CrowdTangle Codebook

Matthew Garmur, Gary King, Zagreb Mukerjee, Nathaniel Persily, and Brandon Silverman

Version 1.1
October 15th, 2019

This document describes the CrowdTangle API and user interface being provided to researchers by Social Science One under its collaboration framework with Facebook. CrowdTangle is a content discovery and analytics platform designed to give content creators the data and insights they need to succeed. The CrowdTangle API surfaces stories, and data to measure their social performance and identify influencers. This codebook describes the data's scope, structure, and fields.

All questions about this codebook should be directed to help@socialscience.one.

Data Access
To obtain access to these data, see the Request for Proposals process at SocialScience.one.

Requirements. (1) Each team granted access to this API must participate in the Social Science One community Slack channel to answer occasional questions from other users. (2) All publications that result must include this citation: Garmur, Matt; King, Gary; Mukerjee, Zagreb; Persily, Nate; Silverman, Brandon, 2019, "CrowdTangle Platform and API", https://doi.org/10.7910/DVN/SCCQYD, Harvard Dataverse, V2

Unit of analysis
A Facebook "post" -- a link, text post, video, or image, shared by a public page, group, or (possibly) verified public person who chooses to make their profile public.

Scope
All posts from Facebook that are:

1. Made by a public page, group, or (possibly) verified public person, who
    a. Has ever (since 2014) had > 110k likes, OR
    b. Has ever been added to a CrowdTangle list by anyone,
2. AND are posted without the poster aiming at a particular audience using Facebook targeting and gating tools (eg. age-gating for alcohol pages, geo-gating if the content has country rights restrictions, targeting to women, etc.)

There are no explicit restrictions on language or country. This number includes 3.5 million Facebook pages and 1.2 million Instagram pages.

Posts that are deleted by the poster, or posts removed for ToS violations (such as foreign propaganda) are removed from CrowdTangle as well.

