Unveiling the Impact of Within-Content Engagement Information: Evidence from a Natural Experiment on YouTube

Keyan Zhu Vineet Kumar

July 29, 2024

Abstract

Keywords: within-content information, natural experiment, video consumption

1 Introduction

2 Related Literature

3 Institutional Setting and Data

3.1 Setting

With more than 2.5 billion monthly users[1] who collectively watch more than one billion hours of videos every day[2], YouTube is the world's largest video-sharing platform and the second most-visited website in the world.[3] Unlike TikTok, a younger competitor that features short, snappy videos mostly ranging from 15 seconds to a minute, YouTube is known for its long-form content, with videos usually ranging from a couple of minutes to a few hours.

Is there an official name for this? In May 2022, YouTube introduced a feature which we termed the "YouTube Engagement Graph," that depicts the moment-to-moment engagement levels throughout a video's duration (see Figure 1). The height of this graph corresponds to how frequently a particular part of the video has been replayed, indicating its engagement level and popularity among viewers[4]. A taller section of the graph suggests a more frequently revisited video segment.

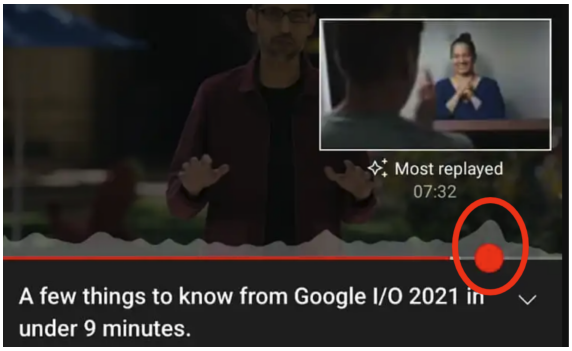


Figure 1: The YouTube Engagement Graph

The YouTube engagement graph feature has been uniformly incorporated across all YouTube platforms, including its web player, Android app, and iOS app. This feature is accessible to *all* users regardless of their subscription status; once a video has an engagement graph, it is visible to everyone. Once videos are endowed with engagement graphs, they

[1] https://www.forbes.com/advisor/business/social-media-statistics/

[2] https://web.archive.org/web/20200806062438/https://blog.youtube/news-and-events/you-know-whats-cool-billion-hours/

[3] https://www.semrush.com/website/top/

[4] https://support.google.com/youtube/thread/164099151/youtube-video-player-updates-most-replayed-video-chapters-single-loop-more?hl=en

are never removed. The translucent engagement graph shows up automatically when a user hovers her mouse over the red progress bar at the bottom of the player.

The engagement graph does not appear immediately upon a video's upload; typically, a couple of days of data accumulation are required before it is activated. Once activated, it remains visible. Some videos may display the engagement graph soon after uploading, while others may not, even after an extended period.

Notably, YouTube retains full control over the activation of engagement graphs for videos, with neither users nor content creators having any control over its visibility. Users and content creators cannot opt out of this feature. The criteria guiding the activation of the engagement graph are not publicly disclosed by the platform. As will be discussed in Section 4, the rules for engagement graph activation witnessed arguably arbitrary alterations during our data collection period. Such policy changes provide us with a unique natural experiment environment, enabling us to identify the causal impact of the engagement graph on user bahaviors.

3.2 Data Collection and Description

To study the impact of the engagement graphs on user response, we compile a novel dataset combining two different data sources. The data collection process spanned from February 4th to April 24th, 2023. collected daily data? or what frequency?

Our first data source is the official YouTube Data API, which we rely on to construct random samples of newly uploaded videos.

Every day, we submitted a search request for videos from all of the 16 official YouTube categories, including "Gaming," "Entertainment," "Science & Technology," "People & Blogs," and more. Within each category, we randomly sampled a maximum of 300 videos that satisfied the following criteria: were of medium length (ranging from 4 to 20 minutes) and had been published within the previous 48 hours.[5] We then filtered out videos with non-English titles or descriptions to maintain a consistent audience appeal. This approach allowed us to add approximately 1,700 new videos to our panel dataset daily.[6]

The official YouTube API does not directly offer engagement graph data. To bridge this gap, we built a web scraper to retrieve the engagement graph data from each video's webpage source code. Each engagement graph consists of 100 data points, one for each percentage of the video timeline, with the graph's height normalized between 0 and 1. What users see

[5]Note that for some niche categories, there might be fewer than 300 videos that met the criteria

[6]Our sample collection operated at a more modest scale before Feb 17th, yielding between 300 to 1,300 new videos daily.

is a line connecting the 100 points (see Figure 2 for an illustration).[7] Our scraper detects whether an engagement graph exists and extracts the graph data points if present.

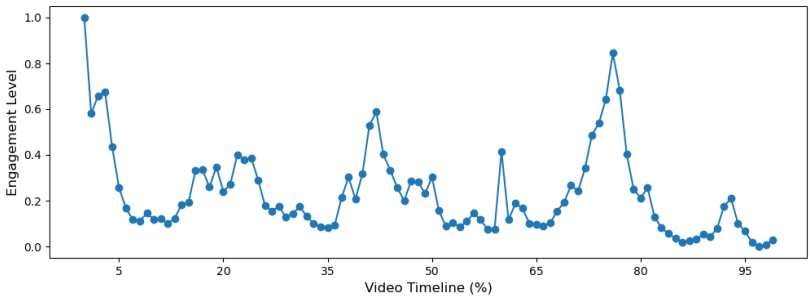


Figure 2: An Example of the YouTube Engagement Graph

Every day, we use the official YouTube Data API to sample newly uploaded videos. We then use our scraper to collect time-series data for videos already incorporated into our panel, capturing detailed information including the video title, description, length, upload date, categories, channel, number channel subscribers, views, likes, and comments, and the engagement graph.

We track each video's data for a period of 25 days.[8]

Our final dataset comprises a total of 126,507 videos. Table 1 presents the number of videos in each category. Table 2 summarizes the video-level statistics for videos with complete observations within the first 14 days after upload. On average, the videos are 628.90 seconds (approximately 10.5 minutes) long, close to the average duration (?) number recorded in industry reports.[9] By the end of the two-week period post-upload, the videos accumulate an average of 186.14 thousand views, 5,958.60 likes, and 217.48 comments. Channels uploading these videos have an average of 2.84 million subscribers. The standard deviations are large for the cumulative metrics, indicating a substantial variation across videos for each of these measures.

Furthermore, 41% of all videos have an engagement graph by the end of two weeks after upload. Among these, the engagement graph is present for an average of 7.13 days on the 14th day post-upload, with a standard deviation of 3.09 days. This variation highlights the differences across videos not only in terms of the presence of the engagement graph but also in the timing of its activation within the video lifecycle.

Table 3 compares the video-level statistics for videos with and without an engagement

[7]Note that our extraction process captures the precise data displayed on the web page, ensuring we mirror exactly what users see, without any loss of information.

[8]The data collection stopped on April 24th, 2023, which is when we ceased adding new videos to the panel. For videos uploaded before April, we tracked their data for 25 days; for videos uploaded in April, we tracked their data until April 24th.

