Research Note

Research note: Examining potential bias in large-scale censored data

We examine potential bias in Facebook's 10-trillion cell URLs dataset, consisting of URLs shared on its platform and their engagement metrics. Despite the unprecedented size of the dataset, it was altered to protect user privacy in two ways: 1) by adding differentially private noise to engagement counts, and 2) by censoring the data with a 100-public-share threshold for a URL's inclusion. To understand how these alterations affect conclusions drawn from the data, we estimate the prevalence of fake news in the massive, censored URLs dataset and compare it to an estimate from a smaller, representative dataset. We show that censoring can substantially alter conclusions that are drawn from the Facebook dataset. Because of this 100-public-share threshold, descriptive statistics from the Facebook URLs dataset overestimate the share of fake news and news overall by as much as 4X. We conclude with more general implications for censoring data.

Authors: Jennifer Allen (1), Markus Mobius, (2) David M. Rothschild, (2) Duncan J. Watts (3,4,5)
Affiliations: (1) Sloan School of Management, Massachusetts Institute of Technology, USA, (2) Microsoft Research, USA, (3) Department of Computer and Information Science, University of Pennsylvania, USA, (4) Annenberg School of Communication, University of Pennsylvania, USA (5) Operations, Information, and Decisions Department, University of Pennsylvania, USA
How to cite: Allen, J., Mobius, M., Rothschild D. M., & Watts, D. J. (2021). Research note: Examining potential bias in large-scale censored data. *Harvard Kennedy School (HKS) Misinformation Review*, *Volume 2(*4).
Received: May 3rd, 2021. Accepted: July 1st, 2021. Published: July 26th, 2021.

Research questions

- How do estimates of the prevalence of "fake news" on Facebook differ when measured in a large, censored dataset compared to a smaller, representative one?
- In general, how does censorship of large-scale data bias the key conclusions drawn from the data?

Research note summary

- We estimate the percentage of clicks on Facebook in the U.S. that are not-fake news, fake news, and not-news using the Facebook URLs dataset, which censors URLs with fewer than 100 public shares. We then compare these results to those from an uncensored, representative dataset from Nielsen's desktop web panel. We find that Facebook's URLs dataset overestimates the relative share of news clicked on the platform by nearly 2X and fake news on the platform by 4X.
- Further matching between Facebook and Nielsen's data, a CrowdTangle investigation, and an internal

---

Facebook investigation show this overestimation is likely due to the 100-public-share threshold.
- This work demonstrates how "big data" can still lead to biased estimates if it is censored. In particular, this note shows that researchers working with the Facebook URLs dataset should be aware that the 100-public-share threshold can dramatically affect even basic descriptive statistics.

Implications

Private companies increasingly control vertically integrated platforms, which include both the production and consumption of information within their walled garden. Their data are critical to understanding news, advertisement, and purchase behavior; thus, academic researchers and data journalists are eager for these companies to release data. Social Science One has piloted a new program that provides approved researchers access to such datasets (King & Persily, 2018). Their first dataset includes all URLs on Facebook from January 2017 to July 2019 that were shared with the "public" setting by 100 or more users as well as accompanying interaction metrics like clicks, views, and shares. This article explores the value of that data: how effective it is for key social science questions, and what we learn about the impact of its privacy measures.

For privacy and business reasons this data is altered from the full scope of URLs shared on Facebook in two respects: first, by applying methods from differential privacy to the URL engagement counts; and second, by censoring the data to include only URLs that have been publicly shared at least 100 times. Much has been made of the first of these precautions (Evans & King, 2020; Gibney, 2019; Mervis, 2019; Messing et al., 2020), which was one of the first empirical applications of what was previously largely a theoretical idea (Dwork, 2008). In short, Facebook added noise to key values in order to ensure that while aggregated statistics are correct, it is impossible to identify individual user-level behavior from the data. However, the second precaution, censoring URLs below the 100-public-share threshold, has drawn less attention.

We find that censoring can substantially alter conclusions that are drawn from that data. First, we show that the percent of news and fake news on Facebook, key findings of interest to social scientists (for example, Guess et al., 2021), are both inflated compared to estimates from a representative dataset. We then provide evidence that this distortion likely comes from censoring. We take a random selection of URLs from the representative dataset and show that fake news URLs are much more likely to exist in the censored Facebook dataset than other URLs. Data from CrowdTangle confirm that fake news is more likely to be shared publicly than other types of URL content, perhaps due to its novelty or the distribution strategies of fake news producers. An internal investigation by Facebook supports our finding, showing that false news URLs are more likely to be shared publicly compared to other types of URLs, potentially introducing bias in the dataset (see Appendix 1).

