Censorship on Twitter: An Initial Exploration

# Issues & Policy Discussion

On July 26, 2018, President Trump tweeted a complaint about something he termed “shadow banning”:

[[1]](#footnote-1)

In the days that followed, many news organizations[[2]](#footnote-2) set out to address what this term means, and to answer the question of whether this was actually happening. There was broad agreement that certain prominent Republicans had been affected, but there was little consensus about the source of this, or whether it had been intentional.

Prior to Trump’s tweet, the website Vice published an article detailing similar allegations to those Trump appeared to be referencing.[[3]](#footnote-3) Twitter responded with a blog post, providing a detailed response and strongly asserting that the allegations were false.[[4]](#footnote-4)

“We do not shadow ban. You are always able to see the tweets from accounts you follow (although you may have to do more work to find them, like go directly to their profile). And we certainly don’t shadow ban based on political viewpoints or ideology.We do rank tweets and search results…. These rankings take many signals into consideration…. We must also address bad-faith actors who intend to manipulate or detract from healthy conversation.”[[5]](#footnote-5)

While Twitter’s denial of any kind of shadow ban seems clear on its face, their answer also underscores the nuance involved in differentiating purported “shadow banning” from other kinds of content manipulation and curation.

What is shadow banning? Twitter uses this definition in their blog post: “[D]eliberately making someone’s content undiscoverable to everyone except the person who posted it, unbeknownst to the original poster.”[[6]](#footnote-6) Other definitions vary, with some employing looser definitions that point to different kinds of “deboosting.”[[7]](#footnote-7) While Twitter categorically denied shadow banning in their July blog post, this claim seems to rest largely on the semantic explanation that individuals were not categorized on ideological or political identity; the actual underlying algorithmic process of reducing visibility is clearly present. Furthermore, Twitter has recently updated its Terms of Service to explicitly acknowledge that they may restrict visibility of any content at any time.[[8]](#footnote-8) Differentiating such activity from “shadow banning” requires fine grain distinctions, and underscores the interpretive framing challenges inherent in discussing these issues.

On December 11, 2019, Twitter founder Jack Dorsey published a series of tweets announcing a special project called “Blue Sky” that Twitter was launching to explore creating an underlying open source protocol for social media, to help separate the functions of network connectivity from content selection and algorithmic processing.[[9]](#footnote-9) Dorsey referenced a detailed proposal by Stephen Wolfram to separate the universality of network / platform connectivity from the Artificial Intelligence selection algorithms:

“Social networks get their usefulness by being monolithic: by having ‘everyone’ connected into them. But the point is that the network can prosper as a monolithic thing, but there doesn’t need to be just one monolithic AI that selects content for all the users on the network. Instead, there can be a whole market of AIs, that users can freely pick between.”[[10]](#footnote-10)

Wolfram envisions a world where “everyone” is still on the same network everyone else, without necessarily being tied to the same algorithmic selection processes. Dorsey suggests in his tweets that he sees Twitter’s real value to users in the content selection / recommendation service, rather than as a core service provider.[[11]](#footnote-11)

The broader issues of content selection and extent to which individuals have the right to post content on social media present new and unique legal challenges in a changing world. Yale Law Professor Jack Balkin has noted that the evolution of technology has created a new paradigm of speech regulation for individuals.

“In the digital age, individuals do not face the familiar dyadic model of speech regulation. In a dyadic model, there are two central actors: the power of the state threatens the individual’s right to speak. Instead, the digital age features a pluralist model of speech control. In the pluralist model individuals may be controlled, censored, and surveilled both by the nation state and by the owners of many different kinds of private infrastructure, who operate across national borders in multiple jurisdictions…. The practical ability to speak in the digital age is shaped by the results of these struggles for control and cooptation.” [[12]](#footnote-12)

These evolving dynamics, Balkin argues, require a significant change in policy in order to effectuate the broader principles of free speech in a vibrant public sphere.[[13]](#footnote-13) As interaction between individuals occurs increasingly in the digital space, it becomes increasingly difficult to square private speech discretion by providers with the core principles of free speech. As Balkin discusses, attempts to fit social media jurisprudence within existing First Amendment doctrine create systemic challenges that undermine the underlying first principles.