[9]https://www.statista.com/statistics/1026923/youtube-video-category-average-length/

Table 1: Number of Videos per Category

Category	**Number of Videos**
Sports	11,594
Education	11,443
Science & Technology	11,288
Gaming	11,111
News & Politics	8,804
Howto & Style	8,708
Autos & Vehicles	8,456
Entertainment	8,288
Comedy	7,212
People & Blogs	7,075
Pets & Animals	7,022
Travel & Events	6,996
Music	6,785
Film & Animation	6,688
Nonprofits & Activism	4,969

Notes: The "Anime/Animation" category is omitted as it contains only 69 videos. Although we submitted the same request capacity of 300 videos per day for each category, the final number of videos in our dataset varies. This variation is due to two primary reasons: (1) the 300-video limit, as some categories did not have enough videos meeting the criteria, resulting in fewer than 300 videos being returned; and (2) the exclusion of non-English videos, which may disproportionately affect different categories.

graph on the 14th day post-upload. On average, videos with engagement graphs are longer (656.45 seconds) than those without engagement graphs (609.49 seconds). Additionally, videos with engagement graphs have significantly higher engagement metrics, including more views (357.38 thousand versus 65.48 thousand), likes (11,188.42 versus 2,224.3), and comments (364.13 versus 128). Moreover, these videos originate from more popular channels, with an average of 4.23 million subscribers compared to 1.86 million for channels without engagement graphs. All group differences are statistically significant at the 0.1% level.

4 Empirical Approach

4.1 Empirical Challenges

Our goal is to identify the causal impact of the engagement graph, which represents within-content engagement information on user consumption and engagement metrics. However, we face two key empirical challenges.

First, there is concern of selection bias at the video level. Although the platform has not disclosed the criteria guiding the activation of engagement graphs, videos are likely selectively chosen to receive them. As described in Section 3.2, we find systematic differences between

Table 2: Descriptive Statistics (Video-Level)

	Variable	**Mean**	**SD**	**N**
Static	Video length (seconds)	628.90	254.23	97,369
Cumulative	Views (thousand)	186.14	689.57	97,369
Cumulative	Likes	5,958.60	26,666.57	96,024
Cumulative	Comments	217.48	254.40	84,553
Cumulative	Channel subscribers (thousand)	2,837.66	9,390.32	97,361
EngGraph Metrics	Has EngGraph (binary)	0.41	0.49	97,369
EngGraph Metrics	Days Post EngGraph	7.13	3.09	40,248

Notes: The reported sample is restricted to the 97,369 videos with no missing data in key variables for the first 14 days post-upload. The unit of observation is a video-day. I thought it was 2 days, right? *Static* variables are constant at the video level. *Cumulative* and *EngGraph Metrics* variables are measured on the 14th day post-upload. The number of comments is capped at 999; less than 0.01% of videos exceed this value. Missing data for various variables may occur due to factors such as access permissions, comment availability restrictions, and technical issues.

Table 3: Descriptive Statistics (Video-Level) by Graph Presence

Variable	**With Graph**		**Without Graph**	
	Mean	**SD**	**Mean**	**SD**
Video length (seconds)	656.45	254.82	609.49	252.01
Views (thousands)	357.38	979.44	65.48	315.31
Likes	11,188.42	36,450.23	2,224.30	15,385.59
Comments	364.13	285.44	128.00	182.67
Channel subscribers (thousand)	4,229.31	10,614.92	1,856.95	8,281.85

Notes: The reported sample, unit of observation, and variable measurement are the same as in Table 2. The statistics are reported separately for two groups: videos with and without an engagement graph on the 14th day post-upload.

videos with and without engagement graphs. For instance, videos with engagement graphs come from more popular channels, suggesting they may have intrinsically higher quality or appeal. Consequently, a simple comparison between videos with and without engagement graphs would be misleading.

One way to address this challenge is to adopt a Difference-in-Differences (DiD) strategy, where we compare changes in user consumption and engagement metrics for videos with engagement graphs (treated videos) before and after graph activation against those of videos without engagement graphs (control videos). This approach accounts for any unobserved time-invariant video-specific disparities between treated and control videos. Under the parallel trends and no anticipation assumptions, the causal effect of the engagement graph can be identified.

However, there is a second challenge concerning the potentially selective timing of en-

gagement graph activation (treatment). This manifests in two ways: firstly, the timing of treatment is not evenly spread over the calendar dates. As illustrated in Figure 3(a), the number of videos receiving engagement graphs varies significantly by date. For example, no videos started receiving treatment on April 4, while 6,934 videos began receiving treatment on April 5. Secondly, as discussed in Section 3.2, the timing of treatment varies within individual video lifecycles. Some videos receive engagement graphs shortly after upload, while others experience a longer wait before graph activation. Both aspects pose concerns that may invalidate a standard DiD strategy. For instance, if the platform only activates the engagement graphs during holidays and weekends that disproportionately affects the treated and control videos, or alternatively, if it activates the graph for a video when it detects a certain downward trend in engagement metrics, then there exist unobserved, time-varying, video-specific factors correlated with both the treatment and the engagement metrics outcome, violating the identification assumptions of the DiD strategy.

Both empirical challenges essentially stems from our lack of knowledge about the factors that influence the platform's algorithm for activating the engagement graphs.

To address these challenges, we employ a two-step approach. First, we reverse engineer the criteria used by the platform for activating the engagement graphs.

Next, we exploit a natural experiment that introduces arguably exogenous changes in the activation criteria, with variation across videos. This allows us to construct fully comparable treated and control groups of videos, enabling us to estimate the causal impact of engagement graphs using a DiD strategy.

4.2 Engagement Graph Activation Criteria

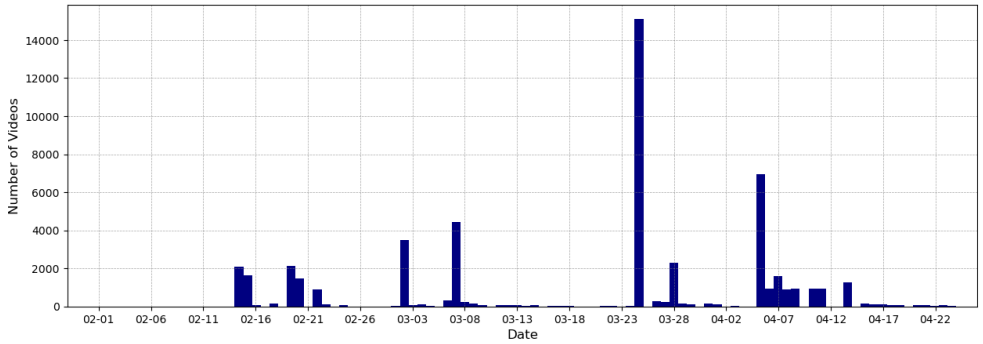
To address the empirical challenges outlined in the previous section, we first seek to reverse-engineer YouTube's criteria for activating engagement graphs. This process is crucial for understanding the platform's decision-making process and forms the foundation for our identification strategy.

In this section, we first describe the activation criteria we uncovered, and then validate their performance through out-of-sample testing and comparison with other benchmarks.

