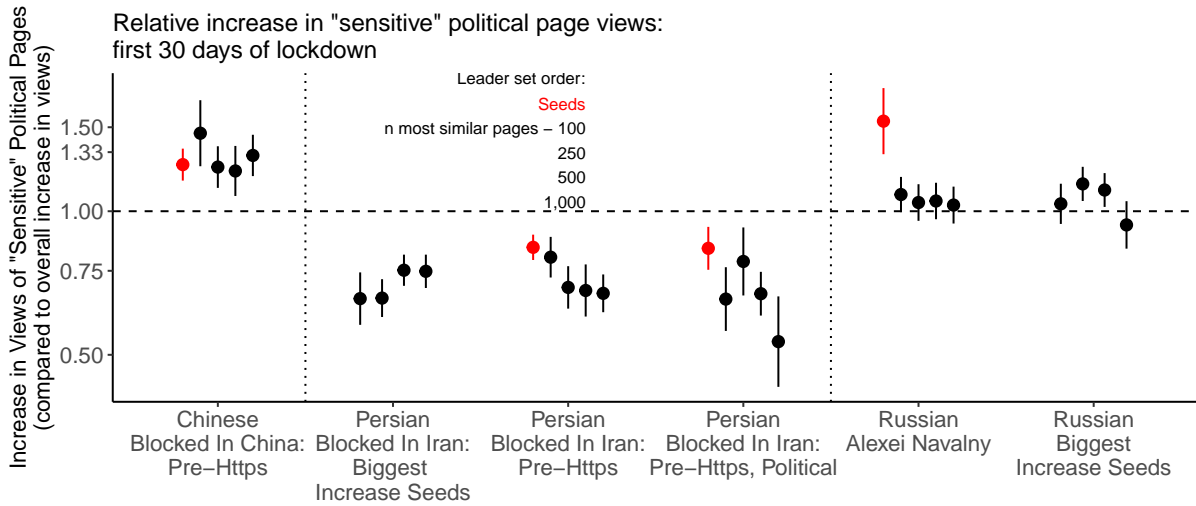


Figure A.13: Changes in views of 'politically sensitive' Wikipedia pages (expanded set of pages)

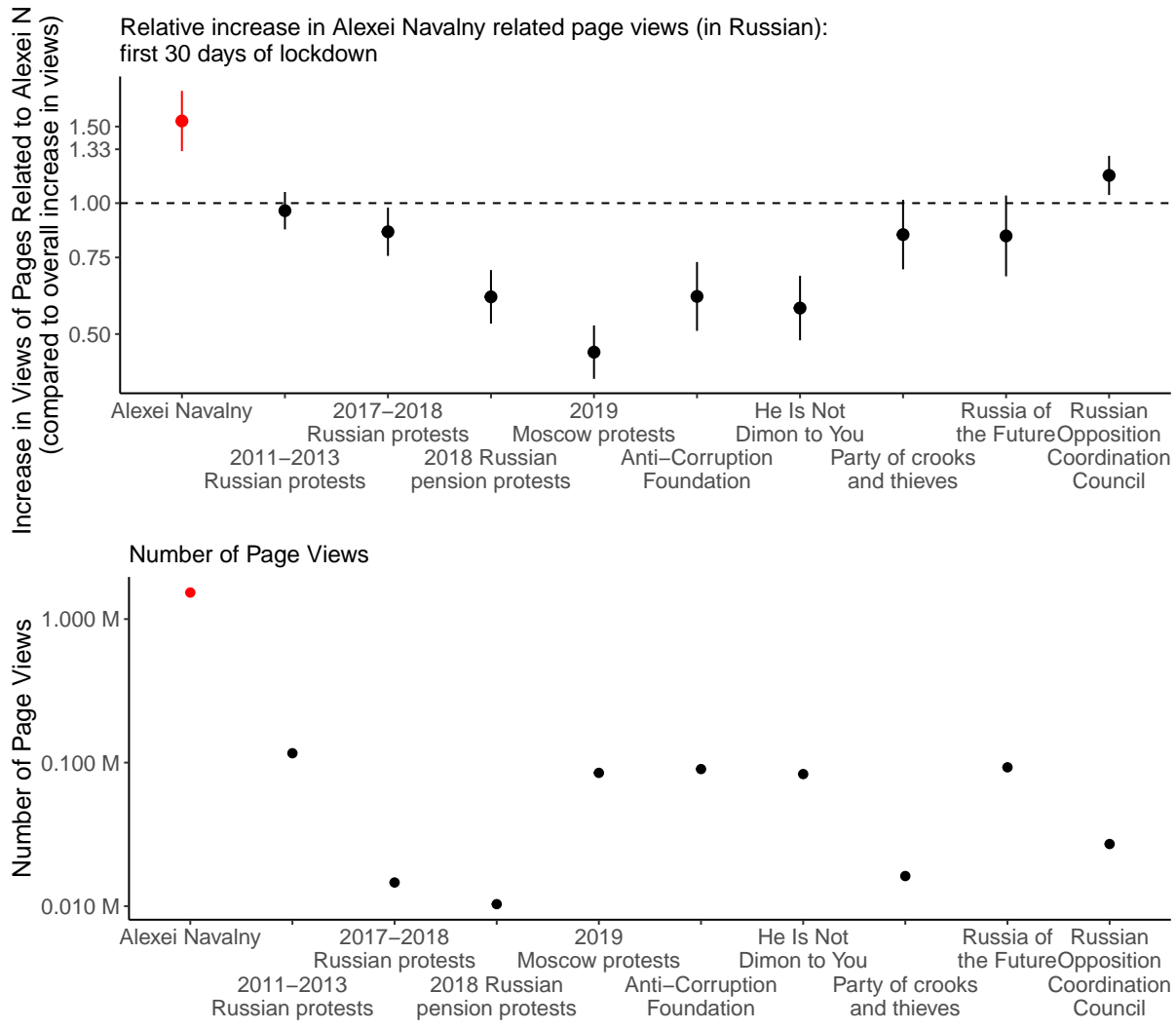


Figure A.14: Changes in views of Alexei Navalny related Wikipedia pages

Note: The Alexei Navalny-related pages in this figure are listed in alphabetical order.

Table A.5: List of opposition-related pages in Russian that were checked for significant increases during lockdown.

2011-2013 Russian protests	He Is Not Dimon to You
2014 anti-war protests in Russia	Human rights in Russia
2017-2018 Russian protests	List of journalists killed in Russia
2018 Russian pension protests	Media freedom in Russia
2019 Moscow protests	Mikhail Khodorkovsky
Alexander Litvinenko	Novaya Gazeta
Alexei Navalny	Open Russia
Anna Politkovskaya	Opposition to Vladimir Putin in Russia
Anti-Corruption Foundation	Party of crooks and thieves
Assassination of Anna Politkovskaya	Pussy Riot
Assassination of Boris Nemtsov	Russia of the Future
Boris Berezovsky (businessman)	Russian Opposition Coordination Council
Boris Nemtsov	Sergei Magnitsky
Corruption in Russia	Sergei Yushenkov

Note: Russia did not block specific Wikipedia pages prior to Wikipedia’s introduction of https. Because of this, we do not have a government-provided list of politically sensitive or objectionable content. As an alternative, we mine a manual list of government opposition-related pages, and then check whether for those increases were narrow and perhaps random (i.e. only occurred for those specific pages) or represented broad increases similar to those seen for historical and previously blocked pages in China. This table lists those Wikipedia pages (translated) that were checked for significant associations during the Russian lockdown period when compared to December 2019. Pages with statistically significant increases ($p < 0.05$) after a Bonferroni multiple testing were used as seeds when expanding with Wikipedia2vec. These “biggest increase seeds” are in bold above.

A.7 Text Analysis of Tweets

A.7.1 Hand labels

To better understand the range of political content that Chinese social media users may have encountered on Twitter, we hand labeled a random sample ($n = 500$) of tweets that 1) mentioned China or Chinese provinces and 2) mentioned any country ¹ between December 1st,

¹We identified country mentions using the Unicode Common Locale Data Repository: <https://www.unicode.org/Public/cldr/39/> (Chinese language country names in core/common/main)

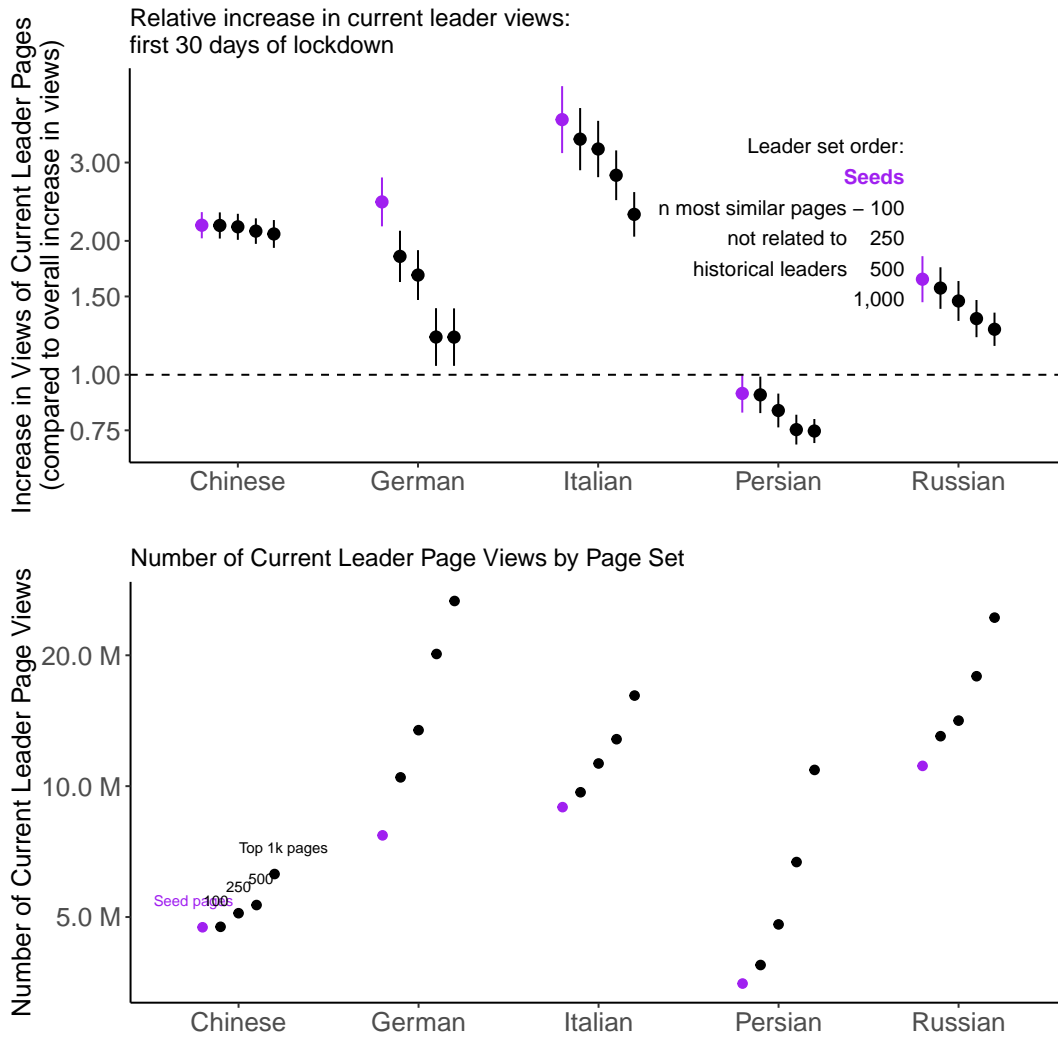


Figure A.15: Changes in views of current leader Wikipedia pages (expanded set of pages)

2019 and December 1st, 2020.

We evaluate the tone of each tweet (positive, negative, or neutral) and the main country being referred to in the tweet. These labels primarily evaluate whether to what extent Chinese social media users may have been exposed to negative content about China and negative content about other countries, including other countries' pandemic responses. However, we also label the main topic of the tweet (reading the text in Chinese to identify what appear to be the most frequent topics). They include the handling of COVID in various countries, Hong Kong protests, US elections, and other recent developments between US-China relations. We provide a comparison to these topics using the output of an automated topic model below.

Overall, we find that the account types that saw a disproportionate increase in followers at the start of the pandemic tended to cover both China and other countries in a negative way (Figure A.16). Coverage of both US and China are mostly negative in international news agencies, while in Chinese state media or Chinese officials (which did not see the same increase in followers) the coverage of China is strictly positive and the coverage of US is only negative. Furthermore, we also find that the tone on different topics was dissimilar between different types of popular accounts (Figure A.17). For international news agencies and activists/citizen journalists, tweets about the U.S. election, COVID in the U.S., Hong Kong protests, and U.S.-China disputes are all relatively negative. In tweets by Chinese state media or officials, the coverage about COVID in China and Chinese economic development are positive.

Notably, the United States was mentioned much more than other countries, suggesting that general international comparisons on COVID-19 pandemic responses might have been relatively rare compared to content about disputes with the United States and, perhaps, with the U.S. presidential administration.