Stepping away from broader policy discussions, this paper offers a brief sampling and analysis of Twitter statuses on the issues of shadow banning and censorship. Such analysis, it is hoped, can serve as a methodological starting point for further gauging how social media users perceive their own speech is being regulated. In light of the news articles above, this study begins with the hypothesis that individuals who utilize social media have a strong interest in the policies that determine how their speech is regulated. Further extensions would likely explore the impact of different news events on overall user sentiment.

# Data and Methods

The data for the corpora used is taken directly from the Twitter API[[14]](#footnote-14) using the rtweet package for R.[[15]](#footnote-15) The analysis begins by first creating a corpus of all English-language tweets from each of the term “censorship”.[[16]](#footnote-16) All retweets are removed from the corpus. The corpus is then analyzed through a variety of computational analysis methods, including a word count frequency analysis, sentiment analysis, and structural topic modeling. The relevant implementation details of each method are detailed below.

Because of Twitter API restrictions, the data is limited to a 9-day window. Access to the full archive is available, but is limited to 5,000 tweets / month for free accounts. The codebase used, however, can easily be adapted to analyzing a much broader window of time. This would be an essential step to understanding broader trends across time. As such, the findings presented represent a prototype for potential future expansion. Because the small time window and limited sample size place inherent limits the robustness of any findings, this paper focuses instead on exploring a variety of methods that can later be extended to more comprehensive data sets.

# Findings and Discussion

We first perform some basic sentiment analysis to get a sense of overall tone and general associations:

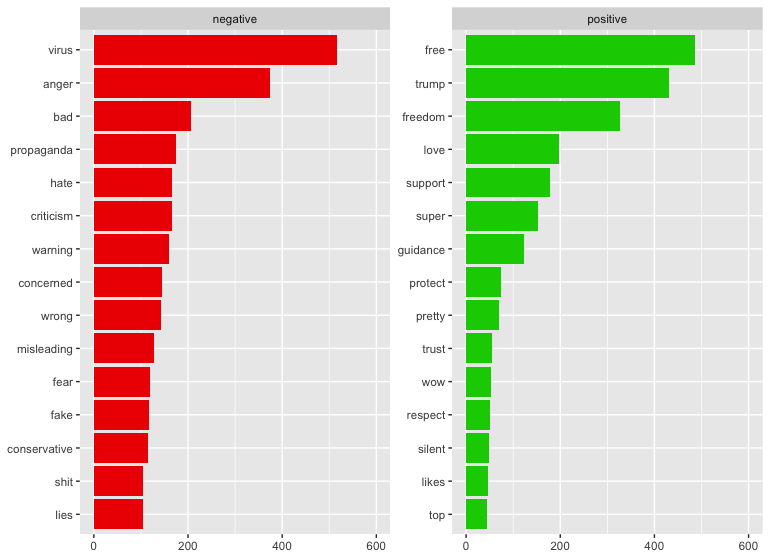


Figure 1: Top words plot. Standard tidytext stopwords are not included.

Note that “trump” is classified by the standard sentiment dictionary as a positive term. As part of an iterative research process, this suggests that a more finely-tailored custom dictionary could be extremely useful to better contextualize sentiment.

To further visualize the data, we can also utilize word clouds:

