Private Sharing of Public Content on Facebook

Kiran Garimella

Rutgers University
kiran.garimella@rutgers.edu

Abstract

In many social networks, there are public and private spaces (e.g. public Facebook pages/groups, and private groups). Content could start spreading on private spaces and spill over to the public parts (or vice versa). Previous studies have acknowledged the potential for content to transition between public and private domains, but a quantitative assessment of such spillover has been lacking. In this paper, we obtain aggregate counts of shares of news content in public and private spaces on Facebook. By comparing the ratio of private shares to the total shares of a piece of content, we provide a first look at information dynamics in private spaces, and to what extent researchers might be missing data if focusing only on public parts of a social network. Through data collected from a diverse set of sources, our study reveals that a significant portion of prominent news content is shared within private networks. We also demonstrate significant heterogeneity in private sharing patterns, and influenced by the topics covered. This study offers valuable insights into the quantitative dynamics of private content sharing for the first time.

1 Introduction

In social media, the dichotomy between public (e.g. pages from news organizations/politicians) and private (e.g. private groups, sharing with friends) spaces presents a compelling and yet largely unexplored frontier for research. Platforms such as Facebook, with its sprawling public pages and groups, offer an enormous treasure trove of data, rich in its potential to inform everything from marketing strategies to policy decisions. However, these public spaces are merely the tip of the iceberg. Beneath the surface lies an intricate web of private sharing of content produced in these public spaces–closed groups, direct messages, and personalized feeds–that are largely inaccessible to external researchers but are equally, if not more, influential in shaping public opinion and discourse.

Imagine the lifespan of a news article on social media platforms like Facebook. Initially, the article is shared by a major news organization on a public Facebook page. From there, users engage with it—liking, commenting, and sharing it further, both in public groups and privately with their friends. Given a news article, it is currently challenging to determine how many of its shares occur in public spaces versus private ones. Furthermore, how does this distribution of public and private shares vary based on the quality of the news source, or the topic of coverage? This paper seeks to address these questions.

The absence of a robust, quantitative understanding of these private spaces is not merely a gap in academic literature; it's a gaping hole in our understanding of contemporary society. The dynamics of private sharing are critical in shaping real-world outcomes, from elections to public health. For instance, while public forums are instrumental in the mass dissemination of information, the discussions and content sharing that happen in private spaces often form the bedrock of individual opinions. Whether it's the spread of misinformation, the organization of social movements, or the propagation of hate speech, the private spheres play an indispensable role.

Currently, the platforms are the sole entities with access to data on private sharing patterns. As a result, little is known about how public information spreads and the extent to which it is shared privately. This limitation is understandable, as such data is rightly inaccessible to researchers due to privacy concerns. Moreover, platforms have little to gain and much to lose by opening up their private spaces to public scrutiny. Whether it's the risk of exposing vulnerabilities that could be exploited by malicious actors or simply the fear of bad press, the gatekeepers of these platforms have historically been reluctant to provide the level of access that a comprehensive study would necessitate.

The limitations of existing research tools have further constrained the scope of prior work. Tools like CrowdTangle,[1] for example, focus exclusively on public interactions on Facebook, providing a skewed representation that fails to account for the complexity and nuance of human interaction on social media platforms. Previous academic focus has been disproportionately aimed at more transparent platforms like Twitter or Reddit, where public data is more readily accessible. This has resulted in a lopsided, incomplete understanding of social dynamics online. There are probably thousands of studies using Twitter and Reddit data and hundreds

[1] https://crowdtangle.com/ - a tool which provided public Facebook data for researchers. CrowdTangle was shut down in August 2024.

using CrowdTangle, the only source to study Facebook.[2]

Our research aims to break new ground by studying the prevalence of public content shared in private spaces. Utilizing CrowdTangle, we obtain aggregated metrics on private interactions of public content, such as URLs of news articles. We furnish, for the first time, quantitative metrics elucidating the scale at which a news URL is shared in both public and private networks. Public counts are sourced directly from CrowdTangle and represent interactions on public pages and groups. Private counts, on the other hand, encompass all other Facebook interactions, even extending to private shares via Facebook Messenger. These private counts are accessed via CrowdTangle's browser extension, CrowdTangle Link Checker (Facebook 2023).