This bias in the dataset is not limited to just news domains. An analysis of domains across a variety of other categories also found that while news and entertainment domains were more likely to be overrepresented in the Facebook URLs dataset, other types of URLs—e.g., retail, social-media, and gaming—were likely to be *underrepresented* (see Appendix 2). These biases offer insight into the relationships between viewing, clicking, and sharing (either publicly or privately) URLs on Facebook and how those relationships might differ depending on content type. Content that is likely to be shared conditional on being viewed or clicked, and conditional on being shared, be shared publicly, is overrepresented in the Facebook URLs dataset. Some content, like fake news, is optimized both to be clicked, since it is likely to be novel and have click-bait headlines, and to be shared publicly, since it is likely meant to draw others to engage. On the other hand, retail ads are optimized to be clicked due to personalized targeting, but not to be shared publicly, since they might contain private information that users would not want to disseminate. For example, micro-targeted Facebook ads containing links to

political fundraising would likely be underrepresented in the Facebook URLs dataset, despite being of interest to researchers, because those types of URLs are unlikely to be shared publicly. Beyond just the questions of censorship in the Facebook URLs dataset, this research opens up many more questions into the relationships between viewing, clicking, and sharing URLs that are important avenues for future research.

We offer some potential suggestions to address the pitfalls of these biases. A better approach would be to eliminate censoring and rely entirely on differential privacy methods to avoid reidentification of individuals in the dataset. If the mere inclusion of a URL might violate privacy, e.g., if it is a link to a private photo album, another approach would be to provide the domain name but censor the full path of URLs below a certain threshold. This workaround would prevent the size of the dataset from exploding, while still allowing for detailed analysis at the domain level. Lowering the threshold may not be as effective of a strategy, since the CrowdTangle data suggested a big jump in fake news as a share of content from 0 to 1 (although the internal Facebook evaluation with false news URLs showed a jump closer to 25). Thus, while we appreciate that any lower threshold reduces bias, it is possible that bias would not be significantly reduced until the threshold was removed entirely.

When using the Facebook URLs dataset, researchers should provide compelling evidence that they are not comparing consumption across URL types that are likely to have different rates of public sharing. For example, comparing rates of fake news consumption across countries is likely to be biased, since fake news URLs in larger countries like India are much more likely to reach the 100-share threshold for inclusion than fake news URLs in smaller countries. Additionally, content that is mostly distributed through ads—like retail content, political donations, or inorganically-distributed disinformation—is not well-represented in the Facebook URLs dataset due to low likelihood of public sharing. On the other hand, organically shared hyperpartisan and fake news in the U.S. is likely to be captured in the Facebook URLs dataset and can be examined in isolation.

Despite the limitations we identify, the URLs dataset produced by the Facebook and Social Science One teams is unique in its size and richness. This article is not a critique of the larger effort, but rather is intended to probe the limits of the current methodology and raise awareness of potential bias to practitioners handling not only the Facebook dataset, but any large dataset with similar censorship issues.