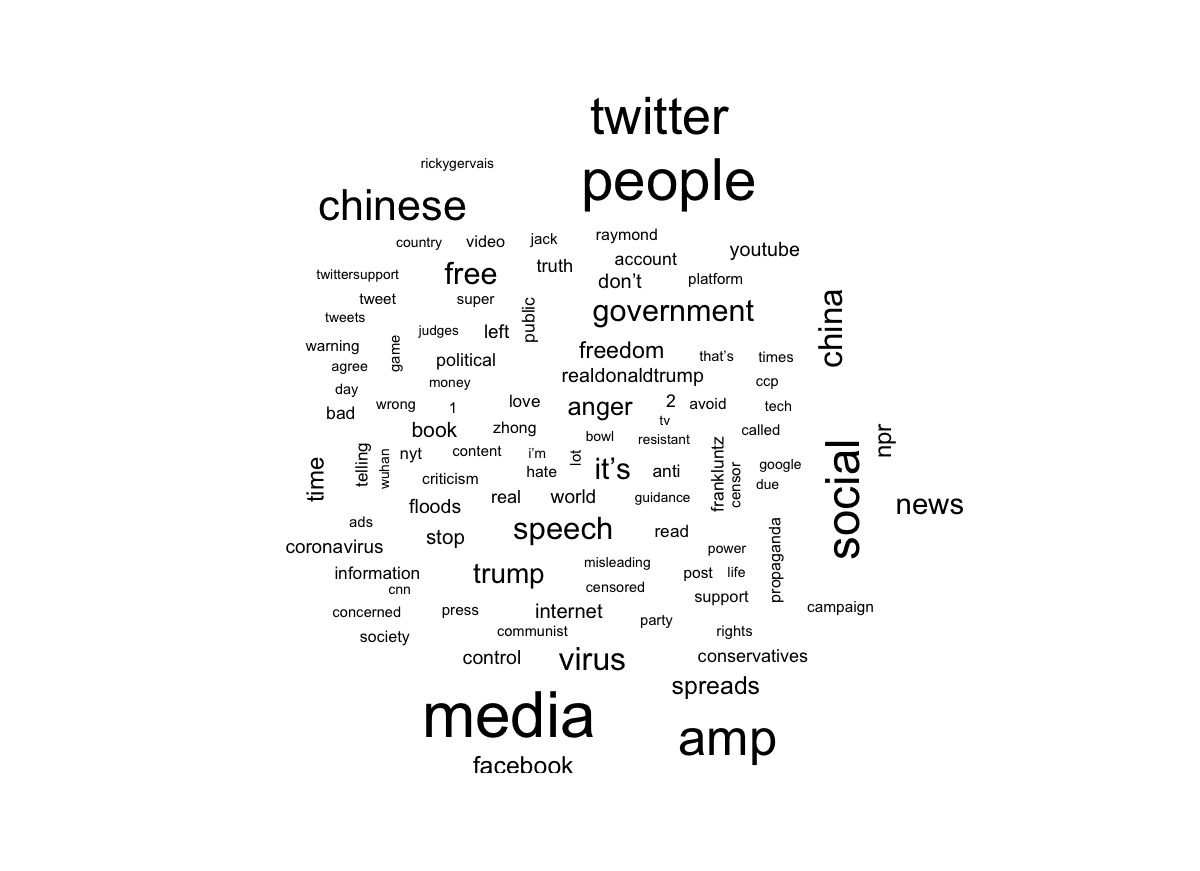


Figure 2: Wordcloud view. Removed words "censorship", "https", "t.co", and standard tidytext stopwords.

