Proceedings of the Twentieth International AAAI Conference on Web and Social Media (ICWSM 2026)

NewsDB: An Automated Approach to Build an Extensive Database of Self-Proclaimed News Providers

Salim Chouaki,[1] Minh-Kha Nguyen,[2] Laura Edelson,[3] Oana Goga,[1] Tobias Lauinger,[4] Damon McCoy[4]

[1]CNRS, INRIA, Institut Polytechnique de Paris, France
[2]Université Grenoble Alpes, CNRS, Inria, Grenoble INP, France
[3]Northeastern University, USA
[4]New York University, USA

Abstract

The credibility of news obtained online has become a concern due to the ease with which individuals or groups can claim to be news publishers and share news-related content. Unfortunately, research on monitoring misleading information in the online news ecosystem is hindered because the community lacks a comprehensive and up-to-date list of social media pages and domains claiming to be news media. This paper employs an automated approach that uses Google's GNews API and Meta's CrowdTangle API to identify self-proclaimed news providers. Our method was able to discover 19k self-proclaimed news providers in the United States active in June 2022 and 23k active in October 2020. Additionally, we retrieve the posting history (totaling 191,182,320 posts) of discovered pages. Among others, our analysis reveals that, on average, 300 new self-proclaimed news pages are created every four months, 56% of them do not declare a managing organization, 15% of the identified news pages are news aggregators, and 57% declare to be local news.

Introduction

Social media platforms have changed how users consume news and stay updated on current events, with nearly half of U.S. adults turning to social media, especially Facebook, as their primary news source (Walker and Matsa 2022). This reliance on Facebook for news brings both advantages and concerns. On one hand, it enables effortless news dissemination and democratizes access to information. On the other hand, many organizations have raised concerns about the platform facilitating exposure to misinformation (Barthel, Mitchell, and Holcomb 2016; Van Green, Ted 2020). One key enabling mechanism is the ease with which anyone can claim to be a news provider and share news content without verification. For instance, recent reports showed the emergence of organizations aiming to influence voters during elections by claiming to be news providers (Bengani 2019).

Fostering a healthy news environment requires constant monitoring and auditing of content shared by both known and less-known self-proclaimed news providers. *Unfortunately, having a comprehensive view remains challenging, as the community lacks a comprehensive and up-to-date list of self-proclaimed news providers*. Several studies (Edelson et al. 2021; Weld, Glenski, and Althoff 2021; Guess, Nyhan, and Reifler 2020; Levy 2021; Scharkow et al. 2020) rely on lists manually aggregated by (Media Bias Fact Check 2026) and (News Guard 2026), giving them a limited view of the news ecosystem. Other works build lists of Facebook news pages by expanding a known seed list using platform-provided similarity signals, such as "related pages" suggestions (Le Quéré, Chiang, and Naaman 2022) or Facebook advertising interests (Guimarães et al. 2020; Ribeiro et al. 2018). Specifically, for each page in a given seed list of news outlets, "related pages" that share similar interests, as defined by the platform, are added to the list.

These approaches (1) depend on the availability of a seed list, which may not exist in all countries, (2) expand the list based on page similarity defined by the platform, rather than on evidence that pages actively publish news content, and (3) are not designed for retrospective analysis and cannot be applied to study specific past periods, such as elections.

In this work, *we build an automated methodology to discover self-proclaimed news providers*. Our approach relies on the assumption that social media pages claiming to be (and wanting to appear as) news sources typically post about current events. Therefore, our key idea is to perform a daily crawl that: (1) exploits the GNews API (GNews 2021) to get a sample of news articles published by established media in the past 24 hours and extract a set of corresponding *keywords*; (2) uses CrowdTangle (CrowdTangle 2024a), an API provided by Meta, to search for Facebook posts mentioning these keywords in the past 24 hours; and (3) filters only Facebook pages that self-identify as news media. Note that the methodology aims to identify pages that *claim* to be news providers without judging their legitimacy according to journalistic standards. This is unnecessary from an auditing perspective, as a page does not need to be legitimate to influence public opinion.

We mainly focus in this paper on discovering news providers primarily based in the U.S. We deployed our methodology at two periods. First, we conducted a *live daily* detection in June 2022, identifying 19,590 Facebook news pages (corresponding to 8,920 unique domains). Then, we conducted a retrospective detection of news pages active in October 2020 (during the U.S. presidential election), discovering 23,992 pages (corresponding to 10,798 unique domains), including pages that have since stopped posting.

Overall, our data collection enabled the discovery of *26k+ U.S.-based self-proclaimed news providers on Facebook*.

To assess coverage and timeliness, we compare this dataset with existing lists. First, we find that our list of self-proclaimed news providers covers over 95% of the pages listed by MBFC and NG, while identifying ten times more U.S.-based news providers than these lists combined. Second, examining the rate of new discoveries per day, we observe that more than 8,000 pages are detected on the first day, over 2,000 on the second day, and approximately 250 new pages per day after the first two weeks, suggesting that one month of crawling is sufficient to reach high coverage. In addition, 90% of pages are detected within ten active days, which is important for capturing actors that are active only during specific periods.

Moreover, we collect the full posting history of each discovered page between July 2017 and July 2022, along with engagement statistics for each post. In total, this we have collected 191,182,320 posts. This constitutes a valuable dataset of Facebook news providers, and provides a unique empirical view of the Facebook news ecosystem. We analyze organizational affiliations, posting behavior, page dynamics, and engagement patterns. Our results show:

(1) *Organizational affiliation*: Our analysis reveals that 44% of identified pages mention a managing organization. We retrieved 3,043 organizations, with 406 owning multiple Facebook pages.

(2) *Posting behavior*: We find that 15% of analyzed pages are *news aggregators*. This category is crucial to scrutinize as such pages can easily be automated to promote specific agendas by re-sharing only information aligned with their motives. Furthermore, we find that 56% of analyzed self-proclaimed news pages are focused on *local news*. Local news are less likely to reach the radar of journalistic auditors and are more likely to be trusted by users (Sands, John 2019), making this an important aspect to consider.

(3) *Page dynamics*: We find that, on average, 300 new self-proclaimed news pages are created every four months in the past 15 years. Moreover, we see two prominent peaks in page creations in 2016 and 2019, potentially linked to the U.S. presidential elections.

The methodology presented in this paper was initially developed using CrowdTangle. Meta shut down CrowdTangle on August 14, 2024 (Euronews 2024) and introduced Meta Content Library and API (Meta Platforms, Inc. 2024). Therefore, we updated our implementation to use the Meta Content Library and deployed the revised methodology over one week, from August 15 to 21, 2025, across nine countries, discovering 8,503 self-proclaimed news providers: Australia (1,064), Belgium (296), France (1,528), Hungary (415), India (191), Italy (2,838), the Netherlands (337), Poland (1,092), and Romania (742).

Overall, our findings offer an empirical view of self-proclaimed news providers on Facebook, capturing their diverse levels of visibility, organization, and activity. The scale of the dataset underscores the importance of automated approaches for studying news content on social media.

Background

This section provides an overview of various aspects related to Facebook news pages, including their creation, verification, and association to web domains. Additionally, it introduces the Facebook News Page Index, an archive for pages predominantly sharing news-related content on Facebook.

Facebook Pages

Facebook pages provide a platform for individuals, businesses, and organizations to build and manage their presence on the platform.

Creation of Facebook pages. Creating a Facebook page is a simple process that only requires a personal Facebook account (Meta Business Help Center 2026c). Users can initiate page creation by visiting a designated URL (Facebook 2026a). They are required to provide a *name*, select a *category* aligning with the page's purpose, and can add additional information to enhance the page's description, including contact details, website links, and profile and cover photos. After completing these steps, the Facebook page becomes active and available for posting *without verifying the accuracy of the provided information, enabling users to claim any category, including news media.*

Domain verification. Facebook pages can link to external domains to claim they represent the corresponding domain. Additionally, they can provide strong proof for this claim by verifying the associated domain. This verification process involves adding a meta tag or uploading an HTML file to the website's root directory (Meta Business Help Center 2026a). Facebook does not mandate it and there is no visible distinction between verified and unverified domains.