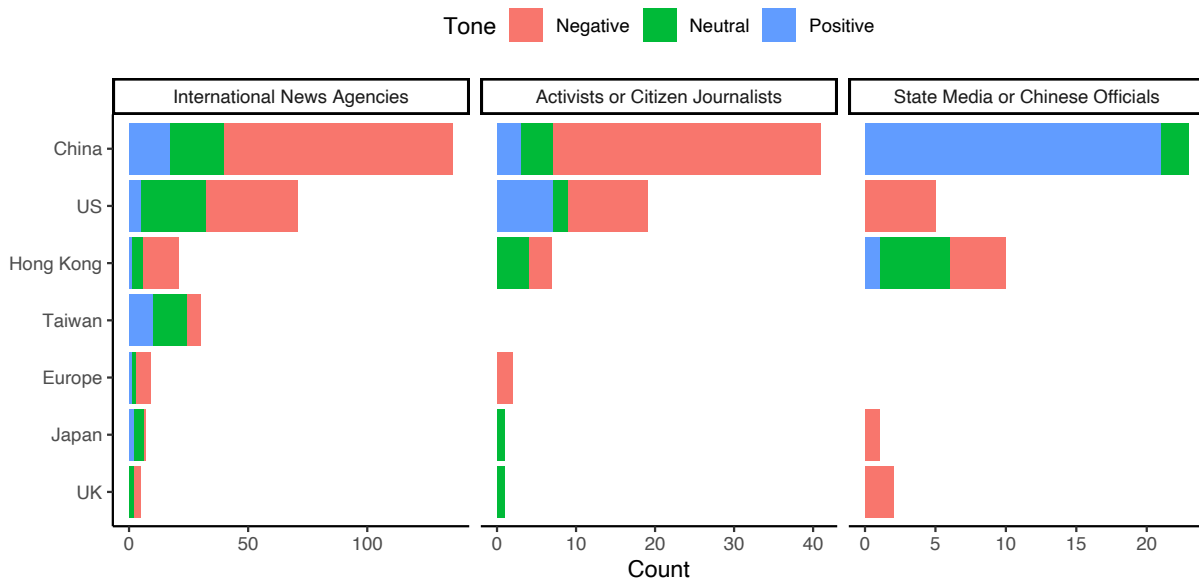


Figure A.16: Tone of tweets by popular Twitter accounts across main countries mentioned

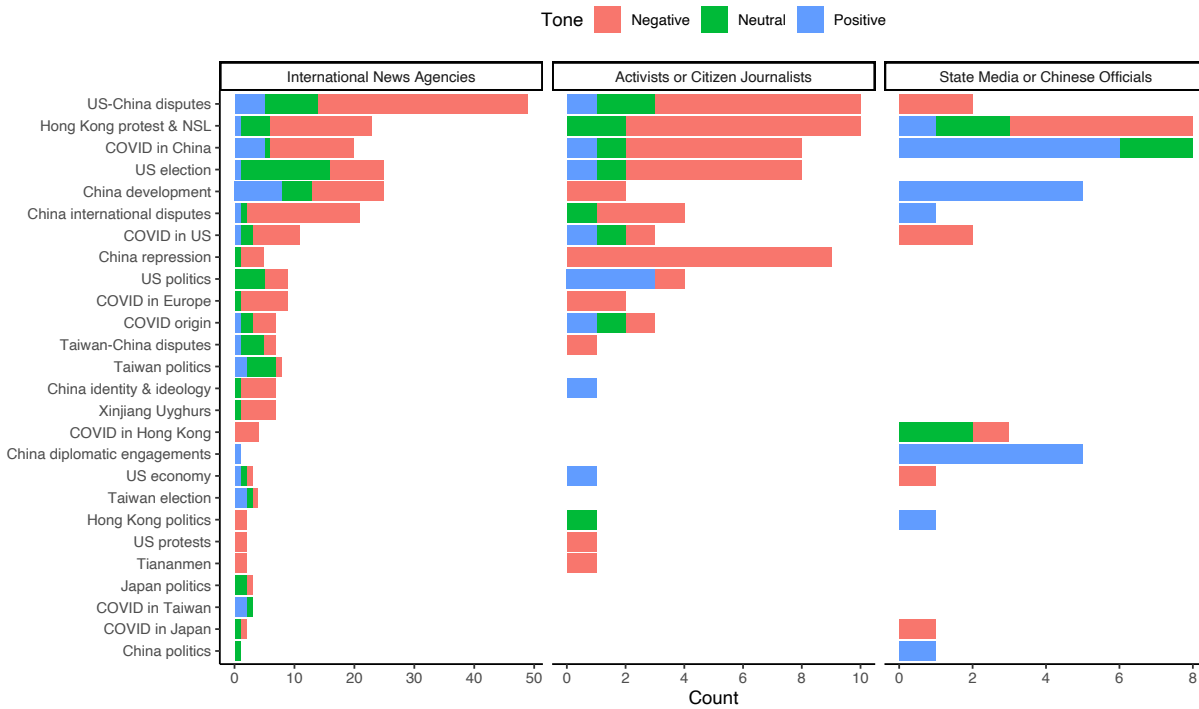


Figure A.17: Tone of tweets by popular Twitter accounts across topics of tweets

A.7.2 Topic models

To supplement the hand labels, we ran a topic model on tweets that mentioned China (Table A.6) or any country (Table A.7). In combination with the hand labels, this evaluates whether coverage of China was uniformly negative or more mixed, and also whether other countries and/or their disputes with China may have been covered in ways potentially favorable to the Chinese government and Communist Party. In international comparisons, for example, we might find favorable comparisons in other countries' handling of the COVID pandemic, coverage of international disputes thought to drive nationalist sentiment in China, or news on anti-Asian racism in the United States and elsewhere.

These topic models were run on 10 thousand tweet samples. Like the hand labels, these simple random samples were drawn from tweets that mentioned China or Chinese provinces (or any country for the second model) in Chinese from December 2019 through December 6, 2020². We also restricted the samples to content from the account categories “International News Agencies”, “Citizen Journalists / Political Bloggers”, and “Activists or US / Taiwan / Hong Kong Politics” – the categories which saw increases in views during the pandemic, and there were both political and not associated with Chinese state media or officials.

Each of topic models was estimated using the structural topic model R package (Roberts, Stewart and Tingley, 2016)³ with the number of topics set to 50, and the structural topic models were estimated without including covariates.

Several topics are related to politically sensitive issues in China (e.g. Tiananmen Square, the Hong Kong national security law, Xinjiang and human rights), but also many topics are related to international disputes with China (especially between China and the United States/President Trump) and the COVID-19 pandemic around the world. We hope this exploratory analysis will

²This slightly differs from the December 1 period above because the hand label and topic model analyses were conducted separately by different members of the research team. We decided that a closer alignment in the time frames would have no meaningful influence on these exploratory analyses.

³<https://www.structuraltopicmodel.com/>

help guide future work.

Table A.6: Topic model on tweets that mention China.

Topic label	Highest probability keywords	Proportion of corpus
US-China trade	美国, 中国, 美, 普, 朗, 奥, 中, 佩, 贸易, 总统 America, China, Trump, Austria, China, trade, President	0.07
Chinese economy	中国, 经济, 公司, 亿, 企业, 美元, 全球, 市场, 投资, 商 China, economy, company, billion, enterprise, USD, global, market, investment, business	0.05
COVID in China (1)	疫, 情, 病毒, 中国, 新冠, 武汉, 卫, 肺炎, 冠状, 爆发 epidemic, emotion, virus, China, COVID-19, Wuhan, health, pneumonia, coronavirus, outbreak	0.05
Chinese media	中国, 媒体, 记者, 报道, 推, 信, 新闻, 网络, 微, 媒 China, media, reporters, reports, tweets, letters, news, internet, micro, media	0.04
HK National Security Law	香港, 法, 时间, 国安, 港, 英国, 日, 版, 会议, 北京 Hong Kong, France, time, national security, Hong Kong, UK, Japan, edition, conference, Beijing	0.03
COVID in China (2)	病例, 诊, 确, 日, 肺炎, 感染, 死亡, 例, 新增, 湖北 case, diagnosis, confirmation, day, pneumonia, infection, death, case, new, Hubei	0.03
Tiananmen Square protests	自由, 政府, 中国, 政治, 批评, 四, 言论, 人士, 行动, 六 freedom, government, China, politics, criticism, four (June 4 incident), speech, people, action, six	0.03
Xi Jinping & Donald Trump	习, 近平, 川, 普, 新闻, 马, 陈, 今天, 李, 中国 Xi Jinping, Trump, news, Ma, Chen, today, Li, China	0.03
China's international disputes	中国, 印度, 军, 海, 日本, 南, 军事, 机, 中, 冲突 China, India, army, sea, Japan, South, military, aircraft, China, conflict	0.03
Human rights in China (1)	维, 权, 律师, 人士, 遭, 罪, 警方, 案, 公民, 当局 protection, rights, lawyers, persons, victims, police, cases, citizens, authorities	0.03

Continued on next page

Table A.6: Topic model on tweets that mention China. (Continued)

Topic label	Highest probability keywords	Proportion of corpus
Chinese Communist Party	中共, 人民, 政权, 世界, 中国, 反, 民众, 统治, 共, 内部 CCP, people, regime, world, China, anti, people, domination, communist, internal	0.03
COVID in China (3)	中國, 武漢, 中共, 美國, 習, 新聞, 奧, 對, 診, 確 China, Wuhan, CPC, United States, Xi, news, and, right, medical, confirmation	0.03
COVID in China (4)	武漢, 封, 肺炎, 城, 醫院, 隔離, 湖北, 人員, 疫, 感染 Wuhan, closure, pneumonia, city, hospital, quarantine, Hubei, personnel, epidemic, infection	0.03
Taiwan-China relationship (1)	党, 台湾, 共产, 民主, 中国, 台, 人民, 蔡, 英文, 代表 Party, Taiwan, Communist, democracy, China, Taiwan, people, Tsai Ing-wen, English, representative	0.03
Chinese people	中国人, 现在, 应该, 没有, 很多, 这种, 都是, 知道, 没, 世界 Chinese, now, should, no, a lot, this kind, all, know, no, world	0.03
Human rights in China (2)	国, 中国, 人权, 国际, 联合, 组织, 声明, 国家, 中华, 世界 country, China, human rights, international, union, organization, statement, country, China, world	0.02
Chinese diplomacy	中国, 希望, 德国, 欧洲, 欧盟, 澳洲, 毅, 王, 国家, 一个 China, hope, Germany, Europe, EU, Australia, Wang Yi, country, one	0.02
China's medical aid	中国, 口罩, 澳大利亚, 文化, 国家, 一周, 日本, 援助, 医疗, 法国 China, mask, Australia, culture, country, one week, Japan, aid, medical, France	0.02
US-China disputes (1)	中国, 美国, 局, 签证, 调查, 州, 政府, 间谍, 情报, 名 China, United States, bureau, visa, investigation, state, government, spy, intelligence, name	0.02
Chinese society	中国人, 一个, 吃, 思想, 没有, 里, 国家, 中国, 社会, 你的 Chinese, one, eat, thought, no, inside, country, China, society, yours	0.02

Continued on next page

Table A.6: Topic model on tweets that mention China. (Continued)

Topic label	Highest probability keywords	Proportion of corpus
Xinjiang interment/reeducation' camps	新疆, 维吾尔, 人权, 营, 民族, 族, 教育, 穆斯林, 集中, 宗教 Xinjiang, Uyghur, human rights, camp, ethnic, education, Muslim, concentration, religion 大学, 学生, 教授, 学院, 中国, 留学生, 蒙古, 教育, 学, 万	0.02
Chinese education	university, student, professor, college, China, international student, Mongolia, education, learning, million 中国, 问题, 视, 国家, 认为, 更, 发展, 春, 研究, 库	0.02
China's economic development	China, issues, view, country, think, more, development, spring, research, library 上海, 张, 店, 一个, 冯, 家, 朋友, 虎, 买, 民	0.02
Shanghai	Shanghai, Zhang, shop, one, Feng, home, friends, tiger, buy, people 长, 副, 委, 强, 中央, 书记, 马, 组, 王, 志	0.02
Chinese politics	Chief, Deputy, committee, strong, Central, Secretary, Ma, group, Wang, Zhi 网, 民, 主义, 联, 中国, 互, 爱国, 爱, 中, 一个	0.02
China's online patriotism	net, people, doctrine, union, China, mutual, patriotic, love, China, one 王, 警察, 全, 北京, 璋, 车, 派出所, 文, 黎, 李	0.01
Wang Quanzhang	Wang Quanzhang, police, full, Beijing, car, police station, Li, Li 北京, 离开, 台北, 台湾, 中共, 五, 破, 高层, 日, 空	0.01
Taiwan-China relationship (2)	Beijing, departure, Taipei, Taiwan, CCP, broken, high-rise, day, empty 文, 李, 医生, 亮, 发, 中国, 福建, 吹哨, 医, 找	0.01
Li Wenliang	Li Wenliang, doctor, bright, hair, China, Fujian, whistling, medical, find 山东, 杨, 家, 喜, 丁, 母, 监狱, 号, 戴, 省	0.01
Shandong	Shandong, Yang, home, Ding, mother, prison, number, Dai, province	0.01

Continued on next page

Table A.6: Topic model on tweets that mention China. (Continued)