Findings

Finding 1: Facebook URL dataset analysis

We matched our set of not-fake (i.e., credible) news and fake news domains to the Facebook URLs dataset and aggregated the number of clicks by U.S. users to not-fake news, fake news, and not-news domains. From 1/2017–12/2018, an average of 12% of clicks per month went to fake news content, 32% went to not-fake news content, and 56% went to other, not-news content. The results are shown in Figure 1. The estimates for both non-fake news and fake news are much higher than other prevailing estimates (Allen et al., 2020; Grinberg et al., 2019; Guess et al., 2019).
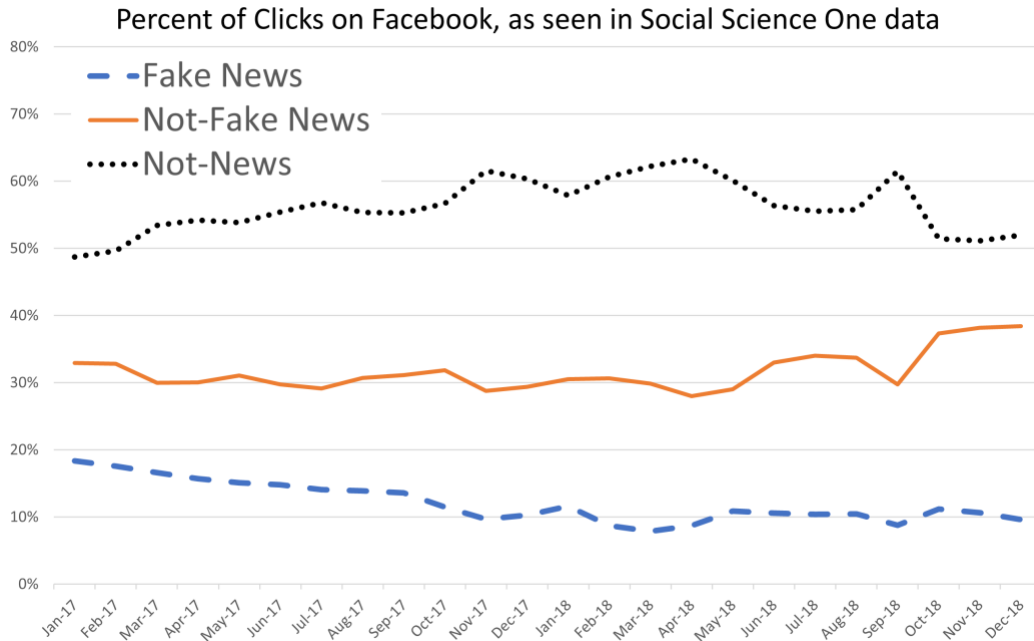
Figure 1. Estimated percentage of clicks on Facebook for fake news, not-fake news, not-news using the raw Facebook URLs data from January 2017 through December 2018.

To investigate this discrepancy, we picked the month of December 2018 to do cross-dataset analysis with the Nielsen dataset (we chose this month because it was the last month for which we had access to both datasets, although our high-level findings are robust to the month picked). In that month of the Facebook dataset, we found that 10% of clicks went to fake news, 38% of clicks went to not-fake news, and 52% of clicks went to not-news sites (shown in Column B in Table 1). We then took a random sample of 1,000 URLs from Nielsen visited by our panelists, clicked from Facebook newsfeed, during the same month. Unlike the Facebook dataset, we observed all the URLs that people visit rather than a censored version of URLs with greater than 100 public shares. We then processed those URLs in the same format as the Facebook dataset and categorized them using the same domain lists that we used for the Facebook URLs. Following this procedure, we found that 2.5% of clicks were to fake news, 24% of clicks were to not-fake news, and 74% of clicks were to not-news sites (shown in Column A in Table 1). These results are starkly different from those estimates found using the Facebook URLs data. In particular, the Facebook URLs data seem to be overcounting news: an almost 4X difference in the amount of fake news and 1.7X difference in the amount of not-fake news.

We hypothesized that this difference in estimates was due to the 100-public-share threshold for inclusion in the Facebook URLs dataset. To test our hypothesis systematically, we attempted to find matches for all the URLs in our Nielsen dataset in the Facebook URLs dataset, with the assumption that those URLs with matches had met the 100-public-share threshold and those without matches did not. We found that 84% of fake news URLs in our Nielsen sample were also present in the Facebook dataset, but only 50% of not-fake news URLs and 23% of not-news URLs were. Fake news was thus 4X more likely to be found in the Facebook dataset than not-news URLs, and 1.7X more likely than not-fake news. When we recalculated the proportion of fake news, not-fake news, and not-news using this matched dataset— dropping the Nielsen URLs without Facebook matches—we found that 7% were fake news, 39% were not-fake news, and 55% were not-news. These estimates, in Column C of Table 1, are much closer to those of the Facebook URLs dataset shown in Column B of Table 1, providing support for our hypothesis that the 100-public-share threshold adds bias to the estimates.

Finding 2: CrowdTangle analysis

To further examine whether these differences in match rates were the result of differential public sharing rates vs. other behaviors, we used the CrowdTangle API to query the number of CrowdTangle public shares for each URL in our dataset since January 2017. Unfortunately, CrowdTangle has a more restrictive definition for public shares than Facebook does. CrowdTangle share counts are limited to only those articles shared by public pages, groups, or users with "public" profiles, who usually are public figures like politicians or celebrities. However, we can still use this data as a benchmark, since the number of CrowdTangle public shares is a lower bound for the number of total public shares.

