Facebook URLs-Light Codebook

Christina DeGregorio, Gary King, Solomon Messing, Zagreb Mukerjee, Chaya Nayak,
Nathaniel Persily, Bogdan State, Arjun Wilkins

24 April 2019

This codebook results from a collaboration between Facebook and Social Science One, originally prepared for Social Science One grantees. It describes the first URLs dataset, including its scope, structure, and fields.

Citation

DeGregorio, Christina; King, Gary; Messing, Solomon; Mukerjee, Zagreb; Nayak, Chaya; Persily, Nathaniel; State, Bogdan; Wilkins, Arjun, 2019, "Facebook URL Shares-Light", https://doi.org/10.7910/DVN/ZT0WZY, Harvard Dataverse, V1

Status

These data may only be accessed through a special privacy-preserving computational infrastructure being built by Facebook. No data may leave the system, but researchers will be able to run their code on the system and make research results available for publication.

Data Access

To obtain access to these data, see 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 through to and including February 19, 2019. URLs are included if shared by more than on average 100 (+ Laplacian noise (mean = 0, scale = 5) to minimize information leakage) unique accounts with public privacy settings. We have also post-processed the URLs (as detailed below) to remove potentially private and/or sensitive data.

The unit of analysis for these data is the URL. All aggregates attached to the URL are taken over the entire time period (in this initial data set). The data set is about 7 gigabytes, comprising approximately 32 million URLs, and about 544 million cell values.

The data set was created from the Facebook platform by running scripts that took several weeks and finished February 19, 2019. User accounts deleted prior to this date do not appear in the dataset; information will be shared with researchers after signing the Research Data Agreement. This dataset includes content that has been taken down due to Community Standards violations, meaning that some URLs that do not work from inside of Facebook may work outside. Other URLs, which worked when the dataset was created, may no longer resolve to a webpage that still exists. Finally, we have attempted to remove URLs and associated engagement statistics that link to known child exploitative imagery from our dataset. We have also tried to remove URLs, the 'Title' and 'Blurb' in our dataset for known non-consensual intimate imagery, suicide and self-harm, although the associated engagement statistics with these links remain in the dataset.

We are developing procedures for data sharing and replication of researcher results following a version of the Replication Standard (King, 1995) and will update this document with details before researchers receive access to the data.

Infrastructure. Facebook will provide approved researchers access to a computer system on which they will conduct all analyses in a manner that protects user privacy and ensures the integrity of research results. Aside from certain fields (URL, Domain, Title and Main blurb), data will be accessible to researchers only via "differentially private" statistical results. The specific differential privacy algorithms applied, their inferential consequences, and how to avoid or document the statistical biases and properly represent uncertainty will be generated, documented, and shared with researchers prior to data access beginning. A simulated sample data set will be provided to researchers when they obtain access to the computational infrastructure.

The interface will require the use of a special software library, similar to SparkML, but which only allows analysts to execute queries only in a differentially private way. Differential privacy also means that each research team will have a "privacy budget," which limits the amount of information (i.e., a combination of the granularity and total number of unique queries run) that analysts can extract. Facebook and Social Science One will work with researchers so they understand the consequences of their analyses and to give them the budget they need to conduct their research.

Recommended capabilities. Research teams should have experience working with data sets that do not fit into memory. Researchers will need the capability to query SparkML and understand what kinds of queries are expensive and likely to exhaust the resources of our computing cluster. They will also 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 experience with SQL/HQL, Python, and Linux is highly recommended.

Fields to be included in data release

URLs will be included in the data if they have been shared publicly - meaning a user chose to share a URL in a post accessible to everyone on Facebook - more than on average 100 times + laplacian noise (with mean = 0, scale = 5), added to minimize information that an attacker could use by exploiting a hard cutoff.

An example dataset can be found here.

- **Url_rid [text]**: a unique URL id created specifically for this data set.