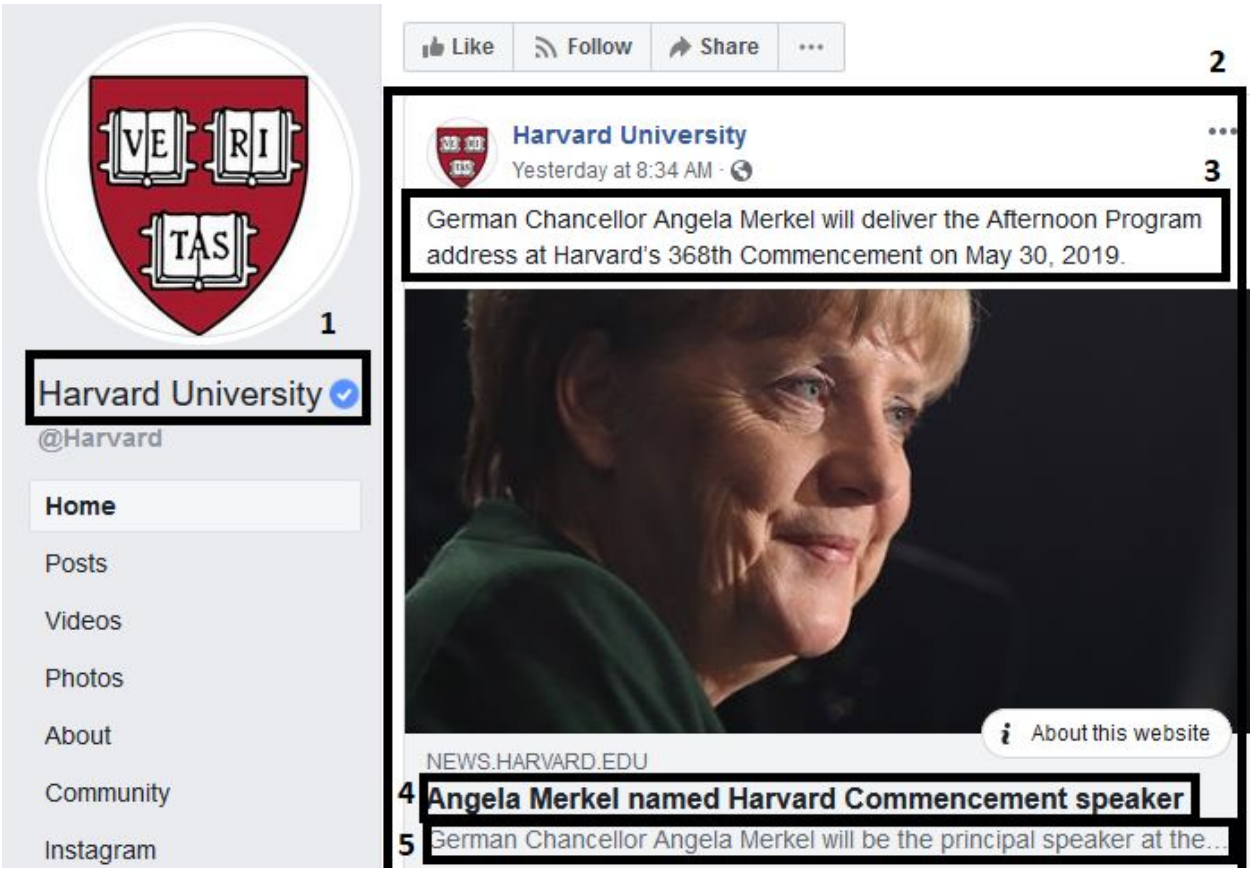
Variables

A data example can be found here (Google Sheet), which is helpful to follow along when understanding the variables below. This particular example is based on US General News, from 9/16/2018 to 10/16/2018. It excludes benchmark information.

Summary Information

Example Post:

1. Name
2. A Post
3. Message
4. Link Text
5. Link Description



Name: The page's unique visible and searchable name

Number of Likes: The size of the page (in terms of Likes on Facebook, not Facebook "followers") at the time the page posted a specific post

Created: The date and time a post was officially posted, UTC time zone. Example: 2018-09-21 05:22:27 EDT

Type: The format of the post. For Facebook this includes links, photo, native video, non-native video (i.e. YouTube links), and live video. For Instagram this includes photos and videos in the post stream. Will be one of these text strings: Photo, Native Video, Link, Status, Live Video

URL: The URL of a post on Facebook.

Message: The blurb of the post, written when the post is uploaded.

Link: The link the publisher uploaded, which could be a link-shortened URL.

Final Link: The unfurled link, if a URL has been shortened.

Link Text: The headline of a link URL or the title of a native video. For example, this will often be a news article title.

Description: For link posts, the sub-header of a link URL: the text that shows up under a link, which is set in the HTML of the linked page (by the author of that page, not by the author of the post)

Sponsor ID: For branded content, the page ID of the marketer, not the page poster.

Branded content - aka a "handshake" is a special feature available to certain brands and pages, where a post on a page can be sponsored by a brand for native advertising. This will show the ID of the brand. ID is a number. It correlates to an address -- eg Nike is 15087023444, and facebook.com/15087023444 is a redirect to facebook.com/nike.

Sponsor Name: For branded content, the Marketer page name.

Score: Based off of CrowdTangle's "overperforming" metric, this is the level at which a post overperformed. The overperformance is computed relative to similar posts from the same page in similar timeframes - high overperformance from a New York Times video posted in the last 15 minutes would mean that the post got more interactions than previously posted New York Times videos in their first 15 minutes of posting.

The score can be computed with the following equations:

	$Intr. \geq Threshold$	$Intr. < Threshold$	$Interactions = 0$
$Interactions \geq Benchmark$	$\frac{Interactions}{Benchmark}$	$\frac{Interactions}{Threshold}$	$-2^{*}Benchmark$
$Interactions < Benchmark$	$-1 * \frac{Benchmark}{Interactions}$	$-1 * \frac{Ben. - Int.}{Benchmark}$	N/A

- *Interactions* is the total number of interactions (like, share etc.). The default behavior is that all of these are simply added together.
- The *threshold* is a minimum set to avoid high variance for small numbers of interactions. For Facebook posts it is 5 likes, 2 of comments/shares/non-like reactions, 100 total page views, and 2 post views. For Instagram it is 5 likes, 2 views, and 2 comments.
- *Benchmark* is the smoothed average of interactions for that page for the last 100 similar posts (so if the post is a Fox News video, the last 100 videos from Fox News). For more details on benchmark calculation see below.
- The setup for different cases is intended to give a relatively smooth curve for different values of interactions and benchmark, without asymptotic behavior for small values of either.

A more detailed description of scores, along with descriptions of the reasoning behind the logic, can be found here.

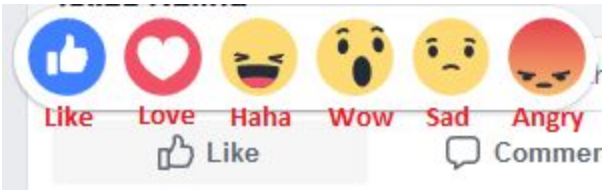
Reactions and Interactions:

Each column in this set of columns is a total number of interactions with posts, of different kinds (likes, comments, shares etc.) The interactions include summary statistics on that post gathered between when the post is posted and the time the API call is made. The summary numbers here do not include either interactions with shares or interactions with comments.