The results from this analysis are shown in Figure 2 and summarized in Table 1, Column D. In short, despite the difference in the definition of public shares, we find results that are very similar to those presented using the Facebook dataset. Above the 100 CrowdTangle share threshold, 7% of clicks were to fake news sites, 42% of clicks were to not-fake news sites, and 51% of clicks were to non-news sites. As Figure 2 shows, this difference in results was driven by a large spike in not-news URLs with 0 CrowdTangle shares. These include the long tail of ad clicks, personal links, and approximately 30% of not-fake news URLs.

While most of the change occurred from 0 to 1 share, the percent of all URL content that was fake news also increased from 5.1% to 7.4% –a 45% increase–as the threshold changed from 1 CrowdTangle share to 100. Similarly, the percent of *news* that was fake went from 11.5% to 15%, a 36% increase. These results show that fake news URLs have a disproportionately high percentage of public shares from pages or verified profiles, which supports the anecdotal evidence suggesting that fake news peddlers use public pages as a distribution channel (Legum, 2019; Roose, 2020).
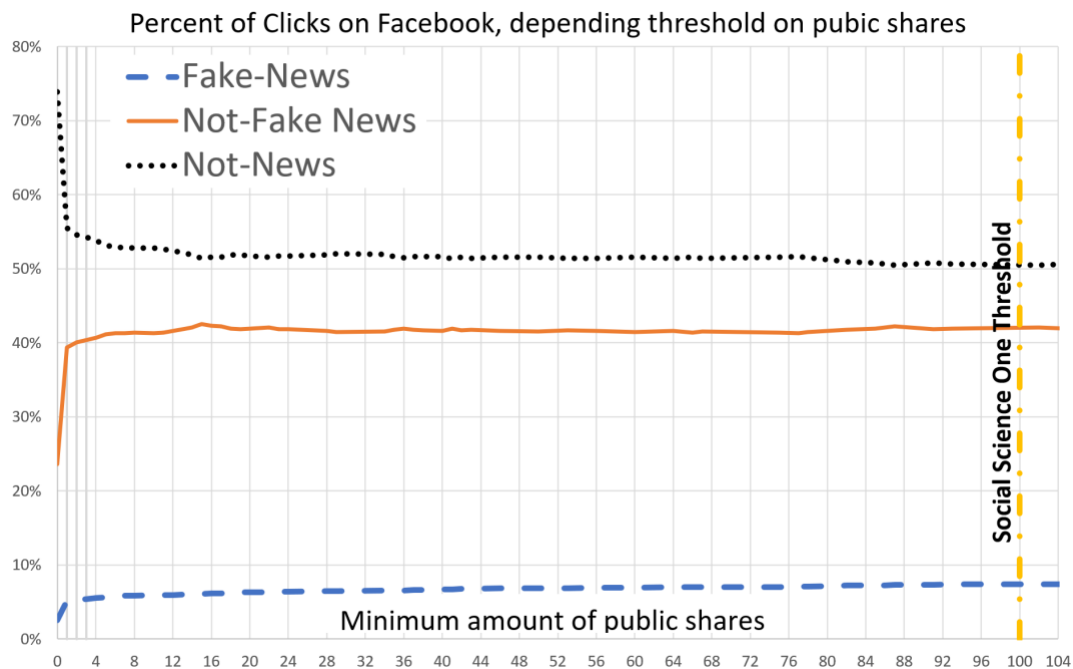

Figure 2. Estimated percentage of clicks on Facebook for fake news, not-fake news, and not-news using the Nielsen data, as threshold moves from 0 to 100 CrowdTangle public shares. Note that CrowdTangle public shares only include shares by public pages and profiles, a more restrictive definition than Facebook employs.

Table 1. *Estimated percentage of clicks on Facebook for fake news, not-fake news, and not-news under various data and assumptions in December 2018.*

	(a) Nielsen: *from Facebook*	**(b) Facebook URLs** dataset	**(c) Nielsen:** *in Facebook URLs Dataset*	**(d) Nielsen:** *>100 CrowdTangle Public Shares*
Fake-News	2.5%	10%	7%	7%
Not-Fake News	24%	38%	39%	42%
Not-News	74%	52%	55%	51%

Methods

To define not-fake news and fake news, we used the same domains as in Allen et al. (2020). The list of not-fake news sites was composed of more than 9,000 websites that primarily covered "hard" news topics like politics, business, and U.S. and international affairs. The list of fake news sites was composed of the 624 websites previously identified by researchers, professional fact checkers, and journalists as sources of fake, deceptive, low-quality, or hyperpartisan news. All other domains were classified as not-news domains.

