Paying Attention

Karthik Srinivasan
University of Chicago,
Booth School of Business
October 31, 2023

Abstract

Humans are social animals. Is the desire for attention from other people a quantitatively important non-monetary incentive? I consider this question in the context of social media, where platforms like TikTok and Reddit successfully attract a large volume of user-generated content without offering financial incentives to most users. Using data on two billion Reddit posts, I estimate the elasticity of content production with respect to attention, as measured by the number of comments and upvotes that a post receives. I isolate plausibly exogenous variation in attention by studying posts that go viral. After going viral, producers create 70% more posts for two weeks. I replicate this result on a sample of TikTok producers: virality causes a 90% increase in production. I complement these reduced form estimates with a large-scale, preregistered field experiment on Reddit. I randomly allocate attention by adding three or six comments to posts. I use generative AI to produce responsive comments in real time, and distribute these comments via a network of bots. Adding three comments causes a 10% increase in production, but I find a null effect for six comments. This difference can be explained by attention quality: comments in the six comment treatment arm are more likely to be downvoted, and downvoted comments decrease production. Overall, I find that the attention labor supply curve is concave. Producers value initial units of attention highly, but the marginal value of attention rapidly diminishes. Motivated by this fact, I propose a model of a social media platform which manages a two-sided market composed of content producers and consumers. The key trade-off is that consumers dislike low-quality content, but including low-quality content provides attention to producers, which boosts the supply of high-quality content in equilibrium. If the attention labor supply curve is sufficiently concave, then the platform includes some low-quality content, though a social planner would include strictly more. This wedge provides a new rationale for the regulation of social media platforms.

[CLICK HERE FOR LATEST VERSION]

JEL Codes: D01, D12, D21, D22, D47, D62, D91, J22, J46, L82

Note: This draft is work-in-progress and is updated frequently. Please do not distribute.

Attention plays an increasingly important role in the economy. Some of the most influential companies of the last two decades can be understood as *attention platforms*, firms which broker attention markets between consumers, content producers, and advertisers. Concretely, one commonality between Facebook, Google, Spotify, TikTok, and The New York Times is that each of these firms offers consumers access to content, and profits by auctioning off the attention of consumers to advertisers. Moreover, each of these firms use algorithms to influence which pieces of content consumers pay attention to, thereby shaping the allocation of consumer attention across posts, websites, songs, videos, and news articles.

Among attention platforms, social media firms have rapidly gained share in the global market for attention. The first social media platform was founded less than thirty years ago.[1] Now, 60% of people are social media users, and the average user spends 15% of their waking hours browsing content on these platforms.[2] Social media platforms capture attention by offering access to billions of pieces of new content each day, generated primarily by users who do not face direct financial incentives.[3]

How do the economies of social media platforms work? How should they be regulated? A literature in economics considers these questions, but tends to focus on the *financial* value of consumer attention to advertisers. In this paper, I provide new insights by revisiting an old idea: people value the attention of others *inherently*. This idea has important implications for how attention platforms work, because it means that the way that platforms allocate attention across content producers alters the incentives for producers to supply content.

The object of interest in this paper is the attention labor supply curve. This curve captures the relationship between the amount of attention (views, likes, comments) that a content producer on social media receives and the supply of posts that they produce. In the empirical sections

[1]Contact: ks@chicagobooth.edu. I thank my advisors Alex Frankel, Devin Pope, Eric Zwick, and Eric Budish for their mentorship. I thank participants in Booth's Student Research in Economics Seminar and Behavioral Economics Lab for valuable comments. I thank Walter Zhang, Benedict Guttman-Kenney, Lucy Msall, Pauline Mourot, Olivia Bordeu, Kevin Lee, Lillian Rusk, Scott Behmer, and Michael Galperin for friendship and advice. I thank my family Shanthi Srinivasan, Muthayyah Srinivasan, Arjun Srinivasan, Anand Srinivasan, and Mochi for their love and support. Finally, I thank my dear friends Gabi Hirsch, Alex Duner, Jaclyn Zhou, Sam Osburn, Anna Cormack, Hayley Hopkins, James Kiselik, Janani Nathan, Gia, Reuben Bauer, Alexi Stocker, Claire Bergey, Mara Zinky, Bill Batterman, and Maggie Berthiaume.

[1]Boyd and Ellison (2007) identify SixDegrees.com, founded in 1997, as the first site that resembles modern social media platforms, though they caveat that this designation rests on how social media is defined.

[2]The DataReportal (2023b) report indicates that there are 4.76 billion active social media users, which is 59.4% of the global population of 8.01 billion (See slide 10). Slide 26 indicates that the average user spends 2 hours and 31 minutes on social media each day, which is 15.1% of the average number of waking hours (16 hours and 42 minutes (Thomas, 2019)).

[3]Counting only posts on Instagram, Twitter, Facebook, and Snapchat, there are over 4 billion posts per day (Domo, 2022). Direct revenue sharing varies by platform, with some platforms not offering any direct compensation for regular users (e.g., Reddit, Facebook, Instagram), some platforms offering revenue shares only to very successful users (e.g., Snapchat, TikTok), and some offering revenue shares widely (e.g., Twitch, YouTube). Many content producers may face indirect financial incentives (e.g. brand deals for sucessful Instagram influencers).

of the paper, I use reduced form and experimental methods to estimate the elasticity of content production with respect to attention at various points along the attention labor supply curve. In the theoretical section of the paper, I derive implications of the shape of the attention labor supply curve on the optimal design and regulation of social media platforms.

I primarily study Reddit, the seventh most visited website in the world.[4] Reddit is a large and rapidly growing social media platform based around interest-driven forums. Its traffic has quadrupled since 2018, and it hosts over 430 million monthly active users, making it comparable in size to LinkedIn, Twitter, Snapchat, and Pinterest.[5] Content producers on Reddit can post text, links, images, and videos. The primary advantage of Reddit as a setting is a prevailing norm of anonymity, which helps to isolate the attention incentive by reducing the presence of confounding social and financial incentives.

I study the near-universe of Reddit posts from 2005-2022, which amounts to over two billion posts. I isolate plausibly exogenous variation in attention by focusing on content producers who "go viral." I define a post as viral if it reaches the 90^{th} percentile of the engagement distribution. I estimate a difference-in-differences design comparing content production around viral and non-viral posts. Producers who go viral produce 70% more posts that are 10% better, as measured by average net upvotes, over the next two weeks.

I replicate these reduced form results on TikTok, a video-based social media platform with over 1 billion monthly active users. I put together a new dataset that follows 9,000 TikTok content producers who produce 750,000 TikToks. After going viral, TikTok producers create 90% more posts that are 15% better, as measured by likes per view, over the next two weeks.

This large volume of observational data allows me to go a step further than prior work on non-monetary incentives. Rather than reporting a coefficient benchmarked to a specific treatment size, I report causal treatment effects for varying amounts of attention by estimating the difference-in-differences design on posts that go viral to varying degrees. This exercise allows me to trace out the attention labor supply curve on Reddit and TikTok. The key empirical finding is that the shape of the attention labor supply curve is increasing and sharply concave. While the first units of attention strongly encourage producers to supply content, marginal units quickly become much less influential.

There are at least three plausible identification concerns with this difference-in-differences design. First, those who go viral may be selected. I show that pre-trends in the supply of posts around viral and randomly selected non-viral posts are similar in both level and trend on Reddit and TikTok, mitigating this concern.

Second, differences in content production after going viral could reflect changes in posting

[4] According to SEMRush (2023b), Reddit falls behind Google, YouTube, Facebook, Twitter, Wikipedia and Instagram. The exact ranking depends on the source: SimilarWeb claims that Reddit is the eighteenth most visited website (SimilarWeb, 2023c).

[5] According to SimilarWeb, Reddit was getting 282 million visits per month in 2018 compared to 1.9 billion in 2023 (SimilarWeb, 2019, 2023b). Reddit claimed to have 430 billion monthly active users in 2019, and has not released this statistic publicly since (Murphy, 2019). This was larger than the monthly active userbases of Twitter, LinkedIn, Snapchat, and Pinterest in 2019 (DataReportal, 2019).

ability, which could be an underlying cause of virality and increased production. Here, I appeal to the shape of the treatment effects, which exhibit a spike-and-fade pattern that is inconsistent with a story of steadily increasing ability.

Third, going viral could confer non-attentional rewards. The institutional features of Reddit diminish this concern. Reddit offers no financial rewards to producers, and a strong norm of anonymity among producers restricts the ability to accrue external social or financial benefits. However, this concern is warranted on TikTok. While most TikTok users do not face direct financial incentives and garner engagement primarily from users they do not know, I cannot rule out that producers anticipate some social or financial returns from success on TikTok. Given this, results on TikTok should be interpreted as capturing the causal effects of engagement rather than mere attention. Nevertheless, the concavity of labor supply with respect to engagement is sufficient for my theoretical results to apply, even if the reduced form results reflect a mixture of attentional, social and financial incentives.

I complement the reduced form analysis with a field experiment on Reddit. I randomly allocate attention to content producers by using Reddit bots to add three or six comments to their posts, and measure changes to the supply of posts over the next week. Comments are generated with a natural language processing pipeline built on top of the OpenAI chat completion API, the engine that powers ChatGPT. This method allows me to generate relevant, plausibly-human comments in real time as posts appear.

The experiment is large in scale. I pilot over a thousand Reddit accounts which post over 10,000 comments on Reddit, and I follow the production decisions of 200,000 content producers. The primary, preregistered outcome is a quality-weighted measure of the number of posts produced by treated users.

The top-level results of the experiment are mixed: adding three comments causes a 10% increase in posting, while I find a null effect for six comments. The positive treatment effect of adding three comments is robust to alternative, pre-specified measures of output including the probability of posting again and the average score of posts, as well as to the inclusion of controls for prior posting frequency.

The null treatment effect of six comments is counterintuitive given the rest of the results in the paper. I show that it can be explained by an unintended form of heterogeneity in treatment. The six comments treatment is more likely to be negatively received by the Reddit community. Comments in the six comments treatment have fewer upvotes, more downvotes, and are more likely to be accused of being bots. I decompose the treatment effect into the effect of high and low quality attention, based on whether comments were upvoted or downvoted. Upvoted comments have positive treatment effects, while downvoted comments have negative treatment effects. I show that the entire difference between the three and six comment treatment arms can be accounted for by the differential quality of attention provided.

In the theoretical section of this paper, I take the concavity of the attention labor supply curve as a starting point, and ask what it can teach us about the optimal design of social media

platforms. I propose a model of a social media platform that manages a two-sided market composed of consumers and content producers. As is standard in two-sided markets, consumers value the size of the content producer side of the market, and content producers value the size of the consumer side of the market. However, in a departure from canonical models, markets clear in attention rather than prices.

Producers decide whether to create content depending on the amount of attention that they expect to receive. Attention depends endogenously on the number of consumers who choose to join the platform, and on a simple content recommendation algorithm that the platform selects. Attention also depends on the quality of content that producers create. Quality is binary, and content realizes as good or bad exogenously. The platform decides how much good and bad content to offer to consumers, selecting from the content that was supplied by producers. Consumers decide whether or not to join the platform based on the quantity and quality of content that is available. Consumers derive positive utility from the inclusion of good content on the platform, and negative utility from the inclusion of bad content on the platform. The central trade-off in the model is that consumers dislike bad content, but showing bad content provides additional attention to content producers, which boosts the aggregate supply of good content in equilibrium.

The first result of the model is that the concavity of the attention labor supply curve determines the extent to which the platform should include bad content: if the supply curve is concave enough, then the platform should show a positive percentage of bad content. The intuition for this result is that if showing bad content boosts supply enough, then consumers' taste for additional good content can dominate their distaste for bad content.

The second result of the model is that a social planner who cares about producer utility would show a larger percentage of bad content than a profit maximizing platform. The intuition for this finding is that the platform only compensates producers to the extent that additional attention raises the value of the platform to consumers. In contrast, a social planner values the utility that content producers derive from attention directly. This finding is interesting because it provides a novel rationale for the regulation of content recommendation algorithms, which are a feature of most attention platforms. The attention incentive generates a wedge between the profit and welfare maximizing algorithms, which implies that mandating a kind of "attention redistribution" can be welfare improving.

The empirical portion of this paper contributes to a large literature in economics which documents the effectiveness of various non-monetary incentives. Status concerns, social pressure, peer comparisons, awards, identity and purpose have all been shown to motivate people to exert effort.[6]

An interdisciplinary literature evaluates the efficacy of non-monetary incentives in the context of online spaces. An early causal contribution to this literature is Chen et al. (2010), who find that

[6]The literature on non-monetary incentives is extensive, and a complete review is beyond the scope of this paper. For status concerns, see Kuhn et al. (2011). For peer pressure, see DellaVigna et al. (2012, 2016); Perez-Truglia and Cruces (2017); DellaVigna and Pope (2018). For peer comparisons, see Kolstad (2013); Ager et al. (2022). For awards, see Delfgaauw et al. (2013); Ashraf et al. (2014); Neckermann et al. (2014). For identity, see Akerlof and Kranton (2000); Atkin et al. (2021). For purpose, see Ariely et al. (2008); Khan (2020).

providing information on the median contribution rate encourages below-median users to supply additional reviews to an online movie review website. The efficacy of social comparisons and status as incentives online has since been demonstrated in a wide variety of contexts (Goes et al., 2016; Sun et al., 2017; Burtch et al., 2018; Kuang et al., 2019; Ke et al., 2020; Zhang et al., 2020; Ma et al., 2022).[7] I study the same setting and use a similar experimental method to Burtch et al. (2022), who document that awards on Reddit increase content production.

Within the empirical literature on non-monetary incentives, this paper is most closely related to work which evaluates the role of audience and engagement as incentives online. Zhang and Zhu (2011) find that when users in mainland China were blocked from Wikipedia, non-blocked users reduced their contributions. Wang et al. (2019) replicate this effect on Douban, a product review website. Other studies emphasize the role of follower networks. Toubia and Stephen (2013) experimentally add Twitter followers to accounts, and find heterogeneous treatment effects. Goes et al. (2014) find that product reviewers with more subscribers produce more and better reviews. Wei et al. (2021) find that followers increase content production on Twitter and Tencent Weibo. Content production responds to engagement as well. Lindström et al. (2021) show that posting behavior on Instagram and in a lab experiment are consistent with users valuing likes. The fact that people value social interaction online dovetails with work in the neuroscience literature which shows that likes on social media cause blood to flow to the area of the brain associated with pleasure (Eisenberger et al., 2003; Davey et al., 2010; Meshi et al., 2013).

I make two contributions to this large, interdisciplinary literature on non-monetary incentives. First, I provide evidence for the role of *mere* attention as an incentive in-and-of-itself. While prior work has established that social interactions can incentivize effort, this evidence comes from platforms where creators' identities and successes are public. In these contexts, higher engagement could bestow social, attentional, and (future) financial rewards. The norm of anonymity on Reddit allows me to better isolate the role of attention. Second, my large sample (two orders of magnitude larger than prior work in this literature) allows me to identify a treatment effect curve rather than a point estimate. This matters because I show that the shape of the attention labor supply curve affects optimal platform design.

The model relates to the theoretical literature on multi-sided platforms.[8] The model borrows structure from canonical models in this literature which study platforms that manage two-sided markets with network externalities (Rochet and Tirole, 2003; Caillaud and Jullien, 2003; Parker and Van Alstyne, 2005; Armstrong, 2006). Within this literature, the model is closest to a strand which focuses on platforms that can choose quality (Weyl, 2010; Veiga et al., 2017; Chan, 2023).

The model also relates to a theoretical literature which focuses on attention platforms specifically, using a wide variety of modeling techniques. Chen (2022) provides a general equilibrium model of the market for attention. Jain and Qian (2021) and Bhargava (2022) consider platforms with consumers and content producers, but focus on financial incentives.

[7] There is also earlier descriptive evidence of these ideas in the computer science literature (Lampe and Johnston, 2005; Arguello et al., 2006; Burke et al., 2009).

[8] For a recent review of this literature, see Jullien et al. (2021) or Sanchez-Cartas and León (2021).

My contribution to the theoretical literature is to introduce the notion that a platform can influence quality by algorithmically manipulating the way that the two sides of the market interact. This leads me to study a different object than the vast majority of the literature: rather than focusing on optimal pricing, I focus on optimal curation. While this kind of algorithmic matchmaking is not a feature of all canonical two-sided markets, it is a salient feature of the attention platforms that I study.[9]

The rest of the paper proceeds as follows. Section 1 covers some relevant institutional details of Reddit. **??** presents reduced form evidence on the shape of the attention labor supply curve using data from Reddit and TikTok. Section 2 presents the experimental strategy and results. Section 3 provides a theoretical analysis of how social media platforms should optimally allocate attention. Section 4 concludes.

1 Institutional Details of Reddit

Reddit is a large social media platform where users can post, vote on, and discuss a diverse array of content including text, links, images and video. In the United States, Reddit is the fourth largest social media platform by traffic and the ninth largest by userbase, with around 3 in 10 adults reporting that they use the platform.[10]

The majority of the empirical work in this paper is devoted to understanding content production on Reddit. Reddit differs from other social media platforms in many ways. In this section, I will focus on institutional details that are relevant for the interpretation of my results.

First, Reddit is structured around interest-based forums called subreddits. For example, r/gardening is a forum where 5.8 million users subscribe to discuss gardening. Similarly, there are subreddits devoted to most interests and topics: world news (r/worldnews), cute pictures (r/aww), media properties (r/DunderMifflin), online humor (r/memes), and questions (r/AskReddit) each have dedicated forums. Overall, there more than 130,000 active subreddits.

Every post must be submitted to a specific subreddit. Submissions may require approval by the subreddit's volunteer moderation team, and typically must follow certain subreddit-specific stylistic rules. This interest-based division of the website is different from social networks like Facebook and Twitter which provide users with one go-to location to post top-level content.[11]

Second, Reddit users are typically anonymous. This norm is acknowledged by the company

[9]For example, credit cards are a canonical two-sided market, but credit card companies do not meaningfully control whether consumers choose to shop at particular businesses within their network.