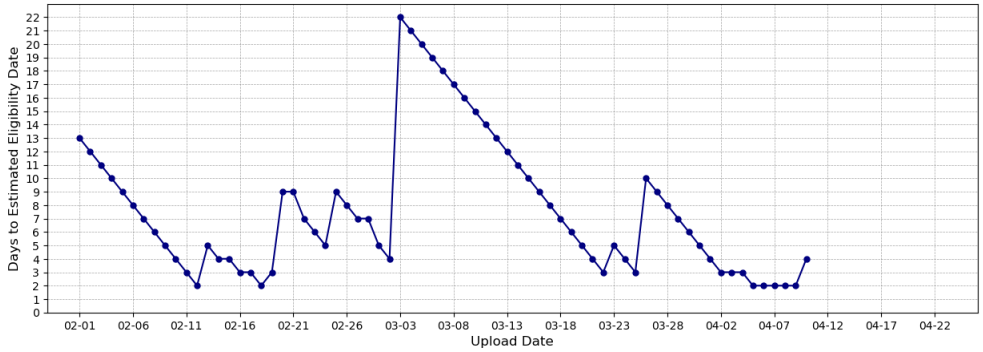
4.2.1 Reverse Engineered Engagement Graph Activation Criteria

The activation rules we uncovered are surprisingly simple yet highly accurate in predicting engagement graph appearance, with a 97.5% out-of-sample test accuracy (see Section 4.2.2). The key elements of the criteria are as follows:

i. **Activation Determinants**: Engagement graph activation only depends on video pop-



(a) Distribution of Treatment Timing



(b) Time Span Between Upload Date and Estimated Eligibility Date (Over Time)

Figure 3

ularity, measured by *views*, and a specific calendar date after upload, which we term the *eligibility date*.

ii. **Activation Condition**: For a video i, its engagement graph status on calendar date t is determined as follows:

$$EngGraph_{it} = \begin{cases} 1, & \text{if } t \geq T_i \text{ and } Views_{it} > 50,000 \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

where T_i is the video's eligibility date.

iii. **Eligibility Date**: The eligibility date T_i is *exclusively* a function of the video's upload date d_i:

$$T_i = f^{eligibility}(d_i)$$

This means all videos uploaded on the same date share the same eligibility date.

iv. **Batch Processing**: Videos don't gain eligibility in isolation. Instead, they're processed in batches. Videos uploaded within a certain timeframe are grouped together

and share a common eligibility date, suggesting a systematic approach by YouTube to roll out the engagement graph feature.

We start by discussing criterion (ii). This criterion implies that if we plot the view trajectories over time for a video, where the x-axis represents the calendar date and the y-axis represents the total number of views, an eligibility region exists in the upper right zone. A video gains an engagement graph if and only if it enters this region. Figure 4(a) illustrates this concept.

Criterion (iii) implies that all videos uploaded on the same date share the same eligibility region. As an informal validation, Figure 4(b) plots the view trajectories for all videos uploaded on March 14. The x-axis shows the calendar date, and the y-axis shows the total number of views (in thousands) on a logarithmic scale. Each line represents the view-date trajectory of a single video, with orange lines indicating videos that received an engagement graph within their first 25 days after upload, and blue lines representing those that did not. Red dots on the orange lines indicate when and at what view level a video first received the engagement graph. The red dots form a clear triangular pattern that closely resembles the conceptual eligibility region depicted in Figure 4(a). Moreover, it visually demonstrates how the eligibility date (the vertical edge of the triangle) and the view threshold (the horizontal edge) jointly determine engagement graph activation for videos uploaded on the same date.

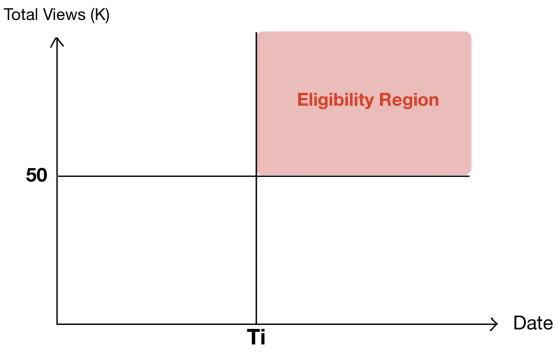
The true eligibility dates in criterion (iii) are not observed but can be estimated. For each upload date d, we estimate the eligibility function as:

$$\hat{f}^{eligibility}(d) = \min\{G_i \,|\, i \in V_d\} \tag{2}$$

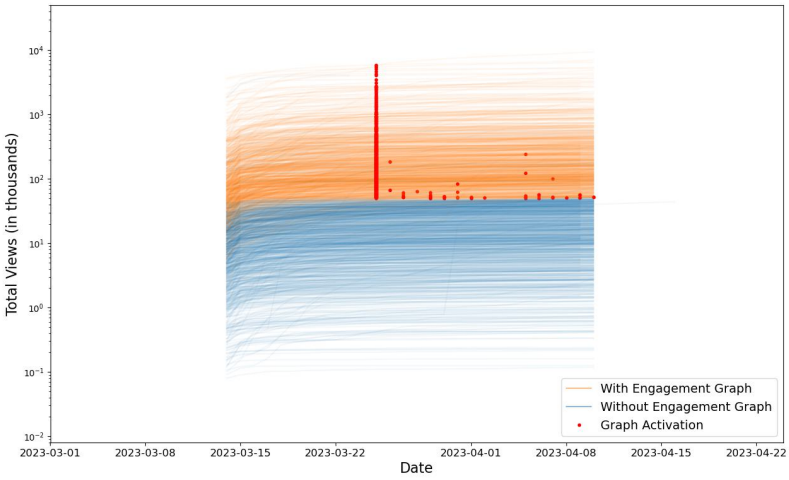
where V_d is the set of videos uploaded on date d, and G_i is the earliest period at which video i received an engagement graph. In other words, we estimate the eligibility date using the earliest date when any of the videos uploaded on that date ever become treated.

In our sample, the time span between the upload date and the eligibility date is always positive ($\hat{f}^{eligibility}(d) - d > 0, \forall d$). Figure 3(b) illustrates how this span varies over time. Each point represents the time to eligibility for videos uploaded on a particular date. For instance, videos uploaded on March 3rd took 22 days to become eligible.

The batch processing described in criterion (iv) is represented in Figure 3(b) by a series of 135-degree lines. These lines appear at 135 degrees because a one-day decrease in the time to eligibility corresponds to a one-day increase in the upload date, creating a diagonal pattern. Each line segment represents a batch of videos sharing the same eligibility date. For example, the first line segment shows that videos uploaded from February 1st to February 12th form a batch with the same eligibility date (February 14th). Notably, the batch sizes



(a)



(b)

Figure 4: Activation Conditions: Illustration

vary over time: some batches are larger, such as the one spanning March 3rd to March 22nd (20 days), while others are smaller, like the one from March 23rd to March 24th (2 days). This variation, possibly as part of their constant A/B testing for any similar platform-wide features, introduces a natural experiment setting where similar videos are assigned to drastically different waiting periods until treatment solely because they were uploaded on different dates. Such a setting is crucial for our identification strategy, as it provides quasi-random variation in treatment timing.