Topic label	Highest probability keywords	Proportion of corpus
China stability	中国, 进入, 不断, 稳, 更加, 维, 度, 期, 月, 更 China, enter, continue, stabilize, more, dimension, degree, period, month, more	0.01
Hubei	黄, 一个, 湖北, 中, 琦, 锋, 写, 先生, 儿子, 两 Huang, one, Hubei, middle, Qi, Feng, write, Mr, son, two 中国, 非洲, 猪, 粮食, 危机, 到底, 国家, 猪肉, 大豆, 粮 China, Africa, pig, food, crisis, end, country, pork, soy, grain	0.01
Africa-China relationship	疫苗, 中国, 实验, 试验, 巴西, 新冠, 测试, 使用, 室, 今日 vaccine, China, experiment, trial, Brazil, COVID-19, test, use, laboratory, today	0.01
Vaccines and China	广东, 太, 带, 场, 云, 私人, 中, 钱, 市, 镇 Guangdong, Tai, with, field, cloud, private, middle, money, city, town	0.01
Guangdong	洪水, 洪, 湖北, 江西, 灾, 村, 暴雨, 三峡, 安徽, 南方 flood, flood, Hubei, Jiangxi, disaster, village, heavy rain, Three Gorges, Anhui, south	0.01
China flood	重庆, 安, 市, 庆, 工程, 街头, 一个, 两, 桥, 万 Chongqing, An, city, Qing, project, street, one, two, bridge, million	0.01
Chongqing	许, 志, 实, 章, 失, 润, 强, 秋, 陈, 男 Xu Zhangrun, Zhi, actual, chapter, lost, run, strong, autumn, Chen Qiushi, male	0.01
Xu Zhangrun & Chen Qiushi	出, 朝鲜, 恩, 传, 中国, 中, 金正, 瑞, 师, 金 out, North Korea, En, biography, China, China, Kim Jong-un, Rui, Division, Kim	0.01
North Korea	四川, 成都, 地震, 教会, 家庭, 恰, 牧师, 网, 线, 日 Sichuan, Chengdu, earthquake, church, family, pastor, net, line, day	0.01
Sichuan earthquake		

Continued on next page

Table A.6: Topic model on tweets that mention China. (Continued)

Topic label	Highest probability keywords	Proportion of corpus
China legal system	获, 高, 法官, 律师, 法院, 审, 权, 智, 企业, 晟 gain, high, judge, lawyer, court, trial, right, wisdom, enterprise, Sheng 西藏, 宗教, 奖, 喇嘛, 藏人, 藏, 达赖, 波, 拉, 流亡 Tibet, religion, award, Lama, Tibetan, Tibet, Dalai Lama, Poland, exile	0.01
Tibet & the Dalai Lama	周, 郭, 江, 河南, 脸, 宝, 勇, 洋, 上街, 克 Zhou, Guo, Jiang, Henan, face, Bao, Yong, Yang, Going to the street 馆, 领, 关闭, 领事, 驻, 中的, 休, 中, 图书, 洛杉矶 library, collar, closing, consul, consulate, Chinese, close, Chinese, books, Los Angeles 复, 工, 非常, 功, 法轮, 迫害, 学员, 分, 吴, 中共 recovery, Falun Gong, extraordinary, persecution, practitioners, points, Wu, CCP 中国, 没有, 亚太, 已经, 视频, 政府, 接受, 大陆, 海外, 一个 China, no, Asia Pacific, already, video, government, accepted, Mainland, Overseas, one 成, 天津, 种, 盼, 历史, 肉, 两, 权, 元, 红 Cheng, Tianjin, kind, hope, history, meat, two, Quan, Yuan, red 引起, 中国, 患者, 续, 关注, 中, 梵蒂冈, 隔离, 协议, 治疗 arouse, China, patient, continued, attention, China, Vatican, quarantine, agreement, treatment 路, 广州, 捷克, 罕见, 深圳, 返回, 移动, 大会, 日, 铁 road, Guangzhou, Czech Republic, rare, Shenzhen, return, mobile, convention, Japan, iron 抵制, 北京, 中共, 中, 人權, 美, 日, 奧, 洁, 佩 boycott, Beijing, CCP, China, human rights, U.S., Japan, Austria, Yang Jiechi, Mike Pompeo	0.01
US-China disputes (2)		0.01
Henan		0.01
Falun Gong		0.01
Overseas Chinese		0.01
Tianjin		0.01
the Vatican		0.01
China's railroads		0.01
US-China disputes (3)		0.01

Table A.7: Topic model on tweets that mention any country.

Topic label	Highest probability keywords	Proportion of corpus
China's diplomacy	中国, 政府, 问题, 北京, 政策, 认为, 关系, 官员, 称, 全球 China, government, issues, Beijing, policy, think, relationship, officials, scale, global	0.05
Hong Kong National Security Law	香港, 法, 国安, 中, 版, 民主, 送, 派, 港人 Hong Kong, law, Hong Kong, national security, China, version, democracy, send, send, Hong Kong people	0.05
COVID	疫, 病毒, 情, 武汉, 新冠, 肺炎, 冠状, 卫生, 卫, 世界 epidemic, virus, emotion, Wuhan, novel coronavirus, pneumonia, coronavirus, health, health, world	0.04
China security	公司, 华, 协议, 机构, 安全, 报告, 政府, 中, 提供, 部 company, China, agreement, agency, security, report, government, China, provision, ministry	0.04
US & Taiwan & China	中國, 中共, 美國, 世界, 對, 武漢, 台灣, 與, 將, 國家 China, Chinese Communist Party, the United States, the world, right, Wuhan, Taiwan, and, will, countries	0.03
Chinese economy	经济, 亿, 美元, 万, 企业, 市场, 投资, 银行, 金融, 公司 economy, billion, USD, million, enterprise, market, investment, bank, finance, company	0.03
China human rights (1)	国际, 国, 人权, 联合, 组织, 声明, 欧盟, 日, 调查, 呼吁 international, national, human rights, union, organization, statement, EU, Japan, investigation, appeal	0.03
Mike Pompeo	中共, 美, 奥, 佩, 蓬, 馆, 制裁, 中, 国务卿, 中美 CPC, United States, Mike Pompeo, Pei, Peng, pavilion, sanctions, China, Secretary of State, Sino-US	0.03
Communism	社会, 主义, 世界, 反, 西方, 化, 种族, 历史, 共产, 一个 society, doctrine, world, anti, Western, transformation, race, history, communism, one	0.03
US election	党, 民主, 选举, 大选, 议员, 共产, 州, 投票, 共和, 两 Party, democracy, election, general election, congressman, communist, state, vote, republic, two	0.03

Continued on next page

Table A.7: Topic model on tweets that mention any country. (Continued)

Topic label	Highest probability keywords	Proportion of corpus
Taiwan election	台湾, 台, 英文, 蔡, 军, 美, 大陆, 总统, 日, 中 Taiwan, Taiwan, Tsai Ing-wen, military, United States, Mainland China, president, Japanese, Chinese	0.03
COVID	日, 诊, 确, 病例, 死亡, 感染, 例, 人数, 新增, 新冠 day, diagnosis, confirmed, case, death, infection, case, number, new, novel coronavirus	0.03
Weibo	媒体, 推, 视频, 文, 发, 信, 微, 信息, 纽约, 博 media, Twitter, video, text, post, letter, Weibo, information, New York, blog	0.02
Chinese immigrants	现在, 一个, 没有, 都是, 很多, 生活, 里, 华人, 移民, 想 now, one, none, all, many, life, Chinese, immigrants, thinking	0.02
Trump	普, 川, 总统, 朗, 白宫, 支持, 表示, 顾问, 日, 竞选 Trump, president, Lang, White House, support, show, advisor, Japan, election	0.02
China human rights (2)	律师, 权, 维, 中国, 当局, 公民, 刘, 许, 人士, 王 lawyers, rights, maintenance, China, authorities, citizens, Liu, Xu, personalities, Wang	0.02
Canada-China trade	宣布, 限制, 加拿大, 贸易, 产品, 制造, 出口, 禁止, 关税, 晚 announce, restrict, Canada, trade, products, manufacturing, export, prohibition, tariff, late	0.02
Xi Jinping	习, 近平, 陈, 时间, 黄, 新闻, 北京, 今天, 中国, 节目 Xi Jinping, Chen, time, Huang, news, Beijing, today, China, programs	0.02
Russia	斯, 俄罗斯, 德, 马, 尼, 利, 罗, 纳, 克, 部长 Slovakia, Russia, Germany, Malaysia, Nigeria, Li, Romania, Croatia, Minister	0.02
US human rights	美国, 已经, 点, 线, 接受, 公民, 人的, 清楚, 不能, 迫使 America, already, point, line, accept, citizen, human, clear, cannot, forced	0.02

Continued on next page

Table A.7: Topic model on tweets that mention any country. (Continued)

Topic label	Highest probability keywords	Proportion of corpus
COVID in France	法国, 封, 隔离, 航, 医院, 班, 人员, 域, 名, 两 France, seal, quarantine, aviation, hospital, class, personnel, city, name, two	0.02
COVID in UK	英国, 疫苗, 英, 新冠, 德, 约翰, 逊, 政府, 瑞, 药物 UK, vaccine, UK, COVID-19, Germany, Boris Johnson, son, government, Switzerland, drugs	0.02
Chinese news reports	记者, 新闻, 月, 媒, 报道, 传, 林, 名, 报, 官 reporter, news, month, media, report, biography, Lin, name, newspaper, official	0.02
Tiananmen Square protests	自由, 中国人, 四, 不会, 六, 大学, 纪念, 学生, 言论, 周年 freedom, Chinese, four (June 4 incident), no, six, university, commemoration, student, speech, anniversary	0.02
Chinese factories, return to work	出现, 复, 工, 州, 中国, 状态, 症状, 城, 危机, 区 appearance, recovery, work, state, China, state, symptom, city, crisis, district	0.02
China's diplomacy	国家, 一个, 非洲, 利益, 亚, 中, 澳洲, 女性, 发展, 已经 country, one, Africa, interest, Asia, China, Australia, female, development, already	0.02
The Epoch Times	人民, 全世界, 看到, 爆, 支持, 纪元, 平台, 完整, 訂閱, 正義 people, worldwide, see, burst, support, The Epoch Times, platform, complete, subscription, justice	0.02
Japan & China	研究, 日本, 事, 科学, 日, 中, 大学, 恩, 发生, 蒙古族 research, Japan, events, science, Japan, China, university, En, occurrence, Mongolian	0.02
Black Lives Matter	没, 做, 全, 老, 黑, 命, 中国, 真, 文革, 骂 no, do, full, old, black, life, China, true, Cultural Revolution, curse	0.02
US election	拜, 总统, 大选, 候选, 福, 辩论, 副总统, 当选, 团队, 提名 Biden, president, general election, candidate, fortune, debate, Vice President, elected, team, nomination	0.02

Continued on next page

Table A.7: Topic model on tweets that mention any country. (Continued)

Topic label	Highest probability keywords	Proportion of corpus
Iran & Soleimani	伊朗, 莱, 核, 导弹, 至少, 美军, 袭击, 苏, 曼, 尼 Iran, Lebanon, nuclear, missile, at least, U.S. military, attack, Soviet, Soleimani, Nepal	0.02
China academia	谈, 中国, 实, 大学, 教授, 学者, 文件, 时事, 出版, 之音 talk, China, reality, university, professor, scholar, document, current affairs, publication, voice	0.01
Chinese history	世界上, 你的, 鸡, 一个, 没有, 两, 历史, 中, 清, 地方 in the world, your, chicken, one, none, two, history, middle, Qing, place	0.01
China's internet	网, 联, 民, 中国, 视, 功, 红, 失, 一定, 互 net, united, people, China, vision, power, red, loss, certain, mutual	0.01
Abe & Japan	日本, 一周, 热门, 东京, 安倍, 回顾, 旅游, 抵制, 亚洲, 热点 Japan, week, popular, Tokyo, Abe, review, travel, boycott, Asia, hot	0.01
Wolf warrior diplomacy	欧洲, 留学生, 战, 女, 狼, 新西兰, 国, 总理, 新, 摆 Europe, student, war, female, wolf, New Zealand, country, prime minister, new, pendulum	0.01
Merkel & Germany	德国, 默, 瑞典, 克, 之声, 政府, 柏林, 中, 书, 出 Germany, Merkel, Sweden, gram, voice, government, Berlin, Chinese, book, out	0.01
India-China disputes	印度, 冲突, 中印, 蒙古, 印, 边境, 边界, 士兵, 军, 量 India, conflict, China-India, Mongolia, India, border, border, soldier, army, quantity	0.01
US-China disputes (1)	驻, 外交部, 发言, 中方, 大使, 捷克, 室, 王, 使馆, 美方 China, Ministry of Foreign Affairs, speech, Chinese, Ambassador, Czech, office, King, Embassy, US	0.01
Singapore	知道, 新加坡, 标准, 一个, 不同, 检测, 需要, 承担, 后果, 次 know, Singapore, standard, one, different, test, need, bear, consequence, time	0.01

Continued on next page

Table A.7: Topic model on tweets that mention any country. (Continued)

Topic label	Highest probability keywords	Proportion of corpus
Wang Liqiang	韩国, 朝鲜, 瑜, 韩, 都在, 一个, 王立, 强, 百分, 心 Korea, North Korea, Yu, Korea, all in, one, Wang Liqiang, percent, heart	0.01
Ant Group	钱, 中, 上市, 蚂蚁, 中国, 诈骗, 局, 一名, 大学, 调查 money, China, listing, Ant, China, fraud, bureau, one, university, investigation	0.01
Masks	口罩, 防疫, 戴, 建议, 瑞士, 越, 措施, 指南, 防护, 民众 masks, epidemic prevention, wear, recommendations, Switzerland, Vietnam, measures, guidelines, protection, people	0.01
Hong Kong protests	香港, 革命, 感谢, 免费, 中, 抗议, 艺术, 百, 时代, 支持 Hong Kong, revolution, thanks, free, China, protest, art, hundred, times, support	0.01
Apple Daily Taiwan	蘋果, 新聞, 網, 台灣, 乘客, 冠狀, 輪, 回, 郵, 公主 Apple, news, web, Taiwan, passenger, coronavirus, round, back, post, princess	0.01
Poland	控, 監, 动物, 恶, 杀, 波兰, 语, 野生, 盛, 监狱 control, prison, animal, evil, kill, Polish, language, wild, Sheng, prison	0.01
Hu Xijin	重, 启, 梦, 米, 锡, 一直, 中国, 没有, 酒, 世界 Hu Xijin, dream, rice, always, China, no, wine, world	0.01
US-China trade	重, 启, 马, 审, 头条, 主, 判, 聚, 六度, 中美 Re, horse, trial, headline, main, judgment, convergence, six degrees, Sino-US	0.01
Tibet	保护, 宗教, 爱国, 西藏, 信仰, 自由, 文化, 尊重, 续, 意识 protection, religion, patriotic, Tibet, faith, freedom, culture, respect, continued, consciousness	0.01
Mexico	级, 毒, 墨西哥, 解决, 墨, 菲, 注意, 问题, 广播, 锅 grade, poison, Mexico, solve, Mexico, Philippines, attention, problem, broadcast, pot	0.01

“disrupt critical communications” between the US and the Asia Pacific region in the event of a future US-China crisis, Microsoft warned on Wednesday.