[10]The 'fourth largest by traffic' statistic comes from SimilarWeb (2023a) which reports a Top Social Media Networks category. SEMRush (2023a) reports that Reddit is the third largest website by visits as of July 2023 across all websites, trailing only Google and YouTube but outpacing Facebook and Amazon. The 'ninth largest by userbase' statistic comes from slide 57 of the DataReportal (2023a)'s US Digital Report which cites a GWI survey of US adults. The eight larger social media platforms by userbase are Facebook, Instagram, Youtube, TikTok, Twitter, Snapchat, Pinterest, and LinkedIn, which have monthly active users that range from over 3 billion to around 450 million. Reddit's last publicly reported monthly active userbase is 430 million, as of 2020. I get to the claim 'ninth' by adding YouTube (which was not asked about, but is larger than Reddit) and by excluding iMessage and Facebook Messenger (which are typically understood as messaging clients, not social media platforms).

[11]This is a prompt at the top of the feed asking the Tweeter 'What is happening?!' On Facebook, the original

which states that "the vast majority of redditors choose a name that represents them, without revealing who they are" (Reddit, 2023). Reddit encourages anonymity by providing new users with options for auto-generated usernames that are random combinations of words and numbers.

Anonymity is crucial to my research designs. At a conceptual level, anonymity helps rule out alternative stories where attention proxies for social or financial returns. At a practical level, the ability to credibly provide attention to users with bot accounts depends on the norm of anonymity which allows the bot accounts to blend in and provide plausibly human interactions.

Third, posts are distinct from comments on Reddit. Like Facebook, each post on Reddit has a dedicated comments section. This construction differs from Twitter, where replies to tweets are themselves tweets. This distinction is important for understanding the variation I study: typically, I look at how a change in the number of upvotes or comments on a post changes the number of subsequent posts that a Reddit user produces. I do not count subsequent comments made by a Reddit user as a measure of output. This means my results are not driven by users responding to comments on their popular posts, which may have been a concern on another platform like Twitter.

Fourth, Reddit allows users to both 'upvote' and 'downvote' posts. This dual directional feedback is somewhat atypical, and is not found on Facebook, Instagram, TikTok or Twitter. Upvotes correspond roughly to likes and hearts on Facebook and Twitter, signifying that the user had a positive interaction with the content. Downvotes express the opposite. Upvotes and downvotes get their names because upvotes move posts towards the top of Reddit and downvotes move posts towards the bottom of Reddit for users who view the website using the default sorting algorithm.[12] The existance of downvotes matters for my paper due to a measurement issue: I observe the net of upvotes minus downvotes, but I do not observe the count of upvotes and downvotes separately. When I refer to upvotes in the results section, I am always referring to net upvotes.

Finally, the combination of Reddit's content sorting algorithm along with typical Reddit user content consumption patterns serves to de-emphasize the importance of profiles and user-following networks. Reddit uses a small number of content sorting algorithms that are publicly known. Posts can be ordered by recency (new), the ratio of upvotes to downvotes (best), the absolute number of upvotes minus downvotes within a fixed window of time (top), and the absolute number of upvotes minus downvotes plus a time deflator that penalizes older posts (hot). Users can choose to browse the website sorted by these algorithms as a whole (the "frontpage") or within subreddits. The important thing to notice is that these methods for sorting and browsing the website do not depend on follower networks at all.

That being said, profiles and follower networks do exist on Reddit. Users can navigate to a profile and view all of the content from that profile. Following a user causes their content to show up in the algorithmically curated Reddit feeds that are available to Reddit users with accounts (those without accounts can browse Reddit in all of the ways outlined above, but cannot vote or

prompt which sat at the top of the feed read 'What's on your mind?,' though more personalized prompts have since been introduced.

[12]Both 'top' and 'hot' sorting methods have the property that upvotes and downvotes move posts in the natural direction. These algorithms can be applied to sort content across the whole website as well as any subreddit.

comment). However, even the algorithmically curated Reddit feed depends heavily on subreddit following decisions, and users are required to follow subreddits upon creation of an account.

The deemphasis of follower networks and profiles matters for the interpretation of the reduced form results, because it means that producers should not anticipate large changes to the popularity of future content driven by changes in their follower network after they go viral.

1.0.1 Viral Difference-in-Differences Design

Going viral causes content producers to increase production on Reddit. Figure 2 graphs the effect of virality on a quality-weighted measure of producer output. Panel A plots the event study coefficients for the treated group. Each point represents a 1 day bin, and event time 0 is the day that the viral post was created.

The treatment effect is large. Comparing the treatment coefficients in the month after going viral to the month before, output increases by 0.21 units per day. This amounts to a 373% increase over the pre-period mean of 0.06 units per day.

The pre-period estimates provide some evidence regarding the parallel trends assumption. In the pre-period, the event study coefficients are relatively stable and lie near zero. This corresponds to the idea that the treatment and control groups are similar not only in trend, but also in level, which is reassuring.

The event study coefficients also provide some evidence on the second identifying assumption, which is that the timing of virality is not correlated with other changes or events that could affect production. The pattern of treatment effects is distinctive: there is a sharp spike in production at time 1, and then the effect fades slowly over time. A primary ex-ante concern for the interpretation of these treatment effects is that going viral could be correlated with changes in producer ability. This observed pattern of treatment effects is qualitatively consistent with an event whose influence fades, which is not consistent with a static change in producer ability.

Panel B of Figure 2 plots heterogeneity in the treatment effect by the degree of virality. Each point estimates the increase in quality-weighted output in the 30 days after going viral in comparison to the event study coefficients in the 30 days prior. Each point is estimated on a subset of viral posts that go viral within a two-percentile band of the upvotes distribution. Posts in the first viral point received between 21-26 upvotes (80th-82nd percentile), while posts in the tenth viral point received more than 531 upvotes (98th-100th percentile). The placebo point is the treatment effect of posting a non-viral post, estimated using a standard (Callaway and Sant'Anna, 2021) difference-in-differences design around random non-viral post.

Panel B is an estimate of the attention labor supply curve.

The primary difference-in-differences specifications are presented in **??** and **??**. Before interpreting the results of these figures, I start by connecting the graphs to the identifying assumptions of the difference-in-differences design.

A particularly convincing piece of evidence that the placebo does a good job of capturing trends is the point at event time -1. At event time -1, content production spikes for all groups including

the placebo. This is an artifact of selecting the sample on points in time when producers are active on Reddit. Since each author is by construction active on Reddit at time 0, producers are likely to be more active in the days immediately before and afterwards, creating a spike pattern in the middle of these figures. The placebo does a good job of capturing and controlling for this artifact of sample selection.

Now, I turn to interpreting the results.

Plot A of **??** graphs how quantity changes in response to going viral. We see a sharp increase in the quantity of posts produced which fades back towards baseline over the next 30 days.

Plot B of **??** graphs how quality changes in response to going viral. Like quantity, mean score spikes before fading back to baseline over the next 30 days for viral posters.

?? graphs $\sum log(\text{score} + 1)$, which is a quality-weighted measure of quantity. As we would expect from the prior two results, this graph two shows a spike-and-fade pattern.

Figure **??** treats each series in **??** as an event study, and compares how posting behavior differs in the two weeks after going viral as compared to the pre-period month.

This figure gives a sense of the large scale of treatment effects. In the two weeks after a post goes viral, creators produce approximately 0.075 more posts per day off of a baseline of 0.06 posts per day, which represents a 125% increase in the rate of post production. The mean baseline score per day is 0.5 score/day, and this increases by 10 net upvotes among the viral group.

Figure **??** also provides evidence that the attention labor supply curve is concave. Treatment effects are essentially flat after the 90^{th} percentile in terms of quantity, though they continue to rise in terms of quality.

Taking stock, these results show that a discontinuous change in the amount of attention that a content producer receives creates a sharp change in the quantity and quality of content produced. Moreover, this incentive is rapidly diminishing: getting 50 upvotes produces essentially the same effect on the quantity of posts as getting 500 upvotes, though effects on quality are less concave.

2 Experiment

In the experiment, I study the effects of randomly allocating attention to Reddit producers. I view the experiment as complementary to the reduced form analysis. Random assignment mechanically prevents selection on ability, one of the primary identification concerns of the difference-in-differences design. However, technical constraints mean that I can only deliver small amounts of attention to content producers, so the experiment cannot be used to trace out the entire attention labor supply curve.

In order to generate experimental variation, I set up a system that monitors subreddits for posts, randomizes posts into treatment or control, generates responsive comments, and adds these comments to treated posts via a network of servers and Reddit bots. I then collect data on the posting behavior of treated and control users over the thirty days following randomization in order to document any changes to posting behavior.

The top-level findings are mixed: adding three comments causes a 10% increase in the probability of posting again, while I find a null effect for six comments. I reconcile this result with the rest of the evidence in the paper by documenting that the six comments treatment induced an unintended form of heterogeneity in the quality of attention. Specifically, comments in the six comments treatment are less well received by the Reddit community: they are more likely to be downvoted, less likely to be upvoted, and replies are more likely to mention the word bot. Since these comments are generated in an identical way to the three comments treatment, this heterogeneity likely reflects community suspicion of the volume of comments. After accounting for this quality dimension, I find that the effect of attention on production is positive and increasing in a way that is consistent with the reduced form evidence.

2.1 Overview of the Experimental Design

In this subsection, I describe the experimental design in chronological order. The experiment starts with an AWS server which monitors a set of subreddits for new posts. Each time a new Reddit submission is posted to one of these subreddits, the server is pinged.

When a post arrives, I check if the post's author has already entered the sample. If so, the author-post pair is skipped and nothing happens. If not, the author-post pair enters the sample.

With 97.5% probability, the post is randomized into the control group, and with 2.5% probability, the post is randomized into treatment. Among treated posts, 50% are randomized into the "three comments" treatment condition, and 50% are randomized into the "six comments" treatment condition.

If a post is randomized into treatment, I generate candidate comments using a natural language processing pipeline built on top of the OpenAI Chat Completion API, the large language model that powers ChatGPT. I provide the API with information on the subreddit and title of the post. If available, I provide the API with information on the first hundred words of the post and the post flair. I prompt the API to provide a short, positive comment. I query the API six to twelve times to create candidate comments.

I then post three or six comments on treated posts, depending on the treatment group. I do this using a network of over a thousand Reddit accounts that I create for this experiment and that I pilot programatically. I refer to these accounts as 'Reddit bots.' I randomly select Reddit accounts from the network, and post the generated comments with a delay of five to ten minutes between comment postings.

Finally, I set up a second server to track the posting behavior of treated and control Reddit users. I do this by repeatedly querying the official Reddit API each day to see the history of posts by each user, and collect information on each new post produced. I keep track of the scores of each new post separately, collecting scores only after twenty-four hours have passed since the post was created in order to give each post a natural lifecycle with which to collect upvotes. I continue to collect information on posting behavior for thirty days after the moment of randomization.

2.2 Choice of Treatment Subreddits

I execute the experiment on very small number of hand-selected subreddits.

I exclude subreddits from the experiment for three independent reasons.

First, there are many subreddits which would be ethically dubious to interact with given the fact that the comments I post are generated randomly using a large language model. I do not post on subreddits that involve advice seeking (relationship, legal, or otherwise), and I do not interact with posts that are tagged as 'serious.' I also avoid interacting with any subreddits that are concerned with mental or physical health as well as any subreddits that engage in the discussion of news or political discourse.

Secondly, there are many subreddits which I believe are be ethical to interact with in principle, but that are not included due to the fact that I do not believe that I am able to produce 'credible' responsive comments to the posts that are involved. The subreddits in this category tend to be highly specific fandom communities (sports, television, video games, and other media properties) as well as subreddits with content that cannot be easily undestood and commented on with the information available from the title and subreddit.

Third, there are subreddits that I excluded because I believed that treatment would be functionally ineffective. These are subreddits where very few posts are made, and nearly all posts get a large degree of engagement. Given the already light-touch nature of treatment, my belief was that it would be infeasible to detect effects when additional comments were a drop in the ocean relative to baseline engagement.

For all three reasons, the subreddits included in the experiment are highly selected. Given that the goal of the experiment is to provide tightly identified causal evidence of the effect of attention on production, I view this selection trade-off as acceptable. The experiment can be thought of as an 'existance' argument, showing that, at least in some cases, attention does incentivize production. I do not claiming that this is a representative set of subreddits, or that attention incentivizes posting in all subreddits. However, this is one reason why the reduced form evidence is complementary to the experiment, as the reduced form strategy can be estimated on all subreddits without running into the same ethical concerns.

I hand check each subreddit for the conditions described above before the inclusion in the experiment. In the end, I include a small number of subreddits in the experiment. These subreddits are r/Awww, r/cats, r/pics, r/mildlyinteresting, r/NoStupidQuestions, r/whowouldwin, r/Satisfyingasfuck, r/RandomThoughts, r/wildlifepohtography, r/TwoSentenceHorror, r/crochet, r/OldSchoolCool.

2.3 Discussion of Treatment Conditions

There are two treatment conditions: adding three comments and adding six comments. The decision to include two treatment conditions was motivated by the theoretical model, which generates findings based on the shape of the attention labor supply curve. Because the control condition

is equivalent to adding 0 comments, two treatment conditions identifies the concavity of the labor supply curve in principle.

The choice of including *only* two treatment conditions reflects power concerns as the experimental intervention is light-touch.

The choice of three and six comments as the two treatment conditions was somewhat arbitrary. I wanted to choose enough comments for the treatment to be noticeable to content creators, but I did not want to choose so many comments as to raise red flags that the attention might be spam.

2.4 Discussion of Credibility of Treatment

From the perspective of Reddit users, bot accounts appear like regular Reddit accounts. They have profile pages and exhibit a relatively low commenting frequency so as to be consistent with human commenting patterns. I include a filter in the natural language processing pipeline that excludes all comments which make reference to large language models and related terms.

However, there are some aspects of these accounts that might have caused users to be suspicious. First, all accounts were created newly for the experiment, so each account is a few months old at the time that it commented (this knowledge is available to users on each accounts profile page). Second, accounts have randomly generated names, no profile pictures, and make no posts. All of these traits are reasonably comment for regular Reddit accounts, but Reddit users who were familiar with the distribution of profile characteristics on Reddit could plausibly have been suspicious that these accounts were more likely to be bots than the average account.

Ultimately, I cannot definitively say that users perceived the posted comments to be written by humans, though that was the intention of treatment.

2.5 Outcome Data Collection Method

I collect data on the activity of treated and control Reddit users using the Reddit API. Each day, I queried the post history of each treated and control user. I use this information to update a dataset of posts by each user with any new posts that have been created during the intervening day. I continue to collect data on the posting behavior of Reddit users in my sample for thirty days after the time of randomization.

I only update the count of upvotes for posts in my dataset after twenty-four hours has past since the time of posting. The decision to wait twenty-four hours to collect upvote data reflects the fact that Reddit posts typically receive the majority of their overall engagement within the first twenty-four hours of their existence. I view the upvote count after twenty-four hours as a reasonably good measure for the overall success of the post.

One limitation of this data collection method is that I do not observe posts that are created and deleted within a twenty-four hour period. Additionally, I do not follow the success of posts after twenty-four hours, so I do not capture any success that posts garner after this moment.

2.6 Preregistred Outcome Variables

I pre-registerer one primary measure and three secondary measures of output.

I will refer to the primary outcome as 'quality-weighted quantity.' This measure is computed by taking the $\sum log(\max(\text{Upvotes}, 1))$ for posts produced in the seven days after randomization. The max function ensures that a post cannot contribute less than zero to output.[13] The log function is a somewhat arbitrary functional form choice that is meant to offset the winning-begets winning nature of upvotes, an institutional feature that results in a long tailed distribution for upvotes. Specifically, a small number of posts on Reddit receive a very large number of upvotes. These posts are helped by the fact that upvotes move posts towards the top of the website, causing more people to view the post, which in turn can result in more upvotes. While I believe that a post which gets a thousand upvotes is certainly better than one that gets ten, due to this winning-begets-winning feature, I do not want to say that a one hundred upvote post represents one hundred times more output than a post that gets one upvote. Interpreting the measure as a whole now, the $\sum log(\max(\text{Upvotes}, 1)$ is a function which increases for each post with upvotes > 1 with larger increases for posts with more upvotes.

I also pre-register three secondary measures of output that are simpler.

The first secondary measure of output is the count of posts. This is a simple count of the number of posts made by the Reddit account in the seven days after randomization.

The next secondary measure is 'posting again': this is an indicator variable for whether the Reddit account posts in the seven days after randomization. This measure is intended to capture the extensive margin, and is deliberately coarse. It throws additional variation that could be gained from studying the count of posts, but ensures that no one person who posts very frequently influences the result too much.[14]

The final measure of output is the mean upvotes conditional on posting. This outcome is meant to be a measure of whether treatment changes the quality of posts. Mean upvotes is an imperfect measure of quality, precisely because of the winning-begets-winning dynamic described previously. However, I believe that average upvotes represents a reasonable measure that is at least positively correlated with true quality.

In addition to the seven day versions of these outcomes, I also preregister looking at the thirty day versions of the same quantities. Again, the primary preregistered outcome is the seven day version of quality-weighted quantity.

2.7 Preregistered Analyses

For all analyses that I describe in this section, I preregister separating analysis into two samples: users who have at least 50 submissions, and those who have less than 50 submissions. The rationale

[13] This outcome is possible because my data includes some posts with 0 or -1 upvotes. The net upvotes distribution is censored at -1 in the Reddit API, so I do not observe posts that are heavily downvoted. They instead show up as having -1 upvotes.

[14] In principal, there is no upper limit to how many times that an account may post, and if bots or superstar Reddit

for this division of sample is that the treatment is much less likely to be influential to experienced Redditors, as a few comments is a drop in the bucket relative to their experienced history of comments. Treatment effects were anticipated for the < 50 submissions sample.

I preregister one primary analysis as well as a variety of secondary analyses that are meant to improve precision.

The primary analysis is a simple comparison of means for the control group and each treated group.

In order to get more precise estimates, I propose two strategies. First, I include controls for past posting behavior. The motivation for this strategy is to reduce noise by accounting for heterogeneity across posting frequency within the samples.