Reactions:

Users can "react" to posts on Facebook to communicate a range of emotional responses. By clicking the "Like" button displayed beneath each post, the user can "Like" the post - a default reaction. By hovering or long-pressing the Like button, a user can access a variety of other reaction types, called: Like, Wow, Sad, Angry, Love, Haha.

Each person can give only one of the reaction types, and give it only once.



Likes: The total number of likes on a Facebook post, created by users clicking the thumbs-up "Like" icon. This does not include other reaction forms.

Love: On Facebook, the total number of love reactions on that post, created by users clicking the heart-shaped icon.

Wow: On Facebook, the total number of wow reactions on that post, created by users clicking the Wow face

Haha: On Facebook, the total number of haha reactions on that post.

Sad: On Facebook, the total number of sad reactions on that post.

Angry: On Facebook, the total number of angry reactions on that post.

Other interactions:

Comments: Users can create a top-level comment on a post by clicking the comment button beneath the post. Users can also click the “Reply” button on a comment to create a second-level comment as a reply to the original, and react to each comment in the same way they would to a post.

This field is the total number of top-level comments on a Facebook or Instagram post. "Top-level" means it does not include “threaded comments,” or replies to comments: the Facebook comments are in a two-tier hierarchy, with comments on the post and replies to comments. For privacy reasons, only the first tier is included. See image below for more details.

Shares: Users may "share" a post to push the post to their own friends and followers. This field represents the total number of shares off of that post, not including shares of shares.

Video Share: On Facebook, if a native video was originally uploaded as this post, or cross-posted from another page. Can be: "original", "share" or blank.

Crossposting is a special Facebook feature available to certain brands/media entities, and means posting from a central video library to several pages. For example, a parent media company might control several local media station pages. It would then be able to crosspost a news video from its central library to local station pages across a state or region. The total number of views for that video across pages is shown here.

Post Views: On Facebook, the number of views a native video accumulated directly from that particular post. This does not include video views accumulated from shares of that post.

Total Views: The combined views for a native Facebook video of both the views from the parent post, and the shares of that parent post.

Total Views for all Crossposts: The total number of views, across all crossposts, for a Facebook native video that has been cross-posted on Facebook. See description of crossposting under "Video Share".

Benchmarks: The CrowdTangle benchmarks are computed for each post and each interaction type. The benchmarks are used in showing over/underperformance, and roughly correspond to the average number of interactions of that type on similar posts by the same page.

Benchmarks are calculated from the last 100 posts across 3 dimensions:

- Account (New York Times, Nike, etc.)
- Post Type (photo, video, link, etc.)
- Age of post (broken into buckets that increase in size as the post ages -- 0-15 minutes old is a bucket, as is 12-15 hours old, as is 6-7 days old)

Within the last 100 posts that share 3 particular dimensions, we sort by each metric (likes, comments, shares, etc.) and then delete the top 25 and bottom 25 to try and account for power law. We then average the middle 50 to get a benchmark for that metric (eg. likes) for that account (eg. NYT) for that type (eg. photo) for that age (eg. 0-15 minutes old). We do this for every iteration and then compare a post's actual data against the benchmark that matches its profile (eg. 10 actual likes vs. 5 expected/benchmarked likes).

As an example, suppose the New York Times posted a photo 12 minutes ago and we want to compute benchmark Likes. We will consider the last 100 photos posted by the NYT and count how many Likes each one got in the first 15 minutes of posting. Then we will throw out the top and bottom 25 photos by Likes, and average the Likes of the remaining 50.

These benchmarks are referred to below as "expected number of likes" etc.

Benchmark Likes: The expected number of likes a post should have for a certain type after a given amount of time.

Benchmark Comments: The expected number of comments a post should have after a given amount of time.

Benchmark Shares: On Facebook, the expected number of shares a post should have after a given amount of time.

Benchmark Love: On Facebook, the expected number of love reactions a post should have after a given amount of time.

Benchmark Wow: On Facebook, the expected number of wow reactions a post should have after a given amount of time.

Benchmark Haha: On Facebook, the expected number of haha reactions a post should have after a given amount of time.

Benchmark Sad: On Facebook, the expected number of sad reactions a post should have after a given amount of time.

Benchmark Angry: On Facebook, the expected number of angry reactions a post should have after a given amount of time.

Benchmark Post Views: For Facebook native videos, the expected number of post-level video views a post should have after a given amount of time.

Benchmark Total Views: For Facebook native videos, the expected number of post-level plus shared video views a post should have after a given amount of time.

Benchmark Total Views for all Crossposts: For crossposted videos on Facebook, the expected number of crossposted video views a post should have after a given amount of time. ee description of crossposting under “Video Share”.