The Chinese hackers have been active since mid-2021 and targeted critical infrastructure organizations in the US territory of Guam and in other parts of the US as part of a stealthy spying and information gathering campaign, Microsoft said in a new report. In a separate advisory released Wednesday, Western security agencies said they believe the Chinese hackers could apply the same stealthy techniques against critical sectors “worldwide.”

B.1.3 (T2) Biden Administration Delayed Sanctions over Spy Balloon to ‘Limit Damage’ to China Ties

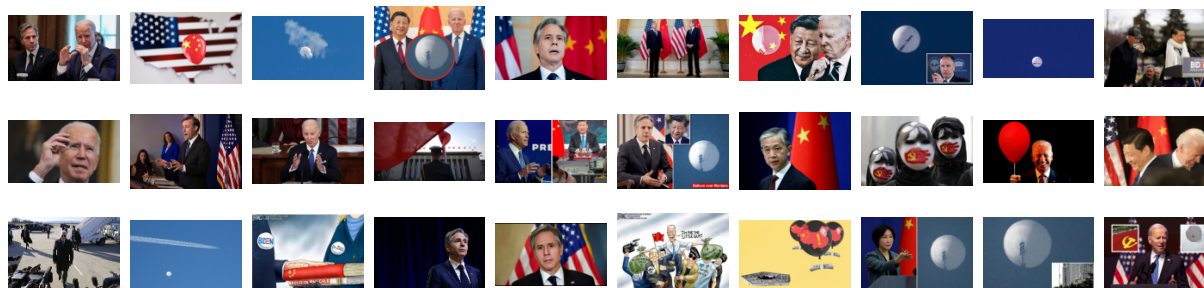


Figure B.2: Treatment Images for Treatment T2.

When an alleged Chinese spy balloon traversed the United States in February, some U.S. officials were confident the incursion would galvanize the U.S. bureaucracy to push forward a slate of actions to counter China. Instead, the U.S. State Department held back human rights-related sanctions, export controls and other sensitive actions to try to limit damage to the U.S.-China relationship, according to four sources with direct knowledge of U.S. policy.

Former diplomats and members of Congress from both parties have argued that the U.S. must keep channels of communication open with Beijing to avoid misunderstandings and navigate crises. But the sources said the current policy hews too closely to an earlier strategy of engagement that enabled China to extract concessions in exchange for high-level dialogues that

often yielded few tangible results.

B.1.4 (F1) China threatens to shoot Nancy Pelosi’s plane down if she visits Taiwan

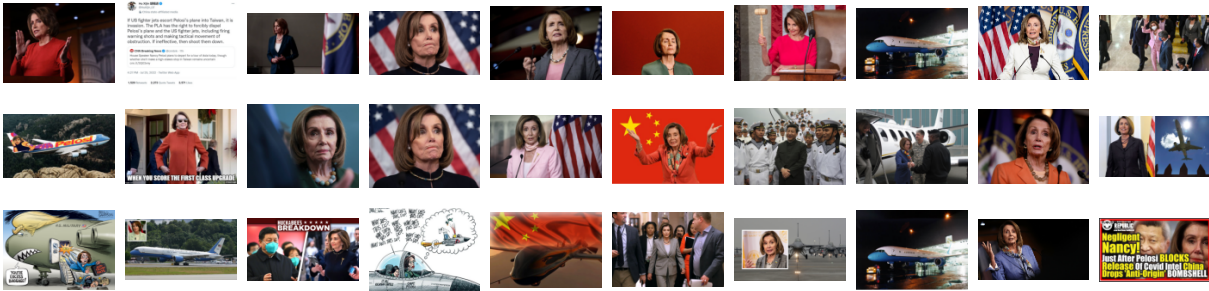


Figure B.3: Treatment Images for Treatment F1.

In a recent statement, Chinese Ministry of Defense spokesman Tan Kefei expressed strong concerns over the possibility of Speaker Nancy Pelosi visiting Taiwan. The spokesman warned that such a visit would pose a severe threat to China’s sovereignty and territorial integrity.

“The Chinese military will never sit idly by and will certainly take strong and resolute measures to thwart any interference by external forces and secessionist attempts for ‘Taiwan independence,’ and firmly defend China’s national sovereignty and territorial integrity,” emphasized Tan Kefei.

B.1.5 (F2) Soviet and Chinese communists have grabbed control of U.S. entertainment, movies, television, music, academia, K-12 education and the news media

In a recent interview, Monica Crowley, former assistant secretary for public affairs at the Treasury Department, made an assertion suggesting that overseas communists have exerted control over key institutions of American life for several decades. Crowley argued that these foreign forces

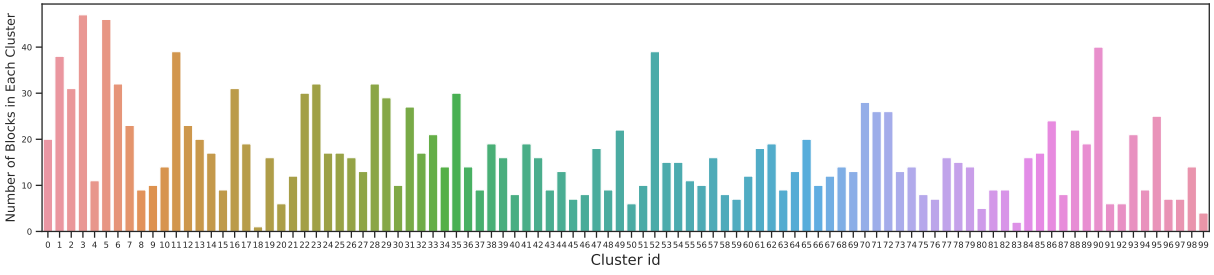


Figure B.6: Distribution of Clusters from K-Means.

B.3 Additional Tables

Table B.1: Average Treatment Effects

	(1)	(2)	(3)	(4)
Has Image	-0.372 (1.210)	-0.341 (1.244)		
Has Image x Z1			-0.276 (2.179)	-0.097 (2.070)
Has Image x Z2			5.418** (2.023)	5.298 (3.836)
Has Image x Z3			4.530+ (2.497)	4.337 (1.546)
Has Image x Z4			-1.169 (2.201)	-1.168** (0.018)
Has Image x Z5			1.246 (1.949)	1.206 (1.390)
Has Image x Z6			-8.140*** (2.375)	-7.958 (2.324)
Has Image x Z7			-7.526** (2.470)	-7.548+ (0.928)
Has Image x Z8			0.813 (2.709)	0.631 (3.158)
Has Image x Z9			3.994 (2.702)	3.602 (1.705)
Has Image x Z10			5.635* (2.466)	4.988+ (0.439)
Fixed Effects: Party, Gender, Race, Age, Edu, Region		Yes		Yes
Num.Obs.	2628	2628	2628	2628
R2	0.000	0.036	0.013	0.048
R2 Adj.	0.000	0.029	0.009	0.038
RMSE	29.22	28.70	29.03	28.52

+ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

B.4 Survey Questionnaire

Survey Flow

Block: Consent (1 Question)
EmbeddedData Treatment_id = \${rand://int/0:1} img_id_T = \${rand://int/1:30} img_id_F = \${rand://int/1:30} participantIdValue will be set from Panel or URL. assignmentIdValue will be set from Panel or URL. projectIdValue will be set from Panel or URL.
Standard: Demographics_Political (8 Questions)
Branch: New Branch If If Help us keep track of who is paying attention to the survey. Please select "Somewhat disagree" fr... Somewhat disagree Is Not Selected
EndSurvey: Advanced
Standard: Instructions (1 Question)
BlockRandomizer: 4 - Evenly Present Elements
Standard: Treatment_T_1 (12 Questions) Standard: Treatment_T_2 (12 Questions) Standard: Treatment_F_1 (12 Questions) Standard: Treatment_F_2 (12 Questions)
Standard: Demographics_General (6 Questions) Standard: Attention_2 (1 Question)
Branch: New Branch If If Help us keep track of who is paying attention to the survey. Please select "Disagree" from the op... Disagree Is Not Selected
EndSurvey: Advanced
EndSurvey: Advanced

Page Break

Start of Block: Consent

consent_form **Take this survey to let your voice heard!** We are researchers at the University of California, San Diego interested in understanding your opinion. A full description of the study is available here: [Consent](#) Please read this document and download or print a version for your records. If you wish to participate in this study, please click the arrow below to continue.

End of Block: Consent

Start of Block: Demographics_Political



party_id Generally speaking, do you usually think of yourself as a Democrat, a Republican, an independent, or what?

- Democrat (1)
 - Republican (2)
 - Independent (3)
 - Something else, please specify: (4)
-

Page Break

Display This Question:

If Generally speaking, do you usually think of yourself as a Democrat, a Republican, an independent,... = Democrat

party_id_dem Would you call yourself a strong Democrat or a not very strong Democrat?

- Strong (1)
- Not very strong (2)

Display This Question:

If Generally speaking, do you usually think of yourself as a Democrat, a Republican, an independent,... = Republican

party_id_rep Would you call yourself a strong Republican or a not very strong Republican?

- Strong (1)
- Not very strong (2)

Display This Question:

If Generally speaking, do you usually think of yourself as a Democrat, a Republican, an independent,... = Independent

Or Generally speaking, do you usually think of yourself as a Democrat, a Republican, an independent,... = Something else, please specify:



party_id_ind Do you think of yourself as closer to the Republican Party or to the Democratic Party?

- Closer to the Republican Party (1)
- Closer to the Democratic Party (2)
- Neither (3)

Page Break

pol_ideology Here is a 7-point scale on which the political views that people might hold are arranged from extremely liberal (left) to extremely conservative (right).

Where would you place yourself on this scale?

	1 (1)	2 (2)	3 (3)	4 (4)	5 (5)	6 (6)	7 (7)	
Extremely Liberal	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Extremely Conservative

pol_attention How often do you pay attention to what's going on in government and politics?

- Always (1)
- Most of the time (2)
- About half the time (3)
- Some of the time (4)
- Never (5)

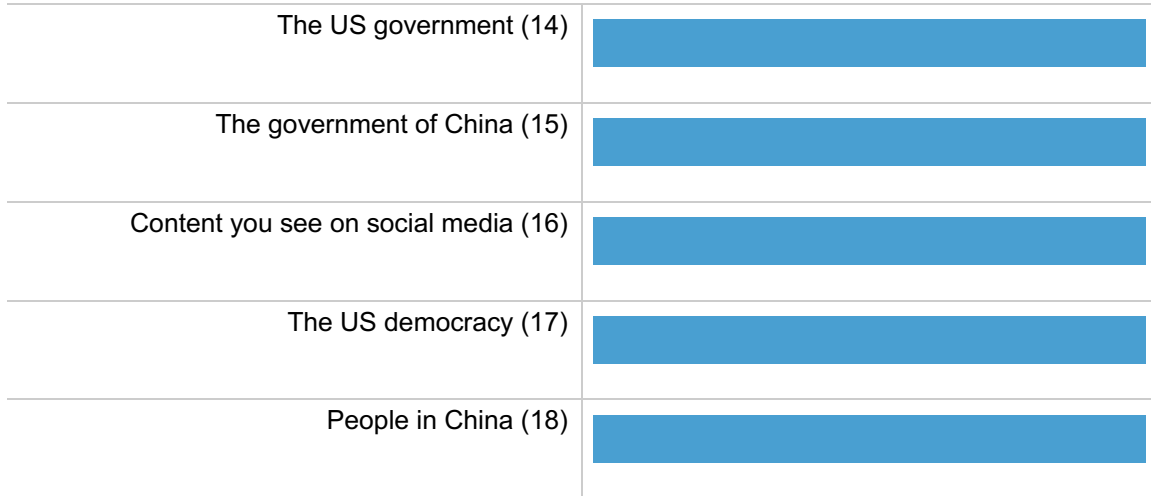
Page Break



trust_pre On a 0-100 scale, how much trust and confidence do you have in the following:

← None at all Not very much A fair amount A great deal →

0 10 20 30 40 50 60 70 80 90 100



Page Break

attention_1 Help us keep track of who is paying attention to the survey. Please select "Somewhat disagree" from the options below.