Further analysis of the estimated eligibility dates reveals that they are distributed reasonably evenly across different days of the week and are not systematically more likely to fall on holidays. Figure 5 shows the histogram of eligibility dates across days of the week, indicating a relatively spread-out distribution. To formally test for any systematic patterns, we regress

an indicator variable for whether a date is an estimated eligibility date ($IsEligibilityDate_t$) on day-of-week dummies and holiday indicators. Table 8 in Appendix A.1 presents the results of this regression. As evident from the regression results, neither specific days of the week nor holiday indicators show statistically significant associations with the likelihood of a date being an eligibility date. This finding supports the notion that YouTube's engagement graph activation process is not tied to particular days or holiday periods, emphasizing the platform's systematic and potentially randomized approach to feature rollout.

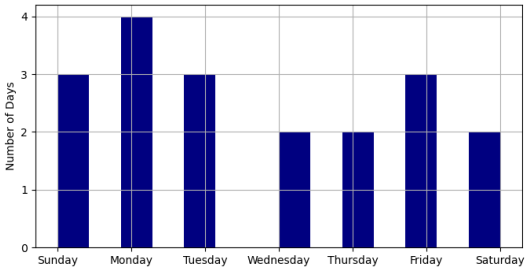


Figure 5: Distribution of Estimated Eligibility Dates Across Days of the Week

4.2.2 Model Validation

To validate our reverse-engineered rules and assess their predictive power, we conduct rigorous out-of-sample testing and compare our results with other benchmarks, including more sophisticated machine learning models.

Train-Test Split and Models Our dataset consists of $\{(x_{it}, y_{it})\}_{t=1,\ldots,T;i=1,\ldots,N}$, where $x_{it} \in \mathbb{R}^D$ represents a D-dimensional feature vector, and the outcome variable $y_{it} \in \{0, 1\}$ indicates the engagement graph status for video i at date t. We randomly divide all videos into a training set (90%, $N_{train} = 113,857$ videos) and a test set (10%, $N_{test} = 12,650$ videos).[10] The split is at the video level, ensuring no video in the test data appears in the training data.[11]

Our reverse-engineered rules use only two features ($D = 2$): *upload date* and *views*. Using the training set data $\{(x_{it}, y_{it})\}_{t=1,\ldots,T;i=1,\ldots,N_{train}}$, we estimate the eligibility function $\hat{f}^{eligibility}(\cdot)$ according to equation (2). We then apply this function to the test set to predict

[10] Our results are robust to other sizes of random train-test split.

[11] For this exercise and subsequent analysis in this paper, we drop the very few videos whose engagement graph status is not absorbing (less than 4.5% of all videos).

engagement graph status:

$$\hat{y}_{it} = \begin{cases} 1, & \text{if } t \geq \hat{f}^{eligibility}(d_i) \text{ and } Views_{it} > 50,000 \\ 0, & \text{otherwise} \end{cases}$$

where d_i is video i's upload date.

To benchmark our approach, we compare it with more sophisticated machine learning models. We implement two random forest models, each comprising 100 decision trees without depth restrictions:

1. RF-limited: Uses essentially the same features as our reverse-engineered rules (views, upload date, current date).

2. RF-extended: Uses an extensive set of features, including video length, categories, channel followers, upload date, current date, views, comments, likes, and daily changes in views, comments, and likes.

To assess the individual importance of *upload date* and *views*, we compare the reverse-engineered rules with the following single-variable dependent rules that make decisions based on a single criterion:

1. Eligibility Date Only:

$$\hat{y}_{it} = \begin{cases} 1, & \text{if } t \geq \hat{f}^{eligibility}(d_i) \\ 0, & \text{otherwise} \end{cases}$$

2. View Threshold Only:

$$\hat{y}_{it} = \begin{cases} 1, & \text{if } Views_{it} > 50,000 \\ 0, & \text{otherwise} \end{cases}$$

Results and Comparison All model predictions are at the video-day level, but we evaluate model performances with two accuracy metrics: (1) video-day level accuracy, which measures the proportion of correct predictions across all video-day observations in the test set; and (2) video-level accuracy, which assesses whether the model correctly predicts the first engagement graph activation date for each video - specifically, for each video, we check if $min\{t \mid y_{it} = 1\}$ matches $min\{t \mid \hat{y}_{it} = 1\}$.[12]

Table 4 summarizes the test accuracies for all models:

[12]We allow for a 1-day buffer between the actual and predicted graph activation date to account for potential discrepancies in graph status update timing. That is, we check if $|min\{t \mid y_{it} = 1\} - min\{t \mid \hat{y}_{it} = 1\}| \leq 1$.

Table 4: Model Performance Comparison

Model	Video-Day Level Accuracy	Video-Level Accuracy
Reverse-Engineered Rules	97.54%	91.54%
RF-limited	97.99%	94.01%
RF-extended	98.84%	96.85%
Eligibility Date Only	63.92%	52.06%
View Threshold Only	83.83%	51.54%

Notes: Video-day level accuracy measures the proportion of correct predictions across all video-day observations in the test set. Video-level accuracy assesses whether the model correctly predicts the first engagement graph activation date for each video; specifically, for each video, we check if $min\{t \mid y_{it} = 1\}$ matches $min\{t \mid \hat{y}_{it} = 1\}$.

Our reverse-engineered rules achieved a remarkable 97.54% accuracy at the video-day level on the test set. The rules correctly predict the first engagement graph activation date for 91.54% of all videos. In fact, over 96% of all videos received the engagement graph within three days of the predicted activation date. The RF-limited model, using the same set of features, only marginally outperformed our rules with 97.99% video-day accuracy and 94.01% video-level accuracy. Even the RF-extended model, with its comprehensive feature set, achieved only minor improvements: 98.84% video-day accuracy and 96.85% video-level accuracy.

The minimal improvement gained from the more sophisticated random forest models, despite their ability to capture complex relationships and interactions, strongly suggests that our parsimonious, interpretable rules closely approximate YouTube's actual criteria for engagement graph activation. The high performance of our reverse-engineered rules and the RF-limited model confirms that views and upload date are indeed the most crucial factors in determining engagement graph activation. The marginal gains from the RF-extended model imply that additional factors play only a negligible role, if any, in the activation decision. The remaining 2-3% misclassified cases likely represent random variations or results of smaller-scale experiments by the platform that cannot be explained by observable factors. Both single-variable dependent rules perform poorly, with much lower test accuracies, confirming that both views and upload date are crucial components of the activation criteria.

4.3 Empirical Strategy

In this section, we describe our empirical strategy for identifying the causal impact of within-content engagement information on user consumption and engagement metrics.

Identification strategy Our identification strategy builds upon the insights gained from reverse-engineering YouTube's engagement graph activation criteria. In Section 4.2, we have established that engagement graph activation (treatment) only depends on two factors: views and eligibility date, with the eligibility date being exclusively determined by the upload date. Our approach leverages the natural experiment that introduces arguably exogenous variation in the time span between the upload date and the eligibility date. For instance, eligible videos uploaded on March 2 need to wait only 4 days to be treated, while eligible videos uploaded on March 3 won't be treated until 22 days after upload.