Second, I use the treatment strategy as an IV for responding to the treated comments, and estimate the treatment effect of response. The rationale for t his estimation strategy is that many producers may not see the treatment comments, and this strategy estimates effects for a subset of producers who are guaranteed to see the treatment comments. The randomly assigned added comments are treated as an instrument for responding to a treated comment, so the instrumental variables assumption is mechanically valid. The key downside of this strategy is that it essentially estimates a distinct parameter, which is the treatment effect of a comment that is observed and merits a response. This is distinct parameter from the treatment effect of adding a comment, and so this estimate is not directly comparable to the rest of the estimates in the paper.

Finally, I attempt to estimate the concavity of the attention labor supply curve by estimating a second degree polynomial on the treatment effects. A negative coefficient on the second degree term indicates concavity.

2.8 Deviations from Preregistration

Due to technical issues with the experiment, I made two significant changes to the way that the experiment was run relative to the preregistration document written on July 31, 2023.

The first change is that I abandon the treatment arms which involve adding upvotes to posts. These arms were dropped after I found that it was technically infeasible to implement this treatment at scale. Reddit has a relatively sophisticated system intended to prevent vote manipulation, and I was not able to have accounts engage in random upvoting without them being flagged and banned by this system.

The second deviation from preregistration is sample size. I initially preregistered a sample size of 100,000 treated units. In the preregistration, I also anticipated that technical issues may result in a smaller sample size, and committed to reporting the results on whatever sample I am able to collect.

The proposed 100,000 sample size became infeasible due to issues with scaling. As the comments arm of the experiment was scaled up, bots were quickly getting flagged and banned. Some subreddits

posters end up unequally randomized, this could bias results. This kind of unequal randomization is possible because bots and superstars are a low fraction of the population relative to the sample size.

that I had initially factored in when making back-of-the-envelope sample size calculations turned out to have strong moderation polices that resulted in account bans. For this reason, I had to run the experiment at a more moderate pace on a smaller number of subreddits, which resulted in a substantially smaller final sample size.

2.9 Results

Figure 7 plots the primary results for all four outcome variables measured over the week after treatment. Overall treatment effects are provided, as well as treatment effects for experienced users (with above 50 submissions) and inexperienced users (with 50 or fewer submissions).

The three comments treatment causes producers to create more content. It increases the extensive margin (posting again) and increases the total number of posts created. Conditional on production, this content is worse on average. Taken together, this results in an increase in the quality-weighted quantity of content produced, but this effect is estimated somewhat imprecisely.

In contrast, I find no significant treatment effects in the six comment treatment arm across all four treatments. This is surprising, given the rest of the findings in the paper. In order to investigate this null result, I document a particular kind of unintended treatment heterogeneity. I find that treatment comments were less well received in the six comments treatment compared to the three comments treatment. One measure of this is shown in **??**, which documents that the six comments treatment is 10% more likely to have an above-average number of net downvotes per comment. Since these comments are generated in an identical way to the three comments treatment, this heterogeneity likely reflects community suspicion of the volume of comments.

This turns out to be an instructive split for understanding heterogeneity in treatment effects. **??** shows that downvoted comments decrease production while non-downvoted comments increase production, in terms of count of posts, probability of posting again, and a quality-weighted measure of posts (sum log score of posts).

3 A Theoretical Model of an Attention Platform

In **??** and Section 2, I document that attention causes content producers to increase their output. Moreover, I show that the labor supply of content producers is concave with respect to the attention incentive. Initial units of attention increase the amount of content supplied, but this increase levels off as more and more attention is received.

In this section, I take this empirical pattern as a starting point, and develop a model with the goal of understanding how the attention incentive informs the optimal design and regulation of social media platforms. In the model, a social media platform manages a two-sided market composed of content producers and consumers. The model builds on classic models of two-sided markets (Rochet and Tirole, 2003; Armstrong, 2006), but incorporates the idea that markets "clear in attention" rather than prices. That is, in equilibrium, the amount of attention that consumers supply must justify the quantity of producers who choose to produce content, and the amount of content produced must justify the amount of attention that consumers supply.

I study a platform whose profits scale with the number of consumers that choose to join. The platform acts as a curator, choosing which pieces of content to serve to consumers among those that have been created by producers. That is, the platform chooses the quantity and quality of content available to consumers subject to feasibility constraints. I interpret this choice as a simple content recommendation algorithm, the kind of algorithm that all major social media platforms use to generate personalized feeds. I use the model to derive results regarding the relationship between the shape of the attention labor supply curve and the optimal profit and welfare maximizing content recommendation algorithms.

The model delivers two key results. First, if the attention labor supply curve is sufficiently concave, then the platform maximizes consumer demand by showing some "bad" content. In the context of the model, bad content is content that provides consumers with negative utility. The intuition for this result is that showing bad content provides producers with additional attention, which boosts aggregate content supply in equilibrium. If consumers value a marginal unit of good content enough, then a large supply response justifies the inclusion of bad content on the platform.

Second, the percentage of bad content which maximizes consumer demand is a lower bound on the percentage of bad content which would maximize consumer welfare, producer welfare, social welfare, and the aggregate number of impressions. If the labor supply curve is sufficiently concave, then maximizing any of these objectives requires showing a strictly positive percentage of bad content.

An implication of the second result is that the algorithm which maximizes social welfare shows more bad content than the profit-maximizing algorithm. The intuition for this wedge is that the platform values producers only insofar as content supply allows them to attract consumers. In contrast, the social planner values consumer and producer utility. The planner trades-off some consumer utility for producer utility by showing more bad content to consumers in order to provide more attention to producers. This wedge implies that regulating content recommendation algorithms can be welfare improving.

All proofs are left to the theoretical appendix in order to simplify exposition.

3.1 Model Setup

Overview. Before formalizing the model, I provide a brief overview. The model is static, but it may be helpful to think about the model as occurring in three stages.

1. First, the platform promises content producers a certain amount of attention. Observing this promise, potential content producers decide whether or not to produce content. The platform observes the quantity of good and bad content that producers have created.

2. Second, the platform curates the content. The platform chooses the quantity of good and bad content that is available to consumers from among the content that has been produced.

3. Third, consumers observe the content that the platform offers, and decide whether to join the platform. Those who join the platform consume the content that is available, generating attention for content producers.

In equilibrium, the amount of attention that consumers produce must be equal to the promise made by the platform in the initial stage.

Producers. Producers decide whether or not to produce content for the platform. Producers value the number of impressions i that their content receives. The number of impressions i is an equilibrium object which depends on the decisions of consumers and the platform, as well as on the quality of the content that the producer creates. Content is good with exogenous probability $q \in (0, 1)$ and bad otherwise. Good content receives i_g impressions while bad content receives i_b impressions. The utility that producers derive from attention is captured by $V(i)$. The attention utility function V is assumed to be positive, increasing, and concave, which corresponds to the empirical findings of **??**.

Content producers face a heterogeneous cost of effort for producing content δ. Effort cost δ is distributed according to the probability density function $k(\delta)$ with cumulative density function $K(\delta)$. Producers decide to produce content if their expected attention returns outweigh their effort cost:

$$\mathbb{E}[V_P] = \begin{cases} qV(i_g) + (1-q)V(i_b) - \delta & \text{Create Content} \\ 0 & \text{Otherwise} \end{cases}$$

Content producer supply S is given by

$$\begin{aligned} S := S(i_g, i_b) &= \int \mathbb{I}\{qV(i_g) + (1-q)V(i_b) > \delta\} k(\delta) d\delta \\ &= K(qV(i_g) + (1-q)V(i_b)) \end{aligned}$$

Since content is good with probability q, we can compute the supply of good and bad content,

denoted S_g and S_b respectively:

$$S_g(i_g, i_b) := qS(i_g, i_b)$$
$$S_b(i_g, i_b) := (1-q)S(i_g, i_b)$$

Producer welfare is given by

$$W_P = \int \max\{qV(i_g) + (1-q)V(i_b) - \delta, 0\}k(\delta)d\delta$$

The Platform's Curation Choice. The platform chooses the number of pieces of good and bad content available to each consumer, subject to the constraint that it cannot show more content than has been supplied by producers. Denote the number of good and bad pieces of content available to each consumer the platform by N_g and N_b. The platform must choose $N_g \leq S_g, N_b \leq S_b$. When $N_b < S_b$, the platform selects a random set of N_b pieces of bad content to show each consumer out of the pool of S_b pieces of content, so the aggregate consumer impressions of bad content are spread evenly across all pieces of bad content (and likewise for good content).

Consumers. Consumers choose whether or not to join the platform. To make this decision, they evaluate the platform as a whole, with their platform consumption utility $U(N_g, N_b)$ increasing in the number of good pieces of content on the platform N_g and decreasing in the number of bad pieces of content on the platform N_b. Each consumer faces an idiosyncratic fixed cost of joining the platform $\epsilon \sim l(\epsilon)$ with cumulative density function $L(\epsilon)$.

Define the consumer utility function $U_C(N_g, N_b)$

$$U_C(N_g, N_b) = \begin{cases} U(N_g, N_b) - \epsilon & \text{Join Platform} \\ 0 & \text{Otherwise} \end{cases}$$

Consumer demand D for the platform is given by

$$\begin{aligned} D := D(N_g, N_b) &= \int \mathbb{I}\{U(N_g, N_b) > \epsilon\}l(\epsilon)d\epsilon \\ &= L(U(N_g, N_b)) \end{aligned}$$

Consumer welfare is given by

$$W_C = \int \max\{U(N_g, N_b) - \epsilon, 0\}l(\epsilon)d\epsilon$$

Market Clearing Conditions. I make an assumption about the way that consumers behave in order to create a tight relationship between the amount of content offered to each consumer (N_g, N_b) and the number of impressions that producers receive (i_g, i_b).

Suppose the platform offers each consumer N_g pieces of good content and N_b pieces of bad content. Then, there are $D(N_g, N_b)$ consumers on the platform. The key assumption is that each

consumer "consumes the platform." That is, each consumer views all N_g pieces of good content and N_b pieces of bad content.

Under this assumption, each of the D consumers views N_g pieces of good content to provide a total of DN_g impressions of good content. The platform distributes these impressions equally across the S_g pieces of good content, so each piece of good content gets $\frac{DN_g}{S_g}$ impressions. This idea yields two market clearing conditions.

For each $\theta \in \{g, b\}$,

$$\underbrace{S_\theta(i_g, i_b)}_{\text{Supply of Content}} \times \underbrace{i_\theta}_{\text{Impressions per Content}} = \underbrace{D(N_g, N_b)}_{\text{Consumer Demand}} \times \underbrace{N_\theta}_{\text{Impressions per Consumer}} \tag{1}$$

These two conditions express the idea that the number of impressions supplied to producers must equal the number of impressions provided by consumers for both good and bad content.

The Platform's Problem. The platform chooses the amount of good and bad content available to consumers in order to maximize profit, subject to the constraint that it cannot show more content than it has available. The supply of content depends endogenously on the platform's choices through the market clearing conditions.

Formally, the platform's problem is

$$\begin{aligned} \max_{N_g, N_b} \quad & \Pi(N_g, N_b) \\ \text{subject to} \quad & N_g \leq S_g(i_g, i_b) \\ & N_b \leq S_b(i_g, i_b) \\ & S_g(i_g, i_b) i_g = D(N_g, N_b) N_g \\ & S_b(i_g, i_b) i_b = D(N_g, N_b) N_b \end{aligned} \tag{2}$$

I start by assuming that $\Pi = D(N_g, N_b)$, meaning that profit scales with the number of consumers who choose to join the platform. This objective function reflects the advertising-based profit model of social media platforms.

Observation. *The platform will show all available good content.* To see this, notice that showing good content both increases the objective function and loosens the constraints. This is because consumers like good content, so increasing the amount of good content on the platform increases the number of consumers on the platform, which increases the number of impressions, which increases supply. Formally, $N_g = S_g$. Applying market clearing, $i_g = D$.

Since the platform's decision about good content is trivial, the primary choice of interest is how the platform handles bad content. This decision can be summarized by a parameter β which is defined as the percentage of bad content that the platform chooses to show, out of the total amount of bad content that was supplied by producers.

$$\beta := \frac{N_b}{S_b}$$

This definition, along with market clearing, simplifies the expressions for supply and demand.

$$D(N_g, N_b) = D(qS, \beta(1-q)S)$$
$$S(i_g, i_b) = S(D, \beta D)$$

The platform's problem can be rewritten as

$$\max_{\beta} \quad \Pi = D(qS, \beta(1-q)S) \tag{3}$$
$$\text{subject to} \quad 0 \leq \beta \leq 1$$

Observation. *If the supply of content is exogenously fixed, then the platform should show no bad content.* More formally, if the supply of content $S(D, \beta D)$ is fixed to some level $\overline{S} > 0$, then $\beta = 0$.[15]

The point is that if we shut down endogenous content supply concerns in this model, then there is no incentive for the platform to show any bad content. For a fixed supply of content, the platform maximizes profits by showing all of the good content ($N_g = S_g$) and none of the bad content ($N_b = 0$).

Assumption. For the rest of the model, assume that $\frac{\partial D}{\partial S} > 0$. Recall that $D(qS, \beta(1-q)S)$. This assumption means that, for any fixed ratio of good to bad content, having more content on the platform is desirable to consumers.[16]

3.2 The Platform's Profit Maximizing Strategy

Recall that $\beta \in [0, 1]$ is the percentage of bad content that the platform shows each consumer out of the supply of bad content S_b.

Definition. Let β_C^* denote the value of β that maximizes the number of consumers on the platform.

Definition. Let D_0 denote the equilibrium value of D when $\beta = 0$. Let D_1 denote the equilibrium value of D when $\beta = 1$.

Proposition 1. *If the producer attention utility function V is sufficiently concave and consumers value good content enough, then the platform shows consumers a positive percentage of bad content.*

- *More formally, suppose that $\frac{\partial D}{\partial N_g} > -\frac{1}{2}\frac{\partial D}{\partial N_b}$ when evaluated at D_0. For fixed values of $V(0)$ and $V(D_0)$, if $V'(0)$ is large enough and $V'(D_0)$ is small enough, then $\beta_C^* > 0$.*

If the producer attention utility function V is sufficiently concave, then the platform does not show consumers all bad content.

[15] For a proof, see the appendix.

[16] This assumption could be justified by imagining some unmodeled consumer heterogeneity, so that larger pools of content allow for better targeting. In this case, consumers are not literally consuming the platform, but instead are consuming fixed fraction of the platform, in order to allow room for search while still allowing for some importance

- *For fixed values of $V(0)$ and $V(D_1)$, if $V'(D_1)$ is small enough, then $\beta_C^* < 1$.*

Discussion. This proposition relates the way that producers value attention V to the optimal content recommendation algorithm β. All results center around β_C^*, which is the percentage of bad content that the platform should choose to show in order to maximize the number of consumers on the platform.

If the producer attention utility function is sufficiently concave, then the platform should show some, but not all, of the bad content that was supplied by producers. The intuition for this proposition comes from the total derivative of consumer demand with respect to β.

$$\frac{dD}{d\beta} = \frac{\partial D}{\partial \beta} + \frac{\partial D}{\partial S}\frac{\partial S}{\partial \beta}$$

This expression showcases the two forces of the model. First, consumers dislike the inclusion of bad content on the platform, which corresponds to the partial derivative $\frac{\partial D}{\partial \beta} < 0$. Second, consumers like additional content supply ($\frac{\partial D}{\partial S} > 0$), and including additional bad content on the platform may increase content supply ($\frac{\partial S}{\partial \beta} > 0$). Whether the platform should choose to show bad content depends on which of these two forces dominates.

The reason that $\frac{\partial S}{\partial \beta}$ may be positive is because increasing β can increase the amount of attention content producers get when they produce bad content. When $\beta = 0$, producers get no attention in the bad state, and realize utility $V(0)$. If there are large utility returns to the first units of attention (i.e. $V'(0) >> 0$), then it can be worth it for the platform to show some bad content, because the large boost to expected producer utility will lead to a large boost to content supply in equilibrium.

By a similar logic, when $\beta = 1$, producers get attention utility $V(D_1)$ if they produce bad content. If the marginal utility of producers at this positive level of attention D_1 is small (i.e. $V'(D_1) \approx 0$), then the inclusion of the last units of bad content on the platform delivers a small boost to producer expected utility. In this case, the correspondingly small boost to equilibrium supply will not offset the direct costs to consumers.

Informally, we can think about the property of the function $\frac{\partial S}{\partial \beta}(\beta)$ that guarantees an intermediate solution as a kind of 'concavity' of the supply curve. If the derivative of S with respect to β is sharply increasing near zero and flattens out when β is large, then the optimal choice of β is somewhere in the middle. This is because increasing the amount of bad content on the platform β has a direct negative effect on consumers, so if the marginal gains to supply decline quickly as β grows large, then at some point these gains will not offset the direct costs to consumers.

This informal 'concavity' intuition extends down to the producer attention utility function V. Increasing the fraction of bad content β may increase the impressions of bad content i_b, which in turn increases producer utility in the bad state $V(i_b)$. If the marginal returns to the first unit of attention $V'(0)$ are large and the marginal returns to units of attention beyond some positive level of attention D_1 are small (i.e. $V'(D_1) \approx 0$), then the optimal amount of attention to offer to producers in the bad state is intermediate, since offering producers attention for bad content is

for the overall supply of content offered by the platform.

costly to consumers.[17]

3.2.1 Extension: Multiple Consumer Types

In the appendix, I extend the model to accommodate a second kind of consumer, which I call a *light* consumer. Rather than "consuming the platform" (i.e. contributing one impression to each piece of content on the platform $N_g + N_b$), light consumers provide a fixed amount of impressions M. Since M is assumed to be small relative to the overall supply of content, the platform has complete flexibility to choose the fraction of good and bad content that this type of consumer is offered.

If producers attention utility function V is linear, then the platform should only show light consumers good content. However, if V is sufficiently concave, then the platform should show light consumers some bad content.

This extension demonstrates that the intuition for Proposition 1 does not depend on the fact that the platform shows bad content *in addition* to good content. Even when showing bad content directly trades off with showing good content, the platform may still want to show some bad content.