The main metric we are interested is the ratio of counts of content shares privately vs. publicly, i.e., the ratio of private shares to total shares (public + private) for various content types. Through this metric, we offer insights into the dynamics of private sharing of various types of URLs. We compiled a list of URLs from a wide range of news sources, including mainstream outlets such as The New York Times and Fox News, as well as partisan publications like Breitbart and Newsmax. Additionally, our dataset includes topic-specific news from the COVID-19 period and longitudinal data from Huffington Post.

Our findings reveal stark disparities in the sharing behavior across different types of content and sources, from mainstream news outlets to fringe groups. Over 50% of shares from mainstream news outlets occur in private spaces. There is notable variation in the types of topics shared publicly versus privately. Content related to identity groups and conspiracy-prone subjects, especially those framed in a negative tone, are predominantly shared in private spaces. This highlights the critical role of these secluded environments in shaping public opinion on sensitive issues.

Relevance to the workshop: Our paper is well aligned with the DETOX workshop's focus on countering the degradation of social media, particularly through its contribution to understanding hidden dynamics within private digital spaces. By quantitatively analyzing the spillover of content between private and public parts of Facebook, we uncover a critical yet often overlooked layer of information diffusion. This is especially important in the current research landscape, where data from private spaces is largely inaccessible through public APIs, and even public data is increasingly at risk—exemplified by the shutdown of tools like CrowdTangle that once enabled transparency.

2 Background and Datasets

2.1 Private content on Facebook

We first define what private content actually means on Facebook. CrowdTangle is a 'public insights tool from Meta that makes it easy to follow, analyze, and report on what's happening with public content on social media.' According to the FAQ page,[3] "CrowdTangle tracks influential public accounts and groups across Facebook, Instagram, and Reddit, including all verified users, profiles, and accounts like politicians, journalists, media and publishers, celebrities, sports teams, public figures and more. . . . CrowdTangle does not track any private accounts. . . . CrowdTangle's database currently includes (on Facebook): 7M+ Facebook pages, groups, and verified profiles. This includes all public Facebook pages with more than 50K likes (automated via API), all public Facebook groups with 95k+ members, all US-based public groups with 2k+ members, and all verified profiles." On top of this information, users can add public pages and groups (of any size) manually and retrieve all the messages shared in these pages/groups. Once any CrowdTangle user adds a page/group, it becomes indexed and searchable for all other CrowdTangle users. However, there is no catalog of which pages were added by the user and which were included by default according to CrowdTangle's definition of 'public' content (defined in the previous sentence).

So to be conservative, it might be safe to assume that content not present on CrowdTangle, which we define as 'private content' is content shared in public pages which have lesser than 50k likes, public groups which have less than 95k members, private pages/groups, individual accounts, and on direct messages sent through Facebook Messenger. The data we obtain from CrowdTangle link checker (Facebook 2023), which contains private counts is defined as "...the number of interactions (including reactions, comments, and shares) that a link has received in all of the Facebook posts that have linked to it. This number is NOT limited to CrowdTangle data, and includes all Facebook posts that have linked to this URL, whether public or private."[4] Although CrowdTangle was originally designed for public content tracking, we also relied on its Link Checker extension to estimate private share counts. However, this approach does not give us direct access to private messages or closed-group activity and should be interpreted as a coarse approximation.

2.2 Datasets

We collected a wide range of datasets, containing news urls spanning popular news sources, content covering different topics, and spanning a wide time frame. For each dataset, we used the CrowdTangle API[5] to obtain the public share count and the CrowdTangle Link Checker Chrome extension (Facebook 2023) to obtain the total share count (including the public and private count). Our unit of analysis is a web link, primarily because links offer a standardized and easily trackable form of content. Furthermore, using links aligns well with CrowdTangle's API capabilities, thereby ensuring compatibility and ease of data retrieval.