Page Publishing Authorization. In response to the 2016 U.S. presidential election controversies, Facebook introduced the *Page Publishing Authorization* in August 2018 to enhance page accountability and prevent bad actors from using compromised accounts (Facebook 2018). This authorization requires the concerned pages' admins to secure their accounts with two-factor authentication (Facebook Help Center 2026b) and confirm their primary country location, with non-compliance resulting in posting restrictions (Facebook Help Center 2026a; Facebook Business 2026). It is applied for pages with a "high potential reach" in the U.S., India, Indonesia, and the E.U. (Facebook Business 2018).

Unlike verification badges, this authorization does not confirm a page's authenticity but focuses on securing admin accounts and confirming their location. Unfortunately, Facebook does not publicly disclose which pages undergo this process, and there is no public information on what qualifies as "high potential reach."

Facebook News Page Index

The Facebook News Page Index is an initiative introduced by Facebook to identify pages primarily publishing news content (Meta Business Help Center 2026b). Page admins can apply to register their pages in this index. The application requires business and domain verification (Meta Business Help Center 2026d), followed by an internal review to

ensure the page avoids misinformation, adheres to community standards, and refrains from clickbait or engagement bait (Facebook 2019; Meta Business Help Center 2026e). Pages included in the News Page Index are exempted from the ads authorization and disclaimer processes when promoting social issues or political advertisements (Meta Business Help Center 2026b). Participation in the Facebook News Page Index is voluntary. Thus, not all self-proclaimed news pages are included in this index.

Method and Data Collection

This section first describes our methodology to identify self-proclaimed news providers on Facebook. For simplicity, we sometimes refer to them as Facebook news pages. We then describe our deployments and the different data collections performed in this study.

We initially implemented this methodology using CrowdTangle, which we used to collect most of the data analyzed in this paper. We therefore present the CrowdTangle-based version of the methodology. Since Meta shut down CrowdTangle and introduced the Meta Content Library, we discuss the implications of this change and describe how we adapted our implementation at the end of this section.

Page Discovery Method

Our *key assumption* is that news providers need to publish content discussing current events to inform their audiences and maintain interactions. Although there can be specialized news websites focusing on niche topics, news organizations that can potentially influence public opinion, and which we would like to monitor and audit, need (at least at some point) to discuss current affairs.

Our approach involves three steps (Figure 1): (1) collect keywords corresponding to current events, (2) search for Facebook posts that mention these keywords, and (3) filter the resulting Facebook posts to include only those from U.S.-based pages that share content in English and claim to be news providers. Our method is designed to perform daily the following tasks:

(1) Extracting keywords corresponding to daily news. We gather popular news headlines from Google News using the *GNews* API (GNews 2021). This API provides access to top-ranked and top-ranked-by-topic news articles across eight topics: World, Nation, Business, Technology, Entertainment, Sports, Science, and Health. We extract top-ranked and top-ranked-by-topic articles across the eight categories and limit our search to articles published in English within the past 24 hours. The median number of daily news headlines retrieved for each topic is as follows: General (37), World (67), Nation (61), Business (69), Technology (71), Entertainment (67), Sports (65), Science (30), and Health (40).

Next, we employ Yake (YAKE 2025), a Python library for selecting the most important keywords in a text. For every title of a news article we instruct Yake to output the most relevant two tuples made of two or more keywords. For instance, Yake generates the tuples *"investigation into Trump"* and *"social network deal"* for an article in June 2022.[1] This step yields a daily list of tuples made of two or more keywords, with a median of 1,002 tuples generated each day.

(2) Collecting Facebook posts covering daily news. We employ the CrowdTangle API, a tool provided by Meta for academics, to search for content on Facebook (CrowdTangle 2024a). Precisely, we use the posts-search end-point that allows retrieving posts matching given parameters and search terms (CrowdTangle API 2021). For each keyword tuple obtained in the previous step, we send a request to the API and limit the search window to 24 hours. We collect all the returned posts for each request, with a median of 343k posts per day. Note that CrowdTangle returns posts from tracked pages only. The API automatically tracks all pages with more than 25,000 followers and all verified profiles, in addition to pages manually added by users (CrowdTangle 2024b).

For each post, the API returns various attributes such as the *post's text, published time, language,* and *engagement level*, along with information about the publisher, such as the *page's name, ID, verification status, category,* and *country of the page's admin*.

(3a) Category filtering. The prior step provides a list of Facebook pages discussing current news. Many of these pages do not claim to be news providers. We consider a Facebook page to be a self-proclaimed news provider if, on its About page, it has put one of the Facebook categories in Table 1.

While some categories, like "Newspaper" or "News & media websites," are clear indicators that a page claims to be a news provider, others, such as "Media" or "Show," are less specific. We opt for a broader net to ensure high coverage and avoid missing relevant pages, particularly since many news providers listed by Media Bias Fact Check or News Guard have a general "Media" category on Facebook.

(3b) Location and language filtering. We filter U.S.-based pages that share English content. Note that this method can be adapted for pages from different locations publishing in various languages.

We first enhance the attributes describing these pages by leveraging the Facebook Ad Library, a publicly accessible platform listing Facebook ads (Facebook 2026b). This library provides details about advertiser pages such as the *the name and country of the organization that manages the page and the main language used in its posts.* Each page has its dedicated web page within the Ad Library site, accessible via a specific URL format.[2] Importantly, we discover that the Ad Library provides information for all pages, including those that have never promoted ads on Facebook. We verified this with a test using a newly created page, confirming its presence in the Ad Library few days after its creation.

We use Selenium (PyPI 2025a), a Python package for automating browser interactions, to retrieve information from

[1] https://www.axios.com/2022/06/13/government-expands-investigation-trump-social-network-deal

[2] https://www.facebook.com/ads/library/?active_status=all&ad_type=all&country=ALL&view_all_page_id={page_id}

Facebook Category	MBFC & NG	Snapshot June 2022	Snapshot October 2020	Overlap
Broadcasting & media company	283	2864	3480	2561
Media	11	447	614	354
Media/news company	421	6363	7588	5346
Newspaper	448	2580	3238	2349
Newsstand	0	8	12	6
News personality	10	2747	3699	2265
News & media website	379	4412	5124	3788
Show	0	129	168	89
Social Media Agency	1	40	69	24
All pages with a news category	1553	19590	23992	16782
Other categories	1059	0	0	0

Table 1: Facebook categories related to news media and the corresponding number of pages in Media Bias Fact Check and News Guard listings, SNAPSHOT_JUNE_2022, SNAPSHOT_OCTOBER_2020, and in the overlap between the two snapshots. *Top other categories include: Nonprofit Organization, Website, Publisher, Community, Political Organization, Entertainment Website, Magazine, Interest, and Public Figure.*

each page's About section in the Ad Library.[3] We created a dedicated Facebook account for this task and implemented randomized delays of 1 to 4 seconds between each iteration to avoid bot detection.

Then, to identify U.S.-based news pages, we use two attributes: *"topAdminCountry"* provided by CrowdTangle, indicating the page's admin's country, and *"organizationCountry"* from the Ad Library, indicating the page's organization's country. We select pages where either of these attributes has *"U.S."* as a value. Finally, to identify pages sharing in English, we use the *"mainLanguage"* attribute from the Ad Library and select only pages with the value *"en."*

Datasets

We performed our first data collection in June 2022, executing the whole process once every day from June 1st to 30th, resulting in the detection of **43,436** self-proclaimed news pages. Among these, **19,590** pages are U.S.-based and primarily share content in English. We refer to this list as **SNAPSHOT_JUNE_2022**.

We conducted a second *retrospective* data collection to identify active news pages from October 1st to 30th, 2020, close to the 2020 U.S. presidential elections. This is possible as both the GNews and the CrowdTangle API support historical data searches within specific date ranges. We gathered data on **46,758** active news pages, with **23,992** being U.S.-based and mainly using English. We refer to this list as **SNAPSHOT_OCTOBER_2020**. Note that CrowdTangle does not return results for pages or posts that have been deleted. As a result, the dataset we obtained might represent a subset of the pages that were available in October 2020.

Across the two data collections, we compiled a total of **55,941** distinct self-proclaimed news pages, with **26,800** being U.S.-based and mainly publishing content in English.