3.3 The Social Planner's Welfare Maximizing Strategy

In this subsection, I consider a variety of alternative objectives that a platform or a social planner might want to pursue. For an agent choosing the percentage of bad content to show to consumers $\beta \in [0, 1]$, define the following objectives:

- Let β^*_C maximize the number of consumers on the platform $D(N_g, N_b)$.
- Let β^*_P maximize the number of producers on the platform $S(i_g, i_b)$.
- Let β^*_{CW} maximize consumer welfare W_C.
- Let β^*_{PW} maximize producer welfare W_P.
- Let β^*_{SW} maximize social welfare, which is a linear combination of producer and consumer welfare. Formally, for $\alpha \in (0, 1)$, social welfare is $W_S = \alpha W_C + (1 - \alpha)W_P$.
- Let β^*_I maximize the number of impressions, which is given by $D(N_g, N_b)(N_g + N_b)$.

Proposition 2. *The percentage of bad content which maximizes each of the welfare objectives is weakly ordered*

$$\beta^*_P = \beta^*_{PW} \geq \beta^*_{SW} \geq \beta^*_{CW} = \beta^*_C$$

Moreover, the percentage of bad content which maximizes impressions is weakly larger than the

[17] In the proof, I formalize the sense in which this property relates to the concavity of V using Taylor expansion.

percentage which maximizes the number of consumers on the platform:

$$\beta_I^* \geq \beta_C^*.$$

Discussion. This proposition makes two interrelated points. First, β_C^* is a lower bound on the percentage of bad content which maximizes a wide variety of objectives including consumer welfare, producer welfare, social welfare, and the aggregate number of impressions on the platform. Applying Proposition 1, if the attention labor supply curve is concave enough, then maximizing any of these objectives requires showing a strictly positive percentage of bad content on the platform.

Second, this proposition relates the optimal content recommendation algorithms that maximize consumer, producer, and social welfare. The planner should show less bad content to maximize consumer welfare, more bad content to maximize producer welfare, and an intermediate amount of bad content to maximize social welfare.

If we maintain the assumption that a profit-maximizing platform wants to maximize the number of consumers on the platform, then the platform chooses β_C^*. Since $\beta_C^* \leq \beta_{SW}^*$, there is a potential wedge between the profit and welfare maximizing algorithms.

This wedge implies that regulating content recommendation algorithms could be welfare improving. Specifically, a profit-maximizing platform may not be sending enough traffic to low-quality content. That the social planner would want to inconvenience users by showing them low-quality content might seem counterintuitive, but it may help to recognize that content producers on social media are people whose utility the social planner values. If these people value attention enough, then it can be worth it for the social planner to direct attention towards their content, even though users do not want to see it. In this case, the social planner engages in a kind of utility cross-subsidization: the planner includes bad content as a tool to trade-off some consumer utility for producer utility in order to maximize welfare.

4 Conclusion and Policy Implications

The last few decades have been marked by the rise of attention platforms. Search engines, news websites, media platforms, and social media companies each preside over markets where consumers, content producers, and advertisers interact.

In this paper, I analyze the optimal design and regulation of attention platforms through the lens of a classic idea: attention can function as a non-monetary incentive. This incentive matters because attention platforms use content recommendation algorithms to distribute consumer attention across content producers. Since producers value attention, these algorithms affect content supply.

In the empirical portion of this paper, I document that attention is an effective non-monetary incentive. I leverage the institutional features of Reddit and TikTok to help disentangle attention from related social and financial incentives. Using difference-in-differences designs, I find that going viral causes content producers to produce 80% more content which is 13% better over the next two weeks.The rich nature of the observational data allows me to trace out an attention labor supply

curve. I find that this curve is concave: the first units of attention sharply increase content supply, while marginal attention beyond a certain threshold is not influential.

I complement this reduced form evidence with a large scale field experiment. I randomly add comments to Reddit posts that I generate with ChatGPT, and deliver these comments through a network of Reddit accounts. I find that adding three comments to Reddit posts causes content creators to produce 5% more posts, though my estimate of the effect of six comments is an imprecise null. Overall, the field experiment confirms that attention functions as an incentive.

I develop a model of a social media platform that takes the attention incentive seriously. In the model, a social media platform manages a two-sided market between content creators and consumers. If the attention labor supply curve is sufficiently concave, than a platform should show a strictly positive percentage of bad content in order to maximize consumer demand. A welfare-maximizing social planner would show a weakly larger percentage of bad content.

Looking forward, this paper gestures at two ways in which understanding the attention incentive could improve policy. First, accounting for attention can help us design healthier online communities. Given the meteoric rise of social media and its function as a forum that shapes our public discourse, getting the design of these online spaces right is important. Second, the value that people place on attention provides a novel justification for the regulation of social media algorithms. The model demonstrates that attention can create a wedge between profit-maximizing behavior and social welfare, which implies that regulating social media algorithms could have a positive impact.

More abstractly, this paper provides empirical evidence that attention is a psychological commodity which people value inherently. This fact has potentially wide-ranging implications across a variety of public policy areas, because the allocation of attention is a fundamental aspect of life as a social species. All of our relationships are mediated by the ways in which we choose to allocate our attention, and one of the core findings of this paper is that a little bit of attention goes a long way.

References

Philipp Ager, Leonardo Bursztyn, Lukas Leucht, and Hans-Joachim Voth. Killer incentives: Rivalry, performance and risk-taking among german fighter pilots, 1939–45. *The Review of economic studies*, 89(5):2257–2292, 2022.

George A Akerlof and Rachel E Kranton. Economics and identity. *The quarterly journal of economics*, 115(3):715–753, 2000.

Jaime Arguello, Brian S Butler, Elisabeth Joyce, Robert Kraut, Kimberly S Ling, Carolyn Rosé, and Xiaoqing Wang. Talk to me: Foundations for successful individual-group interactions in online communities. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*, pages 959–968, 2006.

Dan Ariely, Emir Kamenica, and Dražen Prelec. Man's search for meaning: The case of legos. *Journal of Economic Behavior & Organization*, 67(3-4):671–677, 2008.

Mark Armstrong. Competition in two-sided markets. *The RAND journal of economics*, 37(3): 668–691, 2006.

Nava Ashraf, Oriana Bandiera, and Scott S Lee. Awards unbundled: Evidence from a natural field experiment. *Journal of Economic Behavior & Organization*, 100:44–63, 2014.

David Atkin, Eve Colson-Sihra, and Moses Shayo. How do we choose our identity? a revealed preference approach using food consumption. *Journal of Political Economy*, 129(4):1193–1251, 2021.

Hemant K Bhargava. The creator economy: Managing ecosystem supply, revenue sharing, and platform design. *Management Science*, 68(7):5233–5251, 2022.

Danah M Boyd and Nicole B Ellison. Social network sites: Definition, history, and scholarship. *Journal of computer-mediated Communication*, 13(1):210–230, 2007.

Moira Burke, Cameron Marlow, and Thomas Lento. Feed me: motivating newcomer contribution in social network sites. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 945–954, 2009.

Gordon Burtch, Yili Hong, Ravi Bapna, and Vladas Griskevicius. Stimulating online reviews by combining financial incentives and social norms. *Management Science*, 64(5):2065–2082, 2018.

Gordon Burtch, Qinglai He, Yili Hong, and Dokyun Lee. How do peer awards motivate creative content? experimental evidence from reddit. *Management Science*, 68(5):3488–3506, 2022.

Bernard Caillaud and Bruno Jullien. Chicken & egg: Competition among intermediation service providers. *RAND journal of Economics*, pages 309–328, 2003.

Brantly Callaway and Pedro HC Sant'Anna. Difference-in-differences with multiple time periods. *Journal of econometrics*, 225(2):200–230, 2021.

Lester T Chan. Quality strategies in network markets. *Management Science*, 2023.

Daniel Chen. The market for attention. *Available at SSRN 4024597*, 2022.

Yan Chen, F Maxwell Harper, Joseph Konstan, and Sherry Xin Li. Social comparisons and contributions to online communities: A field experiment on movielens. *American Economic Review*, 100(4):1358–1398, 2010.

DataReportal. Digital 2019: Global digital overview, 1 2019. URL `https://datareportal.com/reports/digital-2019-global-digital-overview`. Accessed: 2023-09-26.

DataReportal. Digital 2023: The united states of america, 2 2023a. URL `https://datareportal.com/reports/digital-2023-united-states-of-america`. Accessed: 2023-09-17.

DataReportal. Digital 2023: The united states of america, 2 2023b. URL `https://datareportal.com/reports/digital-2023-global-overview-report`. Accessed: 2023-09-26.

Christopher G Davey, Nicholas B Allen, Ben J Harrison, Dominic B Dwyer, and Murat Yücel. Being liked activates primary reward and midline self-related brain regions. *Human brain mapping*, 31 (4):660–668, 2010.

Josse Delfgaauw, Robert Dur, Joeri Sol, and Willem Verbeke. Tournament incentives in the field: Gender differences in the workplace. *Journal of Labor Economics*, 31(2):305–326, 2013.

Stefano DellaVigna and Devin Pope. What motivates effort? evidence and expert forecasts. *The Review of Economic Studies*, 85(2):1029–1069, 2018.

Stefano DellaVigna, John A List, and Ulrike Malmendier. Testing for altruism and social pressure in charitable giving. *The quarterly journal of economics*, 127(1):1–56, 2012.

Stefano DellaVigna, John A List, Ulrike Malmendier, and Gautam Rao. Voting to tell others. *The Review of Economic Studies*, 84(1):143–181, 2016.

Domo. Domo data never sleeps 10.0, 2022. URL `https://www.domo.com/data-never-sleeps`. Accessed: 2023-09-26.

Naomi I Eisenberger, Matthew D Lieberman, and Kipling D Williams. Does rejection hurt? an fmri study of social exclusion. *Science*, 302(5643):290–292, 2003.

Paulo B Goes, Mingfeng Lin, and Ching-man Au Yeung. "popularity effect" in user-generated content: Evidence from online product reviews. *Information Systems Research*, 25(2):222–238, 2014.

Paulo B Goes, Chenhui Guo, and Mingfeng Lin. Do incentive hierarchies induce user effort? evidence from an online knowledge exchange. *Information Systems Research*, 27(3):497–516, 2016.

Sanjay Jain and Kun Qian. Compensating online content producers: A theoretical analysis. *Management Science*, 67(11):7075–7090, 2021.

Bruno Jullien, Alessandro Pavan, and Marc Rysman. Two-sided markets, pricing, and network effects. In *Handbook of Industrial Organization*, volume 4, pages 485–592. Elsevier, 2021.

Zhihong Ke, De Liu, and Daniel J Brass. Do online friends bring out the best in us? the effect of friend contributions on online review provision. *Information Systems Research*, 31(4):1322–1336, 2020.

Muhammad Yasir Khan. Mission motivation and public sector performance: experimental evidence from pakistan. *Work. Pap., Univ. Pittsburgh, Pittsburgh, PA*, 2020.

Jonathan T Kolstad. Information and quality when motivation is intrinsic: Evidence from surgeon report cards. *American Economic Review*, 103(7):2875–2910, 2013.

Lini Kuang, Ni Huang, Yili Hong, and Zhijun Yan. Spillover effects of financial incentives on non-incentivized user engagement: Evidence from an online knowledge exchange platform. *Journal of Management Information Systems*, 36(1):289–320, 2019.

Peter Kuhn, Peter Kooreman, Adriaan Soetevent, and Arie Kapteyn. The effects of lottery prizes on winners and their neighbors: Evidence from the dutch postcode lottery. *American Economic Review*, 101(5):2226–2247, 2011.

Cliff Lampe and Erik Johnston. Follow the (slash) dot: effects of feedback on new members in an online community. In *Proceedings of the 2005 international ACM SIGGROUP conference on Supporting group work*, pages 11–20, 2005.

Björn Lindström, Martin Bellander, David T Schultner, Allen Chang, Philippe N Tobler, and David M Amodio. A computational reward learning account of social media engagement. *Nature communications*, 12(1):1311, 2021.

Dandan Ma, Shuqing Li, Jia Tina Du, Zhan Bu, Jie Cao, and Jianjun Sun. Engaging voluntary contributions in online review platforms: The effects of a hierarchical badges system. *Computers in Human Behavior*, 127:107042, 2022.

Dar Meshi, Carmen Morawetz, and Hauke R Heekeren. Nucleus accumbens response to gains in reputation for the self relative to gains for others predicts social media use. *Frontiers in human neuroscience*, page 439, 2013.

Nicole Murphy. Revealing this year's reddit recap, where we highlight how redditors kept it real in

2022, 2019. URL https://www.redditinc.com/blog/reddits-2019-year-in-review/. Accessed: 2023-09-26.

Susanne Neckermann, Reto Cueni, and Bruno S Frey. Awards at work. *Labour Economics*, 31: 205–217, 2014.

Geoffrey G Parker and Marshall W Van Alstyne. Two-sided network effects: A theory of information product design. *Management science*, 51(10):1494–1504, 2005.

Ricardo Perez-Truglia and Guillermo Cruces. Partisan interactions: Evidence from a field experiment in the united states. *Journal of Political Economy*, 125(4):1208–1243, 2017.

Reddit. How does being anonymous work on reddit?, 2023. URL https://support.reddithelp.com/hc/en-us/articles/7420342178324-How-does-being-anonymous-work-on-Reddit-#:~:text=Reddit%20lets%20you%20overshare%20without,without%20revealing%20who%20they%20are. Accessed: 2023-10-06.

Jean-Charles Rochet and Jean Tirole. Platform competition in two-sided markets. *Journal of the european economic association*, 1(4):990–1029, 2003.

Juan Manuel Sanchez-Cartas and Gonzalo León. Multisided platforms and markets: A survey of the theoretical literature. *Journal of Economic Surveys*, 35(2):452–487, 2021.

SEMRush. Website overview: reddit.com, 2023a. URL https://www.semrush.com/website/reddit.com/overview/. Accessed: 2023-09-17.

SEMRush. Most visited websites in the world, july 2023, 2023b. URL https://www.semrush.com/website/top/. Accessed: 2023-09-26.

SimilarWeb. Orverview: reddit.com, 2019. URL https://web.archive.org/web/20180409082256/https://www.similarweb.com/website/reddit.com. Accessed: 2023-09-26.

SimilarWeb. Top websites in united states - social media networks, 2023a. URL https://www.similarweb.com/top-websites/united-states/computers-electronics-and-technology/social-networks-and-online-communities/. Accessed: 2023-09-17.

SimilarWeb. reddit.com traffic and engagement analysis, 2023b. URL https://www.similarweb.com/website/reddit.com/#ranking. Accessed: 2023-09-26.

SimilarWeb. Top websites ranking, 2023c. URL https://www.similarweb.com/top-websites/. Accessed: 2023-09-26.

Yacheng Sun, Xiaojing Dong, and Shelby McIntyre. Motivation of user-generated content: Social connectedness moderates the effects of monetary rewards. *Marketing Science*, 36(3):329–337, 2017.

Rachel Thomas. What we know about america's healthiest, happiest and best-rested people, 2019. URL https://www.sleepcycle.com/sleep-science/what-we-know-about-americas-healthiest-happiest-best-rested/#:~:text=Americans%20spend%20an%20average%20of,on%20a%20scale%20of%20100. Accessed: 2023-09-26.

Olivier Toubia and Andrew T Stephen. Intrinsic vs. image-related utility in social media: Why do people contribute content to twitter? *Marketing Science*, 32(3):368–392, 2013.

André Veiga, E Glen Weyl, and Alexander White. Multidimensional platform design. *American Economic Review*, 107(5):191–195, 2017.

Yang Wang, Paulo Goes, Zaiyan Wei, and Daniel Zeng. Production of online word-of-mouth: Peer effects and the moderation of user characteristics. *Production and Operations Management*, 28 (7):1621–1640, 2019.

Zaiyan Wei, Mo Xiao, and Rong Rong. Network size and content generation on social media platforms. *Production and Operations Management*, 30(5):1406–1426, 2021.

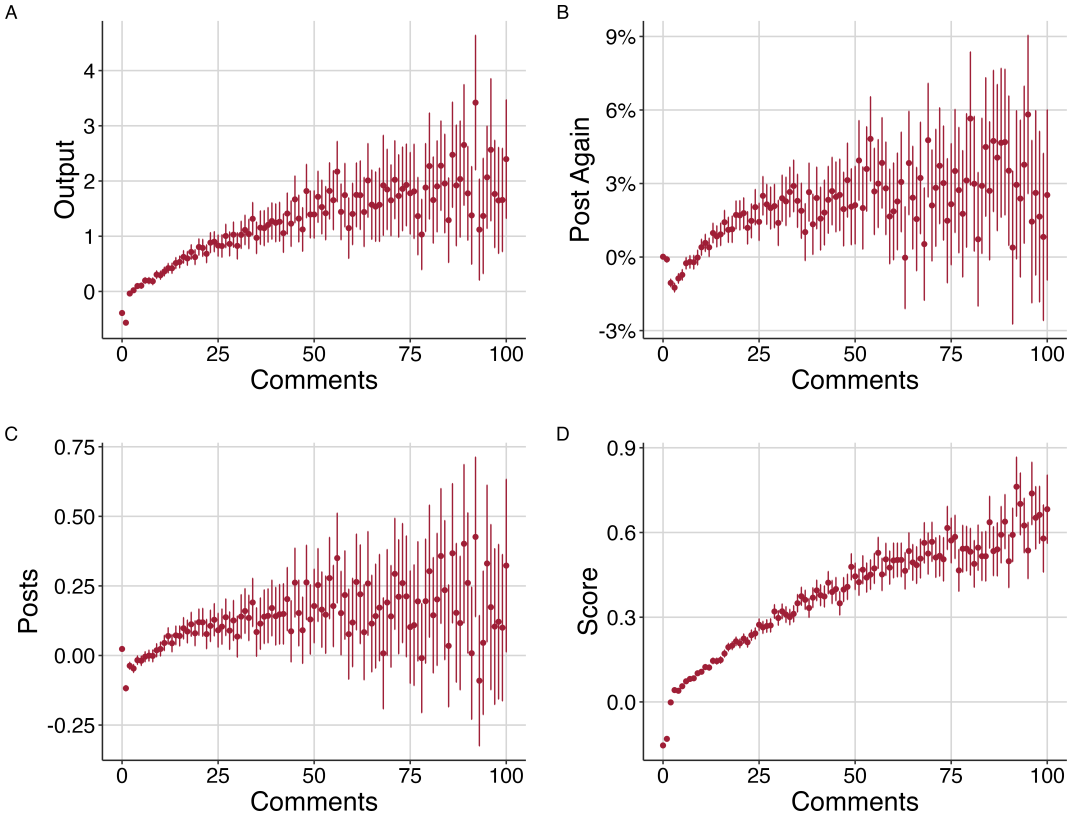
E Glen Weyl. A price theory of multi-sided platforms. *American Economic Review*, 100(4):1642–1672, 2010.