COVID News. Our first dataset contains COVID News obtained from a news aggregation company Aylien,[6] and is a comprehensive collection of articles sourced from an archive of 1,673,353 articles from 440 global news outlets–

[2]The exact number of studies using datasets from these platforms maybe hard to count, but just to get a sense of the scale, using Google Scholar for search terms related to Twitter/Reddit/CrowdTangle return thousands of hits.

[3]https://bit.ly/3VpIl7w

[4]https://bit.ly/4ioSgEj

[5]https://github.com/CrowdTangle/API/wiki

[6]https://aylien.com/resources/datasets/coronavirus-dataset

predominantly from the US and UK during COVID (Nov 2019-Jul 2020). We randomly sampled 20,000 articles from this corpus in this study.

Mainstream News. Next, we collected data from four mainstream news sources spanning the political spectrum: New York Times (center-left), Fox News (center-right), Breitbart (right), and Newsmax (far-right). For each website, in December 2022, we obtained the sitemap and obtained a list of all the URLs published by these websites during that month, which gave us a total of 28,332 URLs.

Huffington Post. We obtained a large dataset of 210,000 URLs from Huffington Post (Misra and Grover 2021), spanning from 2014 to 2021. Apart from the urls, the data includes the categories the news belongs to, such as politics and entertainment, providing us the opportunity to explore how different subject matters exhibit unique sharing behaviors over an extended period. We sampled 20,000 URLs from this dataset at random for the study.

The complete dataset with the URLs and the public private share counts will be released upon the publication of this paper. CrowdTangle was shutdown in August 2024 limiting further research on this topic but we hope our dataset containing aggregated exposure counts will help spur discussion on the importance of such tools for research and public accountability of social media companies.

3 Analysis

The key measure we use in our study is the ratio of public to total shares (public + private) for each link, which we denote with `Ratio`. This ratio serves as a robust indicator of the extent to which content is being disseminated within private networks. `Ratio` nearing zero signals that a piece of content is predominantly shared in private spaces, whereas one approaching one suggests the converse–widespread sharing in the public. By examining this ratio across various topics and content types, we offer a quantitative insight into the often-overlooked realm of private sharing and its implications for both research and public discourse. We first plot the Cumulative Distribution Functions (CDF) of `Ratio` for various datasets. The CDF measures, on the y-axis the fraction of URLs which have a ratio less than a certain x value.

3.1 Mainstream news

Our first line of analysis looks at the sharing patterns of mainstream news outlets. Figure 1 shows that the CDF of mainstream news outlets like The New York Times and Fox News shows a striking similarity: nearly 70% of their URLs have a `Ratio` of less than 0.2. This suggests that only about 20% of shares from these outlets are public, indicating a prevalent trend of private sharing. On the other hand, biased outlets like Breitbart and NewsMax exhibit a substantially different pattern. Contrary to expectations that politically charged content might be shared more privately, we found that a majority of URLs from these sources (70%) are shared publicly at least 50% and 70% of the time, respectively. This counterintuitive could probably be due to Facebook's algorithmic changes preventing the amplification of low quality, biased or polarizing content (Wagner 2017). Additionally,

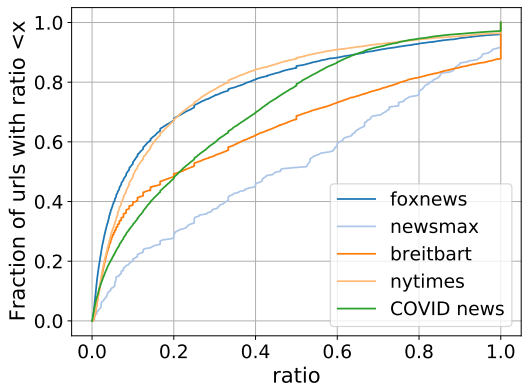

Figure 1: Cumulative Distribution Function (CDF) of `Ratio` for mainstream news sources.

we note that a small but notable percentage of shares for NewsMax (5%) and Breitbart (7%) are exclusively public, confined to posts by their official Facebook pages.