Facebook Social Science One Dataset: In 2020, Facebook released a massive dataset of URLs (Messing et al., 2020). The first component was 38 million+ URLs that were shared from 1/1/2017 to 7/31/2019, as well as various descriptive characteristics like third-party fact-checker ratings. The second component contained over 10 trillion cells with aggregate consumption data like views and clicks for each URL for different combinations of demographic data. Facebook de-duplicated each action such that the engagement metrics reported are not the total number of clicks (or shares, etc.), but rather the total number of users who clicked (or shared) the URL. For example, a given row might describe the number of times a particular URL was clicked, shared, and viewed by women in the U.S., aged 18–24, who lean conservative, in January 2018.

Facebook also added differentially private noise to the engagement-related columns of the dataset. While this noise can change the results of many statistical procedures, the sums of differentially private columns are unbiased estimates of the true sums and thus, we did not do any further corrections in our analysis.

Additionally, Facebook restricted URLs in this dataset to those that have been shared publicly by 100 different users (plus Laplace[5] noise to minimize information leakage). Facebook defines a public share as a user or page post of any link with the privacy set to "Public" as opposed to "Friends," "Only Me," or a custom audience. While Facebook offers documentation about their privacy options, they do not provide statistics about how much content on the site is shared with the "public" setting. However, the default for posting is "Public" until the user changes their settings.

Facebook implemented this 100-public-share threshold in an attempt to further ensure the privacy of users. Many links were created by users to share private material with a small group of friends and not to broadcast information publicly (e.g., a photo album or a personal PayPal link). This could include private links, but also technically public links, which maintain privacy through low engagement (i.e., security through obscurity). Thus, a 100-public-share threshold was a heuristic to ensure that these personal links were not exposed to researchers since these kinds of links were likely to be posted with more restricted sharing settings by a limited number of people.

As a comparison, we used data from a nationally representative desktop web panel from the company Nielsen from 2016 to present. The data included every URL that each individual panelist visited as well as

the time of visit. We used this data to identify URLs that panelists clicked on while on Facebook's newsfeed by examining URLs that included the "fbclid" parameter. In late 2018, Facebook added this parameter to URLs shared on the platform, such that when a user clicks a URL from Facebook (but not Facebook Messenger), the URL they navigate to takes the form "www.domain.com?fbclid=XXXX".

Unlike the Facebook URLs data, the Nielsen data had neither noise nor censorship and had data from a demographically representative random subset of users. There were several more key differences. First, the Facebook URLs dataset did not include clicks from Facebook Messenger, while the Nielsen data captured all activity from Facebook including Facebook Messenger. However, by using the "fbclid" parameter, which does not appear in URLs shared in private messages, we were able to exclude Nielsen referrals that might have come from Facebook Messenger. Additionally, one might be concerned that URLs with the "fbclid" parameter could have originated from sources other than Facebook, such as by users copy-and-pasting URLs from Facebook and sending them via email, but an analysis suggests that this concern is likely unfounded and further restricting the definition of Facebook referrals does not alter results (see Appendix 3). Second, the Nielsen data was not de-duplicated (although results are robust to de-duplicating the Nielsen data, see Appendix 3). Third, Nielsen was desktop only, while the Facebook URLs dataset included mobile.

As part of Social Science One, we gained access to the CrowdTangle API. CrowdTangle, which is owned by Facebook, tracks engagement on public content on Facebook and makes it available to journalists and researchers. Unlike the Facebook URLs dataset, this data is not de-duplicated at the user level. In particular, the API allows researchers to identify the number of shares from public pages, groups, or profiles for a given URL on the Facebook platform.

Bibliography

Allen, J., Howland, B., Mobius, M., Rothschild, D., & Watts, D. J. (2020). Evaluating the fake news problem at the scale of the information ecosystem. *Science Advances, 6*(14), eaay3539. https://doi.org/10.1126/sciadv.aay3539

Dwork C. (2008). Differential privacy: A survey of results. In Agrawal M., Du D., Duan Z., & Li A. (Eds.), *Theory and applications of models of computation (TAMC 2008)* (pp. 1–19). Springer. https://doi.org/10.1007/978-3-540-79228-4_1

Evans, G. & King, G. (2020). *Statistically valid inferences from differentially private data releases, with application to the Facebook URLs dataset* [Manuscript submitted for publication]. Harvard University. https://gking.harvard.edu/dpd