[3]For example, we access the following URL for CNN: https://www.facebook.com/ads/library/?active_status=all&ad_type=all&country=ALL&view_all_page_id=5550296508

Collection of Historical Posts

For each identified page, we get its posting history between July 2017 and July 2022 – i.e., all the content they have published within this timeframe. This step is not essential for discovering Facebook news pages, but is important to analyze pages' posting behavior and users' engagement with their content.

For this, we use the CrowdTangle dashboard's web interface to create lists of the pages for which we want to download the posting histories. Since CrowdTangle only allows downloading files with a maximum of 10,000 posts, we have manipulated the browser to automate the process and select different pages and time ranges for each download, such that we have complete post collections. Each post is characterized by the *posting time, the editing time (if the post was edited), the textual content, the type (link post, text post, image post, video post, or live video post), the post URL, the media URL, the landing URL,* and *the engagement scores of the post*. Moreover, we have additional information about the publisher with each post, such as the *number of followers at posting time*.

We collected historical data for all pages in both the SNAPSHOT_JUNE_2022 and SNAPSHOT_OCTOBER_2020 datasets, covering the period from July 2017 to July 2022. In total we collected **191,182,320** posts. Note that CrowdTangle does not provide posts that have been deleted or made private. Therefore, we might have gaps in the posting history of certain pages.

Transition to the Meta Content Library

Meta shut down CrowdTangle on August 14, 2024 (Euronews 2024) and introduced the Meta Content Library (Meta Platforms, Inc. 2024). We adapted our implementation to use this new library instead of CrowdTangle.

Similar to CrowdTangle, the Meta Content Library supports keyword-based searches and returns Facebook posts matching those keywords (Step 2 of our methodology). However, unlike CrowdTangle, it does not return the cate-

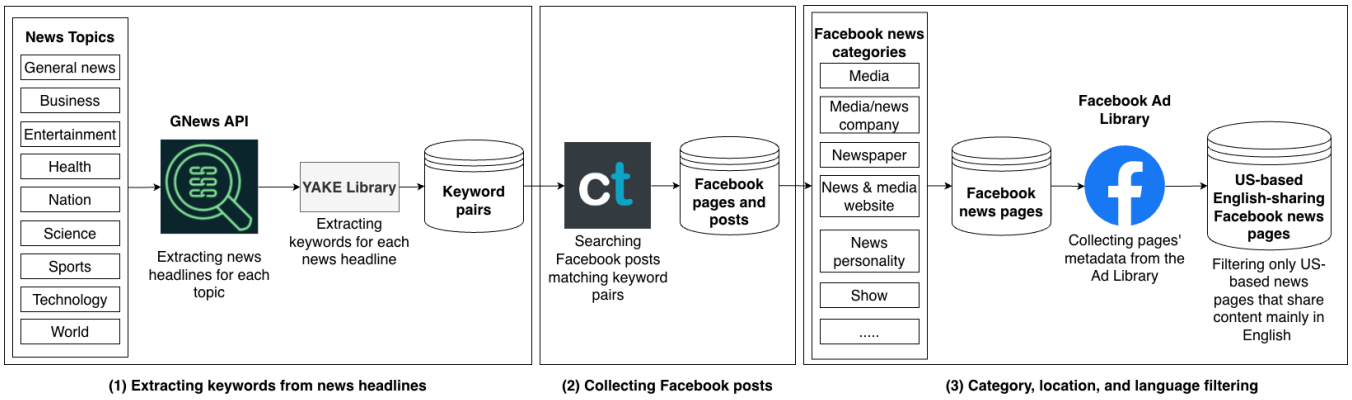

Figure 1: Diagram representing the full methodology for discovering self-proclaimed news providers' Facebook pages.

gory of the page that published each post. As a result, category filtering cannot be applied at this stage.

To address this limitation, we crawl the public Facebook page URL for each page returned by the Meta Content Library and extract the page category. The category is displayed on the main public page and matches the category previously returned by CrowdTangle. For example, CNN is labeled as Media/news company on its public Facebook page.[4] These pages are publicly accessible. We use Selenium to crawl pages with a five-second delay between requests to reduce the risk of bot detection.

Global Deployment

We use the updated version of our pipeline, employing the Meta Content Library, to run a global deployment from August 15 to 21, 2025, across nine countries. We compiled a total of 8,503 news providers: Australia (1,064), Belgium (296), France (1,528), Hungary (415), India (191), Italy (2,838), the Netherlands (337), Poland (1,092), and Romania (742).

We conducted this global deployment to assess our method's ability to generalize beyond the U.S. context. The results show that our approach can be applied across multiple countries with minimal adaptation. However, for the remainder of this paper, we focus exclusively on U.S.-based datasets of self-proclaimed news providers active in 2020 and 2022.

Validation

Our dataset aims to capture self-proclaimed news providers addressing current events in their posts. For this, we rely on external APIs and several imperfect heuristics that can impact the effectiveness of the method. Hence, this section investigates: (1) whether the dataset includes knows news outlets; (2) the speed of capturing active Facebook news pages; and (3) the extent to which the captured news pages are indeed self-proclaimed news media addressing current events.

[4]https://www.facebook.com/cnninternational

Coverage Analysis

We employ two proxies to measure the coverage of our method: (a) the extent to which it captures well-known news media and (b) the rate at which it discovers new unseen pages. A high discovery rate indicates the difficulty in achieving comprehensive coverage since there will inevitably be more pages to discover, while a low rate suggests we may be close to achieving high coverage.

We acquired a list of well-known news providers on Facebook from Edelson et al. (Edelson et al. 2021), a study aggregating news domains listed by Media Bias Fact Check and News Guard and their corresponding Facebook pages. This list was compiled in July 2020 and contains 4,323 news media Facebook pages.

Upon verification, many of these pages are not U.S.-based (e.g., https://www.facebook.com/24urcom/). To ensure a fair comparison, we excluded non-U.S.-based pages by using the *topAdminCountry* and *organisationCountry* fields, provided by CrowdTangle, to retain 2,624 U.S.-based Facebook pages. Furthermore, the MBFC/NG list includes pages that do not claim to be news providers (e.g., https://www.facebook.com/PublicInterestLegal/ is categorized as "Lawyer & Law Firm," and https://www.facebook.com/moneyandmarkets/ as an "Investing Service"). Therefore, we further filter the MBFC/NG list to include only *1,565* pages with one of the news media categories listed in Table 1. Finally, we discarded four pages for which we could not retrieve data from CrowdTangle (due to their deletion) and eight non-English-sharing pages. As a result, we have a list of *1,553 U.S.-based English-sharing Facebook news pages* that we consider to evaluate the coverage of our method.

Our analysis reveals that SNAPSHOT_JUNE_2022 successfully captures 89% of the U.S.-based English-sharing MBFC/NG pages, while SNAPSHOT_OCTOBER_2020 captures 94% of them. The combined scope of both snapshots includes *95%* of the MBFC/NG pages, corresponding to 1,474 out of 1,553 pages. These results show that our method can capture well-known news media and only misses 5% of them (which we further investigate in the next section).

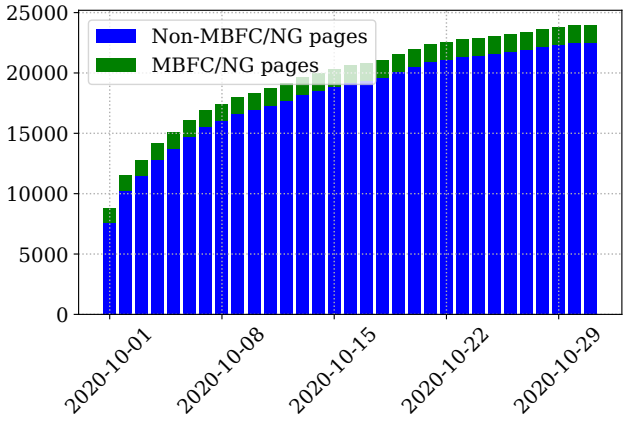

Figure 2: Cumulative number of pages detected by our method each day in October 2020. In green are pages listed by MBFC or NG, and in blue are pages not listed.

To provide an alternative perspective, in Figure 2 we present the cumulative number of Facebook pages detected each day in October 2020. We observe a high discovery rate in the first seven days, with 8,781 pages on the first day and 15,053 pages within the first five days. However, the discovery rate significantly drops afterwards, with an average of 250 discovered pages per day in the last two weeks of data collection. A low rate in the second part of the crawl suggests that our method may be approaching high overall coverage.