The divergent patterns between mainstream and biased outlets may also be influenced by audience size. For instance, The New York Times and Fox News (in May 2024) have more than 18 million likes on their Facebook pages, while Breitbart and NewsMax lag behind with 4.4 million and 3.5 million likes, respectively. Though audience size might play a role, it may not be a reason why we observe this pattern as this is not consistently supported across all types of news sources. This inconsistency indicates that other factors, possibly the nature of the content or the demographic characteristics of the audience or the activity of the news source on Facebook, might also play a role.

3.2 Category of the news

Next, we focus on whether certain categories of news have different `Ratio` values. To study this, we use data from Huffington post (Misra and Grover 2021), where the category of the news article was also included. The category here indicates the section in which the news article was published, e.g. https://www.huffpost.com/news/politics, https://www.huffpost.com/entertainment, https://www.huffpost.com/voices/, etc. Figure 2 shows the relation between news categories and their respective sharing ratios. We can clearly see a wide range of heterogeneity in the `Ratio`distribution with certain categories demonstrating a propensity towards public/private sharing.

3.3 Topics covered vs. `Ratio`

Next, we looked at sharing patterns of URLs and the actual content within the articles. Using the COVID-19 dataset which also contained the text of the articles, we obtained topics from the articles using Latent Dirichlet Allocation (LDA) (Blei, Ng, and Jordan 2003), an unsupervised model that extracts topics from the text of articles. Using the coherence score method (Stevens et al. 2012), we determined that 40 topics was ideal for our dataset and manually annotated the topics. Each URL was associated with a predominant topic based on the LDA results in order to discern any po-

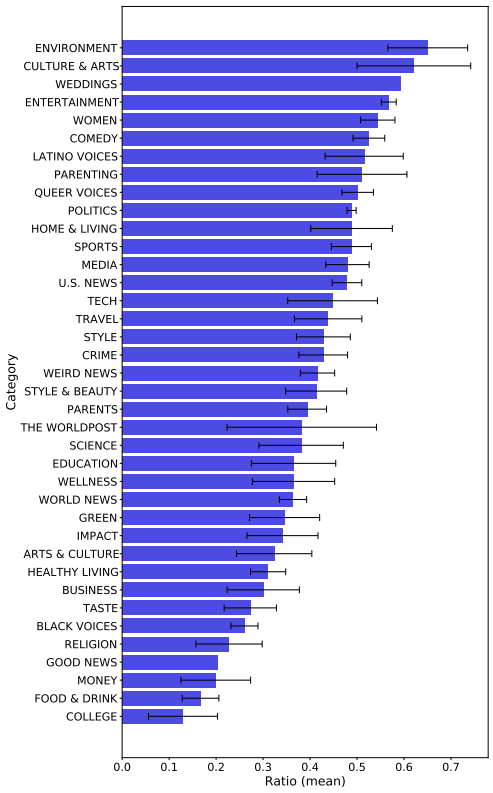

Figure 2: Mean ratio (along with 95% confidence intervals) for various categories of Huffington Post articles.

tential patterns or correlations between the frequency of a topic's occurrence and its `Ratio`. To do this, we first partitioned the ratio into quintiles,[7] and assess whether certain topics were disproportionately represented within each quintile bucket. Particularly, we look at topics which are highly popular in the lower quintiles (highly private) and not so in the higher quintiles (highly public) and vice versa. This helps us get topics which are comparatively more private.