We compare videos that pass the view-threshold criterion but for which the eligibility date criterion is binding. The idea is that, upon meeting the predetermined 50,000 view threshold, two otherwise identical videos with similar view trends would receive treatment at different times in their lifecycle solely due to different upload dates. This variation allows us to estimate the causal impact of the engagement graph using a DiD strategy.

In our DiD setup, the treated group consists of videos with the engagement graph activated during the first two weeks after upload. The control group contains comparable videos that meet all other requirements for receiving treatment but are not treated in the first two weeks only because they have been assigned a longer duration till eligibility. We use the pre-treatment observations of these not-yet-treated videos as our control. To ensure further comparability between the treated and control videos, we perform Coarsened Exact Matching (CEM) on the eligible videos before conducting the DiD analysis.

The key identifying assumption of our DiD strategy is that, in the absence of treatment, the average outcomes would have evolved in parallel for the treated and control videos. To the extent that this parallel trends assumption holds, the causal impact of the engagement graph can be identified.

Our strategy accounts for any unobserved time-invariant video-specific disparities between the treated and control videos and addresses the potentially selective timing of treatment issue mentioned in Section 4.1. Potential threats include time-varying unobservables that might be correlated with both the treatment and the outcome. We address this concern by arguing that since we have identified the exact factors determining treatment, our sample selection and matching procedures effectively control for all variables influencing engagement graph activation. Moreover, we have demonstrated that eligibility dates do not coincide with other confounding factors such as weekends and holidays (Section 4.2). We also conduct robustness checks on our matching procedure and perform event study analyses to test for pre-trends, further validating our strategy.

Sample Construction We group every two consecutive calendar days into one *period* for our analysis.[13] Our empirical analysis focuses on each video's first two weeks post-upload, i.e., $0 \leq p \leq 7$, where p is the period since upload.[14] For each cumulative metric y_p (including views, likes, and comments), we calculate the change $\Delta y_p = y_p - y_{p-1}$ to represent the number of new views, likes, and comments in each period. We clean the data by dropping samples with missing data and negative values in key non-decreasing metrics during the two-week window.[15]

For the treated group, we select videos that meet two criteria: (1) have engagement graphs activated at the 4th or 5th period since upload, and (2) receive treatment on their predicted eligibility dates. The first condition ensures at least two periods both before and after treatment for any treated video in our estimation.[16] The second condition ensures that the duration from upload date to eligibility date is the key binding criterion at play.

For the control group, we select videos that satisfy all the following conditions: (1) without engagement graphs during the first two weeks after upload (i.e., during the $p \leq 7$ window); (2) with an estimated eligibility date at least two weeks away from the upload date; (3) passing the 50,000 view threshold by the 4th period post upload. These conditions ensure that these control videos would have been treated at the 4th or 5th period since upload had they been uploaded on dates with shorter duration till eligibility.

This process yields 6,443 treated videos and 3,615 control videos before matching.

Matching Although we have selected control videos that are equally eligible for engagement graphs as the treated videos by ensuring all videos pass the view threshold, potential imbalances between the treated and control groups may still arise. For instance, unpredictable viral videos that are intrinsically different from the rest of the videos can exist in either group. Therefore, to further refine the comparability between the treated and control videos, we implement a matching procedure based on a rich set of observable characteristics at the first period after upload. Through this procedure, we strive to construct a control group that is as similar as possible to the treated group on all measured dimensions except

[13] While our time-series data was recorded daily, a technical issue with our web scraper caused a 9-hour delay in data collection on March 26. To mitigate this irregularity, we aggregate data from every two consecutive days into a single 48-hour period for our analysis, ensuring consistent time intervals across all observations.

[14] We focus on the first 14 days due to two reasons: first, the majority of engagement metrics accumulation occurs in the first two weeks; second, it allows for a sufficient number of control videos with adequate pre-treatment periods.

[15] Specifically, we exclude videos with missing data for views, likes, or comments. We then remove videos with non-positive values in the number of new views or likes for two reasons: (1) these metrics should normally be non-decreasing, and (2) to allow for log transformation in subsequent analysis. We retain videos with negative new comments as this can occur naturally due to comment deletion.

[16] We use the first-period data ($p = 1$) for matching purposes and exclude it from our estimation and pre-trend tests.

for the presence of the engagement graph.

Specifically, we employ the Coarsened Exact Matching (CEM) method proposed by Iacus et al. (2012). The process involves temporarily coarsening continuous variables into categories and then performing exact matching based on these coarsened values. Unmatched units are pruned from the sample, and the original, uncoarsened values of the matched data are returned for our final DiD analysis. The matching variables include video category, video length, cumulative views and likes at the $p = 1$ period, and new views and new likes gained during the $p = 1$ period.[17]

The matching procedure yields 2,491 treated videos and 1,816 controls. Among the treated videos, 53.8% are videos treated at the 4th period, and the rest are videos treated at the 5th period. After matching, the treated and control groups are indistinguishable in terms of observable characteristics (Table 5).

Table 5: Balance Check on Observable Characteristics

	Variable	**Treated**	**Controls**	**p-value**
Static	Video length (seconds)	710.13 (244.18)	708.19 (246.65)	0.797
Period	New views (thousand)	23.45 (25.72)	24.48 (29.69)	0.984
Period	New comments	25.63 (29.30)	26.86 (31.55)	0.872
Period	New likes	536.56 (552.44)	547.04 (597.96)	0.413
Cumulative	Views (thousand)	129.49 (107.46)	129.67 (110.44)	0.208
Cumulative	Comments	311.38 (216.42)	307.78 (214.03)	0.472
Cumulative	Likes	4698.83 (3789.23)	4550.66 (3755.67)	0.219
Cumulative	Channel subscribers (thousand)	3338.801 (9602.394)	3357.597 (8796.986)	0.948
	Number of Videos	2,491	1,816	

Difference-in-Differences To estimate the effect of an engagement graph on user consumption and engagement metrics, we employ a Difference-in-Differences approach on our matched sample, where we compare changes in user consumption and engagement metrics for videos with engagement graphs before and after graph activation against those of comparable videos without engagement graphs. Specifically, we estimate the main effect using the following two-way fixed effects (TWFE) model:

$$Y_{ip} = \beta EngGraph_{ip} + X_{ip}B + \alpha_i + \eta_p + \epsilon_{ip} \tag{3}$$

[17]In implementation, video category are matched exactly as a categorical variable. Video length is divided into four categories using cutpoints at 4, 8, and 12 minutes. Other continuous variables are first log-transformed and then divided according to the default automatic coarsening algorithm. Our results are robust to different matching variables and coarsening cutpoints (Appendix A.3)

where Y_{ip} represents our dependent variable: the log of new views, log of new likes, or new comments for video i in period p post-upload.[18] Following Lam (2024), we use a *view* as a measure of consumption, which is recorded when a user watches at least 30 seconds of a video, regardless of access method (active selection or autoplay).[19] This definition aligns with YouTube's internal demand measurement. We use *likes* and *comments* as measures of user engagement with the video. These consumption and engagement metrics are crucial as they not only indicate user satisfaction but also play a significant role in content creators' monetization strategies.