- **Clean_url [text]**: The webpage URL after processing. 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). URLs that are no longer reachable will persist in the data. Our processing attempts to consolidate different web addresses that point to the same URL and to remove potentially private and/or sensitive data. This includes the following:
    1. Redirects, including URL shorteners, are followed to the terminal URL.
    2. If the terminal webpage has an "og:url" meta-tag, the associated URL becomes the consolidated URL---often referred to as the "canonical URL." If not, the rel = "canonical" tag is used. If neither tag is provided, the canonical URL is taken from the raw URL address.
    3. The vast majority of obvious personally identifiable information (PII) contained in URLs is already removed by virtue of filtering URLs to those with on average 100 public shares, since less frequently shared URLs contain the bulk of PII.
    4. For URLs with query strings (~21.8% of URLs above), special processing is applied. A query string in a URL passes data to the server when a client requests content, for example the "v=Ipi40cb_RsI" in https://www.youtube.com/watch?v=Ipi40cb_RsI. Sometimes query parameters provide navigation data, which tells the server what content to deliver to the client, as above. However, query parameters can also pass to the server data irrelevant to navigation, such as whether a URL was accessed from Twitter or Reddit, tracking data, and/or PII. We have attempted to remove query parameters unrelated to content navigation by iteratively removing each query parameter and testing the resulting content for differences with original page content (above and beyond the difference introduced by re-loading the page, which can occur due to ads, 'suggested content,' and/or randomized menu options). Note that, for the vast majority of URLs, removing these parameters does not result in content that is different from the original. This is done at the domain level for 100 URLs (unless the domain has fewer than 100 URLs in the data).
    5. We keep query params that result in a different page title AND html content that differs by more than 2%, OR content that is > 95% different from original page. This measure is based on the difflib Python library and is defined as 2.0 * M/T, where M is the number of sequence matches and T is the number of elements in both sequences.
    6. URLs from domains that fail to return a valid response within 120 seconds for more than half of the (up to 100) URLs in each domain (see above), or return a response of on average under 100 characters, are stripped of all query parameters.
    7. Query parameter values that contain common phone number patterns are removed using the phonenumbers Python library.
    8. Any email addresses that appear in any part of the URL string are removed using regular expressions.

Example URLs. Left raw, right processed. Non-essential query values have been altered to protect privacy.

https://media1.tenor.co/images/da7eb8198618472aa82	https://media1.tenor.co/images/da7eb8198618472aa82

151e5d704f521/tenor.gif?itemid=5265827	151e5d704f521/tenor.gif
https://www.pivot.one/app/invite_login?inviteCode=csdfeddshkuyfckyc	https://www.pivot.one/app/invite_login
https://www.youtube.com/watch?v=oXWsoqesw7A&feature=youtu.be	https://www.youtube.com/watch?v=oXWsoqesw7A
https://www.youtube.com/watch?v=oX_fLP191-k&list=RDoX_fLP191-k	https://www.youtube.com/watch?v=oX_fLP191-k
https://news.google.com/newspapers?nid=2478&dat=10260530&id=xFc1AAAAIBAJ&sjid=iiUMAAdFJSIBAJ&pg=1558%2C27085012&hl=en	https://news.google.com/newspapers?id=xFc1AAAAIBAJ&pg=1558%2C27085012

- **Parent_domain [text]:** parent domain name from the URL (eg. foxnews.com).
- **Full_domain [text]:** full domain name from the URL (eg. www.foxnews.com, video.foxnews.com, nation.foxnews.com, insider.foxnews.com).
- **first_post_time [timestamp]** - The date/time when URL was first posted by a user on Facebook, truncated to 10 minute increments. The exact format is YYYY-MM-DD HH:MM:SS, for example: 2015-12-02 18:10:00.
- **first_post_time_unix [unix timestamp]** - The above field translated into unix time---the number of seconds since 1970-01-01 00:00:00, for example: 1449079800.
- **Share_title [text]:** Provided by the author of the URL's content pulled from **og:title** field in original html if possible).
- **Share_main_blurb [text]:** Provided by the author of the URL's content (pulled from **og:description** field in original html if possible).
- **3pfc_rating [text]:** If URL was sent to third-party fact-checkers (3pfc), did they rate it (NULL if not) and if so, how did they rate it? Category values include: 'True', 'False', 'Prank Generator' , 'False Headline or Mixture', 'Opinion', 'Satire', 'Not Eligible, 'Not Rated.' Definitions, and a list of fact checkers, are available here: https://www.facebook.com/help/publisher/182222309230722 and https://www.facebook.com/help/572838089565953?helpref=faq_content. More information on how news that may be false is selected can be found here: https://www.facebook.com/help/1952307158131536. Only available for some stories, and only available in Argentina, Brazil, Cameroon, Canada, Colombia, Denmark, France, Germany, India, Indonesia, Ireland, Italy, Kenya, Mexico, Middle East and North Africa, Netherlands, Nigeria, Norway, Pakistan, Philippines, Senegal, South Africa, Sweden, Turkey, UK, US. When more than one rating is given to a story, we use Facebook's *precedence rules*, described below.
- **3pfc_first_fact_check [timestamp]:** the date-time that article was first fact-checked, if at all. If the article has not been fact checked, this will be NULL. Date-times will be truncated to 10

minute increments. The exact format is YYYY-MM-DD HH:MM:SS, for example: 2015-12-02 18:10:00.