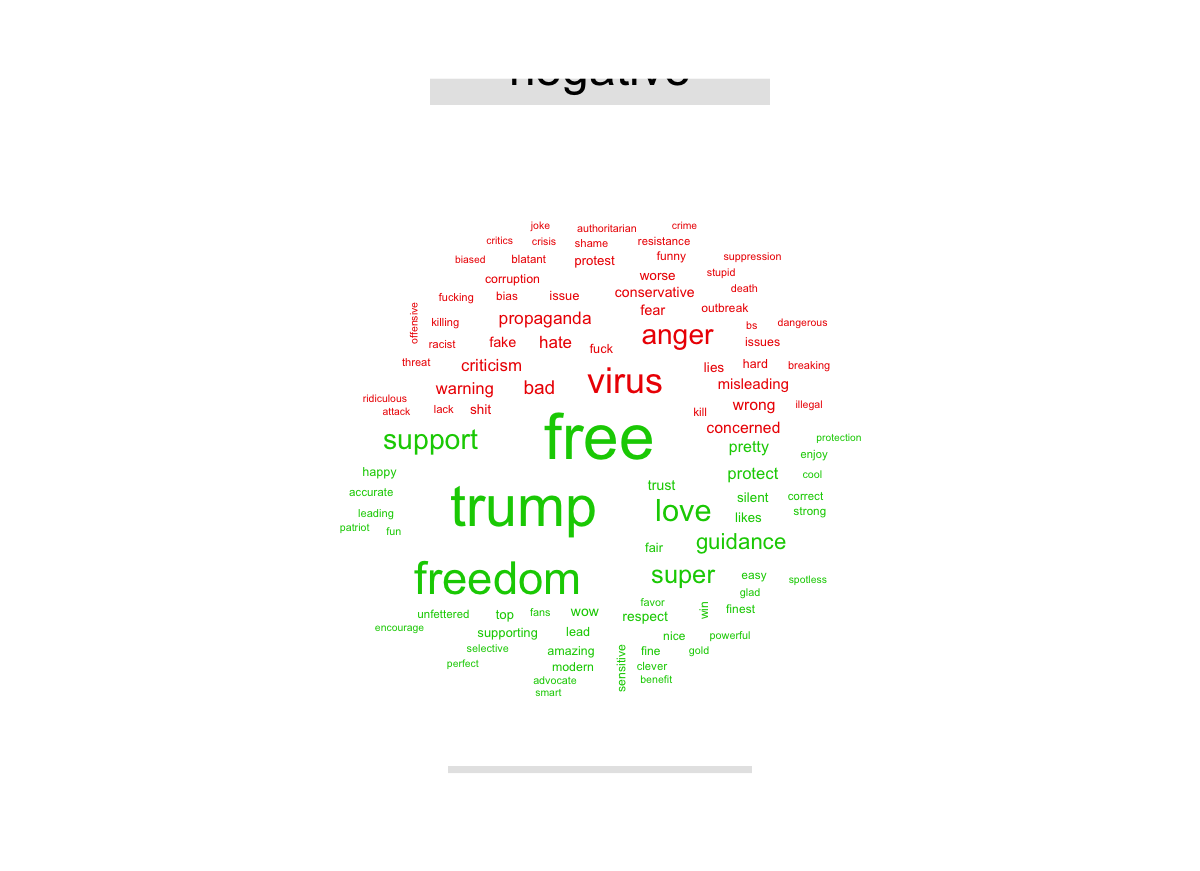


Figure 3: Sentiment word cloud using bing sentiment dictionary.

Again, note that “trump” is classified as strong positive word, while “virus” is classified as a strong negative word. These classifications ignore present political contexts, and may skew the overall positive / negative results.

As a further extension, we can use sentiment analysis to measure the strongest tones present in the tweet sample:

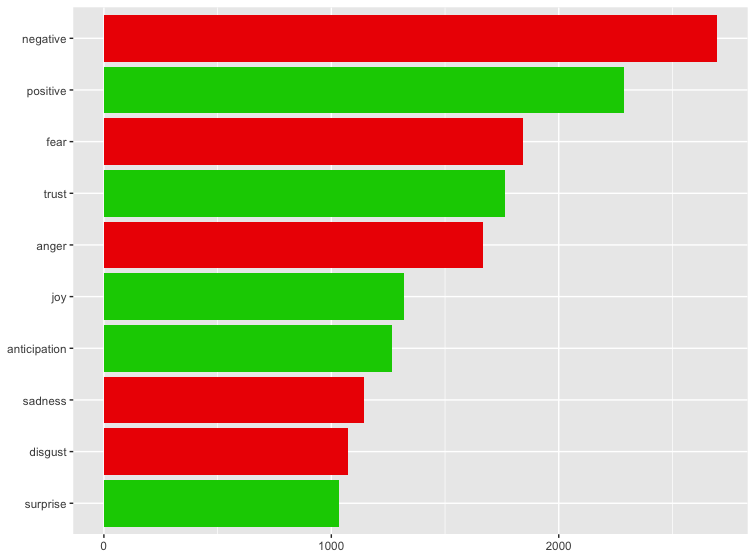


Figure 4: Overall sentiment by category. Using NRC sentiment dictionary.