Benchmarks and Timesteps in Post CSVs:

B	C	D	E	F	G	H	I
Score Date (GMT)	Timestep	Likes	Average Like	Comments	Average Com	Shares	Avg Shares
9/28/19 11:07	44	12758	131	1271	15	13149	24
9/28/19 8:09	43	12343	131	1240	15	12705	24
9/28/19 5:18	42	11932	128	1206	14	12294	23
9/28/19 2:16	41	10301	128	1054	14	10639	23
9/27/19 23:05	40	8103	125	840	14	8375	23
9/27/19 20:17	39	6517	122	646	14	6642	23
9/27/19 19:31	38	6093	121	599	14	6155	22
9/27/19 18:03	37	5310	119	521	13	5243	22
9/27/19 17:11	36	4857	118	477	13	4706	22
9/27/19 16:16	35	4376	116	428	13	4121	22
9/27/19 15:21	34	3890	114	374	13	3525	21
9/27/19 14:04	33	3240	112	303	13	2702	21
9/27/19 13:14	32	2909	110	268	13	2314	20
9/27/19 12:01	31	2505	108	230	13	1912	20
9/27/19 11:16	30	2356	106	215	12	1734	20
9/27/19 10:07	29	2231	104	206	12	1594	19
9/27/19 9:00	28	2186	102	205	12	1551	19

In the Benchmarks section of this codebook, “age of post” is one of the 3 dimensions used to create a benchmark. The dimensions are there to cluster posts based on relevant information, and posts typically gain more engagement the longer they exist, so we chose to bucket posts within comparable ages.

Since posts on many social media platforms tend to display more variability early in their lives than later, our time buckets (“timesteps”) start off very short, and grow as they get older. Comparing a popular post that’s 2 hours old to a post that’s 15 minutes old does not seem terribly relevant, whereas comparing a post that’s 19 days and 2 hours old to a post that’s 19 days and 10 hours old could be an actionable comparison.

The timestep sizes follow a shape that mimics a logarithmic curve, though it’s not actually logarithmic. The first timesteps are each 15 minutes long, then they grow to 30 minutes, and eventually end up at a full 24 hours.

End times specified in the link below are exclusive of the actual final moment. For example, “0-15 minutes” means between 0 and anything just under 15 minutes. Once it hits 15 minutes exactly, that becomes part of the “15-30 minutes” timestep.

Timesteps are listed below:

Timestep	Age of Post (ending value exclusive)	Timestep	Age of Post	Timestep	Age of Post
0	0 - 15 minutes	25	11 - 11.5 hours	50	2 days 18 hours - 3 days
1	15 - 30 minutes	26	11.5 - 12 hours	51	3 days - 3 days 6 hours
2	30 - 45 minutes	27	12 - 13 hours	52	3 days 6 hours - 3 days 12 hours
3	45 - 60 minutes	28	13 - 14 hours	53	3 days 12 hours - 3 days 18 hours
4	60 - 75 minutes	29	14 - 15 hours	54	3 days 18 hours - 4 days
5	75 - 90 minutes	30	15 - 16 hours	55	4 days - 4 days 12 hours
6	1.5 - 2 hours	31	16 - 17 hours	56	4 days 12 hours - 5 days
7	2 - 2.5 hours	32	17 - 18 hours	57	5 days - 5 days 12 hours
8	2.5 - 3 hours	33	18 - 19 hours	58	5 days 12 hours - 6 days
9	3 - 3.5 hours	34	19 - 20 hours	59	6 days - 6 days 12 hours
10	3.5 - 4 hours	35	20 - 21 hours	60	6 days 12 hours - 7 days
11	4 - 4.5 hours	36	21 - 22 hours	61	7 days (as in, 7 days up to just under 8 days)
12	4.5 - 5 hours	37	22 - 23 hours	62	8 days
13	5 - 5.5 hours	38	23 - 24 hours	63	9 days
14	5.5 - 6 hours	39	1 day - 1 day 3 hours	64	10 days
15	6 - 6.5 hours	40	1 day 3 hours - 1 day 6 hours	65	11 days
16	6.5 - 7 hours	41	1 day 6 hours - 1 day 9 hours	66	12 days
17	7 - 7.5 hours	42	1 day 9 hours - 1 day 12 hours	67	13 days
18	7.5 - 8 hours	43	1 day 12 hours - 1 day 15 hours	68	14 days
19	8 - 8.5 hours	44	1 day 15 hours - 1 day 18 hours	69	15 days
20	8.5 - 9 hours	45	1 day 18 hours - 1 day 21 hours	70	16 days
21	9 - 9.5 hours	46	1 day 21 hours - 2 days	71	17 days
22	9.5 - 10 hours	47	2 days - 2 days 6 hours	72	18 days
23	10 - 10.5 hours	48	2 days 6 hours - 2 days 12 hours	73	19 days
24	10.5 - 11 hours	49	2 days 12 hours - 2 days 18 hours	74	20 days or more