The dummy variable $EngGraph_{ip}$ indicates whether video i has an engagement graph on period p since upload. α_i are video-level fixed effects that capture any time-invariant video-specific characteristics, such as the overall content appeal of a video. η_p denotes the period-since-upload fixed effects that capture the overall time trends following the video's release. We also include a set of control variables X_{ip}, such as the day-of-week fixed effects. Standard errors are clustered at the video level to account for any serial correlation (Bertrand et al. 2004).

The coefficient of interest in Equation 3 is β. In a standard multi-period DiD setting with uniform treatment timing, β would identify the average treatment effect on the treated (ATT) of the engagement graph on user consumption and engagement metrics. However, our setting involves staggered treatment, with videos receiving engagement graphs at either $p = 4$ or $p = 5$ periods. Recent econometric literature (De Chaisemartin and d'Haultfoeuille 2020, Sun and Abraham 2021, Goodman-Bacon 2021) has highlighted that TWFE models may produce unreliable estimates in such cases due to negative weight concerns.

To address this issue, we implement the heterogeneity-robust estimator proposed by Callaway and Sant'Anna (2021) (henceforth CS), which produces valid estimates under arbitrary treatment effect heterogeneity. Let $Y_{ip}(g)$ denote video i's potential outcome in period p if first treated at period g, and $Y_{ip}(\infty)$ denote its "never treated" potential outcome in period p. Under the no anticipation assumption, we have $Y_{ip}(g) = Y_{ip}(\infty)$ for all i and $p < g$. The CS approach first identifies and estimates a building block - the average treatment effect at period p for the cohort first treated in period g, i.e., $ATT(g,p) = E[Y_{ip}(g) - Y_{ip}(\infty)|\tilde{G}_i = g]$, where $\tilde{G}_i$ the earliest period at which unit i has received treatment. The estimated $\widehat{ATT}(g,p)$ can then be aggregated into different causal parameters of interest in ways that avoid the negative weighting problem in TWFE models.

[18] New views and new likes per period are non-decreasing and distributed right-skewed. We log-transformed these variables, dropping the few videos with zero outcomes to avoid distortions from taking logarithms after adding 1. Negative or zero values are allowed in new comments as users can delete their comments.

[19] Industry insights suggest that an audience retention above 70% in the first 30 seconds for videos longer than five minutes is considered successful (See https://vidiq.com/blog/post/increase-audience-retention-youtube/).

5 Effects of the Within-Content Engagement Graph

5.1 Main Effects

We estimate the main effects of the engagement graph on video consumption and engagement metrics using both the CS estimator and the TWFE estimator.

For the CS approach, we report the estimated overall treatment effect $\hat{\theta}_0$, which is a weighted average of $\widehat{ATT}(g,p)$ for $p \geq g$ over all groups and time periods:

$$\hat{\theta}_0 = \frac{1}{\kappa_0} \sum_g \sum_{p \geq g} \omega_g \widehat{ATT}(g,p) \tag{4}$$

where ω_g is proportional to the number of videos first treated on period g since upload, and κ_0 is a normalizing constant that ensures the weights sum to one.

Table 6 reports the estimated overall treatment effects. We find that on average, the presence of an engagement graph increases the number of new views of a video by 7.9% ($= exp(.076) - 1$) in the post-treatment periods. As we have discussed, a view is only recorded when a user stays in the video for at least 30 seconds. Hence, the result suggests having the engagement graph successfully retains more viewers and increases video consumption. Moreover, we find that having the engagement graph increases user engagement with the video content: videos with an engagement graph receive 9.2% ($= \exp(.088) - 1$) more new likes per period post-treatment, and the engagement graph leads to an average increase of 3.19 new comments per period post-treatment. All effects are statistically significant at the 0.1% level, indicating that the engagement graph not only enhances content consumption but also promotes user engagement through likes and comments.

Table 6: Effect of Engagement Graph on Video Consumption and Engagement (CS)

	(1) log(new views)	(2) log(new likes)	(3) new comments
Overall treatment effect (θ_0)	0.076***	0.088***	3.19***
	(0.015)	(0.014)	(0.698)
Observations	25,842	25,842	25,842
Number of Videos	4,307	4,307	4,307

Notes: This table presents the CS estimates in Equation (4) on our matched sample. The unit of analysis is a video-period since upload. Standard errors clustered at the video level are in parentheses. $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$

We report the results by the standard TWFE model in Table 7. Unlike the CS estimator, the TWFE model allows us to control for time-varying covariates such as the day-of-week effects. Recently published papers like Ananthakrishnan et al. (2023) also employ both

estimators to provide a comprehensive analysis. Using the TWFE estimates for Equation (3), we find that having the engagement graph increases the number of new views by 4.7% ($= \exp(.046) - 1$) and the number of new likes by 5.7% ($= \exp(.055) - 1$). The engagement graph leads to an average increase of 2.88 new comments per period post-treatment. The findings confirm that providing the within-content engagement information increases overall content consumption and user engagement.

Table 7: Effect of Engagement Graph on Video Consumption and Engagement (TWFE)

	(1) log(new views)	(2) log(new likes)	(3) new comments
EngGraph	0.046**	0.055***	2.875***
	(0.017)	(0.016)	(0.652)
Controls	Yes	Yes	Yes
Video fixed effects	Yes	Yes	Yes
Period since upload fixed effects	Yes	Yes	Yes
R-squared			
Observations	25,842	25,842	25,842
Number of Videos	4,307	4,307	4,307

Notes: This table presents the estimates of coefficient β in Equation (3) on our matched sample. The unit of analysis is a video-period since upload. Standard errors clustered at the video level are shown in parentheses. $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$

5.2 Effects by Length of Exposure

The key identification assumption of our DiD strategy is that the consumption and engagement metrics for the treated videos and the control videos would have evolved in parallel in the absence of the engagement graphs. While the parallel trends assumption is not testable because we do not observe counterfactual outcomes for treated videos, we can test whether there are differential trends in the outcome measures between the treated and control videos in the pretreatment periods. To this end, we report the "event study" parameter by the CS approach that gives the weighted average of treatment effect l periods after treatment across treatment cohorts:

$$\hat{\theta(l)} = \frac{1}{\kappa_l} \sum_g \omega_g \widehat{ATT}(g, g+l) \tag{5}$$

where ω_g is proportional to the number of videos first treated on period g since upload, and κ_l ensures the weights sum to one. Callaway and Sant'Anna (2021) has shown these estimators can completely avoid the pitfalls associated with the typical dynamic TWFE regressions highlighted by Sun and Abraham (2021).

Figure 6 plots the estimates and the 95% confidence intervals of $\theta(l)$ for the three outcome variables: log(new views), log(new likes), and new comments. The analysis spans from $l = -2$ to $l = 3$ periods, with $p = 1$ period data excluded due to its use in matching. Figure 6 yields two crucial insights: first, the estimates are close to zero and statistically insignificant in the pretreatment period ($l < 0$), suggesting no differential trends in outcomes between the treated and control videos, which provides supportive evidence for the parallel trend assumption. Second, after the engagement graph is activated ($l > 0$), the treatment effects become positive and statistically significant. This shift indicates a clear impact of the engagement graph activation on all three outcome variables.

