Facebook URL Shares: Codebook

Solomon Messing, Bogdan State, Chaya Nayak, Gary King, and Nathaniel Persily

11 July 2018

This document describes a **Facebook URL shares dataset**, resulting from a collaboration between Facebook and Social Science One. It was prepared for upcoming Requests for Proposals and describes the dataset's scope, structure, and fields.

Status
This document represents a plan, not an existing data set. Facebook engineers and data scientists will build and run the necessary data pipelines to assemble the data set in the coming months, prior to when data access is granted to researchers through the RFP process. We do not expect major departures from the structure and fields documented below prior to release; any differences will be fully documented.

Additionally, the Facebook team is working to build tooling for data sharing as described in the "Infrastructure and Resources" section of this RFP. The final tool may be different then what is described below.

These data may contain errors. Facebook and Social Science One are committed to working with researchers to correct errors and irregularities as they are uncovered.

Data Access
To obtain access to these data, see the Request for Proposals process at SocialScience.one. No other means of access is allowed.

Summary
The data describes web page addresses (URLs) that have been shared on Facebook starting January 1, 2017 and ending about a month before the present day. URLs are included if shared by at least 20 unique accounts, and shared publicly at least once.

Structure. The unit of analysis for these data is the Cartesian product of the "keys," described below. In other words, the unit of analysis is the URL-country-sub_region-age_bracket-gender-device-time period for non-U.S. regions; further divided into ideology and friend_ideo_dist in the U.S. Some of these cells will therefore contain structural zeros (such as ideology outside the US) and so the corresponding rows are omitted.

There are 3 types of variables in the data structure,: the *keys* that define the cell aggregates; the *url attributes*, which are url-level observations that describe the urls; and the *aggregates*, which are summaries (e.g., sum or mean) within, i.e., conditioned on or grouped by, the keys. The data contain aggregated breakdowns of how many users have shared, viewed, clicked, liked, and commented on each URL in each time period (**7 day windows**).

Estimated size. Facebook is in the process of building the final data set, which will be ready by the time that grants are awarded by Social Science One. We estimate the full data set will contain on the order of 2 million unique urls shared in 300 million posts, per week. Because rows with empty cells will not be included, for rare URLs the data may contain as few as one or two rows; extremely popular URLs may yield up to approximately 300 thousand rows per URL. We estimate that the data will contain on the order of 30 billion rows, translating to an effective raw size on the order of a petabyte.

Data from users who have chosen to delete their accounts are not available due to legal constraints (and availability). Missing data should be expected to be more of a problem for data further in the past.

Infrastructure and Resources. Facebook will provide teams with accepted proposals: (1) a synthetic data set to explore the structure of the data and develop an analysis pipeline based on R and/or Python code; (2) the infrastructure necessary to query the data; (3) a means to execute the analysis pipeline; (4) a means to access results of the analysis; (5) A webinar introducing the data set and useful computational strategies.

Recommended capabilities. Research teams should have experience working with data sets that do not fit into memory. Specifically, teams will need the capability to query large structured databases (e.g., Hive and/or Spark), and will need to write R and/or Python analysis code that does not exhaust system RAM (e.g., on a modern server with around 64GB RAM). Having at least one individual with multiple years of experience using SQL/HQL, Python, and Linux is highly recommended.

Fields to be included in data release:

Keys

- **URL [text]:** The webpage URL. The URLs have been processed in an attempt to consolidate different web addresses that point to the same URL. URLs that are no longer reachable will persist in the data. This is the full URL, not just the domain (e.g., https://www.nytimes.com/2018/07/09/world/asia/thailand-cave-rescue-live-updates.html).
- **Country [text]:** user’s country.
- **Sub_region [text]:** user’s region - in the U.S. this will be state. Not available in all countries.
- **Age_bracket [text, ordered]:** Age brackets include: All, 18-24, 25-34, 35-44, 45-54, 55-64, 65+. Data from users' profiles.
- **Gender [text]:** All, male, female, other. Data from users' profiles.
- **Device [text]:** Including all, mobile, desktop browser.
- **Date_window [integer]:** Date window corresponding to each non-overlapping 7-day period in the data set.

Keys for inclusion in U.S. only data

- **Ideo [integer, ordered]:** ideological affiliation bucket (-2, 2) (U.S. only). This is based on a model of ideology that uses inputs including pages followed and other profile information. Additional detail will be provided in the data release.
- **Avg_friends_ideo_dist [integer, ordered]:** average of friends' ideological affiliation, rounded to whole numbers (-2, 2) (U.S. only). Only available for the set of U.S. friends. This variable is bucketed to whole numbers because it is a key.

URL attributes
These are URL-level fields that are neither keys nor aggregates.

