Count the Interactions, Not the Score: What CrowdTangle's Overperforming Score Measures in Archived Facebook Data

Tristan Hotham*

2026

Preprint, not peer reviewed

Abstract

CrowdTangle closed in August 2024, but many archived exports of Facebook posts remain in use, and each row carries an "Overperforming Score" that compares a post's interactions with an expected value. I examine what this score measures in 856,699 posts from 186 party and leader pages in 20 European countries, 2010 to 2020, exported in October 2020. The score is present for every post and follows the documented sign convention. But the expected value it uses does not reflect the time of posting. Inverting the documented formula, 95% of posts have an implied benchmark within 2% of a single value for their page and post type, and within each page and type the score ranks posts almost exactly as raw interactions do (median rank correlation above 0.999). A constant built from each page's last 100 posts before the export fits the score closely (correlation 0.97 on the log scale), whereas benchmarks built from the posts before each post do not (0.52 to 0.56). As a result, older posts look like heavy underperformers: the median score is -25 for 2010 and -1.1 for 2020. The score removes differences between pages, so populist and leader advantages in raw interactions disappear, while comparisons within pages simply reproduce raw counts. In archived exports the score is best read as raw interactions rescaled by a page and type constant set around the time of export, so researchers should work with raw interaction counts, compared within page and period, or with baselines they build and report themselves.

Keywords: CrowdTangle; overperforming score; Facebook; engagement metrics; measurement; data archives; political parties; research methods

1 Introduction

CrowdTangle, the public insights tool owned by Meta, was the main route by which researchers studied public Facebook pages until it closed in August 2024. Much of what was collected through it survives as CSV exports held by researchers, and these archives will be reused for years. Besides raw counts of likes, comments, shares and reactions, each exported row includes an "Overperforming Score", shown in the dashboard as values such as $+2.5$x or -1.3x. The score looks attractive to researchers because it seems to solve a hard problem: comparing posts across pages of very different sizes and across years in which Facebook's reach changed.

*Independent researcher. ORCID 0000-0003-4931-9349. Correspondence: contact@tristanhotham.com. The views expressed are the author's own and do not represent those of any employer.

CrowdTangle documented the score as a post's interactions divided by an expected value, its "benchmark", computed from the account's last 100 posts of the same type: the top and bottom quarter are dropped and the rest averaged, separately for each post age, such as 15 minutes, 60 minutes or 5 hours after posting (CrowdTangle nd). Underperforming posts are given negative scores, $-1\times$ benchmark/actual, and posts with no interactions $-2\times$ benchmark. Separate rules give scores between -1 and 1 when the benchmark is below a platform minimum, and interaction types can be weighted, by default equally (CrowdTangle nd; Garmur et al. 2019). The codebook describes the comparison as one with similar posts from the same page in similar timeframes (Garmur et al. 2019). Neither source says when "the last 100 posts" are counted for a post that is already old when the data are requested. In a historical export, the answer decides whether the score means anything for a post made years earlier.

This note asks three questions. **RQ1.** How complete is the score in an archived export, and how is it distributed? **RQ2.** Which baseline does it use, and can it be reconstructed from the data? **RQ3.** Does using the score instead of raw or standardised counts change typical substantive comparisons?

2 Data and method

Posts. The posts come from CrowdTangle lists of official party and party leader pages built for my doctoral research and exported between 5 and 28 October 2020 (CrowdTangle Team 2020; Hotham 2021). The analysis file holds 856,699 posts from 112 party pages and 74 leader pages in 20 countries, from January 2010 to 30 September 2020. When a post appeared in more than one export, the file keeps the first, matched by post URL; 23,053 posts (2.7%) come from the earliest export, of 5 October, and the rest from exports of 27 and 28 October. Interactions are the sum of likes, comments, shares and the six other reactions.

Implied benchmark. Inverting the documented formulas, the benchmark implied by a score s for a post with x interactions is x/s if $s \geq 1$, $-sx$ if $s \leq -1$, and $-s/2$ if $x = 0$. It is undefined for 582 posts with at least one interaction and a score strictly between -1 and 1, which follow the separate minimum-benchmark rules. I also convert each score to a signed log ratio of actual to expected interactions, $\log s$ for $s \geq 1$ and $-\log(-s)$ for $s \leq -1$. This is undefined for 615 posts (the 101 zero scores and 514 others between -1 and 1), which are excluded from analyses on the log scale.