Moreover, we compare our discovered pages with the Local News Dataset (Yin 2018). This dataset is mainly focused on U.S. local news outlets, and provides an additional perspective on coverage beyond Media Bias Fact Check and News Guard.

Because this dataset is not restricted to self-proclaimed news providers on social media, we first filter it to retain only outlets whose associated Facebook pages self-declare a news-related category. The Local News Dataset contains 10,257 news domains, of which 7,095 list a corresponding Facebook page ID. Among these, 4,265 pages (60%) self-declare a news-related category on Facebook. We use this filtered subset for comparison and find that 3,453 of these pages (81%) are included in our discovered pages, further supporting the broad coverage of our methodology.

Missed Pages Analysis

The previous section shows that our method failed to detect 79 (5%) MBFC/NG news pages. Among these, 37 pages remained inactive (i.e., did not post any content) during our two data collection periods. Hence, we are left with 42 *active* Facebook news pages we failed to detect. Our analysis of their posting activity during October 2020 reveals these pages displayed significantly lower posting frequency than the pages we successfully identified. Precisely, half of these undetected pages only published a median of one post per active day compared to 13 posts per active day for detected pages (as illustrated in Figure 9 in Appendix). Understandably, pages that produce limited content are less likely to meet our search filters. Thus, our ability to detect such pages is lowered. Furthermore, we manually inspected posts from the ten most active non-detected pages (at least six per active day) to understand why they were not identified. Five of them treat specific niche topics and did not publish content relevant to current news during our data collection. These pages are https://www.facebook.com/BleepingComputer/, https://www.facebook.com/TheScientistMagazine/, https://www.facebook.com/CommunityImpact/, https://www.facebook.com/Face2FaceAfrica/, and https://www.facebook.com/thevintagenews/. Their respective topics, as described in their about sections, are technology, science, hyperlocal news, black history, and vintage. Given the thematic nature of their posts, they are less likely to align with the news headline-based filtering we employ.

Timeliness Analysis

The dynamic nature of the Facebook news ecosystem enables malicious third parties to create several pages, share false or misleading content, and rapidly delete them. It is crucial for a method that aims to identify active news sources to detect such pages before they get deleted. Therefore, we evaluate the timeliness of our method.

To measure the time our method took to detect each page in our dataset, we count the number of active days from a page's first post in our crawling window to its detection time. We only consider days during our data collection period when pages were active, as our method cannot detect pages that do not publish anything (e.g., if a page was active only on "2020-10-01" and "2020-10-10," and we detected it on "2020-10-10," we consider that the duration for detecting this page is 2 days). Figure 3 presents the distribution of the number of (active) days our method required to detect each page within both SNAPSHOT_JUNE_2022 and SNAPSHOT_OCTOBER_2020. The figure demonstrates the rapid detection of most Facebook news pages, with a median detection time of two active days and more than 90% detected in less than ten active days.

Relevancy Analysis

Our method aims to identify self-proclaimed news pages sharing posts related to current events. However, it employs a few imperfect heuristics that can affect the relevancy of the pages returned:

(1) To search pages discussing current events, we rely on a list of keyword tuples. Some keyword tuples may be very general and not necessarily represent current events. For example, some extracted keywords include: arab country[5] and home sales.[6]

(2) To select self-proclaimed news providers, we refer to the categories listed in Table 1. Some of these categories are

[5]Extracted from https://www.cnbc.com/2022/05/31/israel-signs-trade-deal-with-uae-its-biggest-with-any-arab-country.html

[6]Extracted from https://edition.cnn.com/2022/06/12/business/luxury-home-sales-fall-redfin/index.html

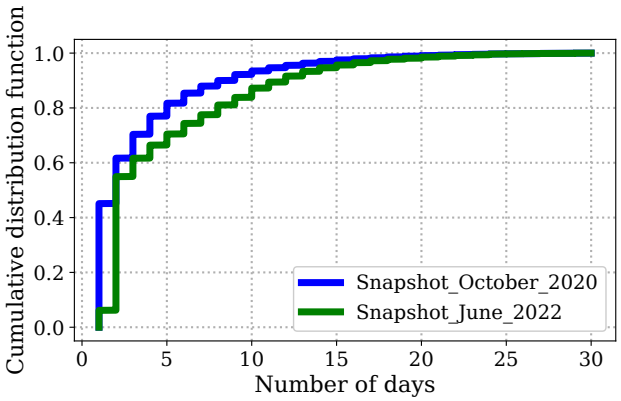

Figure 3: Cumulative distribution of the number of (active) days our method required to detect each Facebook news page in SNAPSHOT_JUNE_2022 and SNAPSHOT_OCTOBER_2020.

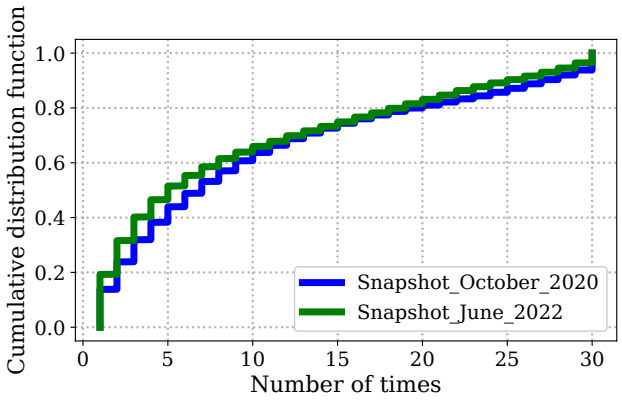

Figure 4: Cumulative distribution of the number of distinct days on which each page was detected by our method in SNAPSHOT_JUNE_2022 and SNAPSHOT_OCTOBER_2020.

broad and can include pages not presenting themselves as news media.

To assess the relevancy of the pages discovered by our method, we randomly sampled 50 pages from our dataset that were not covered by MBFC/NG. We manually annotated the posts shared by each page to determine whether the page consistently posted content related to current news and events. Two annotators independently reviewed 20 random posts from each page, sampled during October 2020 or June 2022, and classified posts as news-related if they discussed current events, regardless of the specific subject. A Facebook page was considered a relevant news source if at least 50% of the inspected posts were news-related. The two annotators agreed on 97% of the pages, and for pages with disagreement, we classified them as non-relevant. Overall, *74% of the examined pages were annotated as relevant news sources*. For transparency, we provide the random sample of 50 pages and our relevancy classification at https://github.com/CHOUAKIsalim/News_discovery/tree/main/Relevancy_analysis.

One way to reduce the number of irrelevant pages is by applying stricter filters. For example, we could consider including a Facebook page only if our method detected it on multiple distinct days, suggesting that the page regularly, rather than occasionally, shares posts about current events. Figure 4 shows the distribution of the number of distinct days on which each page was detected. We found that 19% of SNAPSHOT_JUNE_2022 pages and 14% of SNAPSHOT_OCTOBER_2020 pages were detected only once, while the median number of distinct days was five for SNAPSHOT_JUNE_2022 and seven for SNAPSHOT_OCTOBER_2020.

We conducted a second manual annotation by the same two annotators on 60 randomly selected pages not listed by MBFC or NG. These 60 pages were divided into three categories (20 each): pages detected on one or two days, pages detected on at least five days, and pages detected on at least seven days. As before, the annotators reviewed 20 random posts from each page, sampled during October 2020 or June 2022, and classified posts as news-related if they discussed current events. A page was deemed relevant if at least 50% of the posts were news-related. The results showed that 50%, 80%, and 85% of the pages in each category were relevant, respectively. This indicates that pages detected more frequently are more likely to be relevant news sources.

News Ecosystem Analysis

This section analyzes the characteristics of the self-proclaimed news providers identified in our dataset. We focus on non-listed pages detected in at least five daily crawls to reduce noise. For comparison, we include listed pages without applying the same filter, as Media Bias Fact Check and News Guard have already classified them as relevant news sources. This distinction allows us to compare well-known news providers with lesser-known pages that are identified through automated discovery. Our analysis includes 16,559 pages: 1,474 listed and 15,085 non-listed.