Figure 3 shows the result for a selected set of interesting topics.[8] Our results indicate a clear disparity among topics in relation to privacy. Topics such as the derailment of a train near a Navy ship treating COVID patients,[9] and vaccines –both of which frequently circulate within conspiratorial discussions– along with abortion, and discussions of remote work were substantially more prevalent in the lower ratio buckets indicating that these subjects are predominantly shared within private spaces. On the other hand, topics associated with sports and NBA, news about UK politics, and feel-good content about relationships, were overly represented in the higher ratio buckets, indicating that these subjects are typically shared in a more public context. Though

[7]The results also mostly hold with the analysis done with deciles but to allow for larger ranges of what public and private buckets capture, we decided to use quintiles instead of deciles.

[8]To keep the plot legible, we only pick the top and bottom quintiles and a small set of topics. The full plot with all the topics will be available in an online appendix.

[9]https://archive.is/7mlJh

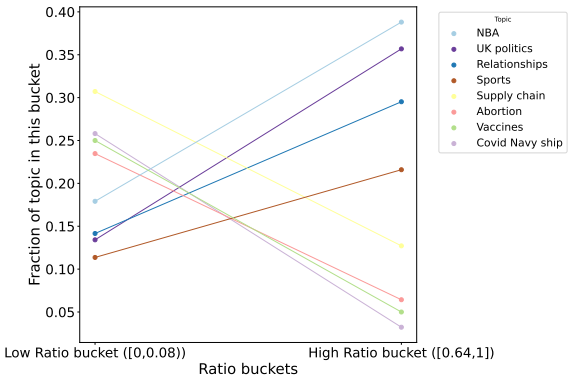

Figure 3: Topics with the highest and lowest prevalence in the top and bottom quintiles of `Ratio` distribution. Topics with high prevalence in the lowest quintile (most private) and low prevalence in the highest quintile (most public) indicate that they are disproportionately shared privately.

the differences among topics is not surprising, the prevalence of highly controversial and conspiratorial news topics in private spaces is concerning and has interesting consequences for research on such topics using public data.

4 Discussion

This research marks the first examination of the prevalence of information shared privately on social media platforms. This novel perspective unlocks new dimensions of understanding the information dynamics within these private spaces. The finding that 80% of mainstream news URLs have at least 50% of their shares privately challenges conventional research methodologies that focus solely on public data (see Fig. 1). There is a significant heterogeneity in the category of news or topics which are shared publicly (see Figs. 2, 3). We show that topics prone to conspiratorial or controversial discussions (such as abortion) are shared mostly privately.

While our study sheds light on the nuances of private and public content sharing on social media, it is constrained by several limitations. Primarily, the research is descriptive and adopts a narrow scope, focusing exclusively on one type of content (URLs) from a specific set of domains (primarily news), which represents only a fraction of Facebook activity. These constraints are further amplified by the capabilities of CrowdTangle, which restricts us from understanding the nature of privately shared content or establishing a clear sampling frame for baseline URL consumption, e.g, we cannot ascertain whether vaccine related content is privately shared due to its conspiratorial nature or for personal privacy.

Our findings, therefore, are better suited for generating hypotheses rather than providing definitive answers. Furthermore, the proprietary nature of comprehensive data limits our–and by extension, the broader research community's–ability to fully understand the dynamics of online information sharing. This underscores the pressing need for more transparent data-sharing practices between social media platforms and researchers to advance the field and foster a more nuanced understanding of these critical issues.

References

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan): 993–1022.

Facebook. 2023. CrowdTangle Link Checker. https://chrome.google.com/webstore/detail/crowdtangle-link-checker/klakndphagmmfkpelfkgjbkimjihpmkh. Accessed: 2024-01-18.

Misra, R.; and Grover, J. 2021. *Sculpting Data for ML: The first act of Machine Learning*. ISBN 9798585463570.

Stevens, K.; Kegelmeyer, P.; Andrzejewski, D.; and Buttler, D. 2012. Exploring topic coherence over many models and many topics. In *EMNLP*, 952–961.

Wagner, K. 2017. Facebook is cutting down on clickbait – again. *The Verge*.