Candidate baselines. I compare two families of baseline. *Post-time* baselines use only posts made before each post, by the same page and of the same type: the trimmed mean (middle half), median or mean of the previous 100 posts, the trimmed mean of the previous 20, and the mean over the previous 30 days; a baseline needs at least 10 earlier posts. *Export-time* baselines use the page's last 100 posts of that type in the data, up to 30 September 2020, and apply the same value to all of that page's posts of that type. As a check, I also try two other page-type constants: the trimmed mean of the page's first 100 posts of the type, and of all of them. For each, I compute the predicted log ratio and correlate it with the one implied by the score.

Substantive comparisons. I estimate three typical comparisons under four outcomes: raw interactions, $\log(1+x)$; the score as a log ratio; a within page-year z-score of $\log(1+x)$; and the log ratio to the post-time trimmed-mean baseline. Populist against mainstream parties (party pages, PopuList coding as in Hotham 2026a) and leader against party pages are compared within country-months. Posts in the last 30 days before a national parliamentary election (ParlGov dates, available for 16 of the 20 countries) are compared with the same page's other posts, with page and calendar-month fixed effects. Standard errors are clustered by page, with country clustering reported where relevant.

Stability across exports. Using the raw export files, I match by post URL the posts that appear both in the export of 5 October 2020 and in an export of 27 October, and posts that appear in both a 27 October and a 28 October export. All matched pairs have the same page and post type, and 99.9% the same posting time to within a minute.

3 Results

3.1 Coverage and distribution

The score is present for every post: none is missing. It is rounded to two decimals, giving 12,377 distinct values, and the 50 most common values cover 12% of posts. 27.6% of posts score 1 or more, 72.3% score -1 or less, and 615 (0.07%) fall between them, including 101 exact zeros. The 514 non-zero scores in this band are concentrated in 17 page-type combinations and 78% have 10 or fewer interactions, consistent with the documented rule for benchmarks below a minimum. The median is -2.17, the quartiles -6.25 and 1.10, and the tails are very asymmetric: the 1st percentile is -270 and the minimum $-30{,}658$, while the 99th percentile is 8.8 and the maximum 279. Of the 11,171 posts with no interactions, 99.7% have negative scores, as the $-2\times$ benchmark rule implies, and 96% share the most common score of their page and type. One value stands out: 3,442 posts score exactly -118. Of these, 3,402 (99%) are the zero-interaction links of a single party page, mostly from 2012 and 2013, whose implied link benchmark is 59: the cluster is what the $-2\times$ benchmark rule gives when one benchmark covers all of a page's posts of a type, not a separate default.

The score falls steeply with the age of the post. The median is -25.1 for 2010, -8.8 for 2012, -1.6 for 2015 and -1.1 for 2020, and the share of posts scoring 1 or more rises from 1% in 2010 to 44% in 2020.

3.2 What the score is relative to

The implied benchmark barely varies within a page and post type. Across 1,367 page-type combinations, 95.5% of posts have an implied benchmark within 2% of the median for their page and type, and 97.4% within 10%. This holds for 95% to 99.5% of posts in each year from 2010 to 2019, but only for 75% in 2020, when many posts were days or weeks old at export and so fell into younger age buckets with their own benchmarks. Because the benchmark is close to a constant, the score within a page and type is close to a monotone transformation of raw interactions. Among the 1,008 page-types with at least 30 posts, the median rank correlation between score and interactions is 0.99997 (10th percentile 0.998). A regression of the score's log ratio on log interactions with page-by-type fixed effects has a slope of 0.99 and an R^2 of 0.99.