Mingyue Zhang, Xuan Wei, and Daniel Dajun Zeng. A matter of reevaluation: incentivizing users to contribute reviews in online platforms. *Decision support systems*, 128:113158, 2020.

Xiaoquan Zhang and Feng Zhu. Group size and incentives to contribute: A natural experiment at chinese wikipedia. *American Economic Review*, 101(4):1601–1615, 2011.

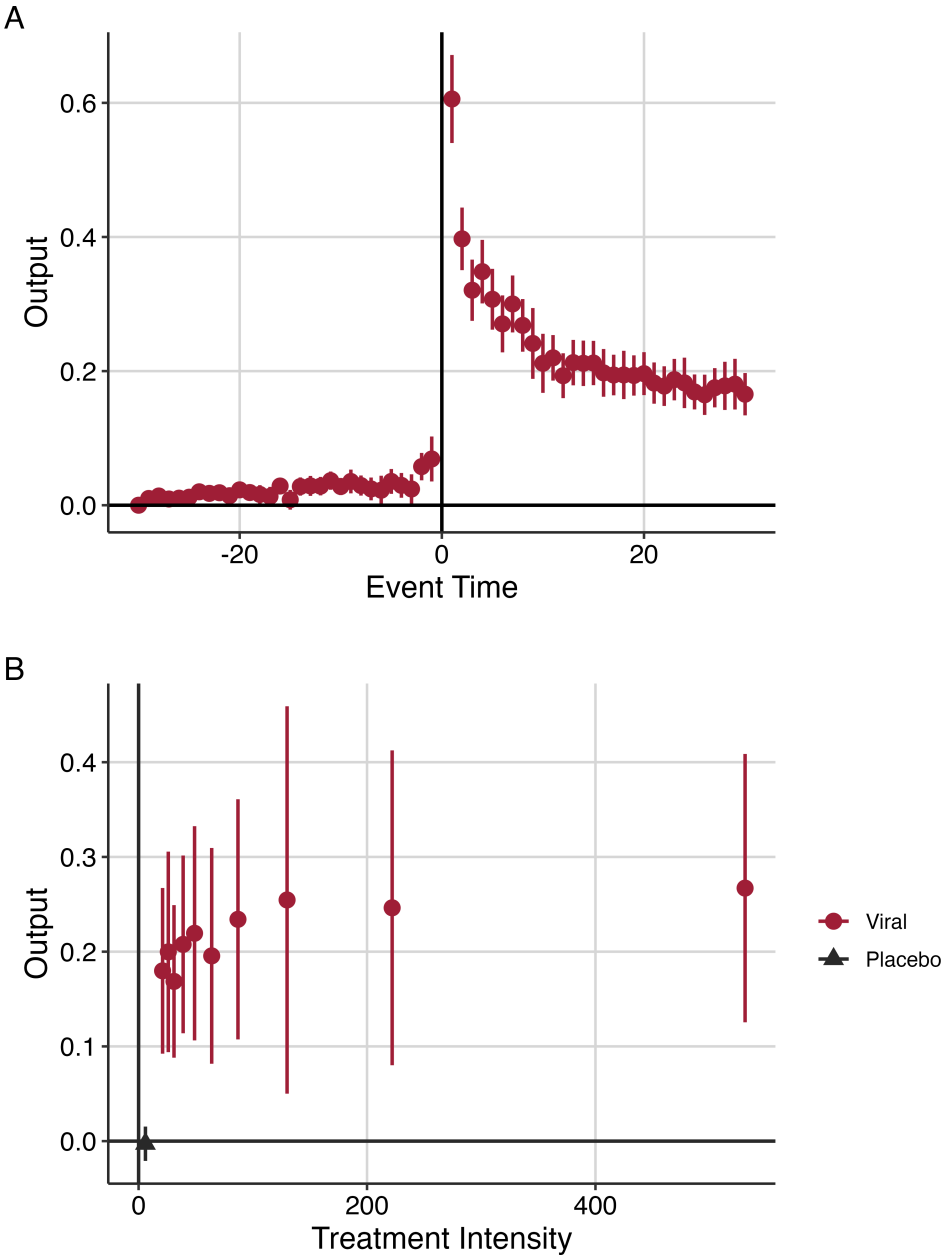
5 Figures

Figure 1: Correlation between Attention and Production



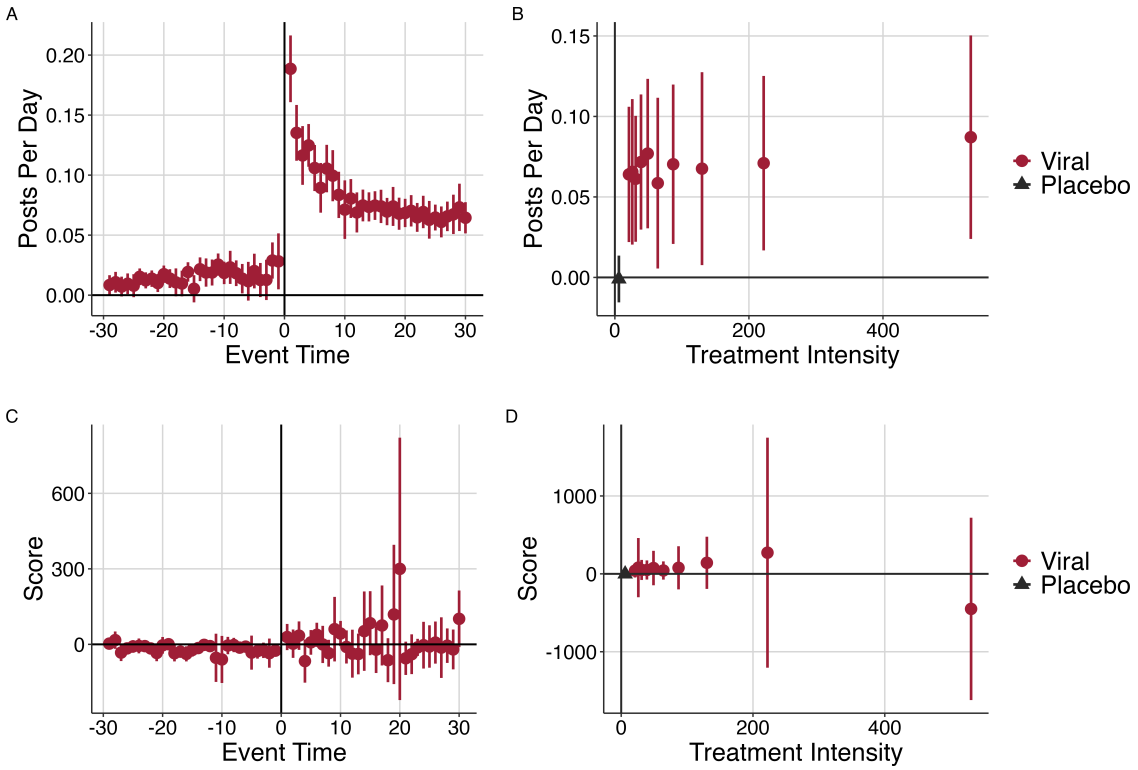
Notes: This figure presents correlations between the attention that a Reddit post receives, as measured by the number of comments, and various measures of content production by the post's author over the next week. Each point represents a one-comment bin, and bars represent 95% confidence intervals. The outcome in Plot A is $\sum log(\text{score}+1)$, which is a quality-weighted measure of output. The outcome in Plot B is an indicator for if any posts are produced in the next week, capturing the extensive margin. The outcome in Plot C is quantity, measured by the count of posts. The outcome in Plot D is quality, measured by the average score of posts. All outcomes are demeaned by subreddit.

Figure 2: The Effect of Virality on Production



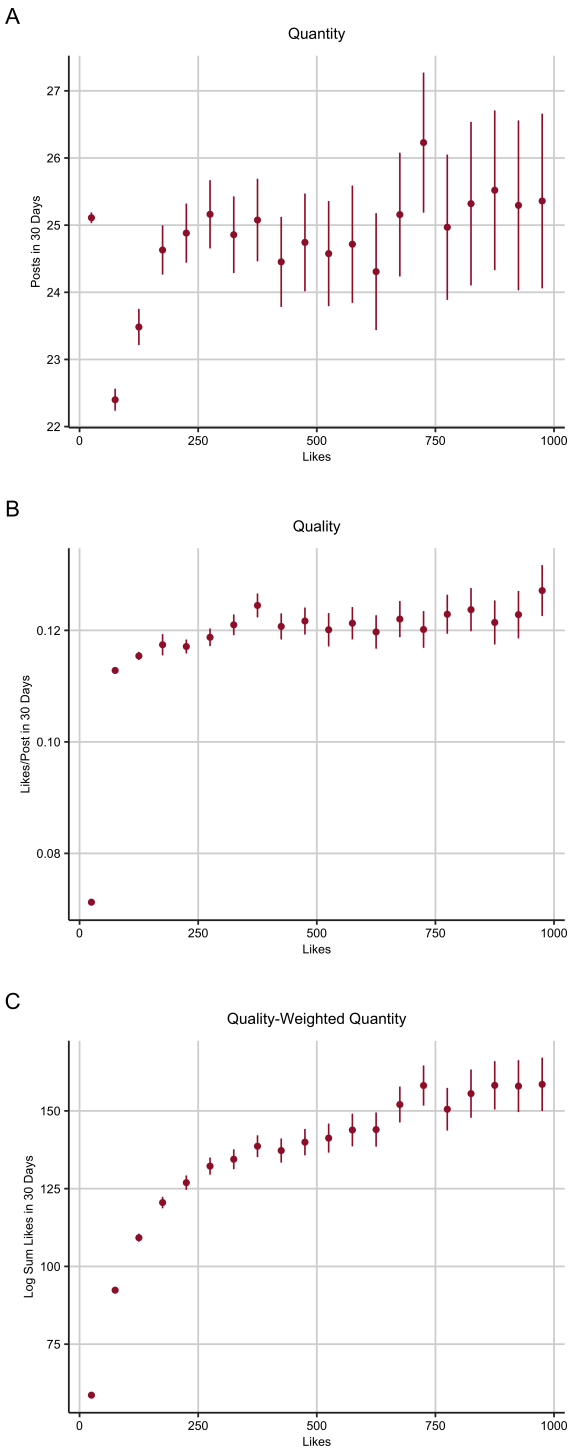
Notes: This difference-in-differences design compares how the production of Reddit posts evolves around viral and randomly selected posts. Posts are viral if they surpass the 80^{th} percentile of the upvotes distribution. The outcome variable is a quality-weighted measure of output: $\sum log(\max(\text{upvotes} + 1, 1))$. In Plot A, each point represents a 1 day bin. Event time 0 is the day that the viral or random post was created, and is excluded from the graph. Output increases by 0.21 units per day in the 30 days following going viral relative to the random baseline, which is 373% increase over the pre-period mean of 0.06 units per day. Plot B estimates the attention labor supply curve by graphing heterogeneity in the treatment effect by the degree of virality. Each point is the output of the difference-in-differences design estimated on the subset of posts that go viral within a two-percentile band of the upvotes distribution. Posts in the first viral point received between 21-26 upvotes (80^{th}-82^{nd} percentile), while posts in the tenth viral point received more than 531 upvotes (98^{th}-100^{th} percentile). The placebo point is the treatment effect of posting a non-viral post, estimated using a difference-in-differences design around random non-viral post. Bars represent 95% confidence intervals.

Figure 3: The Effect of Virality: Quantity vs. Quality



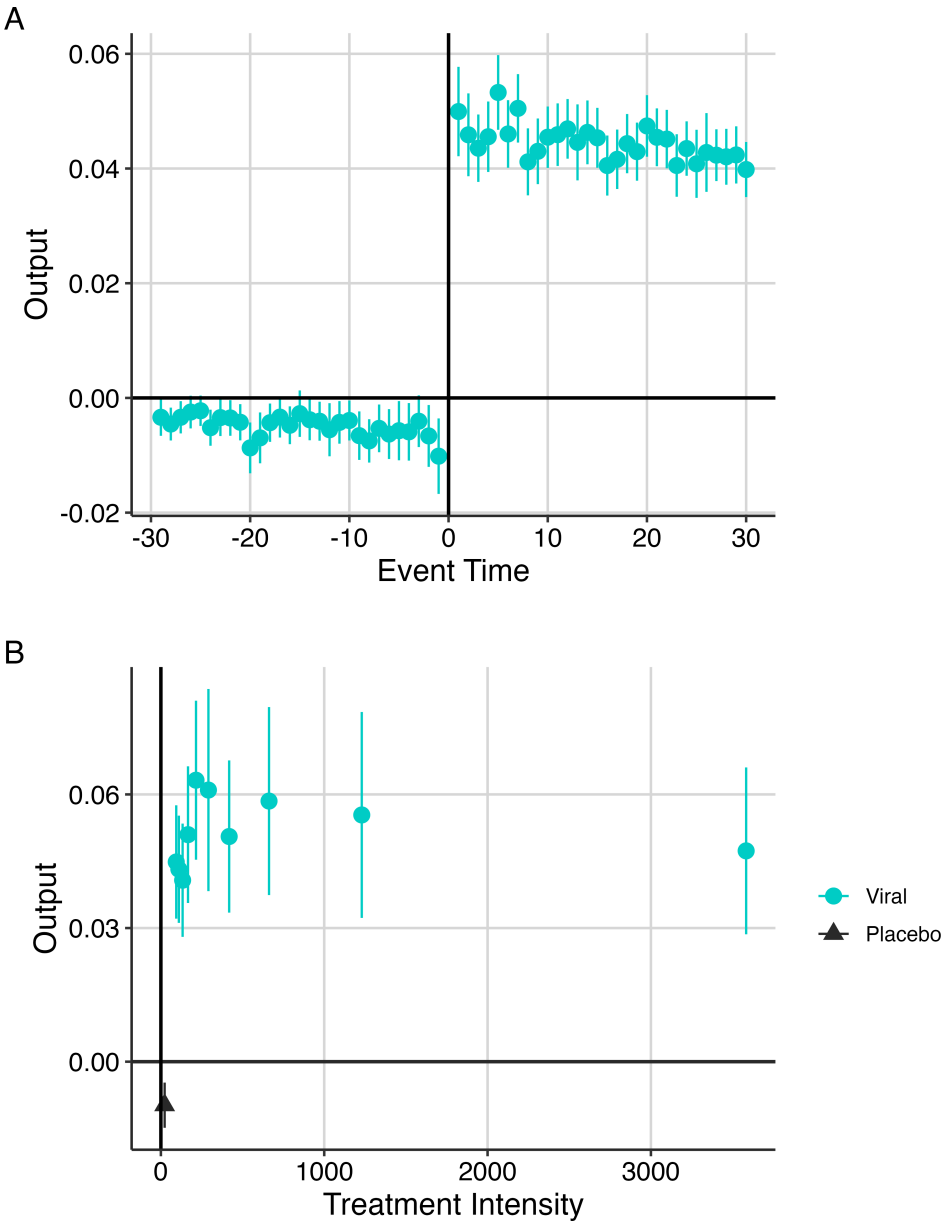
Notes: This figure repeats the difference-in-differences design of Figure 2 for two alternative outcomes. Plots A and B consider the number of posts per day, an interpretable measure of the quantity of output. Viral producers post 0.068 more posts per day, which is 183% of the baseline of 0.037 posts per day. Plots C and D analyze effects on the mean score conditional on posting, which is a measure of post quality. Going viral does not significantly change post quality.

Figure 4: The Effect of Virality: Quantity vs. Quality



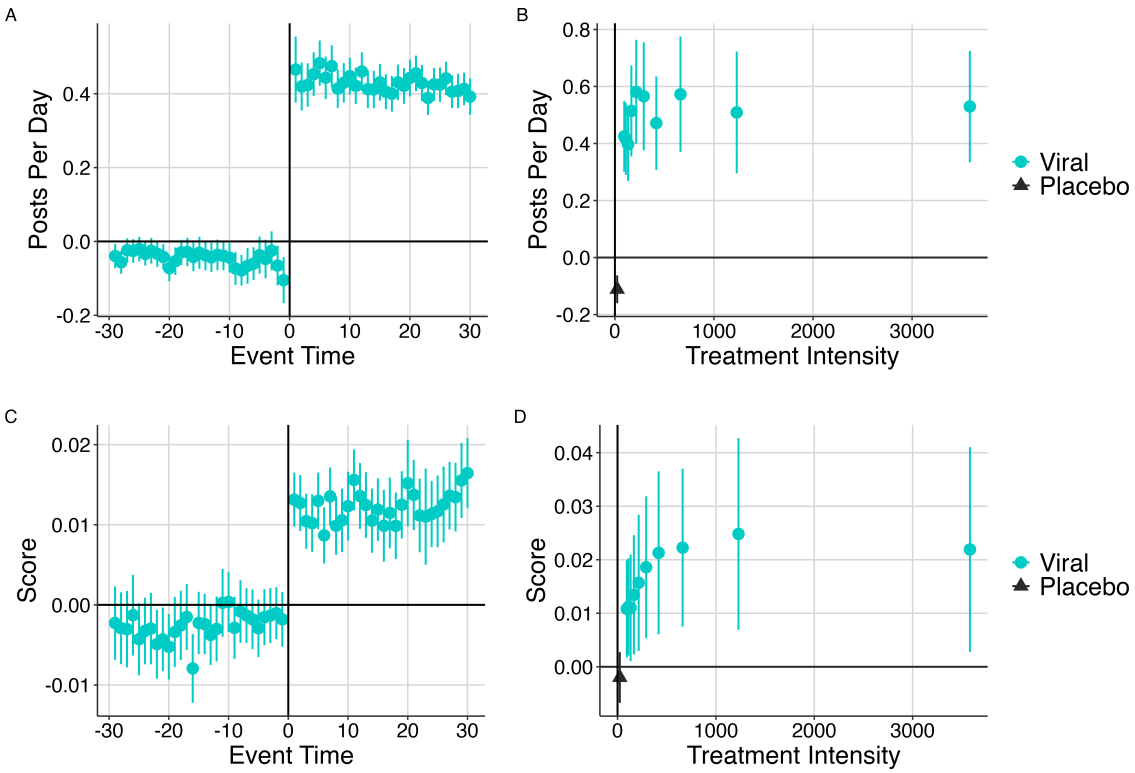
Notes: This figure graphs correlations between the likes that a TikTok post receives and the quality and quantity of content that the post's author produces over the next 30 days. Plot A depicts the number of posts made in the subsequent 30 days. Plot B depicts the quality of posts, measured in terms of likes/view. The outcome in Plot C is $\sum log(likes + 1)$ of posts produced in the next 30 days, which is a quality-weighted measure of output. Posts are grouped into 50-like bins. Bars represent 95% confidence intervals.