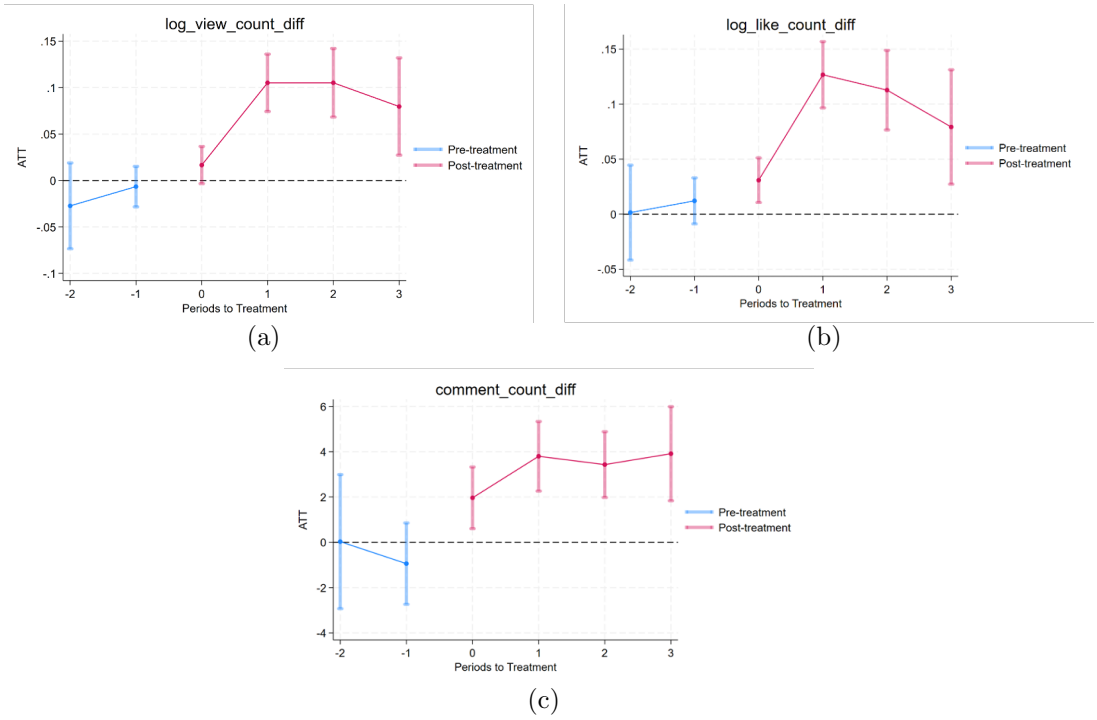


Figure 6: Event Study Plots: Average Treatment Effects by Length of Exposure

5.3 Potential Mechanisms

After demonstrating the effect of the within-content engagement graph on user consumption and engagement metrics, we explore the potential mechanisms driving these effects. We consider two potential mechanisms: improved content navigation and anticipatory savoring. Our analysis indicates that the results are most consistent with the improved content navigation explanation.

5.3.1 Improved Content Navigation

Digital videos can be viewed as experience goods, whose quality or value can only be fully determined after consumption. In such contexts, observational learning can occur, where individuals' behavior is influenced by their observation of other's choices because of the information contained therein (Banerjee 1992; Cai et al. 2009; Tucker and Zhang 2011). The engagement graph provides information on how other users consume and engage with the video content, potentially helping viewers make quality or value inferences. This may lead to more efficient viewing as users more quickly identify the valuable parts of a video. Consequently, the presence of an engagement graph increases the likelihood that users watch for at least 30 seconds (thus counting as a view) and find content they enjoy enough to like or comment on.

If improved content navigation is indeed the underlying mechanism driving the effect of the engagement graph, we would expect to observe: (1) changes in viewing behaviors and time allocation throughout the video in response to the engagement graph, and (2) larger effects when the engagement graph contains richer information about previous users' behaviors.

In the following analysis, we present suggestive evidence supporting this mechanism. We find that (1) over time, engagement graphs become more unequally distributed and have fewer peaks (prominent local maxima), indicating that users increasingly allocate more time to video segments where prominent peaks are observed; (2) effects are stronger for videos with engagement graphs containing more peaks, which are more likely to provide useful visual cues to help users navigate the videos.

Temporal Evolution of Engagement Graphs Do consumers respond to the engagement graph by shifting their attention toward the highlighting moments? To investigate this problem, we examine how the engagement graph distribution evolves over time. Our dataset uniquely captures not only the presence of the engagement graph but also its exact shape. For this analysis, we focus on the first 10 periods post-treatment, using a sample of 14,995 videos with complete engagement graph data for $0 \leq \tilde{p} \leq 10$, where $\tilde{p}$ represents the period since treatment.[20]

We employ two measures to characterize engagement graph shape: the Gini coefficient and the number of peaks. The Gini coefficient quantifies the inequality of the engagement graph distribution, ranging from 0 (perfect equality) to 1 (maximal inequality).[21] We also count the number of peaks (prominent local maxima) in each graph.[22]

[20] Our results remain robust when analyzing videos with available data in the first 6 or 8 periods.

[21] For each engagement graph consisting of 100 data points (one for each percentage of the video timeline), we sort the values into $y_1 \leq ... \leq y_n$ and calculate the Gini coefficient as $G = \frac{1}{n}\left(n + 1 - 2\left(\frac{\sum_{i=1}^{n}(n+1-i)y_i}{\sum_{i=1}^{n} y_i}\right)\right)$.

[22] A peak is defined as a local maximum exceeding a height threshold of 0.2 and separated from neighboring

Non-parametric evidence suggests that over time, engagement graphs become more unequally distributed and exhibit fewer peaks. Figure 7 plots the average Gini coefficient (left panel) and average number of peaks (right panel) across videos over time, along with their respective 95% confidence intervals. Following the activation of the engagement graph, we observe an increase in the average Gini coefficient, indicating that the graph distribution becomes increasingly unequal. Concurrently, the number of prominent peaks decreases, suggesting that existing peaks become more pronounced and concentrated.