- Strongly agree (1)
- Agree (2)
- Somewhat agree (3)
- Neither agree nor disagree (4)
- Somewhat disagree (5)
- Disagree (6)
- Strongly disagree (7)

End of Block: Demographics_Political

Start of Block: Instructions

pre_treatment_block For the next part of the study, you will be shown excerpts from four recent news articles. Please take a look at them, and you will be asked to provide your beliefs about them.

End of Block: Instructions

Start of Block: Treatment_T_1

timing_T_1 Timing

First Click (1)

Last Click (2)

Page Submit (3)

Click Count (4)

Display This Question:

If Treatment_id = 1

T_1_tr Take a look at the following news article that someone has shared on a social media site:

Chinese hackers seeking to disrupt communications between US and Asia in event of crisis, Microsoft says Chinese government-backed hackers are likely pursuing cyber capabilities that could be used to “disrupt critical communications” between the US and the Asia Pacific region in the event of a future US-China crisis, Microsoft warned on Wednesday. The Chinese hackers have been active since mid-2021 and targeted critical infrastructure organizations in the US territory of Guam and in other parts of the US as part of a stealthy spying and information gathering campaign, Microsoft said in a new report. In a separate advisory released Wednesday, Western security agencies said they believe the Chinese hackers could apply the same stealthy techniques against critical sectors “worldwide.”

Display This Question:

If Treatment id = 0



T_1_co Take a look at the following news article that someone has shared on a social media site:

Chinese hackers seeking to disrupt communications between US and Asia in event of crisis, Microsoft says Chinese government-backed hackers are likely pursuing cyber capabilities that could be used to “disrupt critical communications” between the US and the Asia Pacific region in the event of a future US-China crisis, Microsoft warned on Wednesday. The Chinese hackers have been active since mid-2021 and targeted critical infrastructure organizations in the US territory of Guam and in other parts of the US as part of a stealthy spying and information gathering campaign, Microsoft said in a new report. In a separate advisory released Wednesday, Western security agencies said they believe the Chinese hackers could apply the same stealthy techniques against critical sectors “worldwide.”

T_1_share On a 0-100 scale, how likely:

← Extremely Unlikely Extremely Likely →

0 10 20 30 40 50 60 70 80 90 100

You will "like" or "share" this article on social media (8)	
Other people will "like" or "share" this article on social media (10)	

Page Break

timing_T_1_2 Timing
First Click (1)
Last Click (2)
Page Submit (3)
Click Count (4)

Display This Question:

If Treatment_id = 1

T_1_tr_2

Chinese hackers seeking to disrupt communications between US and Asia in event of crisis, Microsoft says Chinese government-backed hackers are likely pursuing cyber capabilities that could be used to “disrupt critical communications” between the US and the Asia Pacific region in the event of a future US-China crisis, Microsoft warned on Wednesday. The Chinese hackers have been active since mid-2021 and targeted critical infrastructure organizations in the US territory of Guam and in other parts of the US as part of a stealthy spying and information gathering campaign, Microsoft said in a new report. In a separate advisory released Wednesday, Western security agencies said they believe the Chinese hackers could apply the same stealthy techniques against critical sectors “worldwide.

Display This Question:

If Treatment_id = 0

T_1_co_2



Chinese hackers seeking to disrupt communications between US and Asia in event of crisis, Microsoft says Chinese government-backed hackers are likely pursuing cyber capabilities that could be used to “disrupt critical communications” between the US and the Asia Pacific region in the event of a future US-China crisis, Microsoft warned on Wednesday. The Chinese hackers have been active since mid-2021 and targeted critical infrastructure organizations in the US territory of Guam and in other parts of the US as part of a stealthy spying and information gathering campaign, Microsoft said in a new report. In a separate advisory released Wednesday, Western security agencies said they believe the Chinese hackers could apply the same stealthy techniques against critical sectors “worldwide.

JS

T_1_perception On a 0-100 scale, how likely:

← Extremely Unlikely Extremely Likely →

0 10 20 30 40 50 60 70 80 90 100

This news is true (5)	
Other people will think this news is true (12)	

Page Break

timing_T_1_3 Timing
First Click (1)
Last Click (2)
Page Submit (3)
Click Count (4)

Display This Question:

If Treatment_id = 1

T_1_tr_3

Chinese hackers seeking to disrupt communications between US and Asia in event of crisis, Microsoft says Chinese government-backed hackers are likely pursuing cyber capabilities that could be used to “disrupt critical communications” between the US and the Asia Pacific region in the event of a future US-China crisis, Microsoft warned on Wednesday. The Chinese hackers have been active since mid-2021 and targeted critical infrastructure organizations in the US territory of Guam and in other parts of the US as part of a stealthy spying and information gathering campaign, Microsoft said in a new report. In a separate advisory released Wednesday, Western security agencies said they believe the Chinese hackers could apply the same stealthy techniques against critical sectors “worldwide.

Display This Question:

If Treatment_id = 0

T_1_co_3

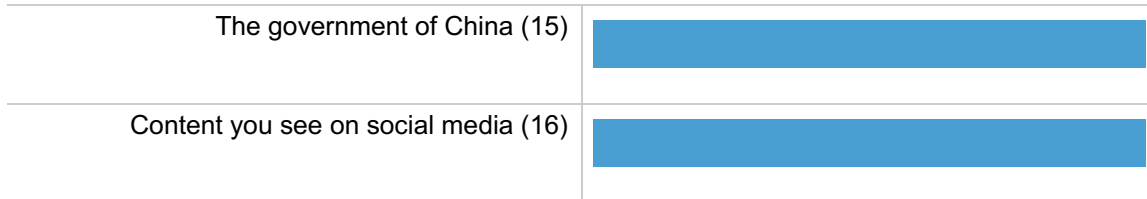
Chinese hackers seeking to disrupt communications between US and Asia in event of crisis, Microsoft says Chinese government-backed hackers are likely pursuing cyber capabilities that could be used to “disrupt critical communications” between the US and the Asia Pacific region in the event of a future US-China crisis, Microsoft warned on Wednesday. The Chinese hackers have been active since mid-2021 and targeted critical infrastructure organizations in the US territory of Guam and in other parts of the US as part of a stealthy spying and information gathering campaign, Microsoft said in a new report. In a separate advisory released Wednesday, Western security agencies said they believe the Chinese hackers could apply the same stealthy techniques against critical sectors “worldwide.

JS

T_1_trust On a 0-100 scale, how much trust and confidence do you have in:

← None at all Not very much A fair amount A great deal →

0 10 20 30 40 50 60 70 80 90 100



End of Block: Treatment_T_1

Start of Block: Treatment_T_2

timing_T_2 Timing

First Click (1)

Last Click (2)

Page Submit (3)

Click Count (4)

Display This Question:

If Treatment_id = 0

T_2_tr Take a look at the following news article that someone has shared on a social media site:

Biden Administration Delayed Sanctions over Spy Balloon to ‘Limit Damage’ to China Ties

When an alleged Chinese spy balloon traversed the United States in February, some U.S. officials were confident the incursion would galvanize the U.S. bureaucracy to push forward a slate of actions to counter China. Instead, the U.S. State Department held back human rights-related sanctions, export controls and other sensitive actions to try to limit damage to the U.S.-China relationship, according to four sources with direct knowledge of U.S. policy. Former diplomats and members of Congress from both parties have argued that the U.S. must keep channels of communication open with Beijing to avoid misunderstandings and navigate crises. But the sources said the current policy hews too closely to an earlier strategy of engagement that enabled China to extract concessions in exchange for high-level dialogues that often yielded few tangible results.

Display This Question:

If Treatment_id = 1

T_2_co Take a look at the following news article that someone has shared on a social media site:

Biden Administration Delayed Sanctions over Spy Balloon to ‘Limit Damage’ to China Ties When an alleged Chinese spy balloon traversed the United States in February, some U.S. officials were confident the incursion would galvanize the U.S. bureaucracy to push forward a slate of actions to counter China. Instead, the U.S. State Department held back human rights-related sanctions, export controls and other sensitive actions to try to limit damage to the U.S.-China relationship, according to four sources with direct knowledge of U.S. policy. Former diplomats and members of Congress from both parties have argued that the U.S. must keep channels of communication open with Beijing to avoid misunderstandings and navigate crises. But the sources said the current policy hews too closely to an earlier strategy of engagement that enabled China to extract concessions in exchange for high-level dialogues that often yielded few tangible results.

T_2_share On a 0-100 scale, how likely:

← Extremely Unlikely Extremely Likely →

0 10 20 30 40 50 60 70 80 90 100

You will "like" or "share" this article on social media (8)	
Other people will "like" or "share" this article on social media (10)	

Page Break

timing_T_2_2 Timing
First Click (1)
Last Click (2)
Page Submit (3)
Click Count (4)

Display This Question:

If Treatment_id = 0

T_2_tr_2

Biden Administration Delayed Sanctions over Spy Balloon to ‘Limit Damage’ to China Ties When an alleged Chinese spy balloon traversed the United States in February, some U.S. officials were confident the incursion would galvanize the U.S. bureaucracy to push forward a slate of actions to counter China. Instead, the U.S. State Department held back human rights-related sanctions, export controls and other sensitive actions to try to limit damage to the U.S.-China relationship, according to four sources with direct knowledge of U.S. policy. Former diplomats and members of Congress from both parties have argued that the U.S. must keep channels of communication open with Beijing to avoid misunderstandings and navigate crises. But the sources said the current policy hews too closely to an earlier strategy of engagement that enabled China to extract concessions in exchange for high-level dialogues that often yielded few tangible results.

Display This Question:

If Treatment_id = 1

T_2_co_2

Biden Administration Delayed Sanctions over Spy Balloon to ‘Limit Damage’ to China Ties When an alleged Chinese spy balloon traversed the United States in February, some U.S. officials were confident the incursion would galvanize the U.S. bureaucracy to push forward a slate of actions to counter China. Instead, the U.S. State Department held back human rights-related sanctions, export controls and other sensitive actions to try to limit damage to the U.S.-China relationship, according to four sources with direct knowledge of U.S. policy. Former diplomats and members of Congress from both parties have argued that the U.S. must keep channels of communication open with Beijing to avoid misunderstandings and navigate crises. But the sources said the current policy hews too closely to an earlier strategy of engagement that enabled China to extract concessions in exchange for high-level dialogues that often yielded few tangible results.

JS

T_2_perception On a 0-100 scale, how likely:

← Extremely Unlikely Extremely Likely →

0 10 20 30 40 50 60 70 80 90 100

This news is true (5)	
Other people will think this news is true (12)	

Page Break

timing_T_2_3 Timing

First Click (1)

Last Click (2)

Page Submit (3)

Click Count (4)

Display This Question:

If Treatment_id = 0

T_2_tr_3

Biden Administration Delayed Sanctions over Spy Balloon to ‘Limit Damage’ to China

Ties When an alleged Chinese spy balloon traversed the United States in February, some U.S. officials were confident the incursion would galvanize the U.S. bureaucracy to push forward a slate of actions to counter China. Instead, the U.S. State Department held back human rights-related sanctions, export controls and other sensitive actions to try to limit damage to the U.S.-China relationship, according to four sources with direct knowledge of U.S. policy. Former diplomats and members of Congress from both parties have argued that the U.S. must keep channels of communication open with Beijing to avoid misunderstandings and navigate crises. But the sources said the current policy hews too closely to an earlier strategy of engagement that enabled China to extract concessions in exchange for high-level dialogues that often yielded few tangible results.

Display This Question:

If Treatment_id = 1

T_2_co_3

Biden Administration Delayed Sanctions over Spy Balloon to ‘Limit Damage’ to China



Ties When an alleged Chinese spy balloon traversed the United States in February, some U.S. officials were confident the incursion would galvanize the U.S. bureaucracy to push forward a slate of actions to counter China. Instead, the U.S. State Department held back human rights-related sanctions, export controls and other sensitive actions to try to limit damage to the U.S.-China relationship, according to four sources with direct knowledge of U.S. policy. Former diplomats and members of Congress from both parties have argued that the U.S. must keep channels of communication open with Beijing to avoid misunderstandings and navigate crises. But the sources said the current policy hews too closely to an earlier strategy of engagement that enabled China to extract concessions in exchange for high-level dialogues that often yielded few tangible results.

JS

T_2_trust On a 0-100 scale, how much trust and confidence do you have in:

← None at all Not very much A fair amount A great deal →

0 10 20 30 40 50 60 70 80 90 100

The government of China (15)	
Content you see on social media (16)	

End of Block: Treatment_T_2

Start of Block: Treatment_F_1

timing_F_1 Timing

First Click (1)

Last Click (2)

Page Submit (3)

Click Count (4)

Display This Question:

If Treatment_id = 0

F_1_tr Take a look at the following news article that someone has shared on a social media site:

China threatens to shoot Nancy Pelosi's plane down if she visits Taiwan In a recent statement, Chinese Ministry of Defense spokesman Tan Kefei expressed strong concerns over the possibility of Speaker Nancy Pelosi visiting Taiwan. The spokesman warned that such a visit would pose a severe threat to China's sovereignty and territorial integrity. "The Chinese military will never sit idly by and will certainly take strong and resolute measures to thwart any interference by external forces and secessionist attempts for 'Taiwan independence,' and firmly defend China's national sovereignty and territorial integrity," emphasized Tan Kefei.