The constant reflects the end of the period, not the time of posting (Table 1). Benchmarks from each page's last 100 posts of the same type before the export fit the score closely: the correlation of predicted with implied log ratios is 0.97, the median absolute error 0.13 log units, and the sign agrees for 94% of posts. Every post-time baseline does much worse, with correlations of 0.52 to 0.56 and sign agreement of 67% to 74%. The pattern is the same on a common sample of 818,687 posts with all baselines defined (0.97 against 0.53 to 0.56). Part of this fit comes from using any constant, but the choice of constant matters: the trimmed mean of all of a page's posts of a type gives a correlation of 0.85, and that of its first 100 posts 0.47. My export-time approximation is still not exact: its benchmark correlates at 0.90 (log scale) with the implied one across 1,366 page-types and is within 25% of it for 55%, against 28%

for the all-posts constant and 22% for the first-100 constant. The gap may reflect posts up to the export date, age-specific values or weights; the data cannot tell these apart. Among post-time baselines the trimmed mean and median of the previous 100 posts fit marginally best, but none comes close.

Table 1. How well candidate baselines reproduce the score. Correlation of the log ratio implied by the score with the predicted log ratio; rank correlation of CrowdTangle-style signed scores; median within-page rank correlation; share of posts with the same sign.

Baseline (same page and type)	Posts	r (log)	Rank r	Within page	Same sign
Previous 100 posts, trimmed mean	832,585	0.55	0.56	0.65	0.69
Previous 100 posts, median	831,896	0.56	0.56	0.66	0.67
Previous 100 posts, mean	833,013	0.53	0.53	0.65	0.74
Previous 20 posts, trimmed mean	832,535	0.54	0.54	0.62	0.69
Previous 30 days, mean	829,078	0.52	0.52	0.58	0.74
Last 100 posts at export, trimmed mean	844,946	0.97	0.96	0.99	0.94
Last 100 posts at export, median	843,429	0.97	0.96	0.98	0.93

3.3 Consequences for substantive comparisons

Because the benchmark reflects each page's typical post in 2020, the score mostly records how far a post falls short of what the page achieved at the end of the period. The median log ratio implied by the score rises from −3.2 in 2010 to −0.1 in 2020, in step with raw interactions, whereas the ratio to a post-time baseline stays close to zero throughout (Figure 1, left). Across posts, the score is barely related to page size (rank correlation with page likes at posting 0.11, against 0.64 for raw interactions).

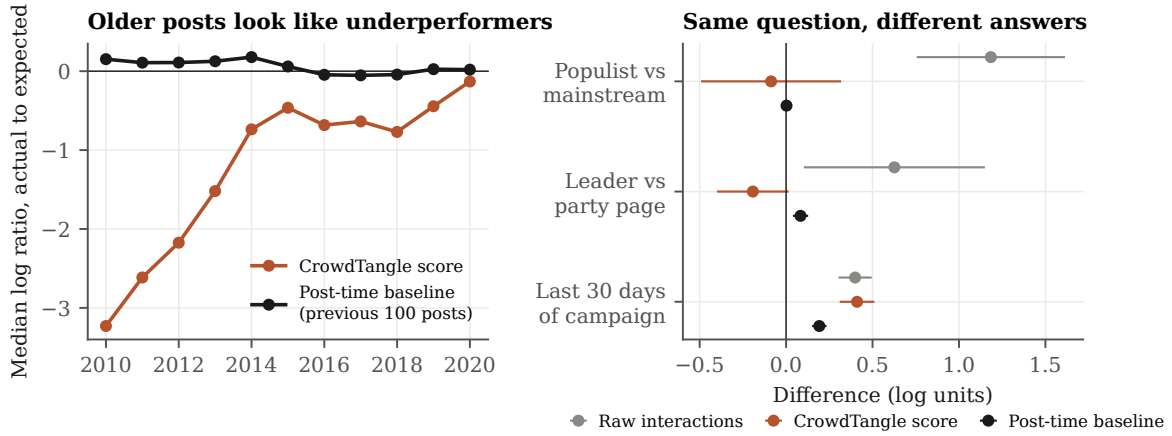


Figure 1. Left: median log ratio of actual to expected interactions by year of posting, as implied by CrowdTangle's score and as computed against the trimmed mean of the page's previous 100 posts of the same type. Right: differences in log units with 95% confidence intervals clustered by page, for three comparisons under three outcomes (specifications as in Table 2).