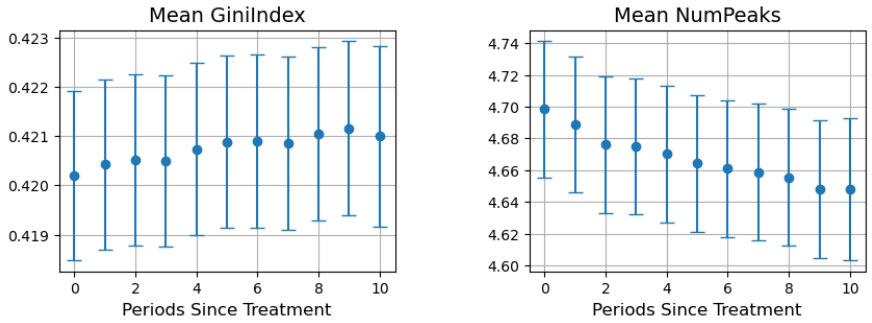


Figure 7: Average Gini Coefficient and Number of Peaks Over Time

More formally, we regress the two engagement graph shape measures on a linear time trend and video fixed effects:

$$ShapeMeasure_{it} = \gamma_1 t + \eta_i + \epsilon_{it} \tag{6}$$

where $ShapeMeasure_{it}$ is either Gini coefficient or the number of peaks for video i on period t post-treatment, and η_i denotes video-level fixed effects. Table 9 in Appendix A.2 reports the estimates for γ_1 in equation 6. We find a positive time trend for Gini coefficient and a negative time trend for number of peaks, both statistically significant at the 0.1% level, corroborating our non-parametric findings.

To visualize the temporal shift in engagement graph shapes, Figure 8 illustrates the average engagement graph for each post-treatment period.[23] The blue lines represent early post-treatment periods, transitioning to red lines for later periods. We observe that engagement levels typically decrease at the beginning and end of videos, with a resurgence often occurring in the latter half.

This visualization confirms that after the introduction of the engagement graph, viewers

peaks by a minimum horizontal distance of 5. Our findings remain consistent across various height thresholds and horizontal distance requirements.

[23]This analysis uses the sample of 14,995 videos. For each post-treatment period, we calculate the average engagement graph shape by taking the mean height at each of the 100 timeline points across all videos. For better visualization, we omitted the first and last 5 points in Figure 8, displaying the graph shape from 5% to 95% of the video timeline.

tend to spend less time on the initial "boring" segments and increasingly allocate their attention to video portions where prominent peaks and higher engagement levels are observed. This pattern is consistent with the idea that users are leveraging the engagement graph to navigate more efficiently through the video content, focusing on segments that previous viewers found most engaging.

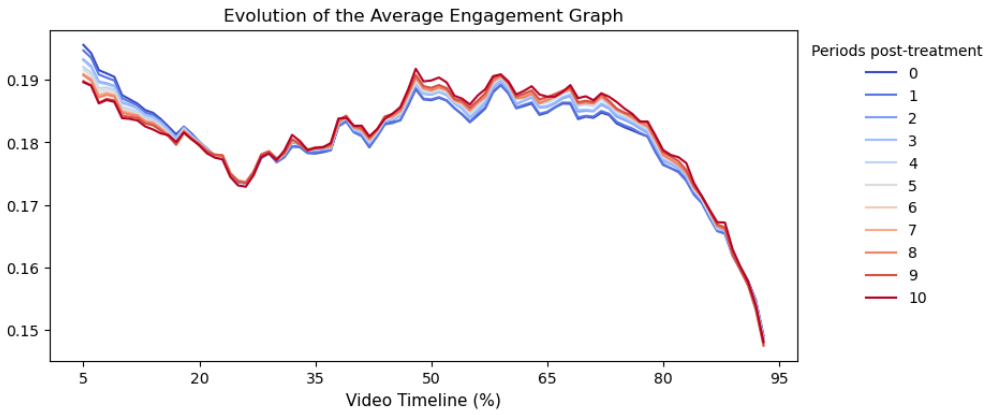


Figure 8: Evolution of the Average Engagement Graph Post Treatment

Moderating Effect of Engagement Graph Shape If improved content navigation is

References

Ananthakrishnan, U., Proserpio, D., and Sharma, S. (2023). I hear you: Does quality improve with customer voice? Marketing Science, 42(6):1143–1161.

Banerjee, A. V. (1992). A simple model of herd behavior. The quarterly journal of economics, 107(3):797–817.

Bertrand, M., Duflo, E., and Mullainathan, S. (2004). How much should we trust differences-in-differences estimates? The Quarterly Journal of Economics, 119(1):249–275.

Cai, H., Chen, Y., and Fang, H. (2009). Observational learning: Evidence from a randomized natural field experiment. American Economic Review, 99(3):864–882.

Callaway, B. and Sant'Anna, P. H. (2021). Difference-in-differences with multiple time periods. Journal of econometrics, 225(2):200–230.

De Chaisemartin, C. and d'Haultfoeuille, X. (2020). Two-way fixed effects estimators with heterogeneous treatment effects. American Economic Review, 110(9):2964–2996.

Goodman-Bacon, A. (2021). Difference-in-differences with variation in treatment timing. Journal of Econometrics, 225(2):254–277.

Iacus, S. M., King, G., and Porro, G. (2012). Causal inference without balance checking: Coarsened exact matching. Political analysis, 20(1):1–24.

Lam, H. T. (2024). Ad-funded attention markets and antitrust: Youtube content economy. Working Paper.

Sun, L. and Abraham, S. (2021). Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. Journal of Econometrics, 225(2):175–199.

Tucker, C. and Zhang, J. (2011). How does popularity information affect choices? a field experiment. Management Science, 57(5):828–842.

A Appendix

A.1 Day of Week and Holiday Patterns of Eligibility Dates

To test whether eligibility dates are systematically more likely to fall on specific days of the week or holidays, we run the following OLS regression:

$$IsEligibilityDatet = \alpha + \sum_{l=1}^{6} \beta^l IsDayofWeek_t^l + \gamma IsFedHoliday_t + \delta IsOtherHoliday_t + \epsilon_t.$$

The unit of analysis is a day from Feb 1 to April 24, 2023. $IsEligibilityDate_t \in \{0, 1\}$ indicates whether a day is an estimated eligibility date. $\{DayofWeek_t^l\}^l$ are dummy variables for days of the week (with Sunday as the omitted category). $IsFedHoliday_t \in \{0, 1\}$ indicates whether a day is a federal holiday in the United States, and $IsOtherHoliday_t \in \{0, 1\}$ indicates whether a day is another significant holiday.

Table 8: Regression Results - Eligibility Date Patterns

	IsEligibilityDate
Constant	0.182
	(0.128)
IsMonday	0.156
	(0.182)
IsTuesday	0.044
	(0.178)
IsWednesday	-0.015
	(0.177)
IsThursday	-0.039
	(0.178)
IsFriday	0.068
	(0.177)
IsSaturday	-0.039
	(0.178)
IsFedHoliday	0.818
	(0.444)
IsOtherHoliday	0.287
	(0.222)
N	83
R^2	0.087

Notes: The unit of analysis is a day (from February 1 to April 24, 2023). Standard errors are in parentheses. $^{*}p < 0.05$,** $p < 0.01$,*** $p < 0.001$

A.2 Temporal Evolution of Engagement Graph Characteristics

Table 9: Regression Analysis of Engagement Graph Shape Measures Over Time

	(1) Gini Coefficient	(2) Num Peaks
t	0.0000908^{***}	-0.00266^{***}
	(0.0000271)	(0.000991)
Video fixed effects	Yes	Yes
Observations	163,224	163,224
Num. of Videos	14,995	14,995

Notes: Standard errors clustered at the video level in parentheses.
$^{*}p < 0.05, ^{**}p < 0.01, ^{***}p < 0.001$

A.3 Robustness Checks for Matching