Display This Question:

If Treatment_id = 1



F_1_co Take a look at the following news article that someone has shared on a social media site:

China threatens to shoot Nancy Pelosi's plane down if she visits Taiwan In a recent statement, Chinese Ministry of Defense spokesman Tan Kefei expressed strong concerns over the possibility of Speaker Nancy Pelosi visiting Taiwan. The spokesman warned that such a visit would pose a severe threat to China's sovereignty and territorial integrity. "The Chinese military will never sit idly by and will certainly take strong and resolute measures to thwart any interference by external forces and secessionist attempts for 'Taiwan independence,' and firmly defend China's national sovereignty and territorial integrity," emphasized Tan Kefei.

F_1_share On a 0-100 scale, how likely:

← Extremely Unlikely Extremely Likely →

0 10 20 30 40 50 60 70 80 90 100

You will "like" or "share" this article on social media (8)	
Other people will "like" or "share" this article on social media (10)	

Page Break

timing_F_1_2 Timing
First Click (1)
Last Click (2)
Page Submit (3)
Click Count (4)

Display This Question:
If Treatment_id = 0

F_1_tr_2

China threatens to shoot Nancy Pelosi's plane down if she visits Taiwan In a recent statement, Chinese Ministry of Defense spokesman Tan Kefei expressed strong concerns over the possibility of Speaker Nancy Pelosi visiting Taiwan. The spokesman warned that such a visit would pose a severe threat to China's sovereignty and territorial integrity. "The Chinese military will never sit idly by and will certainly take strong and resolute measures to thwart any interference by external forces and secessionist attempts for 'Taiwan independence,' and firmly defend China's national sovereignty and territorial integrity," emphasized Tan Kefei.

Display This Question:
If Treatment_id = 1



F_1_co_2

China threatens to shoot Nancy Pelosi's plane down if she visits Taiwan In a recent statement, Chinese Ministry of Defense spokesman Tan Kefei expressed strong concerns over the possibility of Speaker Nancy Pelosi visiting Taiwan. The spokesman warned that such a visit would pose a severe threat to China's sovereignty and territorial integrity. "The Chinese military will never sit idly by and will certainly take strong and resolute measures to thwart any interference by external forces and secessionist attempts for 'Taiwan independence,' and firmly defend China's national sovereignty and territorial integrity," emphasized Tan Kefei.

JS

F_1_perception On a 0-100 scale, how likely:

← Extremely Unlikely Extremely Likely →
0 10 20 30 40 50 60 70 80 90 100

This news is true (5)	
Other people will think this news is true (12)	

Page Break

Timing_F_1_3 Timing
First Click (1)
Last Click (2)
Page Submit (3)
Click Count (4)

Display This Question:

If Treatment_id = 0

F_1_tr_3

China threatens to shoot Nancy Pelosi's plane down if she visits Taiwan In a recent statement, Chinese Ministry of Defense spokesman Tan Kefei expressed strong concerns over the possibility of Speaker Nancy Pelosi visiting Taiwan. The spokesman warned that such a visit would pose a severe threat to China's sovereignty and territorial integrity. "The Chinese military will never sit idly by and will certainly take strong and resolute measures to thwart any interference by external forces and secessionist attempts for 'Taiwan independence,' and firmly defend China's national sovereignty and territorial integrity," emphasized Tan Kefei.

Display This Question:

If Treatment_id = 1

F_1_co_3



China threatens to shoot Nancy Pelosi's plane down if she visits Taiwan In a recent statement, Chinese Ministry of Defense spokesman Tan Kefei expressed strong concerns over the possibility of Speaker Nancy Pelosi visiting Taiwan. The spokesman warned that such a visit would pose a severe threat to China's sovereignty and territorial integrity. "The Chinese military will never sit idly by and will certainly take strong and resolute measures to thwart any interference by external forces and secessionist attempts for 'Taiwan independence,' and firmly defend China's national sovereignty and territorial integrity," emphasized Tan Kefei.

JS

F_1_trust On a 0-100 scale, how much trust and confidence do you have in:

← None at all Not very much A fair amount A great deal →

0 10 20 30 40 50 60 70 80 90 100

The government of China (15)	
Content you see on social media (16)	

End of Block: Treatment_F_1

Start of Block: Treatment_F_2

timing_F_2 Timing
 First Click (1)
 Last Click (2)
 Page Submit (3)
 Click Count (4)

Display This Question:

If Treatment_id = 1

F_2_tr Take a look at the following news article that someone has shared on a social media site:

Soviet and Chinese communists have grabbed control of U.S. entertainment, movies, television, music, academia, K-12 education and the news media In a recent interview, Monica Crowley, former assistant secretary for public affairs at the Treasury Department, made an assertion suggesting that overseas communists have exerted control over key institutions of American life for several decades. Crowley argued that these foreign forces have utilized these institutions as pillars to inflict significant damage over time, ultimately leading the country to a critical juncture. “With those pillars, they have been able to inflict tremendous damage over many decades. And now we are at a tipping point where the useful idiots on the left—the Soviet Union collapsed, the CCP (Chinese Communist Party) stepped in to take over this grand project to destroy the country from within—that’s exactly what’s happening,” stated Crowley during the interview.

Display This Question:

If Treatment_id = 0



F_2_co Take a look at the following news article that someone has shared on a social media site:

Soviet and Chinese communists have grabbed control of U.S. entertainment, movies, television, music, academia, K-12 education and the news media In a recent interview, Monica Crowley, former assistant secretary for public affairs at the Treasury Department, made an assertion suggesting that overseas communists have exerted control over key institutions of American life for several decades. Crowley argued that these foreign forces have utilized these institutions as pillars to inflict significant damage over time, ultimately leading the country to a critical juncture. “With those pillars, they have been able to inflict tremendous damage over many decades. And now we are at a tipping point where the useful idiots on the left—the Soviet Union collapsed, the CCP (Chinese Communist Party) stepped in to take over this grand project to destroy the country from within—that’s exactly what’s happening,” stated Crowley during the interview.

F_2_share On a 0-100 scale, how likely:

← Extremely Unlikely Extremely Likely →

0 10 20 30 40 50 60 70 80 90 100

You will "like" or "share" this article on social media (8)	
Other people will "like" or "share" this article on social media (10)	

Page Break

timing_F_2_2 Timing
First Click (1)
Last Click (2)
Page Submit (3)
Click Count (4)

Display This Question:

If Treatment_id = 1

F_2_tr_2

Soviet and Chinese communists have grabbed control of U.S. entertainment, movies, television, music, academia, K-12 education and the news media In a recent interview, Monica Crowley, former assistant secretary for public affairs at the Treasury Department, made an assertion suggesting that overseas communists have exerted control over key institutions of American life for several decades. Crowley argued that these foreign forces have utilized these institutions as pillars to inflict significant damage over time, ultimately leading the country to a critical juncture. “With those pillars, they have been able to inflict tremendous damage over many decades. And now we are at a tipping point where the useful idiots on the left—the Soviet Union collapsed, the CCP (Chinese Communist Party) stepped in to take over this grand project to destroy the country from within—that’s exactly what’s happening,” stated Crowley during the interview.

Display This Question:

If Treatment_id = 0

F_2_co_2

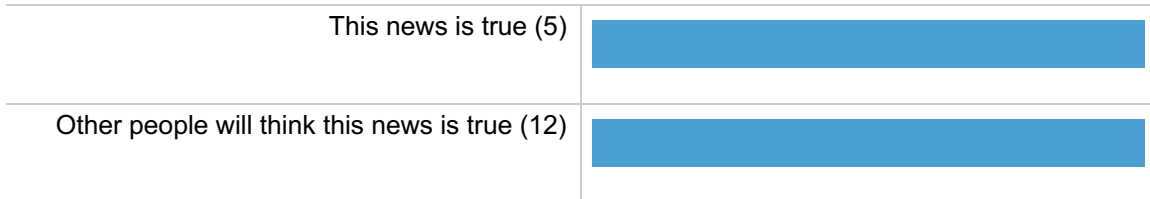
Soviet and Chinese communists have grabbed control of U.S. entertainment, movies, television, music, academia, K-12 education and the news media In a recent interview, Monica Crowley, former assistant secretary for public affairs at the Treasury Department, made an assertion suggesting that overseas communists have exerted control over key institutions of American life for several decades. Crowley argued that these foreign forces have utilized these institutions as pillars to inflict significant damage over time, ultimately leading the country to a critical juncture. “With those pillars, they have been able to inflict tremendous damage over many decades. And now we are at a tipping point where the useful idiots on the left—the Soviet Union collapsed, the CCP (Chinese Communist Party) stepped in to take over this grand project to destroy the country from within—that’s exactly what’s happening,” stated Crowley during the interview.



F_2_perception On a 0-100 scale, how likely:

← Extremely Unlikely Extremely Likely →

0 10 20 30 40 50 60 70 80 90 100



Page Break

timing_F_2_3 Timing
First Click (1)
Last Click (2)
Page Submit (3)
Click Count (4)

Display This Question:

If Treatment_id = 1

F_2_tr_3

Soviet and Chinese communists have grabbed control of U.S. entertainment, movies, television, music, academia, K-12 education and the news media In a recent interview, Monica Crowley, former assistant secretary for public affairs at the Treasury Department, made an assertion suggesting that overseas communists have exerted control over key institutions of American life for several decades. Crowley argued that these foreign forces have utilized these institutions as pillars to inflict significant damage over time, ultimately leading the country to a critical juncture. “With those pillars, they have been able to inflict tremendous damage over many decades. And now we are at a tipping point where the useful idiots on the left—the Soviet Union collapsed, the CCP (Chinese Communist Party) stepped in to take over this grand project to destroy the country from within—that’s exactly what’s happening,” stated Crowley during the interview.

Display This Question:

If Treatment_id = 0

F_2_co_3

Soviet and Chinese communists have grabbed control of U.S. entertainment, movies, television, music, academia, K-12 education and the news media In a recent interview, Monica Crowley, former assistant secretary for public affairs at the Treasury Department, made an assertion suggesting that overseas communists have exerted control over key institutions of American life for several decades. Crowley argued that these foreign forces have utilized these institutions as pillars to inflict significant damage over time, ultimately leading the country to a critical juncture. “With those pillars, they have been able to inflict tremendous damage over many decades. And now we are at a tipping point where the useful idiots on the left—the Soviet Union collapsed, the CCP (Chinese Communist Party) stepped in to take over this grand project to destroy the country from within—that’s exactly what’s happening,” stated Crowley during the interview.



F_2_trust On a 0-100 scale, how much trust and confidence do you have in:

← None at all Not very much A fair amount A great deal →

0 10 20 30 40 50 60 70 80 90 100

The government of China (15)	
Content you see on social media (16)	

End of Block: Treatment_F_2

Start of Block: Demographics_General

Page Break

dem_race How would you describe your race or ethnicity? Please mark all that apply.

- White (1)
- Hispanic or Latino/Latina (2)
- Black or African American (3)
- Asian or Pacific Islander (4)
- American Indian or Alaskan native (5)
- Other (6)

Page Break

dem_birth What is your year of birth?



dem_gender What is your gender?

- Male (1)
- Female (2)
- Prefer to self-describe: (3)

Page Break

dem_edu What is the highest level of education you have completed?

- No degree or diploma earned (1)
- High school diploma or GED (4)
- Some college (2)
- Associate's degree (3)
- Bachelor's degree (5)
- Graduate or professional degree (6)



dem_zip_code What is your US Zip Code?

Page Break



dem_social_media In which social media applications do you have an active account?

By active account, we mean you open this application at least once every week:

- Facebook (1)
 - Instagram (2)
 - Twitter (3)
 - YouTube (4)
 - TikTok (5)
 - Reddit (6)
 - Others, please specify: (7)
-
- None (8)

End of Block: Demographics_General

Start of Block: Attention_2

attention_2 Help us keep track of who is paying attention to the survey. Please select "Disagree" from the options below.

- Strongly agree (1)
- Agree (2)
- Somewhat agree (3)
- Neither agree nor disagree (4)
- Somewhat disagree (5)
- Disagree (6)
- Strongly disagree (7)

End of Block: Attention_2

Bibliography

- Adena, Maja, Ruben Enikolopov, Maria Petrova, Veronica Santarosa, and Ekaterina Zhuravskaya. 2015. "Radio and the Rise of the Nazis in Prewar Germany." *The Quarterly Journal of Economics* 130(4): 1885–1939.
- Albertson, Bethany, and Shana Kushner Gadarian. 2015. *Anxious politics: Democratic citizenship in a threatening world.*: Cambridge University Press.
- Alizadeh, Meysam, Jacob N. Shapiro, Cody Buntain, and Joshua A. Tucker. 2020. "Content-Based Features Predict Social Media Influence Operations." *Science Advances* 6(30), p. eabb5824. <http://dx.doi.org/10.1126/sciadv.abb5824>.
- Allcott, Hunt, and Matthew Gentzkow. 2017. "Social Media and Fake News in the 2016 Election." *Journal of Economic Perspectives* 31(2): 211–36. <http://dx.doi.org/10.1257/jep.31.2.211>.
- Badrinathan, Sumitra. 2021. "Educative Interventions to Combat Misinformation: Evidence from a Field Experiment in India." *American Political Science Review* 115(4): 1325–1341. <http://dx.doi.org/10.1017/S0003055421000459>.
- Bail, Christopher A., Lisa P. Argyle, Taylor W. Brown, John P. Bumpus, Haohan Chen, M. B. Fallin Hunzaker, Jaemin Lee, Marcus Mann, Friedolin Merhout, and Alexander Volfovsky. 2018. "Exposure to opposing views on social media can increase political polarization." *Proceedings of the National Academy of Sciences* 115(37): 9216–9221. <http://dx.doi.org/10.1073/pnas.1804840115>.
- Ball-Rokeach, Sandra J, and Melvin L DeFleur. 1976. "A Dependency Model of Mass-Media Effects." *Communication Research* 3(1): 3–21.
- Barari, Soubhik, Christopher Lucas, and Kevin Munger. 2021. "Political Deepfakes Are As Credible As Other Fake Media And (Sometimes) Real Media." January. <http://dx.doi.org/10.31219/osf.io/cdfh3>.
- Barberá, Pablo. 2016. "Less Is More? How Demographic Sample Weights Can Improve Public Opinion Estimates Based on Twitter Data." <https://www.semanticscholar.org/paper/Less-is-more-How-demographic-sample-weights-can-on-Barber%C3%A1/ecf62dec71b39710ab193f43c35db31457f64446>.