- **Domain [text]:** domain name from the URL.
- **Title [text]:** Provided by the author of the document (pulled from **og:title** field in original html if possible).
- **Blurb [text]:** Provided by the author of the document (pulled from **og:description** field in original html if possible).
- **was_sent_to_3PFC [integer, dichotomous]:** Was the URL was sent to third-party fact-checkers?
- **3pfc_rating [integer, dichotomous]:** If URL was sent to third-party fact-checkers, did they rate as false? See details explained here: https://www.facebook.com/help/publisher/182222309230722 and https://www.facebook.com/help/572838089565953?helpref=faq_content. Only available for some stories, only available in some countries.
- **Og_text_hard_news [integer, dichotomous]**: an indication of whether URL's open-graph description text contains "hard news" topics/vocabulary (U.S. only). This indicator is constructed based on a supervised model that leverages section labels that news websites often provide in the URLs. It is provided for convenience and for research purposes only and is based on Bakshy, Messing, and Adamic (2015). Details can be found in section 1.4 in http://science.sciencemag.org/content/sci/suppl/2015/05/06/science.aaa1160.DC1/Bakshy-SM.revision.1.pdf. This classifier has roughly 97% accuracy, 85% precision, and 50% recall. Any substantial differences between this and the final classifications will be fully documented.

Aggregates:

- **Pct_views_friend [double]:** percent of views from posts by a users' friends compared with non-friends (e.g., posts from pages or posts seen in groups).
- **Number_of_views [integer]:** Number of unique views for the URL during the time window.
- **Number_of_views_pre_fact_check [integer]:** number of views before fact-checking.
- **Number_of_viewers [integer]:** Number of unique users who viewed the URL during the time window.
- **number_of_views_over_[time window]_days [integer array]:** Array consisting of # views for each day during the time window (days in sequential order). Users and pages can share URLs more than once. (This is not available within demographic splits).
- **Number_of_shares [integer]:** Number of times the URL was shared during the time window. Users and pages can share URLs more than once.

- **number_of_shares_over_[time window]_days [integer array]:** Array consisting of # shares per day during the time window. Users and pages can share URLs more than once.
- **Number_of_shares_by_page [integer]:** Number of times the URL was shared during the time window by a page.
- **Number_of_shares_by_person [integer]:** Number of times the URL was shared during the time window by a typical user account.
- **Number_of_clicks [integer]:** Number of times the URL was clicked on during the time window. Users can click on URLs more than once.
- **Number_of_reshares_wo_clicks [integer]:** number of users who reshare a post but who did not click on the link *on the platform* during the time window. It's common for users to share an article without first clicking on it *on Facebook*, though this number could help identify articles that users are particularly likely to share without reading, or URLs used in organized campaigns to spread content.
- **Number_of_likes [integer]:** Number of likes for the URL.
- **Number_of_love [integer]:** Number of 'love' reactions to the URL.
- **Number_of_haha [integer]:** Number of 'haha' reactions to the URL.
- **Number_of_wow [integer]:** Number of 'wow' reactions to the URL.
- **Number_of_sad [integer]:** Number of 'sad' reactions to the URL.
- **Number_of_anger [integer]:** Number of angry reactions to the URL.
- **Number_of_comments [integer]:** total number of comments.
- **Avg_top_slot_in_feed [double]:** all stories appear in news feed in ranked order, in slot 1, 2, 3, etc. That ranking is dynamic, changing multiple times per day. This variable records the top slot a story reached for the average user who saw it.
- **Avg_false_news_usr_feedback [double]:** number of times posts containing URL marked as false news by users, divided by total views.*
- **Avg_hate_speech_usr_feedback [double]:** number of times posts containing URL marked as hate speech by users, divided by total views.*
- **Avg_spam_usr_feedback [double]:** number of times posts containing URL marked as spam by users, divided by total views.*

***Note about user feedback fields:** these fields constitute information provided by users, which like any survey question or coding exercise, may not function as the researcher intends.

For example, for the variable “Avg_hate_speech_usr_feedback”, users may share URLs to endorse or oppose the content. Endorsements of hate speech violate Facebook's community standards policy, while opposing it does not. Users may also flag content as hate speech because they disagree with it, rather than to actually indicate hate speech, resulting in false positive reports of hate speech if taken literally. This makes the hate speech field difficult to interpret. Similar issues apply to other fields; these subtleties should be noted by researcher.

For “Avg_spam_usr_feedback”, in contrast to URLs found to contain hate speech (which Facebook deliberately does not block due to the subtleties above), URLs containing content that violates spam policies are blocked from the platform for future sharing.

To learn how Facebook defines and measures key issues, refer to the Community Standards Enforcement Report. The numbers cited in this report are not comparable to the data presented in this RFP as they reference different underlying data. For additional info: Community Standards | Enforcement Report Guide.