Figure 5: The Effect of Virality on Production on TikTok



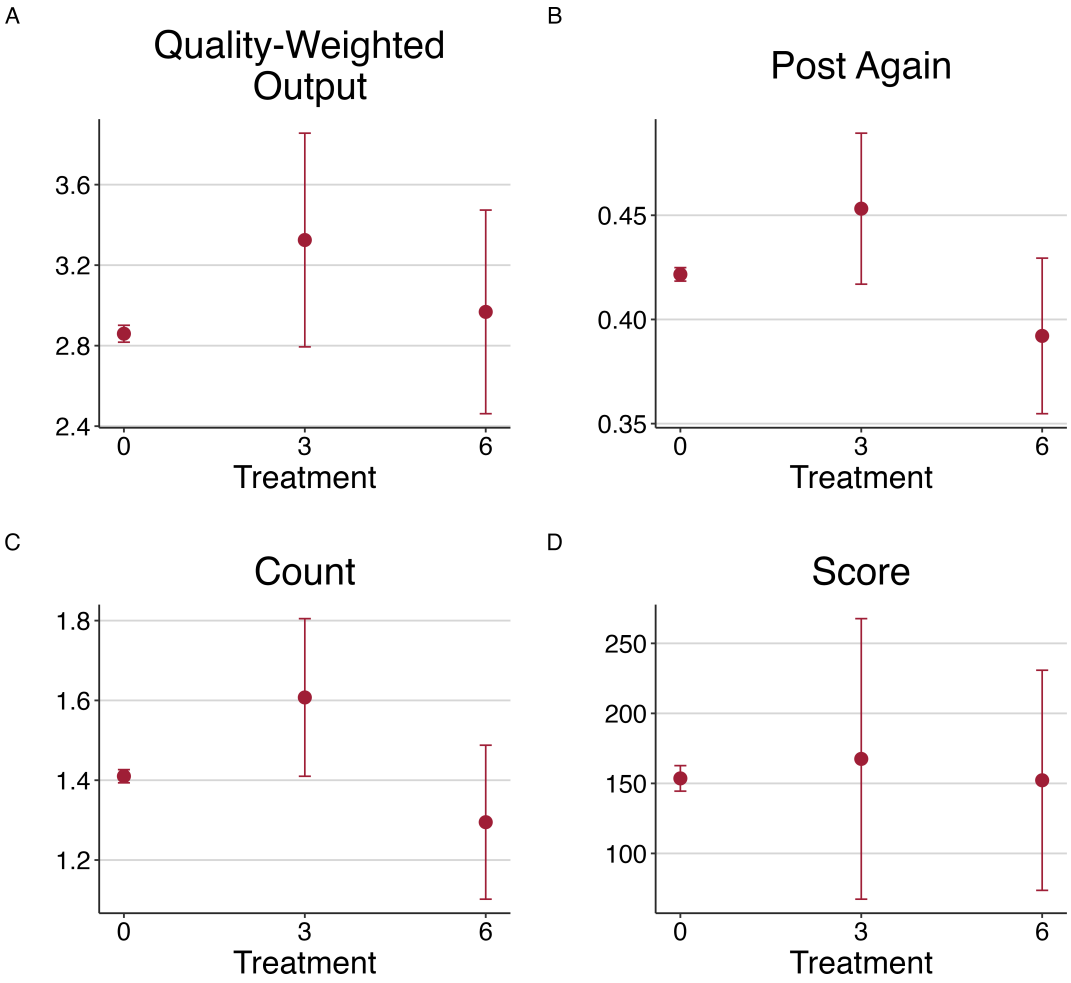
Notes: This figure replicates the difference-in-differences design of Figure 2 on TikTok. The outcome is a quality-weighted of output, where quality is measured by likes per view. Posts are viral if they surpass the 80^{th} percentile of the likes distribution. In Plot A, each point represents a 1 day bin. Event time 0 is the day that the viral or random TikTok is created, and is excluded from the graph. Output increases by 0.049 units per day in the 30 days following going viral TikTik relative to the random baseline, which is 279% increase over the pre-period rate of 0.017 units per day. Plot B graphs heterogeneity in the treatment effect by the degree of virality. Each point is the output of the difference-in-differences design estimated on the subset of posts that go viral within a two-percentile band of the upvotes distribution. Posts in the first viral point received between 94-110 likes (80^{th}-82^{nd} percentile), while posts in the tenth viral point received more than 3,583 upvotes (98^{th}-100^{th} percentile). The placebo point is the treatment effect of posting a non-viral post, estimated using a difference-in-differences design around random non-viral post. Bars represent 95% confidence intervals.

Figure 6: The Effect of Virality on TikTok: Quantity vs. Quality



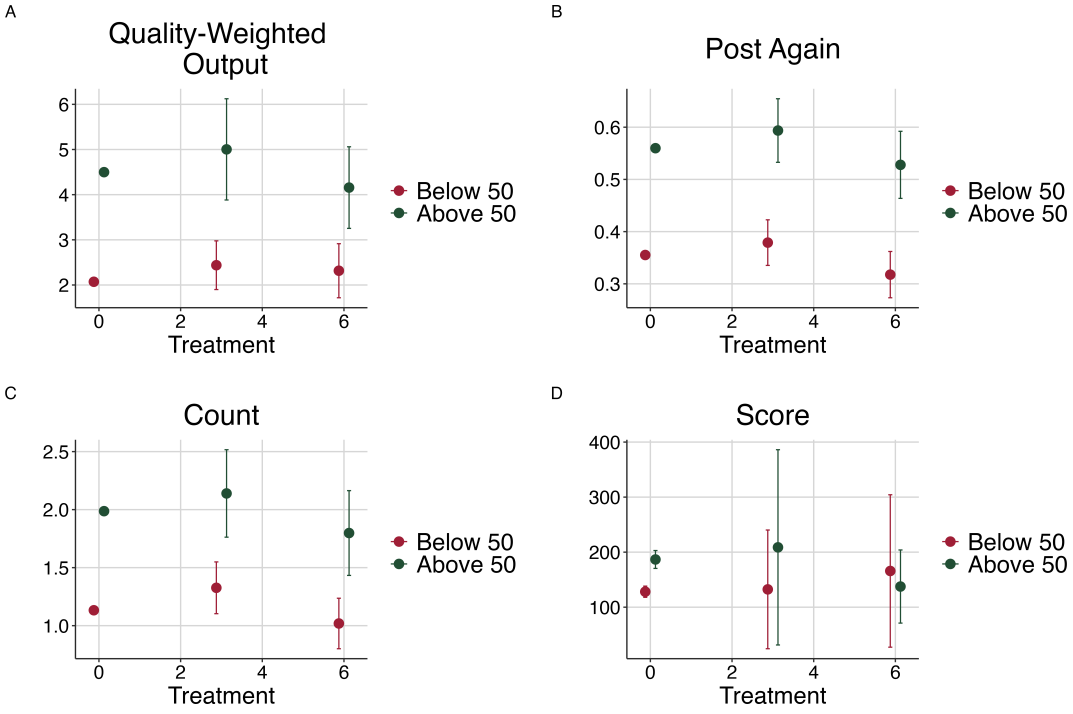
Notes: This figure repeats the difference-in-differences design of Figure 5 for two alternative outcomes. Plots A and B consider the number of TikToks per day, an interpretable measure of the quantity of output. Viral producers post 0.43 more TikToks per day, which is 190% of the baseline of 0.24 TikToks per day. Plots C and D analyze effects on the mean score conditional on posting, which is a measure of post quality. Going viral increases average post quality by 0.014 units which is 20% of the pre-period mean quality of 0.07 units.

Figure 7: The Effect of Attention on Production: Experimental Evidence



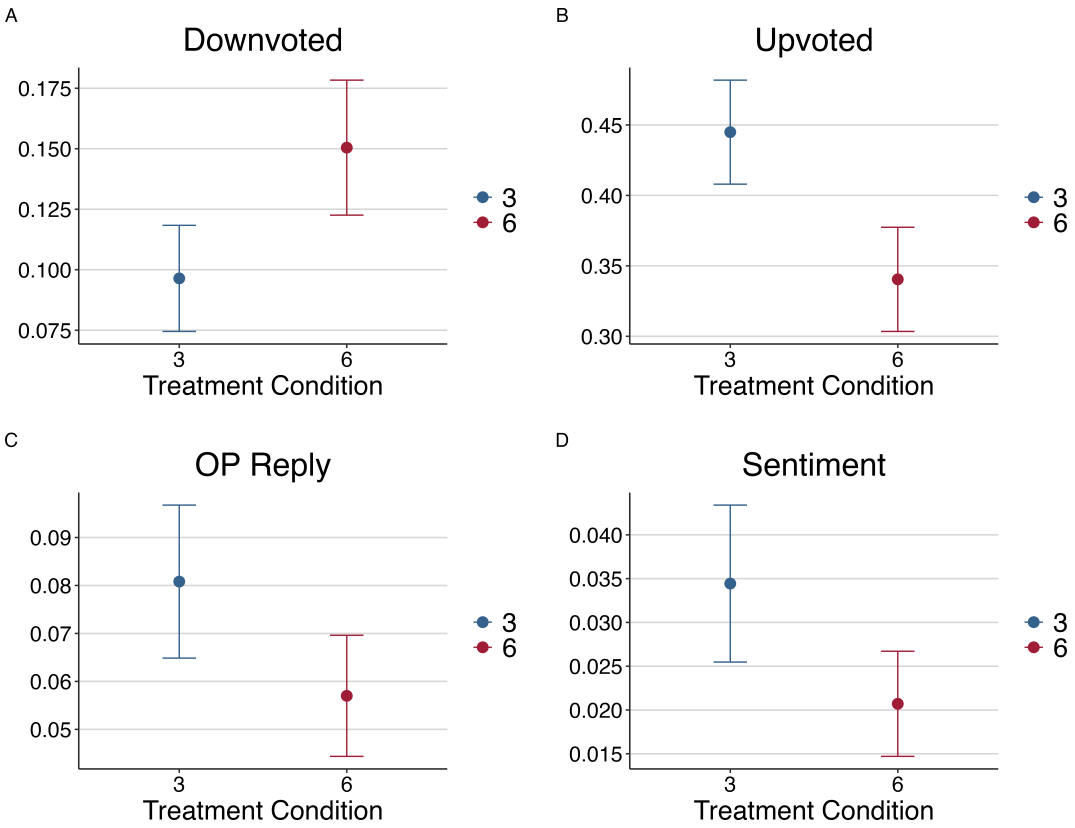
Notes: This figure plots all main outcomes in the experiment. There are four preregistered outcome variables. Panel A plots a quality-weighted measure of output, computed by taking the sum of the log of the score of posts produced. Panel B plots the the probability of posting again, a measure of the extensive margin. Panel C plots a count of posts, an interpretable measure of quantity. Finally, Panel D plots the mean score conditional on posting, a measure of quality. All outcomes are measured in the 7 days after treatment. The 3 comments treatment increases the the quality-weighted measure of output, the probability of posting again, and the count of posts over the next week, but has no effect on average score. I find null effects for the 6 comment treatment across all outcomes.

Figure 8: Heterogeneity by Poster Experience



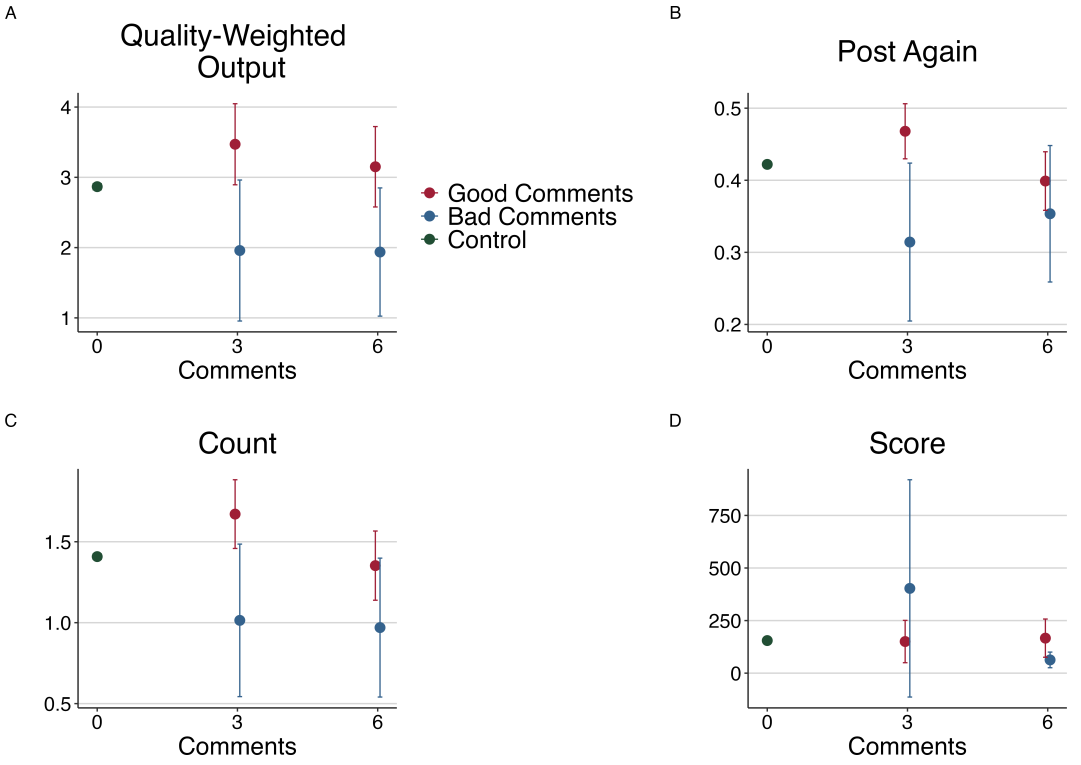
Notes: This figure plots a pre-registered split in the main treatment by user experience. Users are grouped by whether or not they have below 50 prior posts at the time of randomization. The point estimates of this heterogeneity split mirror the pooled results of the experiment for both groups, though the estimates are noisy and null effects cannot be rejected.

Figure 9: Heterogeneity in Comment Quality by Treatment Condition



Notes: This figure plots four measures of comment quality by treatment condition. Panel A shows that comments in the six comment treatment have an above average probability of being downvoted. This result of this split is analyzed in Figure 10. Panel B shows that comments in the three comment treatment have an above average probability of being upvoted. Panel C shows that the original poster is more likely to reply to comments in the three comments treatment. Panel D shows that the average sentiment of replies to the three comments treatment is higher. Sentiment is measured using the VADER sentiment model, and larger numbers reflect a more positive sentiment. Bars represent 95% confidence intervals.

Figure 10: Heterogeneity by Comment Quality



Notes: This figure plots heterogeneity in the four experimental outcomes, splitting the sample by a measure of average comment quality. Specifically, the sample is split on the average rate that a comment was downvoted, which is a sign that a comment was actively disliked. There is a large degree of heterogeneity by comment quality. Quality-weighted output, the count of posts, and the probability of posting again are all significantly larger for high quality comments when compared to low quality comments within the same treatment condition. This heterogeneity provides an explanation for the average null effect for the six comments treatment condition: comments in the six comments treatment are more likely to be downvoted, and downvoted comments have negative treatment effects.

6 Theoretical Appendix

6.1 Recasting the Model in terms of a Single Policy Parameter

Equation 2 represents the platform's problem in terms of four choice variables N_g, N_b, i_g, i_b. In this subsection, I will show that the problem can be represented in terms of a single choice variable β, where β will represent the percentage of available bad content that the platform chooses to show.

Expressing N_g^* and i_g^* in terms of S and D. The first thing to notice is that the platform always wants to show consumers all of the good content it has available. This means each piece of good content on the platform will be seen by all consumers on the platform. Formally,

Lemma 3. $N_g^* = S_g$. $i_g^* = D$.

Proof. In the platform's problem, the objective function $D(N_g, N_b)$ is strictly increasing in N_g. Since N_g increases the objective directly, the only reason not to choose the maximum feasible N_g is if increasing N_g tightens the constraints of the maximization problem. The two inequality constraints are $0 \leq N_g \leq S_g$ and $0 \leq N_b \leq S_b$. S_b and S_g are both strictly increasing functions of $S(i_g, i_b)$, which itself is strictly increasing in both of its arguments. Recall the expressions for i_g and i_b.

$$i_g = \frac{N_g}{S_g} D(N_g, N_b)$$
$$i_b = \frac{N_b}{S_b} D(N_g, N_b)$$

Since both of these expressions are increasing in N_g, we can see that both constraints are strictly loosening in N_g. Then, the optimal choice of N_g must be the maximum feasible choice of N_g since increasing N_g increases the objective function and loosens the constraints. So, we have that $N_g^* = S_g$. This result immediately simplifies the expression for i_g^*:

$$i_g^* = \frac{DN_g^*}{S_g} = D$$

Expressing N_b and i_b in terms of S, D and β. Because the platform's choice of N_g^*, i_g^* is trivial, the primary decision of interest in this model is how the platform handles bad content. It will be notationally convent to think about the amount of bad content shown N_b as fraction of the total amount of bad content available, S_b. Define

$$\beta := \frac{N_b}{S_b}$$

This definition allows us to express i_b in terms of β and D.

$$i_b = \frac{DN_b}{S_b} = \beta D$$

The parameter β is defined in and out of equilibrium, so the expression for N_b in terms of S_b and β holds in and out of equilibrium. The expression for i_b in terms of β, S, and D depends on implementing the equality constraint. Applying these new definitions to the equilibrium quantities i_b^* and N_b^* gives

$$\begin{aligned} N_b^* &= \beta S_b \\ i_b^* &= \beta D \end{aligned}$$

Expressing the Platform's Problem in terms of β. Taking stock of the above definitions and results, we can now to express the equilibrium parameters $N_g^*, N_b^*, i_g^*, i_b^*$ in terms of β as well as the supply and demand functions S and D.