Dynamics

Given the risk of news pages created to disseminate false or biased information, the first question we ask is how dynamic is the news ecosystem: (1) how many new news pages are created each year, and (2) whether they have a stable activity over time or their activity only revolves around important events such as elections.

Creation. Figure 5 presents the timeline of the creation of news pages in our dataset. The figure shows that 297 new news pages in the median were created on Facebook every four months in the past 15 years. Notably, non-listed pages tend to be more recent, with over 50% emerging after 2012, in contrast to the listed pages, where only 18% were created post-2012. We particularly see two prominent peaks in the creation time in 2016 and 2019 that might be linked to the 2016 and 2020 U.S. presidential elections.

Activity. We explore whether identified self-proclaimed news pages exhibit consistent or intermittent posting activ-

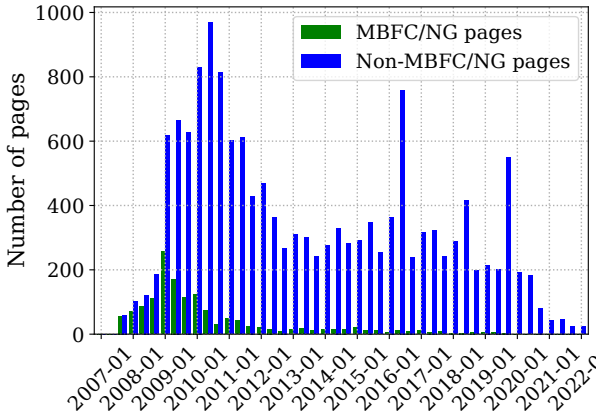

Figure 5: Creation time of Facebook pages: MBFC/NG pages vs. non-listed pages. Each point represents the number of pages created in a four-month period.

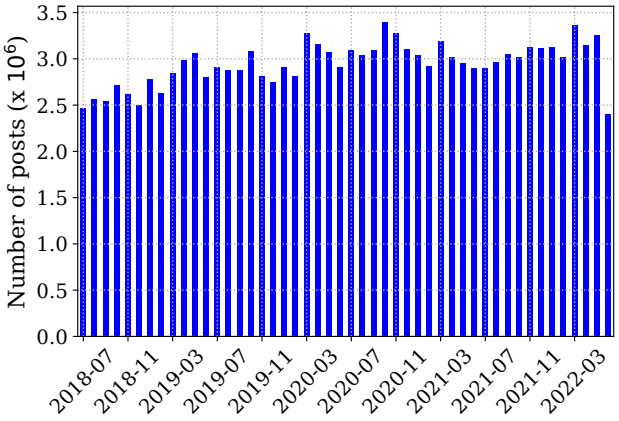

Figure 6: Combined total number of posts over all discovered pages for each month between July 2018 and July 2022.

ity. Figure 6 presents a timeline of the combined number of posts across all pages per month. The figure shows that the total number of posts does not consistently increase despite the continuous creation of pages, suggesting that certain pages stop being active or are only active for specific periods. For instance, we identified 53 news pages that operated only from January 2020 to June 2021 (6 months before and after the U.S. presidential election).

These findings show the ever-changing nature of the Facebook news ecosystem, with new pages regularly emerging. Our method effectively detects these pages, particularly if used continuously.

Affiliations

To promote content associated with political and social issues on Facebook, pages are required to disclose and verify their managing organization (Ben Matthews 2022). We retrieved this information from the Facebook Ad Library for 7,277 (44%) self-proclaimed news provider pages.

We have identified 3,043 distinct organizations, of which 406 own at least two pages. Table 2 presents the organizations with the largest number of Facebook pages. This table uncovers several insights. First, some news organizations possess multiple Facebook news pages, none of which are present on MBFC or NG lists. Examples include "Particle Media, Inc." and "On3 Media, LLP." Second, even organizations audited by News Guard and Media Bias Fact Check, such as "Gatehouse Media LLC" and "Sinclair Broadcast Group Inc.," have numerous Facebook pages that are not listed. For instance, while https://www.facebook.com/TND/ is included in the MBFC/NG list, https://www.facebook.com/KLEWNews/ managed by the same organization (Sinclair Broadcast Group Inc.), is not listed. These findings underscore the relevance of the self-proclaimed news sources identified by our method.

Types

This section explores self-proclaimed news providers that act as news aggregators or as local news sources.

News Aggregators. Recent reports have raised concerns about the rise of news aggregators, pages that republish news from various sources without creating original content, potentially driven by specific agendas and selectively sharing information aligning with their motives (Bengani 2021). This section investigates the prevalence of news aggregators among self-proclaimed news providers.

We analyze landing URLs in pages' posts from July 2017 to July 2022. We first unshorten links to various URL shortening services to obtain the actual landing URL.[7] Then, we extract the distinct domains and compute the proportion of posts leading to each domain. We use the "tldextract" Python package (PyPI 2025b) and the "Public Suffix List" (Mozilla Foundation 2022) for this purpose. Note that we only consider the registered domain, discarding the complete domain name. For instance, if a page shares posts leading to edition.cnn.com and us.cnn.com, we consider the unique registered domain cnn.com. We assume that news creators predominantly share posts leading to a single domain (i.e., their own), while news aggregators share posts with URLs spanning multiple websites, lacking a predominant domain. Therefore, we classify a page as a news aggregator if it does not have a predominant landing domain, meaning no domain accounts for 50% of the page's posts.

We find that 15% of the identified pages (2,508 pages) are news aggregators. The vast majority of these aggregators (97%) were not listed in the MBFC/NG list, like https://www.facebook.com/GodUnderstandPrayers/. Additionally, we find that a median news aggregator has posted URLs from 123 distinct domains, and 1% of aggregators have shared over 1,000 domains, such as https://www.facebook.com/TullyNews/.

These results underscore another dimension of the relevance of the self-proclaimed news pages identified by our method. The method allows discovering news aggregators, most of which are not listed by MBFC and NG.

[7]The list of URL shorteners we consider: bitly.com, cutt.ly, ow.ly, rebrandly.com, shorturl.at, tiny.cc, tinyurl.com, t.ly, trib.al, and usehyperlink.com.

Organization name	# Listed	# Nonlisted
Particle Media, Inc.	0	928
Planck, LLC	1	552
Gannett Satellite Information Network, LLC	88	106
Gatehouse Media LLC	72	56
Townsquare Media, INC.	3	113
Lee Enterprises Incorporated	57	50
Entercom Communications CORP.	2	81
Gray Television, INC.	54	27
BuzzFeed	2	50
Sinclair Broadcast Group Inc.	37	14
TAP Into Local LLC	1	49
Advance Local Media LLC	12	36
College Spun Media INC.	0	44
On3 Media, LLP	0	44
Insider, INC.	2	39
Alpha Media LLC	0	37
Heavy, INC.	1	35
CANTATA MEDIA LLC	1	32
Hearst	18	14
IHEARTMEDIA, INC.	0	32

Table 2: Top organizations and the number of pages they manage: listed by MBFC and NG vs. non-listed.

Local news We explore the geographical coverage of self-proclaimed news providers. We assume that pages focused on news at a city or state level will explicitly mention the corresponding locations in their About sections. Therefore, we analyze all pages' names and About section descriptions and classify them as local if they mention a city or a state.

For this purpose, we utilize the *locationtagger* Python library (PyPI 2020), which employs Named Entity Recognition techniques to extract location information, such as countries, regions/states, and cities, from input text or URLs. This library provides geographical information in three categories: cities, states, and countries. If a city or state is mentioned in the page's name or about section, we classify the page as a local news source.

We extracted geographical data from 57% (9,452) of the self-proclaimed news pages. Pages lacking geographical information in their names or About sections are more likely to be national or global news sources. We find that 56% of analyzed pages (9,367) are dedicated to local news, with 52% focusing on city-level news and 4% on state-level news. Noteworthy, 92% of these local news pages are not listed by Media Bias Fact Check and News Guard.

Engagement

This section analyzes the extent to which users follow and interact with content from self-proclaimed news providers, which are valuable indicators of pages' visibility and potential impact on users.