The choice of outcome changes the answer (Table 2, Figure 1, right). In raw interactions, populist parties' posts drew about 1.18 log units more than mainstream parties' posts in the same country and month (SE 0.22), roughly 3.3 times as many, consistent in direction with earlier work on these data (Hotham 2026a). Measured with the score, the difference is −0.09 (SE 0.21), and with a within-page z-score or a post-time baseline it is close to zero. Leaders' pages drew about 87% more than party pages in raw terms (0.63, SE 0.27) but slightly less on the score (−0.19, SE 0.11), a difference not distinguishable

from zero. These measures remove differences between pages by construction, so they ask whether a page did well relative to itself.

Within pages, the score and raw counts agree. Posts in the last 30 days of a campaign drew about 0.40 log units (49%) more interactions than the same page's other posts (SE 0.05; 0.09 clustered by country), and 0.41 (51%) more on the score. Against a post-time baseline the campaign difference is about half as large, 0.19 (21%, SE 0.02), because the baseline itself rises as the campaign goes on; the within page-year z-score gives 0.24 standard deviations (SE 0.03).

Table 2. Three comparisons under four outcomes. Coefficients with standard errors clustered by page in parentheses. Populist: party pages, country-month fixed effects. Leader: all pages, country-month fixed effects. Campaign: last 30 days before a national election, 16 countries, page and calendar-month fixed effects. The z-score is in standard deviations; other outcomes in log units.

Outcome	Populist	Leader page	Campaign
Raw, $\log(1+x)$	1.18 (0.22)	0.63 (0.27)	0.40 (0.05)
CrowdTangle score, log ratio	−0.09 (0.21)	−0.19 (0.11)	0.41 (0.05)
Within page-year z-score	0.003 (0.006)	−0.008 (0.004)	0.24 (0.03)
Post-time baseline, log ratio	0.003 (0.016)	0.08 (0.02)	0.19 (0.02)
Posts (raw outcome)	669,760	856,695	755,652

3.4 Pitfalls

Export date. The score and the counts change between exports. Of 31,840 posts from five UK pages present in both the 5 October export and a late October export, 65% have the same score and 73% the same interaction count. Among the 31,139 posts more than 60 days old, interactions fell for 26% and rose for 0.2%, by tiny amounts (median absolute change 0.2% among those that changed), and 9% kept the same count but changed score. Implied benchmarks moved for some pages, by up to 34% for one page's videos. Exports a day apart were almost identical (99.6% the same score, 76,864 posts from 13 pages). Rank order is stable (rank correlation 0.99), but scores from different exports are not strictly comparable.

Type-specific baselines. Each post type has its own benchmark, so differences in score between types partly reflect how each type's 2020 benchmark compares with its history, not how well each format worked.

The May 2015 capture change. The number of photo posts in these data jumps in May 2015: 826 photo posts were captured in April 2015, against 1,874 to 3,346 a month from May to December 2015, while link posts stayed flat (3,384 in April, 2,551 to 3,668 a month from May to December). The number of pages with any photo post rose from 112 to 140. This looks like a change in capture rather than in posting, so counts of photo posts are not comparable across the break. Per-post interactions on photos also shift: within pages, the mean interactions per photo post were 1.58 times higher (median across 127 pages; 1.23 times for the median per post) in June to September 2015 than in January to April, against 1.07 times (1.03) for links; this may partly reflect which photos were captured. Because the archived score uses a single benchmark from the end of the period, this per-post shift passes straight through: the median photo log ratio moves from −1.00 in April 2015 to −0.49 in May, while links barely move (−0.25 to −0.27). A post-time baseline absorbs it gradually (median photo ratio 0.33 and 0.22 in May and June 2015, near zero from July).

Ties and zeros. Rounding produces many ties, scores of posts with no interactions depend only on the benchmark (hence clusters such as -118), the band between -1 and 1 follows a separate rule, and the asymmetric scale, from about $-30{,}000$ to 280, makes means of raw scores misleading.