- Belluck, Pam. 2020. "Coronavirus Death Rate in Wuhan Is Lower than Previously Thought, Study Finds." *New York Times*. <https://www.nytimes.com/2020/03/19/health/wuhan-coronavirus-deaths.html>.
- Berinsky, Adam J. 2017. "Rumors and Health Care Reform: Experiments in Political Misinformation." *British Journal of Political Science* 47(2): 241–262. <http://dx.doi.org/10.1017/S0007123415000186>.
- Beskow, David M., Sumeet Kumar, and Kathleen M. Carley. 2020. "The Evolution of Political Memes: Detecting and Characterizing Internet Memes with Multi-Modal Deep Learning." *Information Processing & Management* 57(2), p. 102170. <http://dx.doi.org/10.1016/j.ipm.2019.102170>.
- Bestvater, Sam, Sono Shah, Gonzalo Rivero, and Aaron Smth. 2022. "Politics On Twitter: One-third Of Tweets From U.s. Adults Are Political." Technical Report, Pew Research Center. https://www.pewresearch.org/politics/wp-content/uploads/sites/4/2022/06/PDL_06.16.22_Twitter_Politics_full_report.pdf.
- Blei, David M., Andrew Y. Ng, and Michael I. Jordan. 2003. "Latent Dirichlet Allocation." *The Journal of Machine Learning Research* 3(null): 993–1022.
- Boxell, Levi, and Zachary Steinert-Threlkeld. 2019. "Taxing dissent: The impact of a social media tax in Uganda."
- Buckley, Chris, and Steven Lee Myers. 2020. "As New Coronavirus Spread, China's Old Habits Delayed Fight." *The New York Times*. <https://www.nytimes.com/2020/02/01/world/asia/china-coronavirus.html>.
- Buntain, Cody, Monique Deal Barlow, Mia Bloom, and Mila A Johns. 2022. "Paved with Bad Intentions: QAnon's Save the Children Campaign." *Journal of Online Trust and Safety* 1(2). <http://dx.doi.org/10.54501/jots.v1i2.51>.
- Cao, Qitong. 2020. "The Limitations of Internet Censorship for Authoritarian Information Control: Evidence from China."
- Caron, Mathilde, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. 2019. "Deep Clustering for Unsupervised Learning of Visual Features." *arXiv:1807.05520 [cs]*. <http://arxiv.org/abs/1807.05520>.
- Casas, Andreu, and Nora Webb Williams. 2019a. "Images that Matter: Online Protests and the Mobilizing Role of Pictures." *Political Research Quarterly* 72(2): 360–375. <http://dx.doi.org/10.1177/1065912918786805>.
- Casas, Andreu, and Nora Webb Williams. 2019b. "Images That Matter: Online Protests and the Mobilizing Role of Pictures." *Political Research Quarterly* 72(2): 360–375. <http://dx.doi.org/10.1177/1065912918786805>.

- Chang, Keng-Chi. 2022. "Mapping Visual Themes among Authentic and Coordinated Memes." *PhoMemes Workshop of 2022 International AAAI Conference on Web and Social Media (ICWSM-2022)*. <http://dx.doi.org/10.48550/arXiv.2206.02306>.
- Chang, Keng-Chi, and Cody Buntain. 2023. "Characterizing Image Sharing Behaviors in US Politically Engaged, Random, and Demographic Audience Segments." *PhoMemes Workshop of 2023 International AAAI Conference on Web and Social Media (ICWSM-2023)*.
- Chaudhary, Amit. 2020. "A Visual Exploration of DeepCluster." April. <https://amitness.com/2020/04/deepcluster/>.
- Chen, Yuyu, and David Y Yang. 2019. "The Impact of Media Censorship: 1984 or Brave New World?" *American Economic Review* 109(6): 2294–2332.
- China Data Lab. 2020. "Baidu Mobility Data." *Harvard Dataverse*. <http://dx.doi.org/10.7910/DVN/FAEZIO>.
- Cirone, Alexandra, and William Hobbs. 2022. "Asymmetric Flooding as a Tool for Foreign Influence on Social Media." *Political Science Research and Methods*: 1–12. <http://dx.doi.org/10.1017/psrm.2022.9>.
- Conover, Michael D., Bruno Gonçalves, Alessandro Flammini, and Filippo Menczer. 2012. "Partisan asymmetries in online political activity." *EPJ Data Science* 1(1), p. 6. <http://dx.doi.org/10.1140/epjds6>.
- Deerwester, Scott, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. 1990. "Indexing by Latent Semantic Analysis." *Journal of the American Society for Information Science* 41(6): 391–407. [http://dx.doi.org/10.1002/\(SICI\)1097-4571\(199009\)41:6<391::AID-ASI1>3.0.CO;2-9](http://dx.doi.org/10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9).
- Deibert, Ronald, John Palfrey, Rafal Rohozinski, and Jonathan Zittrain. 2011. *Access Contested: Security, Identity, and Resistance in Asian Cyberspace*. Cambridge: MIT Press.
- DiResta, Renée, Shelby Grossman, and Alexandra Siegel. 2021. "In-House Vs. Outsourced Trolls: How Digital Mercenaries Shape State Influence Strategies." *Political Communication* 0(0): 1–32. <http://dx.doi.org/10.1080/10584609.2021.1994065>.
- Du, Yuhao, Muhammad Aamir Masood, and Kenneth Joseph. 2020. "Understanding Visual Memes: An Empirical Analysis of Text Superimposed on Memes Shared on Twitter." *Proceedings of the International AAAI Conference on Web and Social Media* 14: 153–164. <https://ojs.aaai.org/index.php/ICWSM/article/view/7287>.
- Eckart, Carl, and Gale Young. 1936. "The Approximation of One Matrix by Another of Lower Rank." *Psychometrika* 1(3): 211–218. <http://dx.doi.org/10.1007/BF02288367>.
- Egami, Naoki, Christian J. Fong, Justin Grimmer, Margaret E. Roberts, and Brandon M. Stewart. 2022. "How to Make Causal Inferences Using Texts." *Science Advances* 8(42), p. eabg2652. <http://dx.doi.org/10.1126/sciadv.abg2652>.

- Egami, Naoki, Musashi Jacobs-Harukawa, Brandon M. Stewart, and Hanying Wei. 2023. “Using Large Language Model Annotations for Valid Downstream Statistical Inference in Social Science: Design-Based Semi-Supervised Learning.” June. <http://dx.doi.org/10.48550/arXiv.2306.04746>.
- Enikolopov, Ruben, Maria Petrova, and Ekaterina Zhuravskaya. 2011. “Media and Political Persuasion: Evidence from Russia.” *American Economic Review* 101(7): 3253–3285.
- Flores, Alejandro Quiroz, and Alastair Smith. 2013. “Leader survival and natural disasters.” *British Journal of Political Science*: 821–843.
- Flynn, D. J., Brendan Nyhan, and Jason Reifler. 2017. “The Nature and Origins of Misperceptions: Understanding False and Unsupported Beliefs About Politics.” *Political Psychology* 38 (S1): 127–150. <http://dx.doi.org/10.1111/pops.12394>.
- Fong, Christian J., and Justin Grimmer. 2016. “Discovery of Treatments from Text Corpora.” In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*. 1: 1600–1609, August. <http://dx.doi.org/10.18653/v1/P16-1151>.
- Fong, Christian J., and Justin Grimmer. 2023. “Causal Inference with Latent Treatments.” *American Journal of Political Science* 67(2): 374–389. <http://dx.doi.org/10.1111/ajps.12649>.
- Garimella, Kiran, and Dean Eckles. 2020. “Images and Misinformation in Political Groups: Evidence from WhatsApp in India.” *Harvard Kennedy School Misinformation Review*. <http://dx.doi.org/10.37016/mr-2020-030>.
- Gläbel, Christian, and Katrin Paula. 2019. “Sometimes less is more: Censorship, news falsification, and disapproval in 1989 East Germany.” *American Journal of Political Science*.
- Griffiths, James, Tara John, and Steve George. 2020. “Unprecedented lockdown on 10 cities and 30 million people.” *CNN*. https://www.cnn.com/asia/live-news/coronavirus-outbreak-hnk-intl-01-24-20/h_2587b2ec049c50eb87e75f321f40d2b4.
- Griffiths, James, and Amy Woodyatt. 2020. “80 million people in China are living under travel restrictions due to the coronavirus outbreak.” *CNN*. <https://www.cnn.com/2020/02/16/asia/coronavirus-covid-19-death-toll-update-intl-hnk/index.html>.
- Griffiths, Thomas L., and Zoubin Ghahramani. 2011. “The Indian Buffet Process: An Introduction and Review.” *Journal of Machine Learning Research* 12(32): 1185–1224. <http://jmlr.org/papers/v12/griffiths11a.html>.
- Guang, Lei, Margaret Roberts, Yiqing Xu, and Jiannan Zhao. 2020. “<http://chinadatalab.ucsd.edu/viz-blog/pandemic-sees-increase-in-chinese-support-for-regime-decrease-in-views-towards-us/>.” *China Data Lab Blogpost*.

- Guess, Andrew M., Brendan Nyhan, and Jason Reifler. 2020. "Exposure to Untrustworthy Websites in the 2016 US Election." *Nature Human Behaviour* 4(5): 472–480. <http://dx.doi.org/10.1038/s41562-020-0833-x>.
- Guess, Andrew, Jonathan Nagler, and Joshua Tucker. 2019. "Less than You Think: Prevalence and Predictors of Fake News Dissemination on Facebook." *Science Advances* 5(1), p. eaau4586. <http://dx.doi.org/10.1126/sciadv.aau4586>.
- Hameleers, Michael, Thomas E. Powell, Toni G.L.A. Van Der Meer, and Lieke Bos. 2020. "A Picture Paints a Thousand Lies? The Effects and Mechanisms of Multimodal Disinformation and Rebuttals Disseminated via Social Media." *Political Communication* 37(2): 281–301. <http://dx.doi.org/10.1080/10584609.2019.1674979>.
- Hassanpour, Navid. 2014. "Media disruption and revolutionary unrest: Evidence from Mubarak's quasi-experiment." *Political Communication* 31(1): 1–24.
- He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. "Deep residual learning for image recognition." In *Proceedings of the IEEE conference on computer vision and pattern recognition*: 770–778.
- Hobbs, William R, and Margaret E Roberts. 2018. "How sudden censorship can increase access to information." *American Political Science Review* 112(3): 621–636.
- Hu, Bo, Jun Shao, and Mari Palta. 2006. "Pseudo-R² in logistic regression model." *Statistica Sinica*: 847–860.
- Hu, Tao, Weihe Wendy Guan, Xinyan Zhu, Yuanzheng Shao, Lingbo Liu, Jing Du, Hongqiang Liu, Huan Zhou, Jialei Wang, Bing She, Luyao Zhang, Zhibin Li, Peixiao Wang, Yicheng Tang, Ruizhi Hou, Yun Li, Dexuan Sha, Yifan Yang, Ben Lewis, Devika Kakkar, and Shuming Bao. 2020. "Building an Open Resources Repository for COVID-19 Research." *Data and Information Management* 4(3): 130–147. <http://dx.doi.org/doi:10.2478/dim-2020-0012>.
- Huang, Haifeng. 2015. "Propaganda as Signaling." *Comparative Politics* 47(4): 419–444.
- Huang, Haifeng, and Yao-Yuan Yeh. 2019. "Information from abroad: Foreign media, selective exposure and political support in China." *British Journal of Political Science* 49(2): 611–636.
- Hughes, Adam G, Stefan D McCabe, William R Hobbs, Emma Remy, Sono Shah, and David M J Lazer. 2021. "Using Administrative Records and Survey Data to Construct Samples of Tweeters and Tweets." *Public Opinion Quarterly* 85(S1): 323–346. <http://dx.doi.org/10.1093/poq/nfab020>.
- Huszár, Ferenc, Sofia Ira Ktena, Conor O'Brien, Luca Belli, Andrew Schlaikjer, and Moritz Hardt. 2022. "Algorithmic amplification of politics on Twitter." *Proceedings of the National Academy of Sciences* 119(1), p. e2025334119. <http://dx.doi.org/10.1073/pnas.2025334119>.