Figure 7 in the appendix presents the cumulative distribution of follower counts, and Figure 8 in the appendix presents the cumulative distribution of average weekly interactions for discovered pages. The figures show that non-listed pages generally have significantly fewer followers (7,576 in median) and engagement scores (351 interactions per week in median) than listed pages (86,817 followers and 6,868 interactions per week). However, we find that the total followers and interaction scores across non-listed pages (3,512,253,595 followers and 113,401,864 interactions per week) are higher than those of listed pages (2,651,529,840 followers and 112,974,157 interactions per week). Hence, non-listed pages have slightly greater overall visibility (as measured by followers and engagement) than listed pages, making them important to scrutinize and consider for news and misinformation studies.

Limitations

Our methodology has some limitations caused by the APIs it relies on. First, we depend on GNews for sourcing daily news headlines and extracting search keywords. Consequently, the range of news items we can cover is tied to the news items returned by GNews. Second, to retrieve Facebook posts and pages, we rely on CrowdTangle, which exclusively returns posts from actively tracked pages. The API automatically tracks all pages with over 25,000 followers, verified pages, and pages manually added by users. The Meta Content Library API tracks the same pages (Meta Platforms, Inc. 2024). As a result, our methodology fails to identify news-related, non-verified pages with fewer than 25,000 followers that were not added by users. Third, CrowdTangle and the Meta Content Library do not provide access to posts that have been deleted or set to private. This implies that (a) we may have missed some deleted news pages in our retrospective detection, and (b) we may not have considered the complete posting history of certain pages in our analysis.

Another limitation concerns country identification. We infer a page's country by combining multiple signals, including the admin's country, the managing organization's country, and the main language used by the page. While this reduces reliance on a single signal, these indicators are not always reliable. In particular, language-based filtering can be noisy, and cross-border overlap is expected in regions sharing the same language (e.g., France and Belgium). As a result, country-level assignments should be interpreted as best-effort approximations rather than ground truth.

Related Work

A vast body of literature studies online news, focusing on the prevalence of misinformation (Efstratiou and De Cristofaro 2022; Guess, Nagler, and Tucker 2019; Guess, Nyhan, and Reifler 2020; Allcott, Gentzkow, and Yu 2019; Mosleh and Rand 2022; Amieur et al. 2025) and political polarization (Flaxman, Goel, and Rao 2016; Scharkow et al. 2020; Bakshy, Messing, and Adamic 2015; Levy 2021; Chouaki et al. 2024). A critical element in these studies is the list of news sources used to construct news exposure and interaction data. Some studies used manually compiled lists of widely recognized news sources, thereby providing a limited number of sites to monitor (Levy 2021; Fletcher and Nielsen

2018; Scharkow et al. 2020; Flaxman, Goel, and Rao 2016; Agarwal et al. 2021; Cardenal et al. 2019).

More recent work builds larger lists that include less popular, local, and social-media-focused outlets. (Bakshy, Messing, and Adamic 2015) released a list of 500 news domains; (Grinberg et al. 2019) built a list of 1,250 domains based on editorial practices; (Yin 2018) collected 6,290 domains for state newspapers, TV stations, and magazines; and (Edelson et al. 2021) used 2,551 U.S. news publishers from (Media Bias Fact Check 2026) and (News Guard 2026). (Robertson et al. 2023) combined these datasets into a list of 11,902 unique domains. Another line of work builds large archives of news stories. (Roberts et al. 2021) build Media Cloud, a system that monitors a large list of news websites and regularly collects their articles using RSS feeds and web crawling. (Nwala, Weigle, and Nelson 2019) take a different approach. They search Reddit and Twitter for keywords related to specific events (such as elections) and extract the URLs that users share, then use these links as starting points for archiving. In both cases, the focus is on news web pages rather than social media accounts. Social media platforms are mainly used as sources of links to external news sites. Finally, (Hagar et al. 2025) use large language models to detect newsworthy stories in a large collection of articles gathered from RSS feeds.

Closer to our work, some studies build lists of Facebook news pages. (Le Quéré, Chiang, and Naaman 2022) start from a known list of outlets and use Facebook's "related pages" suggestions to discover additional pages. Similarly, (Guimarães et al. 2020) and (Ribeiro et al. 2018) rely on Facebook's advertising tools to identify pages that share Facebook-defined "interests" with an initial set of news outlets. Specifically, for each seed page, these methods retrieve "related pages" that are deemed similar by the platform. In both cases, the resulting pages are those that resemble the initial seed list according to Facebook's internal criteria, and thus are likely to correspond to news outlets, but may also reflect the platform's biases.

These are valuable approaches. However, our methodology differs in several important ways. First, our approach does not require a seed list and can therefore discover previously unknown news providers without prior knowledge. Second, our method relies on current-events-based keyword filtering, ensuring that the discovered pages actively discuss current news, whereas the advertising-based approach does not enforce this constraint. Third, our methodology can be applied both in live data collection and retrospectively to identify pages that were active during past periods.

Concluding Discussion

Our work analyzes a large-scale dataset of more than 26k self-proclaimed news providers on Facebook and their posting histories, providing a comprehensive empirical view of the platform's news ecosystem. The dataset captures a wide range of news actors that are often overlooked by journalistic agencies, including local news pages and news aggregators. Our analysis shows that news production on social media is highly dynamic, with new pages continuously emerging over time. Importantly, pages with relatively few followers can still reach large audiences and play a significant role in shaping users' news exposure.

Moreover, this work presents a methodology to systematically identify self-proclaimed news providers based on their engagement with current events. The methodology proved effective for constructing a U.S.-focused dataset and was also applicable to identifying news providers in other countries. However, approaches of this kind inherently depend on the data access and APIs provided by online platforms. As a result, they require continuous monitoring and adaptation to account for changes in platform interfaces, policies, and data availability (Van der Vlist et al. 2022; Hogan 2018).

These limitations highlight a broader challenge: the lack of transparent, platform-provided lists of news providers. The European Union recently adopted the Digital Services Act, which requires online platforms to share data with researchers and regulators to assess systemic risks. As the European Commission is still defining how this data access will work, we believe that creating an official index of self-proclaimed news providers should be a priority. Such an index would help improve research on misinformation and manipulation.

Finally, our findings highlight the importance of transparency and accountability for pages that claim to be news providers. The analysis suggests that applying verification mechanisms similar to those used for political advertising—such as the disclosure of the individual or organization managing a page—could improve the reliability of information about news providers. In addition, domain verification for pages listing external websites and the availability of aggregated audience statistics may help limit impersonation and facilitate more robust empirical research on online news ecosystems.

Code and Data Availability

Facebook data were mainly collected using CrowdTangle. Data collection was based on submitting keyword search queries and retrieving Facebook post data. In line with CrowdTangle's terms of service (CrowdTangle Team 2024), we do not publicly release raw post-level exports. Instead, we release (1) the final filtered list of self-proclaimed news providers and (2) aggregated page-level statistics (e.g., number of posts, total likes). We also make our source code available to support reproducing the data collection and filtering procedure. The news discovery code and the lists of self-proclaimed news providers are available at: https://github.com/CHOUAKIsalim/News_discovery/

Acknowledgements

We thank the anonymous reviewers for their helpful comments. This research was supported in part by the French National Research Agency (ANR) through the ANR-22-CE38-0017-02 grant and by the EU 101041223 grant. This work was also supported in part by the US National Science Foundation (NSF) through grant 2344939.

References

Agarwal, V.; Vekaria, Y.; Agarwal, P.; Mahapatra, S.; Set, S.; Muthiah, S. B.; Sastry, N.; and Kourtellis, N. 2021. Under the Spotlight: Web Tracking in Indian Partisan News Websites. In *Proceedings of the International AAAI Conference on Web and Social Media*.

Allcott, H.; Gentzkow, M.; and Yu, C. 2019. Trends in the diffusion of misinformation on social media. *Research & Politics*, 6(2).

Amieur, N.; Chouaki, S.; Goga, O.; and Roussillon, B. 2025. A Comparative Study of News Exposure and Consumption On and Off Facebook. In *Proc. ACM Hum.-Comput. Interact.*, volume 9. New York, NY, USA: Association for Computing Machinery.

Bakshy, E.; Messing, S.; and Adamic, L. A. 2015. Exposure to ideologically diverse news and opinion on Facebook. *Science*.

Barthel, M.; Mitchell, A.; and Holcomb, J. 2016. Many Americans believe fake news is sowing confusion. https://www.journalism.org/2016/12/15/many-americans-believe-fake-news-is-sowing-confusion/. Accessed: 2026-01-05.