Gibney, E. (2019). Privacy hurdles thwart Facebook democracy research. *Nature, 574*(7777), 158–159. https://doi.org/10.1038/d41586-019-02966-x

Grinberg, N., Joseph, K., Friedland, L., Swire-Thompson, B., & Lazer, D. (2019). Fake news on Twitter during the 2016 U.S. presidential election. *Science, 363*(6425), 374–378. https://doi.org/10.1126/science.aau2706

Guess, A., Aslett, K., Tucker, J., Bonneau, R., & Nagler, J. (2021). Cracking open the news feed: Exploring what US Facebook users see and share with large-scale platform data. *Journal of Quantitative Description: Digital Media*, *1*, 1–48. https://doi.org/10.51685/jqd.2021.006

Guess, A., Nagler, J., & Tucker, J. (2019). Less than you think: Prevalence and predictors of fake news dissemination on Facebook. *Science Advances, 5*(1), eaau4586. https://doi.org/10.1126/sciadv.aau4586

King, G., & Persily, N. (2020). A new model for industry–academic partnerships. *PS: Political Science & Politics, 53*(4), 703–709. https://doi.org/10.1017/S1049096519001021

Legum, J. (2019, October 28). *Facebook allows prominent right-wing website to break the rules*. Popular Information Newsletter. https://popular.info/p/facebook-allows-prominent-right-wing

Mervis, J. (2019). Privacy concerns could derail Facebook data-sharing plan. *Science, 365*(6460), 1360–1361. https://doi.org/10.1126/science.365.6460.1360

Messing, S., DeGregorio, C., Hillenbrand, B., King, G., Mahanti, S., Nayak, C., Persily, N., State, Bogdan, & Wilkins, A. (2020). *Facebook privacy-protected full URLs data set* [Data set]. https://doi.org/10.7910/DVN/TDOAPG

Matsakis, L. (2019, January 23). Facebook cracks down on networks of fake pages and groups. *Wired*. https://www.wired.com/story/facebook-pages-misinformation-networks/

Roose, K. (2020, August 27). What if Facebook is the real "silent majority"? *The New York Times.* https://www.nytimes.com/2020/08/27/technology/what-if-facebook-is-the-real-silent-majority.html

Competing interests

Facebook researchers performed the analysis in Appendix Section 1. J.A. has a significant financial interest in Facebook.

Ethics

No data were collected from human subjects specifically for this project.

Data availability

All materials needed to replicate this study are available via the Harvard Dataverse:
https://doi.org/10.7910/DVN/L0ZXYE

Appendix 1: Internal Facebook investigation

An internal analysis performed by Facebook, in response to the above findings, provides further evidence that fake news was over-represented in the URLs dataset due to the 100-public-share threshold. After we shared our findings, researchers at Facebook examined the URLs labeled as false by their third-party fact-checker program from 1/1/2017-12/31/2018. Note that the Facebook team is examining individual URLs marked false, rather than publisher-level designations of fake used in the above analysis.

The Facebook team found that the number of unique false news URLs as a proportion of all unique URLs in the dataset increased by two orders of magnitude when a 100 public share threshold was implemented, going from 0.00023% with no threshold to 0.022% with one (a two-magnitude increase). Additionally, when the researchers focused on the total public URL shares (a closer analog to the click analysis that we did previously, but clicks are more common than shares) rather than the number of unique URLs, they found that the proportion of false news URLs shares doubled with the 100-share threshold, going from 0.091% with no threshold to 0.185% with the 100-share threshold. These data reveal how skewed the distribution of false news sharing really is. There are 30,500,000 public shares of the 37,860 unique false URLs, yet 30,300,000 of those public shares come from the 8,660 unique false URLs that have over 100-shares. That is, just 200,000 of the 30,500,000 million shares are from URLs shared less than 100 times publicly, even though a significant majority of false news URLs are under that threshold. A summary of these findings can be found in Table A1 below.

Table A1. Results of Facebook's internal analysis of how the number and composition of URLs in the URL dataset changes when a 100 public share threshold is applied.

	All URLs		False URLs		False to All URL Ratio	
	No Threshold	**>100-shares**	**No Threshold**	**>100-shares**	**No Threshold**	**>100-shares**
Unique URL Count	16.2 bil.	38.2 mil.	37,860	8,660	0.00023%	0.022%
Public Share Count	33.7 bil.	16.3 bil.	30.5 mil.	30.3 mil.	0.091%	0.185%

The Facebook team also explored how changing the threshold changes the proportion of fake news URLs in the dataset, the results of which are shown in Figure A1. While they did not find the sharp discontinuity from 0 to 1 that we saw in our CrowdTangle analysis (perhaps due to different definitions of "public share" for CrowdTangle vs. the Facebook URLs dataset), their analysis does show that the number of unique URLs drops off much more sharply at low thresholds (around 10 public shares) for non-false news URLs than for false news URLs.