- Imai, Kosuke, and Aaron Strauss. 2011. “Estimation of Heterogeneous Treatment Effects from Randomized Experiments, with Application to the Optimal Planning of the Get-Out-the-Vote Campaign.” *Political Analysis* 19(1): 1–19. <http://dx.doi.org/10.1093/pan/mpq035>.
- Jansen, Sue Curry, and Brian Martin. 2003. “Making Censorship Backfire.” *Counterpoise* 7(3): 5–15.
- Jiang, Junyan, and Dali L. Yang. 2016. “Lying or Believing? Measuring Preference Falsification From a Political Purge in China.” *Comparative Political Studies* 49(5): 600–634. <http://dx.doi.org/10.1177/0010414015626450>.
- Joo, Jungseock, and Zachary C. Steinert-Threlkeld. 2018. “Image as Data: Automated Visual Content Analysis for Political Science.” *arXiv:1810.01544 [cs, stat]*. <http://arxiv.org/abs/1810.01544>.
- Joshi, Amogh, and Cody Buntain. 2022. “Examining Similar and Ideologically Correlated Imagery in Online Political Communication.” January. <http://arxiv.org/abs/2110.01183>.
- Karkkainen, Kimmo, and Jungseock Joo. 2021. “FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age for Bias Measurement and Mitigation.” In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*.: 1548–1558.
- Kiela, Douwe, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. “The Hateful Memes Challenge: Detecting Hate Speech in Multimodal Memes.” In *Advances in Neural Information Processing Systems*. eds. by H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin 33: 2611–2624: Curran Associates, Inc., . https://proceedings.neurips.cc/paper_files/paper/2020/file/1b84c4cee2b8b3d823b30e2d604b1878-Paper.pdf.
- Kinetz, Erika. 2021. “Army of fake fans boosts China’s messaging on Twitter.” *Associated Press*.
- Kraemer, Moritz U. G., Chia-Hung Yang, Bernardo Gutierrez, Chieh-Hsi Wu, Brennan Klein, David M. Pigott, Open COVID-19 Data Working Group, Louis du Plessis, Nuno R. Faria, Ruoran Li, William P. Hanage, John S. Brownstein, Maylis Layan, Alessandro Vespignani, Huaiyu Tian, Christopher Dye, Oliver G. Pybus, and Samuel V. Scarpino. 2020. “The effect of human mobility and control measures on the COVID-19 epidemic in China.” *Science* 368 (6490): 493–497. <http://dx.doi.org/10.1126/science.abb4218>.
- Kumar, Srijan, Robert West, and Jure Leskovec. 2016. “Disinformation on the Web: Impact, Characteristics, and Detection of Wikipedia Hoaxes.” In *Proceedings of the 25th International Conference on World Wide Web*. WWW ’16: 591–602, Republic and Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee, , April. <http://dx.doi.org/10.1145/2872427.2883085>.
- Lazarev, Egor, Anton Sobolev, Irina V. Soboleva, and Boris Sokolov. 2014. “Trial by fire: A natural disaster’s impact on support for the authorities in rural Russia.” *World Politics* 66(4): 641–668. <http://dx.doi.org/10.1017/S0043887114000215>.

- Lee, Juheon. 2021. "The social impact of natural hazards: a multi-level analysis of disasters and forms of trust in mainland China." *Disasters* 45(1): 158–179. <http://dx.doi.org/10.1111/disa.12410>.
- Li, Junnan, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. "BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models." January. <http://arxiv.org/abs/2301.12597>.
- Li, Yiyi, and Ying Xie. 2020a. "Is a Picture Worth a Thousand Words? An Empirical Study of Image Content and Social Media Engagement." *Journal of Marketing Research* 57(1): 1–19. <http://dx.doi.org/10.1177/0022243719881113>.
- Li, Yiyi, and Ying Xie. 2020b. "Is a Picture Worth a Thousand Words? An Empirical Study of Image Content and Social Media Engagement." *Journal of Marketing Research* 57(1): 1–19. <http://dx.doi.org/10.1177/0022243719881113>.
- MacKinnon, Rebecca. 2012. *Consent of the Networked: The Worldwide Struggle For Internet Freedom*. New York: Basic Books.
- MacKuen, Michael, Jennifer Wolak, Luke Keele, and George E Marcus. 2010. "Civic engagements: Resolute partisanship or reflective deliberation." *American Journal of Political Science* 54(2): 440–458.
- Marcus, George E, and Michael B MacKuen. 1993. "Anxiety, enthusiasm, and the vote: The emotional underpinnings of learning and involvement during presidential campaigns." *American Political Science Review*: 672–685.
- Marcus, George E, W Russell Neuman, and Michael MacKuen. 2000. *Affective intelligence and political judgment*.: University of Chicago Press.
- McFadden, Daniel. 2021. "Quantitative methods for analysing travel behaviour of individuals: some recent developments." In *Behavioural travel modelling*.: Routledge, : 279–318.
- Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeffrey Dean. 2013. "Distributed representations of words and phrases and their compositionality." *NIPS*: 3111–3119.
- Morozov, Evgeny. 2011. *The Net Delusion: The Dark Side of Internet Freedom*. New York: PublicAffairs.
- Mozur, Paul. 2018. "China Presses Its Internet Censorship Efforts Across the Globe." *The New York Times*.
- Mozur, Paul. 2019. "Twitter Users in China Face Detention and Threats in New Beijing Crackdown." *The New York Times*.
- Munger, Kevin, and Joseph Phillips. 2022. "Right-Wing YouTube: A Supply and Demand Perspective." *The International Journal of Press/Politics* 27(1): 186–219. <http://dx.doi.org/10.1177/1940161220964767>.

- Nabi, Zubair. 2014. "Censorship is futile." *First Monday* 19(11).
- Nazeri, Nima, and Collin Anderson. 2013. "Citation Filtered: Iran's Censorship of Wikipedia." *Iran Media Program*.
- Pan, Jennifer, and Alexandra A Siegel. 2020. "How Saudi crackdowns fail to silence online dissent." *American Political Science Review* 114(1): 109–125.
- Pennycook, Gordon, Ziv Epstein, Mohsen Mosleh, Antonio A. Arechar, Dean Eckles, and David G. Rand. 2021. "Shifting Attention to Accuracy Can Reduce Misinformation Online." *Nature* 592(7855): 590–595. <http://dx.doi.org/10.1038/s41586-021-03344-2>.
- Pugh, Alex, and Michelle Torres. 2023. "Beyond Prediction: Identifying Latent Treatments in Images." *Working Paper*, p. 49.
- Qin, Amy, and Vivian Wang. 2020. "Wuhan, Center of Coronavirus Outbreak, Is Being Cut Off by Chinese Authorities." *The New York Times*. <https://www.nytimes.com/2020/01/22/world/asia/china-coronavirus-travel.html>.
- Roberts, Margaret E. 2018. *Censored: Distraction and Diversion inside China's Great Firewall*.: Princeton University Press.
- Roberts, Margaret E., Brandon M. Stewart, Dustin Tingley, and Edoardo M. Airoidi. 2013. "The Structural Topic Model and Applied Social Science." In *Advances in Neural Information Processing Systems Workshop on Topic Models: Computation, Application, and Evaluation*.
- Roberts, Margaret, Brandon Stewart, and Dustin Tingley. 2016. "stm: R Package for Structural Topic Models." *Journal of Statistical Software*. http://scholar.google.com/scholar?q=related:1eMVP-zcgFMJ:scholar.google.com/&hl=en&num=20&as_sdt=0,5.
- Robinson, Darrel, and Marcus Tannenberg. 2019. "Self-censorship of regime support in authoritarian states: Evidence from list experiments in China." *Research and Politics* 6(3). <http://dx.doi.org/10.1177/2053168019856449>.
- Sanovich, Sergey, Denis Stukal, and Joshua A Tucker. 2018. "Turning the virtual tables: Government strategies for addressing online opposition with an application to Russia." *Comparative Politics* 50(3): 435–482.
- Shen, Xiaoxiao, and Rory Truex. 2020. "In Search of Self-Censorship." *British Journal of Political Science*: 1–13. <http://dx.doi.org/10.1017/S0007123419000735>.
- Sivic, and Zisserman. 2003. "Video Google: A Text Retrieval Approach to Object Matching in Videos." In *Proceedings Ninth IEEE International Conference on Computer Vision*.: 1470–1477 vol.2, October. <http://dx.doi.org/10.1109/ICCV.2003.1238663>.
- Stasavage, David. 2020. "Democracy, Autocracy, and Emergency Threats: Lessons for COVID-19 from the Last Thousand Years." *International Organization*: 1–17.

- Steinert-Threlkeld, Zachary C. 2017. “Longitudinal Network Centrality Using Incomplete Data.” *Political Analysis* 25: 308–328. <http://dx.doi.org/10.1017/pan.2017.6>.
- Stockmann, Daniela. 2012. *Media Commercialization and Authoritarian Rule in China*. Cambridge: Cambridge University Press.
- Stockmann, Daniela, and Mary E Gallagher. 2011. “Remote Control: How the Media Sustain Authoritarian Rule in China.” *Comparative Political Studies* 44(4): 436–467.
- Sverdrup, Erik, Han Wu, Susan Athey, and Stefan Wager. 2023. “Qini Curves for Multi-Armed Treatment Rules.” June. <http://dx.doi.org/10.48550/arXiv.2306.11979>.
- Tai, Zixue, and Tao Sun. 2007. “Media dependencies in a changing media environment: The case of the 2003 SARS epidemic in China.” *New Media & Society* 9(6): 987–1009.
- Torres, Michelle. 2018. “Give Me the Full Picture: Using Computer Vision to Understand Visual Frames and Political Communication.” *Working Paper*, p. 30.
- Torres, Michelle, and Francisco Cantú. 2021. “Learning to See: Convolutional Neural Networks for the Analysis of Social Science Data.” *Political Analysis* Forthcoming: 1–19. <http://dx.doi.org/10.1017/pan.2021.9>.
- Veit, Andreas, Tomas Matera, Lukas Neumann, Jiri Matas, and Serge Belongie. 2016. “COCO-Text: Dataset and Benchmark for Text Detection and Recognition in Natural Images.” *arXiv:1601.07140 [cs]*. <http://arxiv.org/abs/1601.07140>.
- Wager, Stefan, and Susan Athey. 2018. “Estimation and Inference of Heterogeneous Treatment Effects Using Random Forests.” *Journal of the American Statistical Association* 113(523): 1228–1242. <http://dx.doi.org/10.1080/01621459.2017.1319839>.
- Whyte, Martin. 2010. *Myth of the social volcano: Perceptions of inequality and distributive injustice in contemporary China.*: Stanford University Press.
- Williams, Nora Webb, Andreu Casas, and John D. Wilkerson. 2020. “Images as Data for Social Science Research: An Introduction to Convolutional Neural Nets for Image Classification.” *Elements in Quantitative and Computational Methods for the Social Sciences*. <http://dx.doi.org/10.1017/9781108860741>.
- Wittenberg, Chloe, Ben M. Tappin, Adam J. Berinsky, and David G. Rand. 2021. “The (Minimal) Persuasive Advantage of Political Video over Text.” *Proceedings of the National Academy of Sciences* 118(47). <http://dx.doi.org/10.1073/pnas.2114388118>.
- Wu, Liang, Fred Morstatter, Kathleen M. Carley, and Huan Liu. 2019. “Misinformation in Social Media: Definition, Manipulation, and Detection.” *ACM SIGKDD Explorations Newsletter* 21 (2): 80–90. <http://dx.doi.org/10.1145/3373464.3373475>.

- Yamada, Ikuya, Akari Asai, Jin Sakuma, Hiroyuki Shindo, Hideaki Takeda, Yoshiyasu Takefuji, and Yuji Matsumoto. 2020. “Wikipedia2Vec: An Efficient Toolkit for Learning and Visualizing the Embeddings of Words and Entities from Wikipedia.” *EMNLP*: 23–30.
- Yanagizawa-Drott, David. 2014. “Propaganda and conflict: Evidence from the Rwandan genocide.” *The Quarterly Journal of Economics* 129(4): 1947–1994.
- Yang, Yunkang, Trevor Davis, and Matthew Hindman. 2023. “Visual Misinformation on Facebook.” *Journal of Communication*, p. jqac051. <http://dx.doi.org/10.1093/joc/jqac051>.
- You, Yu, Yifan Huang, and Yuyi Zhuang. 2020. “Natural disaster and political trust: A natural experiment study of the impact of the Wenchuan earthquake.” *Chinese Journal of Sociology* 6(1): 140–165. <http://dx.doi.org/10.1177/2057150X19891880>.
- Zannettou, Savvas, Tristan Caulfield, Jeremy Blackburn, Emiliano De Cristofaro, Michael Sirivianos, Gianluca Stringhini, and Guillermo Suarez-Tangil. 2018. “On the Origins of Memes by Means of Fringe Web Communities.” *arXiv:1805.12512 [cs]*. <http://arxiv.org/abs/1805.12512>.
- Zannettou, Savvas, Tristan Caulfield, Barry Bradlyn, Emiliano De Cristofaro, Gianluca Stringhini, and Jeremy Blackburn. 2019. “Characterizing the Use of Images in State-Sponsored Information Warfare Operations by Russian Trolls on Twitter.” *arXiv:1901.05997 [cs]*. <http://arxiv.org/abs/1901.05997>.
- Zhang, Han, and Yilang Peng. 2021. “Image Clustering: An Unsupervised Approach to Categorize Visual Data in Social Science Research.” May. <http://dx.doi.org/10.31235/osf.io/mw57x>.
- Zhang, Han, and Yilang Peng. 2022. “Image Clustering: An Unsupervised Approach to Categorize Visual Data in Social Science Research.” *Sociological Methods & Research*, p. 00491241221082603. <http://dx.doi.org/10.1177/00491241221082603>.
- Zhou, Laura. 2020. “Chinese officials have finally discovered Twitter. What could possibly go wrong?” *South China Morning Post*. <https://scmp.com/news/china/diplomacy/article/3021310/chinese-officials-have-finally-discovered-twitter-what-could>.