Another approach is to measure popular bigrams. This may be useful with larger sample sizes. However, the results present little information for the studied sample:

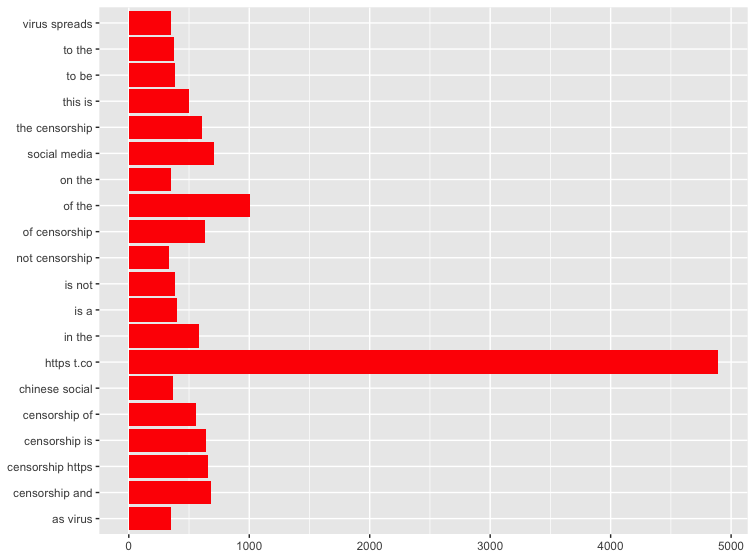


Figure 5: Top bigrams (no words removed).

Another analytical technique that can be employed is Structured Topic Modeling. We explore here both 3 and 5 topic models:

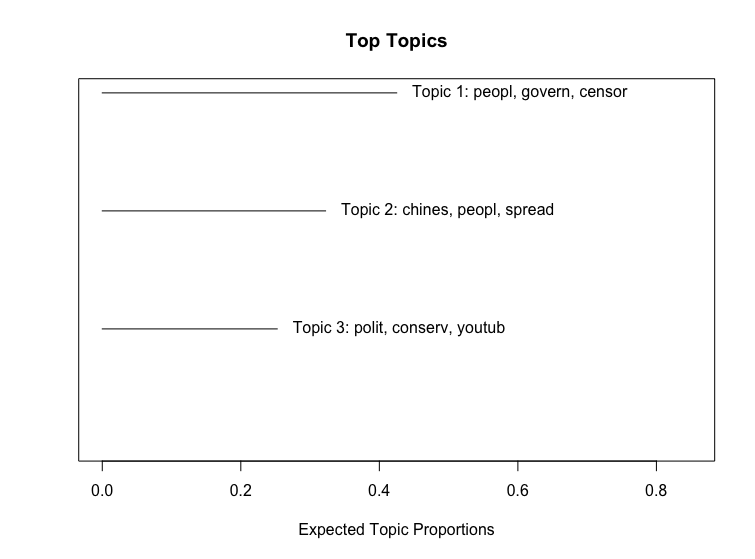


Figure 6: 3-topic STM summary

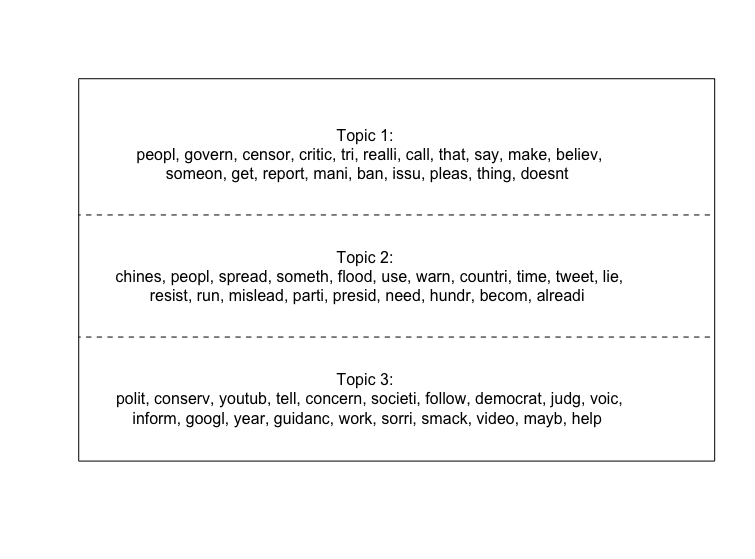


Figure 7: Top words by topic for 3-topic STM

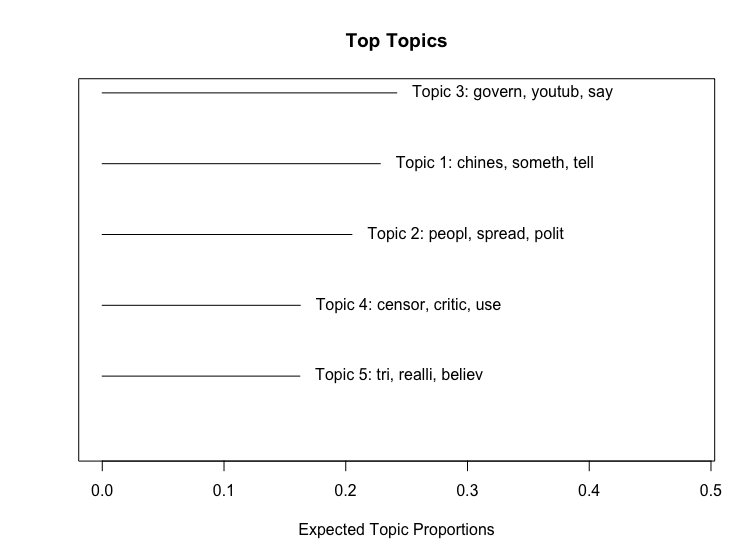


Figure 8: 5-topic STM summary

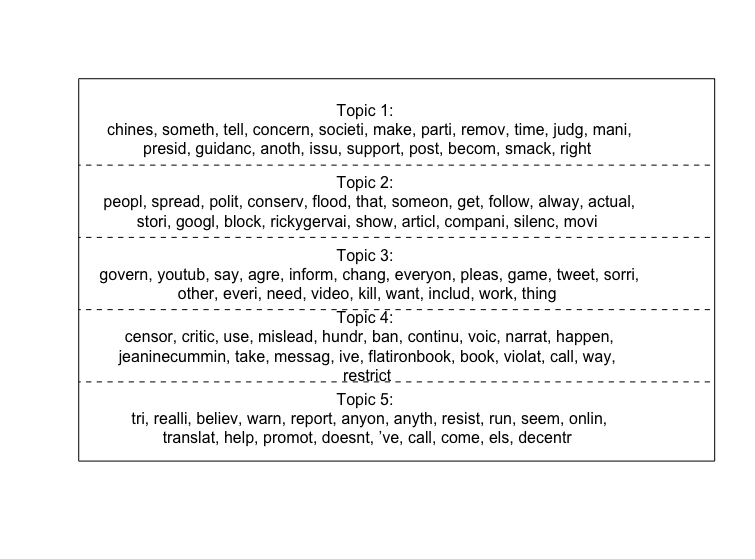


Figure 9: Top words by topic for 5-topic STM

# Discussion and Future Research

The research here provides a starting point for future exploration on larger datasets. However, even on a restricted dataset, both Sentiment Analysis and Structured Topic Modelling seem to provide meaningful categorizations. The Semantic Analysis results show some strong initial sentiment results, and also the prevalence of some context-specific phrases. This suggests that custom dictionary methods could be an excellent next step in providing a fine-grain sentiment analysis to test specific hypothesis. One potential extension: explore the use of the terms “shadowban” and “shadow ban” immediately following Trump’s 2018 tweet on the topic.

The Structured Topic Modeling method also seems promising. Even on a limited dataset, the different topics detected seem to provide useful ways of understanding the different ongoing conversations. With a larger dataset, it would be very interesting to employ map Sentiment Analysis onto each topic, and compare the results across topics.