4 Discussion

In archived exports, the score does not measure how a post did against what was normal for the page when it was posted. It is, to a close approximation, the post's interactions divided by a page and type constant that reflects the page's posts around the time of export, with a sign flip below one. This is consistent with the documented method if "the last 100 posts" were counted when the data were requested rather than when each post was made, although the documentation does not say so and I cannot observe the calculation itself. In the live dashboard, where most posts viewed were recent, the two readings would largely coincide. For a post from 2012 in an export from 2020, they do not.

This has three practical consequences. First, within a page and post type, the score adds nothing beyond raw interactions, and comparisons over time inherit every trend in raw counts, including growth in audiences, the 2018 News Feed change (Hotham 2026b) and capture changes such as that of May 2015. Second, across pages, it removes each page's typical interactions around 2020, not at the time, so it is not a valid size adjustment for earlier years. Third, raw interaction counts compared within page and period, for instance with page and month fixed effects, do the same job more transparently; if a relative measure is wanted, it is better to build one openly, for instance against the trimmed mean of the page's previous 100 posts of the same type, while recognising that such a baseline absorbs slow trends and campaign build-ups, and so answers a narrower question. Researchers should report the baseline, the export date and how many posts lack one.

The analysis has limits. The data cover one set of political pages in Europe and one export period, and the documentation may not describe every version of the method. My export-time reconstruction fits far better than post-time baselines but does not match the level of the benchmark exactly, and I cannot observe age-specific values or interaction weights. The overlap between exports covers only five pages. The findings concern the score as exported in historical reports; scores collected in real time through the API, close to posting, may behave as documented.

5 Conclusion

Across 856,699 posts, CrowdTangle's Overperforming Score in an archived export behaves as interactions relative to each page's typical post around the time of export, not at the time of posting. Within pages it is close to raw counts in disguise; across pages it removes each page's level as it stood near the export. For these data, researchers reusing CrowdTangle archives are better served by counting the interactions: comparing raw counts within page and period, or against baselines they construct and report themselves, rather than treating the score as a time-consistent performance measure.

Data and code

The posts were exported from CrowdTangle in October 2020. CrowdTangle's terms do not permit sharing the raw posts. The analysis code, the list of pages and post-level metric aggregates without text

are available from the author.

Provenance and AI use

The data were exported from CrowdTangle in October 2020 for my doctoral research. This paper builds on my doctoral thesis (Hotham 2021). The analysis and writing were completed in 2026.

I designed the study and am responsible for the analysis and conclusions. I wrote the text, with drafting and editing assistance from Claude (Anthropic), which also helped prepare the analysis code. I reviewed and checked all text and results.

Ethics

Not applicable: the study involves no human participants. The data are public posts by the official Facebook pages of political parties and party leaders, analysed only as counts of interactions and CrowdTangle scores aggregated across posts. No post content is quoted or reproduced, and no data on individual users are used.

Funding

I received no specific funding for this work. It builds on my doctoral research, which was funded by the Economic and Social Research Council (ESRC).

References

CrowdTangle (n.d.). How do you calculate overperforming scores? CrowdTangle Help Center. https://help.crowdtangle.com/articles/2013937-how-do-you-calculate-overperforming-scores, accessed October 2026.

CrowdTangle Team (2020). CrowdTangle. Facebook, Menlo Park, CA.

Garmur, M., King, G., Mukerjee, Z., Persily, N., and Silverman, B. (2019). CrowdTangle platform and API. Harvard Dataverse, V2. https://doi.org/10.7910/DVN/SCCQYD. Codebook version 1.1, 15 October 2019.

Hotham, T. (2021). *How Do Political Parties Use Facebook and What Does It Offer to Their Campaigns*. PhD thesis, University of Bath. https://doi.org/10.5281/zenodo.23300282.

Hotham, T. (2026a). Do populists win on Facebook? The engagement premium of populist parties in Europe, 2010 to 2020. Preprint, Zenodo. https://doi.org/10.5281/zenodo.23270349.

Hotham, T. (2026b). Shared more, liked less? Facebook's 2018 news feed change and the pages of European parties and leaders. Preprint, Zenodo. https://doi.org/10.5281/zenodo.23403683.