Ben Matthews. 2022. How to get authorisation to run political and social issue ads on Facebook and Instagram. https://empower.agency/how-to-get-authorisation-to-run-political-and-social-issue-ads-on-facebook-and-instagram/. Accessed: 2026-01-05.

Bengani, P. 2019. Hundreds of 'pink slime' local news outlets are distributing algorithmic stories and conservative talking points. *Columbia Journalism Review*.

Bengani, P. 2021. Advocacy groups and Metric Media collaborate on local 'community news'. *Columbia Journalism Review*.

Cardenal, A. S.; Aguilar-Paredes, C.; Galais, C.; and Pérez-Montoro, M. 2019. Digital technologies and selective exposure: How choice and filter bubbles shape news media exposure. *The International Journal of Press/Politics*, 24(4): 465–486.

Chouaki, S.; Chakraborty, A.; Goga, O.; and Zannettou, S. 2024. What News Do People Get on Social Media? Analyzing Exposure and Consumption of News through Data Donations. In *Proceedings of the ACM Web Conference 2024*, WWW '24, 2371–2382. New York, NY, USA: Association for Computing Machinery. ISBN 9798400701719.

CrowdTangle. 2024a. A tool from Meta to help follow, analyze, and report on what's happening across social media. https://web.archive.org/web/20240509040952/https://www.crowdtangle.com/. Accessed: 2026-01-05.

CrowdTangle. 2024b. What data is CrowdTangle tracking? https://web.archive.org/web/20240313145103/https://help.crowdtangle.com/en/articles/1140930-what-data-is-crowdtangle-tracking. Accessed: 2026-01-05.

CrowdTangle API. 2021. Post-search endpoint. https://web.archive.org/web/20240315001358/https://github.com/CrowdTangle/API/wiki/Search. Accessed: 2026-01-05.

CrowdTangle Team. 2024. Understanding and Citing CrowdTangle Data. https://web.archive.org/web/20240616221006/https://help.crowdtangle.com/en/articles/4558716-understanding-and-citing-crowdtangle-data. Accessed: 2026-01-05.

Edelson, L.; Nguyen, M.-K.; Goldstein, I.; Goga, O.; McCoy, D.; and Lauinger, T. 2021. Understanding Engagement with U.S. (Mis)Information News Sources on Facebook. In *Proceedings of the 21st ACM Internet Measurement Conference*.

Efstratiou, A.; and De Cristofaro, E. 2022. Adherence to Misinformation on Social Media Through Socio-Cognitive and Group-Based Processes. In *Proceedings of the ACM on Human-Computer Interaction*, CSCW2, 1–35. ACM New York, NY, USA.

Euronews. 2024. Researchers Fear Effects of Meta's CrowdTangle Shutdown. https://www.euronews.com/next/2024/08/02/researchers-fear-effects-of-metas-crowdtangle-shutdown. Accessed: 2026-01-05.

Facebook. 2018. Making Ads and Pages More Transparent. https://about.fb.com/news/2018/04/transparent-ads-and-pages/. Accessed: 2026-01-05.

Facebook. 2019. Introducing Facebook News. https://about.fb.com/news/2019/10/introducing-facebook-news/. Accessed: 2026-01-05.

Facebook. 2026a. Create a Facebook page. https://www.facebook.com/pages/creation/. Accessed: 2026-01-05.

Facebook. 2026b. Facebook Ad Library. https://www.facebook.com/ads/library/. Accessed: 2026-01-05.

Facebook Business. 2018. New Authorization for Pages. https://www.facebook.com/business/news/new-authorization-for-pages. Accessed: 2026-01-05.

Facebook Business. 2026. Get authorized to manage Pages with large audiences. https://www.facebook.com/business/m/one-sheeters/page-publishing-authorization. Accessed: 2026-01-05.

Facebook Help Center. 2026a. Get authorized to post or interact as your Page. https://www.facebook.com/help/1939753742723975. Accessed: 2026-01-05.

Facebook Help Center. 2026b. How two-factor authentication works on Facebook. https://www.facebook.com/help/148233965247823. Accessed: 2026-01-05.

Flaxman, S.; Goel, S.; and Rao, J. M. 2016. Filter Bubbles, Echo Chambers, and Online News Consumption. *Public Opinion Quarterly*.

Fletcher, R.; and Nielsen, R. K. 2018. Are people incidentally exposed to news on social media? A comparative analysis. *New Media & Society*.

GNews. 2021. A Python Package that searches Google News RSS Feed and returns a usable JSON response. https://github.com/ranahaani/GNews. Accessed: 2026-01-05.

Grinberg, N.; Joseph, K.; Friedland, L.; Swire-Thompson, B.; and Lazer, D. 2019. Fake news on Twitter during the 2016 US presidential election. *Science*, 363(6425): 374–378.

Guess, A.; Nagler, J.; and Tucker, J. 2019. Less than you think: Prevalence and predictors of fake news dissemination on Facebook. *Science Advances*.

Guess, A. M.; Nyhan, B.; and Reifler, J. 2020. Exposure to untrustworthy websites in the 2016 US election. *Nature Human Behaviour*.

Guimarães, S. S.; Reis, J. C.; Lima, L.; Ribeiro, F. N.; Vasconcelos, M.; An, J.; Kwak, H.; and Benevenuto, F. 2020. Identifying and characterizing alternative news media on Facebook. In *2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 448–452. IEEE.

Hagar, N.; Silver, E.; Spencer, C.; and Diakopoulos, N. 2025. LLM-Assisted News Discovery in High-Volume Information Streams: A Case Study. https://arxiv.org/abs/2509.25491.

Hogan, B. 2018. Social media giveth, social media taketh away: Facebook, friendships, and APIs. *International Journal of Communication*, 12.

Le Quéré, M. A.; Chiang, T.-W.; and Naaman, M. 2022. Understanding local news social coverage and engagement at scale during the covid-19 pandemic. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 16, 560–572.

Levy, R. 2021. Social media, news consumption, and polarization: Evidence from a field experiment. *American Economic Review*, 111(3): 831–870.

Media Bias Fact Check. 2026. https://mediabiasfactcheck.com/. Accessed: 2026-01-05.

Meta Business Help Center. 2026a. About domain verification in Meta Business Manager. https://www.facebook.com/business/help/286768115176155. Accessed: 2026-01-05.

Meta Business Help Center. 2026b. About News Pages Index. https://www.facebook.com/business/help/377680816096171. Accessed: 2026-01-05.

Meta Business Help Center. 2026c. How to create a new Page on Facebook. https://www.facebook.com/business/help/1199464373557428?id=418112142508425. Accessed: 2026-01-05.

Meta Business Help Center. 2026d. Register your News Page. https://www.facebook.com/business/help/316333835842972. Accessed: 2026-01-05.

Meta Business Help Center. 2026e. Registration guidelines for the news Page index. https://www.facebook.com/business/help/270254993785210. Accessed: 2026-01-05.

Meta Platforms, Inc. 2024. Meta Content Library. https://transparency.meta.com/researchtools/meta-content-library. Accessed: 2026-01-05.

Mosleh, M.; and Rand, D. G. 2022. Measuring exposure to misinformation from political elites on Twitter. *Nature Communications*, 13(1): 7144.

Mozilla Foundation. 2022. The public suffix list. https://publicsuffix.org/. Accessed: 2026-01-05.

News Guard. 2026. https://www.newsguardtech.com/. Accessed: 2026-01-05.

Nwala, A.; Weigle, M.; and Nelson, M. 2019. Using MicroCollections in Social Media to Generate Seeds for Web Archive Collections. In *2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, 251–260.

PyPI. 2020. locationtagger. https://pypi.org/project/locationtagger/. Accessed: 2026-01-05.

PyPI. 2025a. Selenium 4.9.1. https://pypi.org/project/selenium/. Accessed: 2026-01-05.

PyPI. 2025b. The tldextract python package. https://pypi.org/project/tldextract/. Accessed: 2026-01-05.

Ribeiro, F.; Henrique, L.; Benevenuto, F.; Chakraborty, A.; Kulshrestha, J.; Babaei, M.; and Gummadi, K. 2018. Media bias monitor: Quantifying biases of social media news outlets at large-scale. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 12.