Specifically, write demand in terms of β and supply as follows

$$\begin{aligned} D &:= D(N_g^*, N_b^*) \\ &= D(S_g, \beta S_b) \\ &= D(qS, \beta(1-q)S) \end{aligned}$$

Similarly, write supply in terms of demand and β as follows

$$\begin{aligned} S &:= S(i_g^*, i_b^*) \\ &= S(D, \beta D) \end{aligned}$$

Then, the platform's problem is

$$\begin{aligned} \max_{\beta} \quad & \Pi = D(qS, \beta(1-q)S) \\ \text{subject to} \quad & 0 \leq \beta \leq 1 \end{aligned} \tag{4}$$

Lemma 4. *If the supply of content is exogenously fixed, then platform should show no bad content. Let β_Π^* denote the platform's profit maximizing choice of β. Fix the supply of content $S(D, \beta D)$ to some level $\overline{S} > 0$. Then, $\beta_\Pi^* = 0$.*

Proof. Consider the firm's profit maximization problem for a fixed supply of content $\overline{S}$:

$$\begin{aligned} \max_{\beta} \quad & \Pi = D(q\overline{S}, \beta(1-q)\overline{S}) \\ \text{subject to} \quad & 0 \leq b \leq 1 \end{aligned}$$

We know that $D(N_g, N_b)$ is decreasing in its second argument. Once supply is fixed, β only appears in the expression for N_b. Then, to maximize Π, we should choose β to minimize $N_b = \beta(1-q)\overline{S}$. Since $(1-q)\overline{S}$ is a positive constant, this expression is minimized when $\beta = 0$.

6.2 The Platform's Profit Maximizing Strategy

Proposition 5. *If the producer attention utility function V is sufficiently concave and consumers value good content enough, then the platform shows consumers a positive percentage of bad content.*

- *More formally, suppose that $\frac{\partial D}{\partial N_g} > -\frac{1}{2}\frac{\partial D}{\partial N_b}$ when evaluated at D_0. For fixed values of $V(0)$ and $V(D_0)$, if $V'(0)$ is large enough and $V'(D_0)$ is small enough, then $\beta_C^* > 0$.*

If the producer attention utility function V is sufficiently concave, then the platform does not show consumers all bad content.

- *For fixed values of $V(0)$ and $V(D_1)$, if $V'(D_1)$ is small enough, then $\beta_C^* < 1$.*

Proof.

We are interested in whether $\beta_C^* > 0$. Consider the equilibrium where we exogenously fix $b = 0$. Call the equilibrium level of demand $D_0 := D(S_0, 0)$ where $S_0 := (D_0, 0)$.

Consider the profit equation

$$\Pi = D(qS, \beta(1-q)S)$$

Take the total derivative of this equation with respect to β.

$$\frac{d}{d\beta}[\Pi = D(qS, \beta(1-q)S)]$$
$$\frac{dD}{d\beta} = \frac{\partial D}{\partial \beta} + \frac{\partial D}{\partial S}\frac{\partial S}{\partial \beta}$$

The platform should choose $\beta_C^* > 0$ if the total derivative is positive at 0. This condition is

$$0 < \frac{\partial D}{\partial \beta}(0) + \frac{\partial D}{\partial S}(0)\frac{\partial S}{\partial \beta}(0)$$

Note that

$$\begin{aligned}\frac{\partial D}{\partial \beta}(0) &= \frac{\partial D(S, \beta S)}{\partial \beta}(0) \\ &= D^1(S_0)\frac{\partial S}{\partial \beta} + D^2(0)[S + \frac{\partial S}{\partial \beta}]\end{aligned}$$

where $D^1(S_0)$ denotes the derivative of D with respect to it's first argument, N_g, evaluated at the level S_0. Additionally, note that

$$\begin{aligned}\frac{\partial D}{\partial S}(0) &= \frac{\partial D(S, \beta S)}{\partial S}(0) \\ &= D^1(S_0) + \beta D^2(0) \\ &= D^1(S_0)\end{aligned}$$

So, the inequality condition is

$$\begin{aligned}
0 &< D^1(S_0)\frac{\partial S}{\partial \beta} + D^2(0)[S_0 + \frac{\partial S}{\partial \beta}] + D^1(S_0)\frac{\partial S}{\partial \beta}(0) \\
0 &< 2D^1(S_0)\frac{\partial S}{\partial \beta} + D^2(0)[S_0 + \frac{\partial S}{\partial \beta}] \\
-D^2(0)S_0 &< 2D^1(S_0)\frac{\partial S}{\partial \beta} + D^2(0)\frac{\partial S}{\partial \beta} \\
-D^2(0)S_0 &< [2D^1(S_0) + D^2(0)]\frac{\partial S}{\partial \beta}
\end{aligned}$$

By assumption $2D^1(S_0) + D^2(0) > 0$, so we can move this term to the other side without flipping the inequality.

$$\frac{-D^2(0)S_0}{2D^1(S_0) + D^2(0)} < \frac{\partial S}{\partial \beta}$$

where the left hand side of this inequality is a positive term that is constant with respect to V' holding fixed $V(0)$ and $V(D_0)$, which are terms that determine S_0 and D_0 in equilibrium when $b = 0$ exogenously.

Consider the derivative at $\frac{\partial S}{\partial \beta}(0)$. Recall that supply is given by

$$\begin{aligned}
S := S(i_g, i_b) &= \int \mathbb{I}\{qV(i_g) + (1-q)V(i_b) > \delta_j\}k(\delta_j)d\delta_j \\
&= \int \mathbb{I}\{qV(D) + (1-q)V(\beta D) > \delta_j\}k(\delta_j)d\delta_j \\
&= K(qV(D) + (1-q)V(\beta D))
\end{aligned}$$

Taking the first derivative, we have

$$\frac{\partial K(\mathbb{E}[V])}{\partial \beta} = k(\mathbb{E}[V])\frac{\partial \mathbb{E}[V]}{\partial \beta}$$

where

$$\begin{aligned}
\frac{\partial \mathbb{E}[V]}{\partial \beta} &= \frac{\partial}{\partial \beta}[qV(D(\beta)) + (1-q)V(\beta D(\beta))] \\
&= qV'(D(\beta))D'(\beta) + (1-q)V'(\beta D(\beta))[D(\beta) + \beta D'(\beta)]
\end{aligned}$$

Evaluating this derivative at $b = 0$, we have

$$\frac{\partial \mathbb{E}[V]}{\partial \beta}(0) = qV'(D(0))D'(0) + (1-q)V'(0)D(0)$$

Then, the derivative of supply with respect to β evaluated at $b = 0$ is

$$\frac{\partial K(\mathbb{E}[V])}{\partial \beta}(0) = k(qV(D) + (1-q)V(0))[qV'(D(0))D'(0) + (1-q)V'(0)D(0)]$$

Rewriting our expression in D_0 notation,

$$\begin{aligned}\frac{\partial S}{\partial \beta}(0) &= k(qV(D_0) + (1-q)V(0))[qV'(D_0)D_0' + (1-q)V'(0)D_0] \\ &= k(qV(D_0) + (1-q)V(0))[qV'(D_0)D^1(S_0)[D^1(S_0)\frac{\partial S}{\partial \beta} + D^2(0)[S + \frac{\partial S}{\partial \beta}]] + (1-q)V'(0)D_0]\end{aligned}$$

since

$$\begin{aligned}\frac{\partial D}{\partial \beta}(0) &= \frac{\partial D(S, \beta S)}{\partial \beta}(0) \\ &= D^1(S_0)\frac{\partial S}{\partial \beta} + D^2(0)[S + \frac{\partial S}{\partial \beta}]\end{aligned}$$

Isolating $\frac{\partial S}{\partial \beta}(0)$,

$$\begin{aligned}\frac{\partial S}{\partial \beta}(0) &= k(\cdot)[qV'(D_0)[D^1(S_0)\frac{\partial S}{\partial \beta} + D^2(0)[S + \frac{\partial S}{\partial \beta}]] + (1-q)V'(0)D_0] \\ \frac{\partial S}{\partial \beta}(0) &= \frac{\partial S}{\partial \beta}(0)[k(\cdot)qV'(D_0)D^1(S_0) + k(\cdot)D^2(0)] + k(\cdot)D^2(0)S + k(\cdot)(1-q)V'(0)D_0 \\ \frac{\partial S}{\partial \beta}(0) &= \frac{k(\cdot)D^2(0)S + k(\cdot)(1-q)V'(0)D_0}{1 - [k(\cdot)qV'(D_0)D^1(S_0) + k(\cdot)D^2(0)]}\end{aligned}$$

If $k(\cdot)qV'(D_0)D^1(S_0) < 1$, then the denominator is positive. So, if $V'(D_0)$ small enough, this denominator is positive.

The numerator $k(\cdot)D^2(0)S + k(\cdot)(1-q)V'(0)D_0$ can be made arbitrarily large for $V'(0)$ large enough.

So, for $V'(D_0)$ small enough and $V'(0)$ large enough, the derivative $\frac{\partial S}{\partial \beta}(0)$ can be made arbitrarily large. Under the conditions described above, for $\frac{\partial S}{\partial \beta}(0)$ large enough, the total derivative $\frac{dD}{d\beta}(0) > 0$.

Claim 2

Now, consider the total derivative expression at $\beta = 1$

$$\frac{dD}{d\beta}(1) = \frac{\partial D}{\partial \beta}(1) + \frac{\partial D}{\partial S}\frac{\partial S}{\partial \beta}(1)$$

The platform should choose $\beta < 1$ if the total derivative is negative at $\beta = 1$. This condition is

$$0 > \frac{\partial D}{\partial \beta}(1) + \frac{\partial D}{\partial S}(1)\frac{\partial S}{\partial \beta}(1)$$

Note that

$$\begin{aligned}\frac{\partial D}{\partial \beta}(1) &= \frac{\partial D(S, \beta S)}{\partial \beta}(1) \\ &= D^1(S_1)\frac{\partial S}{\partial \beta} + D^2(S_1)[S_1 + \frac{\partial S}{\partial \beta}]\end{aligned}$$

where $D^1(S_1)$ denotes the derivative of D with respect to it's first argument and $D^2(S_1)$ denotes the derivative of D with respect to it's second argument. Additionally, note that

$$\begin{aligned}\frac{\partial D}{\partial S}(1) &= \frac{\partial D(S, \beta S)}{\partial S}(1) \\ &= D^1(S_1) + \beta D^2(S_1) \\ &= D^1(S_1) + D^2(S_1)\end{aligned}$$

Substituting in these expressions, the inequality condition is

$$\begin{aligned}0 &> D^1(S_1)\frac{\partial S}{\partial \beta}(1) + D^2(S_1)[S_1 + \frac{\partial S}{\partial \beta}(1)] + [D^1(S_1) + D^2(S_1)]\frac{\partial S}{\partial \beta}(1) \\ 0 &> \frac{\partial S}{\partial \beta}(1)[D^1(S_1) + D^2(S_1) + D^1(S_1) + D^2(S_1)] + D^2(S_1)S_1 \\ -D^2(S_1)S_1 &> 2\frac{\partial S}{\partial \beta}(1)[D^1(S_1) + D^2(S_1)]\end{aligned}$$

By assumption, $D^1(S_1) + D^2(S_1) > 0$ (from assuming $\frac{\partial D}{\partial S} > 0$), so we can rearrange terms.

$$\frac{-D^2(S_1)S_1}{2[D^1(S_1) + D^2(S_1)]} > \frac{\partial S}{\partial \beta}(1)$$

Now, recall the expression for $\frac{\partial S}{\partial \beta}$. Taking the first derivative, we have

$$\frac{\partial K(\mathbb{E}[V])}{\partial \beta} = k(\mathbb{E}[V])\frac{\partial \mathbb{E}[V]}{\partial \beta}$$

where

$$\begin{aligned}\frac{\partial \mathbb{E}[V]}{\partial \beta} &= \frac{\partial}{\partial \beta}[qV(D(\beta)) + (1-q)V(\beta D(\beta))] \\ &= qV'(D(\beta))D'(\beta) + (1-q)V'(\beta D(\beta))[D(\beta) + \beta D'(\beta)]\end{aligned}$$

Evaluating this derivative at $\beta = 1$, we have

$$\frac{\partial \mathbb{E}[V]}{\partial \beta}(1) = qV'(D(1))D'(1) + (1-q)V'(D(1))[D(1) + D'(1)]$$

Then, the derivative of supply with respect to β evaluated at $\beta = 1$ is

$$\begin{aligned}\frac{\partial S}{\partial \beta}(1) &= k(V(D_1))[qV'(D_1)D_1' + (1-q)V'(D_1)[D_1 + D_1']] \\ &= k(V(D_1))[V'(D_1)D_1' + (1-q)V'(D_1)D_1]\end{aligned}$$

Recall that $D_1' = D^1(S_1)\frac{\partial S}{\partial \beta} + D^2(S_1)[S_1 + \frac{\partial S}{\partial \beta}]$. Then,

$$\begin{aligned}\frac{\partial S}{\partial \beta}(1) &= k(V(D_1))[qV'(D_1)D_1' + (1-q)V'(D_1)[D_1 + D_1']] \\ &= k(\cdot)[V'(D_1)D^1(S_1)\frac{\partial S}{\partial \beta}(1) + D^2(S_1)[S_1 + \frac{\partial S}{\partial \beta}(1)] + (1-q)V'(D_1)D_1] \\ &= \frac{\partial S}{\partial \beta}(1)[k(\cdot)V'(D_1)D^1(S_1) + k(\cdot)D^2(S_1)] + k(\cdot)D^2(S_1)S_1 + k(\cdot)(1-q)V'(D_1)D_1 \\ &= \frac{k(\cdot)D^2(S_1)S_1 + k(\cdot)(1-q)V'(D_1)D_1}{1 - [k(\cdot)V'(D_1)D^1(S_1) + k(\cdot)D^2(S_1)]}\end{aligned}$$

In the denominator, the only negative term is $k(\cdot)V'(D_1)D^1(S_1)$, so if $V'(D_1)$, the denominator will be positive.

In the numberator, $k(\cdot)D^2(S_1)S_1$ is negative and $k(\cdot)(1-q)V'(D_1)D_1$ can be made arbitrarily small as $V'(D_1)$ grows small.

Then, $\frac{\partial S}{\partial \beta}(1)$ can be made arbitrarily close to zero (or negative) as $V'(D_1)$ grows small. In particular, chose $V'(D_1)$ such that the inequality condition derived above holds (which is possible since the left hand side is a positive constant with respect to $V'(D_1)$).

"Sufficiently Concave" To justify the phrasing of the proposition that these results hold if the function V is "sufficiently concave", consider the following Taylor Expansion.

$$V'(D_0) \approx V'(0) + D_0 V''(0)$$

If $V''(0)$ is negative enough (i.e. V is very concave), then $V'(D_0)$ can be made arbitrarily small (by assumption, it cannot be negative, as V is assumed to be increasing). Reversing this expression, write

$$V'(0) \approx V'(D) - D_0 V''(D)$$

Again, for $V''(D)$ negative enough, $V'(0)$ can be made arbitrarily large.

In practice, for these expressions to hold V must be in a class where the Taylor Expansions are 'close enough', but the point of these expressions is just to capture the idea that having a very sharp positive derivative at 0 and a very flat derivative at some positive point D_0 or D_1 can be understood through the lens of concavity.

6.3 The Social Planner's Welfare Maximizing Strategy

Proposition 6. *The percentage of bad content which maximizes each of the welfare objectives is weakly ordered*

$$\beta_P^* = \beta_{PW}^* \geq \beta_{SW}^* \geq \beta_{CW}^* = \beta_C^*$$

Moreover, the percentage of bad content which maximizes impressions is weakly larger than the percentage which maximizes the number of consumers on the platform:

$$\beta_I^* \geq \beta_C^*.$$

Proof. First I will show that maximizing consumer welfare is equivalent to maximizing the number of consumers on the platform, and maximizing producer welfare is equivalenet to maximizing the number of producers on the platform.

Recall that

$$W_C = \int \max\{U(N_g, N_b) - \epsilon, 0\} l(\epsilon) d\epsilon$$

The social planner wants to choose β to maximize W_C. The integrand is weakly increasing in $U(N_g, N_b)$, and β does not enter the problem elsewhere, so the social planner's problem is equivalent to choosing β to maximize $U(N_g, N_b)$.

Now consider the platform's profit maximizing problem. The platform wants to choose β to maximize

$$\Pi = D(\beta) = \int \mathbb{I}\{U(N_g, N_b) > \epsilon\} l(\epsilon) d\epsilon$$

This integrand is also weakly increasing in $U(N_g, N_b)$, so the platform's problem is also equivalent to choosing β to maximize $U(N_g, N_b)$. Then, $\beta_{CW}^* = \beta_C^*$.

Next, consider the content producer welfare function.

$$W_P = \int \max\{qV(D(\beta)) + (1-q)V(\beta D(\beta)) - \delta, 0\} k(\delta) d\delta$$

The social planner wants to choose β to maximize W_P. The integrand is weakly increasing in $qV(D(\beta)) + (1-q)V(\beta D(\beta))$, and β does not enter the problem elsewhere, so the social planner's problem is equivalent to choosing β to maximize $qV(D(\beta)) + (1-q)V(\beta D(\beta))$.

Compare this to the supply function.

$$S := S(i_g, i_b) = \int \mathbb{I}\{qV(D(\beta)) + (1-q)V(\beta D(\beta)) > \delta\} k(\delta) d\delta$$

Again, this function is weakly increasing in $qV(D(\beta)) + (1-q)V(\beta D(\beta))$, and β does not enter

the problem elsewhere, so the problem is equivalent to choosing β to maximize $qV(D(\beta)) + (1-q)V(\beta D(\beta))$. Then, $\beta^*_{PW} = \beta^*_P$.