This analysis differed from the one presented in the main text largely due to privacy and dataset limitations. Due to privacy limitations, the Facebook analysis was conducted at the share rather than click level, since Facebook could not share any data on clicks below the 100-share threshold. Additionally, they defined URLs as false on the URL level rather than domain level, since they as a rule did not want to perform analysis on the de-anonymized data based on an external research team's domain list (which could have violated privacy in unknown ways and would have to be subject to an intensive privacy review). Unfortunately, this prevents apples-to-apples comparison with the Nielsen dataset, since we do not have share data for Nielsen, nor do we have URL-level false news labels for below the 100-share hold level. However, substantively, both analyses—ours and Facebook's—ask the same question about whether fake news is overrepresented in the Facebook URLs dataset. Despite differences in methodologies, we believe that the internal results from Facebook largely replicate the findings in our analysis using external data

sources. In particular, they give credence to the claim that the 100-share threshold can introduce unexpected bias in the URLs dataset and that bias causes the share of fake news to be overestimated in the Facebook URLs dataset.
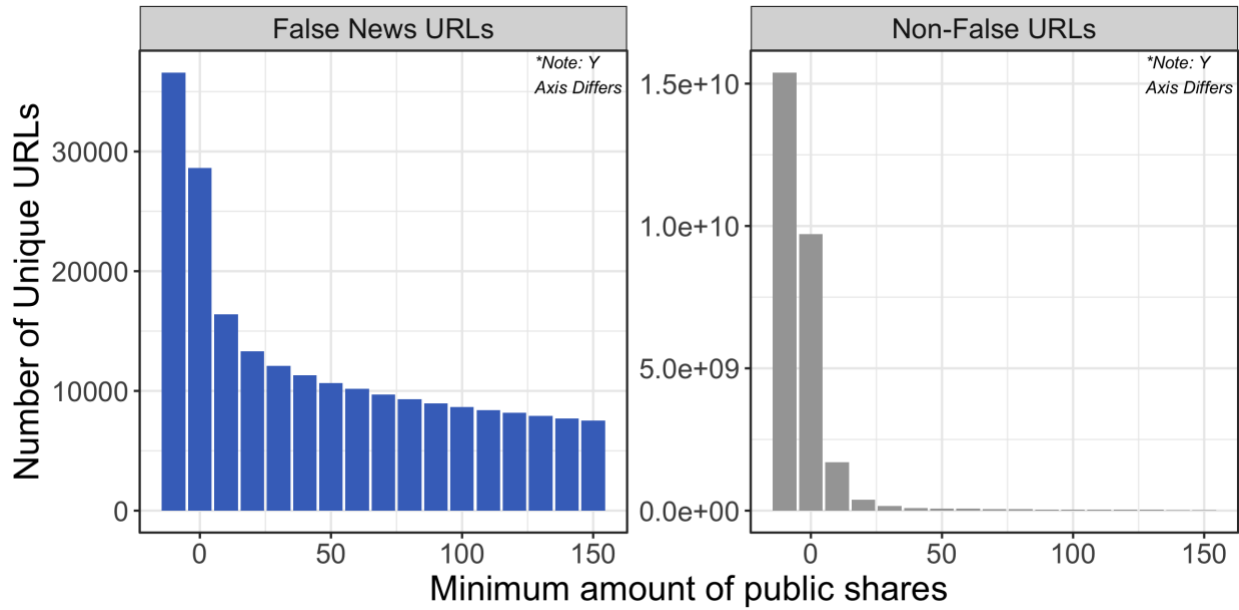

Figure A1. Number of unique URLs that are eligible for inclusion in the Facebook URLs dataset at different public share thresholds. *The threshold starts below 0 due to the Laplacian noise applied to the number of public shares. Note that the scales of the Y axes differ for false and non-false URLs.*