This paper provides a preliminary descriptive overview of methods that could be employed to better understand the public debate about the phenomenon of shadow banning. Computational Text Analysis provides the capability to track, aggregate, and analyze data in ways that help extract meaningful signal. As social media and public discussion continue to evolve in the digital age, these techniques provide powerful tools to gather and understand this incredible amount of data.

1. Donald J. Trump (@realDonaldTrump), Twitter (Jul 26, 2018, 7:46 AM), https://twitter.com/realdonaldtrump/status/1022447980408983552. [↑](#footnote-ref-1)
2. See Jason Wilson, What is 'shadow banning', and why did Trump tweet about it? The Guardian (2018), https://www.theguardian.com/media/2018/jul/26/what-is-shadow-banning-conservatives-twitter-trump (last visited Jan 31, 2020); Philip Bump, Analysis | Trump 'shadow ban' tweet: A F.A.Q. The Washington Post (2018), https://www.washingtonpost.com/news/politics/wp/2018/07/26/trump-shadow-ban-tweet-a-f-a-q/ (last visited Jan 31, 2020); Edgar Alvarez, Why are Trump and sex workers angry about shadow banning? Engadget (2019), https://www.engadget.com/2018/07/26/twitter-shadow-banning-trump-explainer/ (last visited Jan 31, 2020). [↑](#footnote-ref-2)
3. Alex Thompson, Twitter appears to fixed "shadow ban" of prominent Republicans like the RNC chair and Trump Jr.'s spokesman Vice (2018), https://www.vice.com/en\_us/article/43paqq/twitter-is-shadow-banning-prominent-republicans-like-the-rnc-chair-and-trump-jrs-spokesman?utm\_campaign=sharebutton (last visited Jan 31, 2020). [↑](#footnote-ref-3)
4. Vijaya Gadde and Kayvon Bekpour, *Setting the record straight on shadow banning*, Twitter Company Blog (Jul. 26, 2018), https://blog.twitter.com/en\_us/topics/company/2018/Setting-the-record-straight-on-shadow-banning.html. [↑](#footnote-ref-4)
5. Id. [↑](#footnote-ref-5)
6. Id. [↑](#footnote-ref-6)
7. The website https://shadowban.eu allows one to easily check on any given username for four different kinds of “shadow banning”: a search suggestion ban, a search ban, a ghost ban, and reply deboosting. Each of these is a specific methodology that a user’s influence on the network could be disrupted. In a brief, unscientific test involving a few dozen prominent politicians and ideologues, I was unable to find any examples of any person being banned by any of these methods. [↑](#footnote-ref-7)
8. Twitter Terms of Service, <https://twitter.com/en/tos> (last visited Jan. 30, 2020). [↑](#footnote-ref-8)
9. Jack Dorsey (@jack), Twitter (Dec. 11, 2019, 9:13 AM), <https://twitter.com/jack/status/1204766078468911106>. [↑](#footnote-ref-9)
10. Steven Wolfram, *Testifying at the Senate about A.I.-Selected Content on the Internet*, Steven Wolfram Writings (Jun. 25, 2019), <https://writings.stephenwolfram.com/2019/06/testifying-at-the-senate-about-a-i-selected-content-on-the-internet/>. [↑](#footnote-ref-10)
11. Id. [↑](#footnote-ref-11)
12. Jack, M. Balkin, *Free Speech in the Algorithmic Society: Big Data, Private Governance, and New School Speech Regulation*, 51 U. of Cal., Davis L. Rev., 1149, 1153. [↑](#footnote-ref-12)
13. Id. at 1209-1210. See also Jack M. Balkin, *How to Regulate (and Not Regulate) Social Media*, SSRN, Nov. 20, 2019, https://papers.ssrn.com/sol3/papers.cfm?abstract\_id=3484114. [↑](#footnote-ref-13)
14. See <https://developer.twitter.com/en.html>. [↑](#footnote-ref-14)
15. See <https://rtweet.info/>. [↑](#footnote-ref-15)
16. This analysis was also performed using other terms including “shadowban”, “shadow ban”, and “censorship social media”; however, after removing retweets, these terms returned only a few hundred results each. With access to larger samples, the analysis reported here could be easily extended. [↑](#footnote-ref-16)