Next, I will show that the optimal policy to maximize the three welfare objects are weakly ordered. The goal is to show that $\beta^*_{PW} \geq \beta^*_{SW} \geq \beta^*_{CW}$. I will start by showing that $\beta^*_{PW} \geq \beta^*_{CW}$.

For a contradiction, assume that $\beta^*_{PW} < \beta^*_{CW}$.

In order to choose β to maximize producer welfare, we want to choose β that maximizes $qV(D(\beta)) + (1-q)V(\beta D(\beta))$. We know that β^*_{CW} maximizes $D(\beta)$, by definition.

Then, it must be the case that $V(D(\beta^*_{CW})) > V(D(\beta^*_{PW}))$ because V is an increasing function.

Moreover, it must be the case that $V(\beta^*_{CW} D(\beta^*_{CW})) > V(\beta^*_{PW} D(\beta^*_{PW}))$ because we have assumed that $\beta^*_{PW} < \beta^*_{CW}$ and we know that $D(\beta)$ is maximized at β^*_C.

But, this means that $qV(D(\beta^*_{CW}))+(1-q)V(\beta^*_{CW} D(\beta^*_{CW})) > qV(D(\beta^*_{PW}))+(1-q)V(\beta^*_{PW} D(\beta^*_{PW}))$, so β^*_{CW} provides higher producer welfare than β^*_{PW}. This contradicts the definition of β^*_{PW}, because we have identified a value of $\beta \neq \beta^*_{PW}$ that generates more producer welfare, so β^*_{PW} is not optimal. Then, $\beta^*_{PW} \geq \beta^*_{CW}$.

Next, consider the relationship between β^*_{CW} and β^*_{SW}. Recall that social welfare is assumed to be a linear combination of producer and consumer surplus. Consider a social planner trying to maximize W_S with $\alpha \in (0,1)$.

$$\begin{aligned} W_S &= \alpha W_C + (1-\alpha) W_P \\ &= \alpha \left(\int \max\{U(N_g, N_b), 0\} l(\epsilon) d\epsilon \right) \\ &+ (1-\alpha) \left(\int \max\{qV(i_g) + (1-q)V(i_b) - \delta, 0\} k(\delta) d\delta \right) \end{aligned}$$

For a contradiction, suppose that $\beta^*_{SW} < \beta^*_{CW}$. Note that, by definition, β^*_{CW} maximizes consumer welfare. Additionally, we have already shown that choosing $\beta < \beta^*_{CW}$ provides strictly less welfare to producers. Then, consider selecting $\beta = \beta^*_{CW}$. This increases both consumer and producer welfare relative to β^*_{SW}. But, this is a contradiction with the definition of β^*_{SW}, because it means that there exists a $\beta \neq \beta_{SW}$ that provides strictly larger social welfare. Then, we have that $\beta^*_{SW} \geq \beta^*_{CW}$.

Finally, consider the relationship between β^*_{PW} and β^*_{SW}. For a contradiction, suppose that $\beta^*_{SW} > \beta^*_{PW}$. By definition, β^*_{PW} maximizes producer welfare.

Now, consider consumer welfare, which depends on maximizing $U(N_g, N_b) = U(S, \beta S)$. By assumption, this function is increasing in its first argument, and decreasing in its second argument, because consumers like good content and dislike bad content. We're interested in the comparison of consumer welfare at two points β^*_{PW} and β^*_{SW}. Since maximizing producer welfare maximizes the number of producers on the platform, it must be the case that $S_{PW} > S_{SW}$.

Now, by assumption, $\frac{\partial D}{\partial S} > 0$, so we know that $D(S_{PW}, \beta S_{PW}) > D(S_{SW}, \beta S_{SW})$ for a fixed

β. In particular, consider β^*_{PW}, so we have

$$D(S_{PW}, \beta^*_{PW} S_{PW}) > D(S_{SW}, \beta^*_{PW} S_{SW})$$

Moreover, since demand is decreasing in its second argument, it must the case that

$$D(S_{SW}, \beta^*_{PW} S_{SW}) > D(S_{SW}, \beta^*_{SW} S_{SW})$$

since we have assumed that $\beta^*_{SW} > \beta^*_{PW}$. Then,

$$D(S_{PW}, \beta^*_{PW} S_{PW}) > D(S_{SW}, \beta^*_{SW} S_{SW})$$

This means that choosing β^*_{PW} provides higher consumer welfare than choosing β^*_{SW}. But, if β^*_{PW} provides higher consumer welfare and higher producer welfare, then we have found $\beta \neq \beta^*_{SW}$ that provides higher social welfare, which contradicts the definition of β^*_{SW}. So, we have that $\beta^*_{SW} \leq \beta^*_{PW}$.

Finally, consider maximizing the number of impressions on the platform. This is given by

$$\begin{aligned} \Pi &= D(N_g, N_b)(N_g + N_b) \\ &= D(S, \beta S)(S + \beta S) \end{aligned}$$

For a contradiciton, assume that $\beta^*_I < \beta^*_C$. By definition, β^*_C maximizes demand. Moreover, we have already shown that $\forall b < \beta^*_C$, $S(b) < S(\beta^*_C)$. But, this means that every term in the views profit function (D, S, β) is larger for β^*_C than for $\beta^*_I < \beta^*_C$, so we have identified a $\beta \neq \beta^*_I$ that generates a larger number of impressions. This contradicticts the definition of β^*_I, so it must be that $\beta^*_I \geq \beta^*_C$.

6.4 Multiple Consumer Types

Up until this point, I have considered one type of consumer who decides whether or not to join the platform. If this consumer joins the platform, then they "consume the platform" in the sense that they contribute one impression i for every piece of content on the platform ($N_g + N_b$). While this assumption can be relaxed so that consumers views a fixed percentage of the platform, one reasonable objection is that many people consume a negligible fraction of content relative to the size of the platform. In this section, I will extend the model to account for an alternative type of consumer who consumes a fixed number of impressions M. I will refer to the initial type of consumer as a "heavy consumer" and call this new type of consumer a "light consumer."

Because M is assumed to be small in proportion to the supply of content, the platform has complete control over the allocation of M towards good and bad content. Define f as the fraction of a light consumer's impressions that go to good content.

Demand for light consumers $D_L(f, S)$ can depend on f as well as the total amount of content

on the platform S. Intuitively, the reason that supply will still enter the demand function for light consumers who do not consume the whole platform is if there is an unmodeled horizontal differentiation in content, so that having more content implies having more types of content which attract different kinds of casual users.

Equilibrium (Heavy + Light). The market clearing equations must be rewritten to account for the inclusion of light consumers. An equilibrium in this model is a tuple $N_g^*, N_b^*, f^*, i_g^*, i_b^*$ such that:

1. The market for impressions of good content clears.

$$\underbrace{qS(i_g^*, i_b^*)}_{\substack{\text{Supply of} \\ \text{Good Content}}} \times \underbrace{i_g^*}_{\substack{\text{Impressions per} \\ \text{Good Content}}} = \underbrace{D_L(f, S)}_{\text{Light Consumer Demand}} \times \underbrace{fM}_{\substack{\text{Good Impressions per} \\ \text{Light Consumer}}} + \underbrace{D_H(N_g^*, N_b^*)}_{\text{Heavy Consumer Demand}} \times \underbrace{N_g^*}_{\substack{\text{Good Impressions per} \\ \text{Heavy Consumer}}}$$

2. The market for impressions of bad content clears.

$$\underbrace{(1-q)S(i_g^*, i_b^*)}_{\substack{\text{Supply of} \\ \text{Bad Content}}} \times \underbrace{i_b^*}_{\substack{\text{Impressions per} \\ \text{Bad Content}}} = \underbrace{D_L(f, S)}_{\text{Light Consumer Demand}} \times \underbrace{(1-f)M}_{\substack{\text{Bad Impressions per} \\ \text{Light Consumer}}} + \underbrace{D_H(N_g^*, N_b^*)}_{\text{Heavy Consumer Demand}} \times \underbrace{N_b^*}_{\substack{\text{Bad Impressions per} \\ \text{Heavy Consumer}}}$$

3. The composition of content shown on the platform is feasible.

$$N_g^* \leq S_g^* = qS^*(i_g^*, i_b^*)$$
$$N_b^* \leq S_b^* = (1-q)S^*(i_g^*, i_b^*)$$

Platform's Problem (Light Consumers). The platform's problem is

$$\begin{aligned} \max_{N_g, N_b, f, i_b, i_g} \quad & \Pi = D_H(N_g, N_b) + D_L(f, S) \\ \text{subject to} \quad & N_g \leq S_g(i_g, i_b) \\ & N_b \leq S_b(i_g, i_b) \\ & qS(i_g, i_b) \times i_g = fD_L(f, S)M + D_H(N_g, N_b) \times N_g \\ & (1-q)S(i_g, i_b) \times i_b = (1-f)D_L(f, S)M + D_H(N_g, N_b) \times N_\beta \end{aligned} \tag{5}$$

Now I will consider one case where the demand from light consumers increases when they are shown more good content. In particular, suppose that $D_L(S, f) = fS$.

Expressions for i_g and i_b Use the equality constraints to get an expression for i_g.

$$\begin{aligned} qS(i_g, i_b) \times i_g &= f^2 MS + D_H(N_g, N_b) \times N_g \\ i_g &= \frac{f^2 MS + D_H(N_g, N_b) \times N_g}{qS(i_g, i_b)} \\ i_g &= \frac{f^2 M + D_H(N_g, N_b)}{q} \end{aligned}$$

Similarly, use the other equality constraint to get an expression for i_b.

$$\begin{aligned} (1-q)S(i_g, i_b) \times i_b &= (1-f)fMS + D_H(N_g, N_b) \times N_b \\ i_b &= \frac{(1-f)fMS + D_H(N_g, N_b) \times N_b}{(1-q)S(i_g^*, i_b^*)} \\ i_b &= \frac{(1-f)fM + \beta D_H(N_g, N_b)}{(1-q)} \end{aligned}$$

Recasting this problem in β notation, the platform's problem is:

$$\begin{aligned} \max_{\beta, f} \quad & \Pi = D_H(S, \beta S) + fS \\ \text{such that} \quad & 0 \leq b \leq 1 \\ & 0 \leq f \leq 1 \\ & i_g = \frac{f^2 M + D_H(N_g, N_b)}{q} \\ & i_b = \frac{(1-f)fM + \beta D_H(N_g, N_b)}{(1-q)} \end{aligned}$$

Lemma 7. *If the producer attention utility function V is linear, then the platform should only show light consumers good content. Formally, if $V(i) = \beta i + \zeta$ with $\beta > 0$, then $f^* = 1$.*

Proof. Take the derivative $\frac{\partial \Pi}{\partial f}$.

$$\begin{aligned} \frac{\partial \Pi}{\partial f} &= \frac{\partial D_H}{\partial S}\frac{\partial S}{\partial f} + S + f\frac{\partial S}{\partial f} \\ &= \frac{\partial S}{\partial f}(\frac{\partial D_H}{\partial S} + f) + S \end{aligned}$$

By assumption, $\frac{\partial D_H}{\partial S} > 0$, $S > 0$ and $f \geq 0$. Then, if $\frac{\partial S}{\partial f} > 0$, $\frac{\partial \Pi}{\partial f}$ is always positive, so profit will be maximized at $f = 1$.

We want to know the sign of $\frac{\partial S}{\partial f}$. Since $V(i) = \beta i + \zeta$,

$$\begin{aligned} S &= K(qV(\frac{f^2M + D_H(N_g, N_b)}{q}) + (1-q)V(\frac{(1-f)fM + \beta D_H(N_g, N_b)}{(1-q)})) \\ &= K(\beta f^2 M + \beta D_H(N_g, N_b) + \zeta + \beta(1-f)fM + \beta\beta D_H(N_g, N_b) + \zeta) \\ &= K(\beta f M + D_H(N_g, N_b) + \beta D_H(N_g, N_b) + 2\zeta) \end{aligned}$$

The derivative of the inner term with respect to f is βM which is positive since $\beta, M > 0$. Then, supply is increasing in f, so profit is increasing in f, so the platform should choose the maximum feasible f, $f^* = 1$.

Lemma 8. *If the producers attention utility function V is concave and $q < 0.5$, then $f^* < 1$.*

Proof. Recall that

$$\begin{aligned} \Pi &= D_H + fS \\ \frac{\partial \Pi}{\partial f} &= \frac{\partial D_H}{\partial S}\frac{\partial S}{\partial f} + S + f\frac{\partial S}{\partial f} \\ &= \frac{\partial S}{\partial f}(\frac{\partial D_H}{\partial S} + f) + S \end{aligned}$$

We are interested in finding a condition for when this derivative will be negative when $f = 1$.

$$\begin{aligned} 0 &> (\frac{\partial D_H}{\partial S} + f)\frac{\partial S}{\partial f} + S(i_g, i_b) \\ -(\frac{\partial D_H}{\partial S} + f)\frac{\partial S}{\partial f} &> S(i_g, i_b) \\ -\frac{\partial S}{\partial f} &> \frac{S(i_g, i_b)}{\frac{\partial D_H}{\partial S} + f} \\ -\frac{\partial S}{\partial f} &> \frac{S(i_g, i_b)}{\frac{\partial D_H}{\partial S} + 1} \end{aligned}$$

In order for this condition to hold, it must be that $\frac{\partial S}{\partial f}$ is negative, since the right hand side of the inequality is the ratio of two positive terms. Recall the expression for supply in this model:

$$S = K(qV(\frac{fM}{q} + \frac{D_H(N_g, N_b)}{q}) + (1-q)V(\frac{(1-f)M}{(1-q)} + \frac{\beta D_H(N_g, N_b)}{(1-q)})$$

First, consider how the inner term changes with respect to f.

$$y = qV(\frac{f^2M}{q} + \frac{D_H(N_g,N_b)}{q}) + (1-q)V(\frac{(1-f)fM}{(1-q)} + \frac{\beta D_H(N_g,N_b)}{(1-q)}$$

$$\begin{aligned}\frac{\partial y}{\partial f} &= qV'\left(\frac{f^2M}{q} + \frac{D_H(N_g,N_b)}{q}\right) \times \frac{2fM}{q} + (1-q)V'\left(\frac{(1-f)fM}{(1-q)} + \frac{\beta D_H(N_g,N_b)}{(1-q)}\right) \times [\frac{M}{(1-q)} - 2f\frac{M}{(1-q)}] \\ &= 2fMV'\left(\frac{f^2M}{q} + \frac{D_H(N_g,N_b)}{q}\right) + [M-2fM]V'\left(\frac{(1-f)fM}{(1-q)} + \frac{\beta D_H(N_g,N_b)}{(1-q)}\right) \\ &= MV'\left(\frac{(1-f)fM}{(1-q)} + \frac{\beta D_H(N_g,N_b)}{(1-q)}\right) \\ &+ 2fM[V'\left(\frac{f^2M}{q} + \frac{D_H(N_g,N_b)}{q}\right) - V'\left(\frac{(1-f)fM}{(1-q)} + \frac{\beta D_H(N_g,N_b)}{(1-q)}\right)]\end{aligned}$$

Evaluating the derivative of the inner term at $f = 1$, we have

$$\frac{\partial y}{\partial f}(1) = MV'\left(\frac{\beta D_H(N_g,N_b)}{(1-q)}\right) + 2M[V'\left(\frac{M}{q} + \frac{D_H(N_g,N_b)}{q}\right) - V'\left(\frac{\beta D_H(N_g,N_b)}{(1-q)}\right)]$$

If we call $\frac{\beta D_H(N_g,N_b)}{(1-q)} = x$, we know that this expression is equivalent to

$$\begin{aligned}\frac{\partial y}{\partial f}(1) &= MV'(x) + 2M[V'(x+\delta) - V'(x)] \\ &= M[V'(x+\delta) - V'(x)]\end{aligned}$$

Since M is positive, the size of this derivative depends on the relative size of $V'(x+\delta)$ and $V'(x)$. For $q > 0.5$, we know that δ is positive. To see this, consider the inequality

$$\frac{M}{q} + \frac{D_H(N_g,N_b)}{q} > \frac{\beta D_H(N_g,N_b)}{(1-q)}$$

Since $\frac{M}{q} > 0$, this inequality will hold if

$$\begin{aligned}\frac{D_H(N_g,N_b)}{q} &> \frac{\beta D_H(N_g,N_b)}{(1-q)} \\ 1-q &> qb \\ 1 &> q + qb \\ 0.5 &> q\end{aligned}$$

where we use the fact that $b \leq 1$.

Then, V concave and $q < 0.5$ guarantees that $\frac{\partial y}{\partial f}(1) < 0$. Moreover, if $v'(x) >> v'(x+\delta)$, then $\frac{\partial y}{\partial f}$ can be made arbitrarily negative.

Now, thinking about the larger derivative $\frac{S}{f}$, note that

$$\frac{S}{f} = k(\cdot)\frac{\partial y}{\partial f}$$

where we know that $k(\cdot)$ is positive and is constant for fixed values of q, b, D_H, M, as well as fixed values of $V(\frac{M}{q} + \frac{D_H(N_g,N_b)}{q})$ and $V(\frac{\beta D_H(N_g,N_b)}{(1-q)})$. Then, fixing all of these values, but letting $v'(x)$ be large, guarantees an arbitrarily negative $\frac{\partial S}{\partial f}$.

Choose $V'(x)$ large enough to satisfy $-\frac{\partial S}{\partial f} > \frac{S(i_g,i_b)}{\frac{\partial D_H}{\partial S}+1}$, which is possible because the terms on the right hand side of the inequality do not depend on $V'()$.

Then, $f^* < 1$.

Discussion. These two lemmas relate the producer valuation function V to the optimal fraction of good content to show light consumers f. The key takeaway is that for V concave enough, the platform should show light consumers some bad content. This is true even though showing light consumers bad content directly trades off with showing light consumers good content. The intuition for this result is that choosing $f = 1$ means all of the attention from light consumers is going to content producers in the good state. For concave V, producers would prefer it if some attention was redistributed from the good state to the bad state since they are receiving less attention in the bad state, so choosing $f < 1$ will increase supply. If V is concave enough, then choosing slightly lower f will result in a large increase in supply, and it will be optimal to choose $f^* < 1$. The idea is that a large increase in content supply will increase demand of heavy consumers in a way that more than offsets the loss in demand of light consumers from choosing $f < 1$, so total profits will increase.