Roberts, H.; Bhargava, R.; Valiukas, L.; Jen, D.; Malik, M. M.; Bishop, C. S.; Ndulue, E. B.; Dave, A.; Clark, J.; Etling, B.; et al. 2021. Media cloud: Massive open source collection of global news on the open web. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 15, 1034–1045.

Robertson, R. E.; Green, J.; Ruck, D. J.; Ognyanova, K.; Wilson, C.; and Lazer, D. 2023. Users choose to engage with more partisan news than they are exposed to on Google Search. *Nature*, 618(7964): 342–348.

Sands, John. 2019. Local news is more trusted than national news — but that could change. https://knightfoundation.org/articles/local-news-is-more-trusted-than-national-news-but-that-could-change/. Accessed: 2026-01-05.

Scharkow, M.; Mangold, F.; Stier, S.; and Breuer, J. 2020. How social network sites and other online intermediaries increase exposure to news. *Proceedings of the National Academy of Sciences*.

Van der Vlist, F. N.; Helmond, A.; Burkhardt, M.; and Seitz, T. 2022. API governance: The case of Facebook's evolution. *Social Media+ Society*, 8(2): 20563051221086228.

Van Green, Ted. 2020. Few Americans are confident in tech companies to prevent misuse of their platforms in the 2020 election. https://www.pewresearch.org/fact-tank/2020/09/09/few-americans-are-confident-in-tech-companies-to-prevent-misuse-of-their-platforms-in-the-2020-election/. Accessed: 2026-01-05.

Walker, M.; and Matsa, K. E. 2022. News Consumption Across Social Media in 2021. https://www.pewresearch.org/journalism/2021/09/20/news-consumption-across-social-media-in-2021/. Accessed: 2026-01-05.

Weld, G.; Glenski, M.; and Althoff, T. 2021. Political bias and factualness in news sharing across more than 100,000 online communities. In *Proceedings of the International AAAI Conference on Web and Social Media*.

YAKE. 2025. Yet Another Keyword Extractor. https://github.com/INESCTEC/yake. Accessed: 2026-01-05.

Yin, L. 2018. Local News Dataset. https://doi.org/10.5281/zenodo.1345145.

AAAI ICWSM Paper Checklist

(a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes, we believe that our methodology and datasets could provide significant value for future news-related research. Moreover, our approach relies solely on publicly available tools (or tools made accessible to researchers) and does not involve humans. Therefore, it avoids any form of unfair profiling and fully complies with ethical and regulatory standards.

(b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes

(c) Do you clarify how the proposed methodological approach is appropriate for the claims made? Yes

(d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? NA. Our research does not involve users. Hence, no populations-specific distributions).

(e) Did you describe the limitations of your work? Yes, we include a limitations section in the paper. Additionally, we discuss relevant limitations contextually within the methodology section whenever appropriate

(f) Did you discuss any potential negative societal impacts of your work? No, we do not expect the work to have any negative societal impact.

(g) Did you discuss any potential misuse of your work? No, we do not beleive that the work has the potential to be misused.

(h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? No

(i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes

1. Additionally, if your study involves hypotheses testing...

(a) Did you clearly state the assumptions underlying all theoretical results? NA

(b) Have you provided justifications for all theoretical results? NA

(c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? NA

(d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? NA

(e) Did you address potential biases or limitations in your theoretical framework? NA

(f) Have you related your theoretical results to the existing literature in social science? NA

(g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? NA

2. Additionally, if you are including theoretical proofs...

(a) Did you state the full set of assumptions of all theoretical results? NA

(b) Did you include complete proofs of all theoretical results? NA

3. Additionally, if you ran machine learning experiments...

(a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? NA

(b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? NA

(c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? NA

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? NA

(e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? NA

(f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? NA

4. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...

(a) If your work uses existing assets, did you cite the creators? Yes

(b) Did you mention the license of the assets? Yes, we mention that the GNews API we use is publicly available under an open license, and that CrowdTangle and the Meta Content Library are accessible to researchers.

(c) Did you include any new assets in the supplemental material or as a URL? No, except a link to anonymized github repository with the data and the code.

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? NA, we do not involve people in our study.

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? Our work does not involve users. Therefore, the dataset does not include PIIs

(f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR (see FORCE11 (2020))? Our dataset is fully available on the anonymized GitHub repository linked from the paper, provided in CSV format, and is completely free to use under an open license

(g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset (see Gebru et al. (2021))? No, we have included CSV files in the anonymized GitHub repository. We will create a datasheet at a later stage.

5. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...

(a) Did you include the full text of instructions given to participants and screenshots? NA

(b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? NA

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? NA

(d) Did you discuss how data is stored, shared, and deidentified? NA

References

FORCE11. 2020. The FAIR Data principles. https://force11.org/info/the-fair-data-principles/.

Gebru, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J. W.; Wallach, H.; Iii, H. D.; and Crawford, K. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12): 86–92.

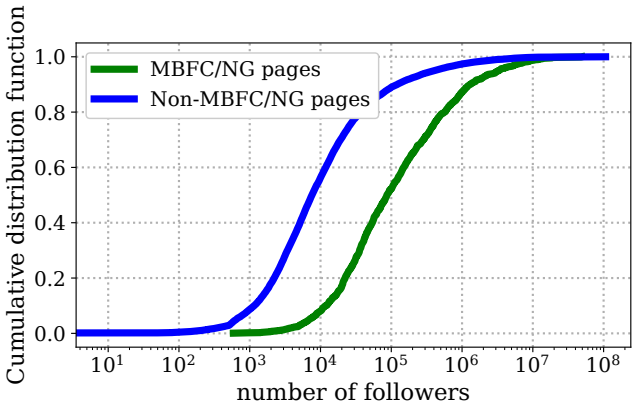

Figure 7: CDF of the number of followers for Facebook pages listed by MBFC/NG and Facebook pages not listed by MBFC/NG but discovered by our method.

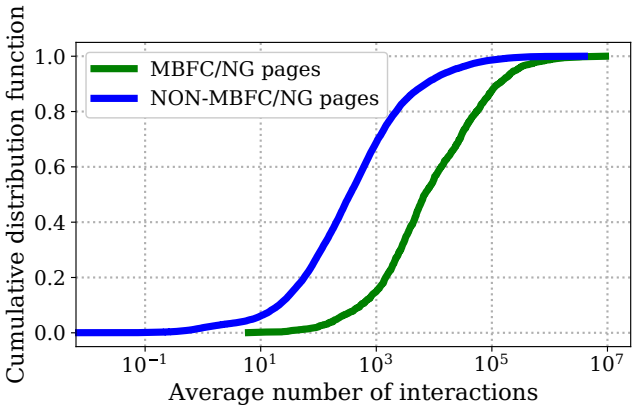

Figure 8: Cumulative distribution of the average number of interactions per active week for each Facebook news page.

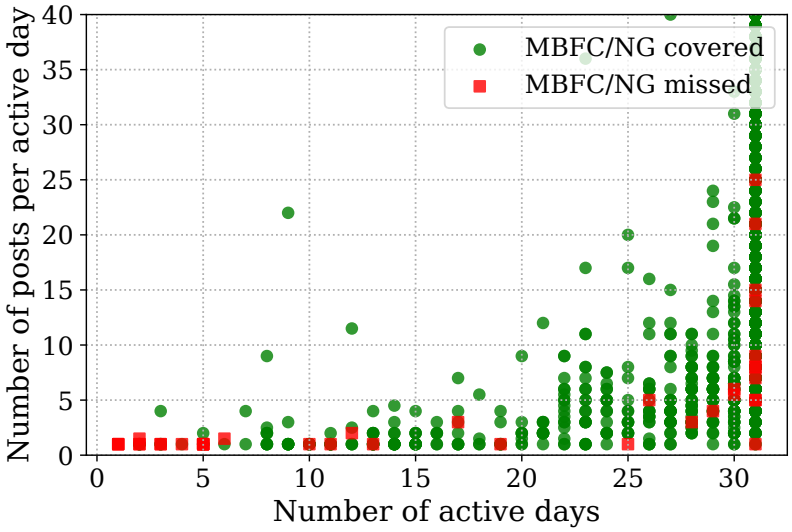

Figure 9: Number of active days and median number of posts per active day for MBFC/NG pages during October 2020.