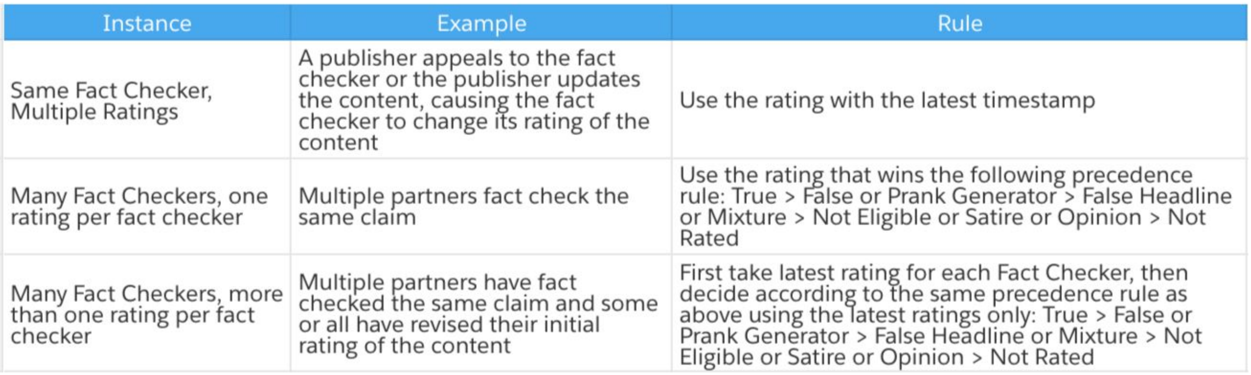
- **3pfc_first_fact_check [unix timestamp]** - The above field translated into unix time---the number of seconds since 1970-01-01 00:00:00, for example: 1449079800.
- **total_public_shares [integer]**: total number of unique accounts that shared the URL with public privacy settings.
- **total_spam _usr_feedback** [integer]**: the total number of unique users who reported posts containing the URL as spam.*
- **total_false_news_usr_feedback** [integer]**: the total number of unique users who reported posts containing the URL as false news.*
- **total_hate_speech_usr_feedback** [integer]**: the total number of unique users who reported posts containing the URL as hate speech.*
- **Prop_share_without_clicks [double]**: number of users who shared a post but did not click on the link, divided by the total number of unique users who shared the post containing the URL during the time window. (Many users share articles without first clicking through to the actual content. Hence, this number may help identify articles that users are sharing without reading, or URLs used in organized campaigns to spread content.)
- **Public_shares_top_country* [text]**: Top user country among users who publicly shared the URL, provided as an ISO 3166-1 alpha-2 letter code.

Third Party Fact Checking (3PFC) Ratings and Precedence Rules

Based on a single fact-check, Facebook reduces the distribution of a specific piece of false content. Facebook also uses similarity detection methods to identify duplicates of debunked stories and reduce their distribution as well. Facebook uses this as a signal to reduce the overall distribution of Pages and web sites that repeatedly share things found to be false by fact-checkers. Facebook gets signals about false content that can feed back into our machine learning model, helping us more effectively detect potentially false items in the future.

Occasionally, multiple fact-checkers apply different ratings to the same piece of content. In these cases, the more definitive rating takes precedence, e.g. ‘False’ or ‘True’ trumps ‘Mixture’. In rare cases where the two most definitive ratings, ‘True’ and ‘False’, are applied to the same piece of content, ‘True’ takes precedence since we refrain from demoting content rated ‘True’ by a fact-checking partner. Our 3pfc_rating incorporates the below precedence rules when decisioning how we handle multiple fact checker ratings for the same URL. It is very rare for multiple fact-checkers to rate the same URL.

For third-party fact-checked content, a fact-checker in a country other than the top public shares country may have rated content if it circulated broadly within their country. For a complete list of our third party fact checkers, please visit this website and this one.



Instance	Example	Rule
Same Fact Checker, Multiple Ratings	A publisher appeals to the fact checker or the publisher updates the content, causing the fact checker to change its rating of the content	Use the rating with the latest timestamp
Many Fact Checkers, one rating per fact checker	Multiple partners fact check the same claim	Use the rating that wins the following precedence rule: True > False or Prank Generator > False Headline or Mixture > Not Eligible or Satire or Opinion > Not Rated
Many Fact Checkers, more than one rating per fact checker	Multiple partners have fact checked the same claim and some or all have revised their initial rating of the content	First take latest rating for each Fact Checker, then decide according to the same precedence rule as above using the latest ratings only: True > False or Prank Generator > False Headline or Mixture > Not Eligible or Satire or Opinion > Not Rated

**User feedback fields: these fields constitute information provided by users, which may not be indicative of actual violations of our Community Standards, and like any survey question or coding exercise, may not be a measure of the concept the researcher intends. For example, for the variable “total_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 (if they perceive the difference or can distinguish if they do), rather than to actually indicate hate speech, resulting in ambiguous or false positive reports of hate speech, if taken literally. Similar issues apply to other fields.

For “total_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.

The data were originally collected or derived from operational information or data sources or otherwise -- and not for research purposes. Features of the dataset may be inaccurate, incomplete, or collected in ways that are not compatible with research goals. Researchers need to adapt their method, research designs, and quantities of interest to the data at hand. Please let us know if you see anything we might be able to adjust in generic ways for everyone.

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.

Reference

Gary King. 1995. “Replication, Replication.” *PS: Political Science and Politics*, 28, Pp. 444-452. http://j.mp/2oSOXJL