Appendix 2: Extension to other types of URLs

In order to expand our understanding of the generalizability of these findings for this paper, we compared the share of clicks to domains in the Nielsen vs. Facebook URLs datasets across a variety of categories, not just news. To do this analysis, we took the list of the most popular top 2000 web domains and their categories ("news," "entertainment," "retail," etc.) from ComScore (adapted from Allen et al., 2020). Note we collapsed across the "news" and "fake news" categories for this analysis and only took the top 2,000 domains, which include mostly not-fake news, so numbers differ from above). We then calculated the percentage of clicks from Facebook to domains in each respective category in the Nielsen data and the Facebook URLs dataset and compared the two. The results are in Table A2 below. We found that with this set of top 2,000 domains, news and entertainment domains have a larger share in the Facebook URLs dataset than in the Nielsen dataset. On the other hand, we found that social media, retail, financial, and gaming links have a larger share in Nielsen than in the Facebook URLs dataset. These differences have face-value validity. For example, retail URLs (e.g., amazon.com, walmart.com) are likely to originate from Facebook ads and thus unlikely to attract many public shares. In fact, it is plausible that retail URLs could attract many clicks and zero public shares, since ads are personalized based on browsing history and behavior that consumers might want to reshare to a wide audience. However, both news (e.g., foxnews.com) and entertainment URLs (e.g., buzzfeed.com, ranker.com) are designed to be shared publicly and conditional on being shared, be clicked on. Thus, it is unsurprising that these kinds of URLs are overrepresented in the Facebook URLs dataset.

Table A2. Estimated percentage of clicks on Facebook for different categories of URLs, comparing the Nielsen and Facebook URLs dataset.

Category	Nielsen	Facebook URLs	Ratio: Facebook / Nielsen
entertainment	5.1%	7.0%	1.382
financial	1.6%	0.4%	0.247
gaming	2.9%	0.1%	0.025
lifestyles	3.8%	3.4%	0.893
news	14.9%	33.3%	2.228
other	51.3%	49.2%	0.961
retail	7.7%	1.0%	0.134
search	1.3%	0.6%	0.467
services	2.3%	1.4%	0.617
social-media	9.3%	3.3%	0.350

Appendix 3: Robustness to deduplication and referral *specifications*

A) Deduplication

One distinction between the Facebook URLs dataset and the Nielsen dataset is that the Facebook URLs dataset deduplicates engagement such the dataset only records a single action per user per URL, regardless of whether the user shared, clicked, or viewed the URL multiple times, whereas the Nielsen dataset does not deduplicate and tracks all clicks to the same URL by a given URL. Thus, the Facebook data is undercounting some of the engagement on URLs in the Facebook URLs dataset, which could affect the way that these numbers are interpreted (e.g., if fake news has a much higher click or share count in reality due to inauthentic activity or click farms).

For robustness, we decided to examine the number of repeated clicks in the Nielsen dataset. 93% of URLs from Facebook are clicked only one time per user. And, most critically, almost all (>99%) of the double-clicks are on non-news (neither fake or real) URLs and are to "evergreen" links gaming or weather. This does not affect the overall results, and convinces us that the reverse issue of Facebook missing sharing of fake news by only counting unique shares, is not an issue. We also redid our primary analysis and found no difference when using the deduplicated URLs (see Appendix 3C).

B) Referral specifications

One potential issue with our analysis is in the way we defined clicks from Facebook as URLs that contained the "fbclid" parameter. However, it is possible that URLs with the "fbclid" parameter are actually URLs that have been copy-pasted and sent to the user via different platforms (e.g., email) and not just Facebook. In order to assess the impact that this definition of Facebook clicks had, we examined the Nielsen browning data and found that 60.6% of URLs with "fbclid" in the title are immediately preceded by a click from "facebook.com". We suspect the clicks not immediately preceded by "fbclid" are more likely due to browser / tab switching away from facebook.com and then back to an earlier tab without reloading rather than users clicking to an "fbclid" from an email or messaging site, since we do not see any consistent pattern of URLs with "fbclid" coming from email or messaging clients.

We also redid our primary analysis and found no difference when using a stricter definition of Facebook clicks (see Appendix 3C).

C) Repetition of main analysis

We re-calculated the percentage of fake news, not-fake news, and real news on Facebook, based on the issues raised above. In particular, we 1) de-duplicated clicks from Facebook at the user-URL level and 2) restricted to the set of URLs that contained the "fbclid" parameter and were immediately preceded by a visit to "facebook.com." We found almost identical results to our main text, with 2.6% of clicks going to fake news, 24% of clicks going to not-fake news, and 73% of clicks going to non-news. This is almost identical to the 2.5% fake news, 24% not-fake news, and 74% not-news reported in our main text.