Personalization, Engagement, and Content Quality on Social Media: An Evaluation of Reddit's News Feed

Alex Moehring*

May 30, 2024

Abstract

Digital platforms increasingly curate their content through personalized algorithmic feeds. Platforms have an incentive to promote content that increases the predicted engagement of each user to lift advertising revenues. This paper studies how ranking content to maximize engagement affects the credibility of news content with which users engage. In addition, I evaluate how the ranking algorithm itself can be designed to promote engagement with high-credibility content. Using data from the Reddit politics community, I exploit a novel discontinuity in the ranking algorithm to identify the causal effect of a post's rank on the number of comments it receives. I use this discontinuity to identify a model of user comment decisions and estimate the credibility of news content that users engage with under a personalized engagement-maximizing algorithm. The personalized engagement-maximizing algorithm exacerbates differences in the credibility of news content with which users engage. I then evaluate a credibility-aware algorithm that explicitly promotes credible news publishers and find the platform can substantially increase the share of engagement with high-credibility publishers for a small reduction in total engagement. These findings suggest algorithmic interventions can be a useful tool for managers to balance engagement quantity and content quality.

*MIT Sloan School of Management, email: moehring@mit.edu. I am especially thankful to Catherine Tucker, Nikhil Agarwal, and Dean Eckles for their guidance and support on this project. I also thank Sinan Aral, Sarah Bana, Seth Benzell, Rahul Bhui, Avinash Collis, Tim Di Silva, Glenn Ellison, David Holtz, John Horton, Madhav Kumar, Jia Liu, Carlos Molina, David Rand, Manish Raghavan, Tobias Salz, Ananya Sen, Sebastian Steffen, Duncan Simester, Hong-Yi Tu Ye, Emma van Inwegen, MacKenzie Walker, Birger Wernerfelt, Michael Zhao, and numerous seminar participants for helpful conversations, comments, and suggestions.

1 Introduction

Social media platforms curate content for their users because of limited user attention and the vast amount of available content. The advertising business model adopted by many platforms creates an incentive to promote content through ranking algorithms that predict what content users are most likely to act on via clicking, liking, or commenting [Thorburn et al., 2022, Narayanan, 2023]. There is an active debate surrounding the benefits and potential risks of personalized rankings that optimize for engagement. Platform managers often contend that ranking algorithms act as agents for users by promoting a user's preferred content and reducing search frictions on the platform [Dorsey, 2022]. Critics, however, raise concerns that optimizing for engagement can incentivize low-quality or problematic content and reduce the diversity of viewpoints to which users are exposed [Pariser, 2011, Orlowski, 2020]. These concerns have led policy makers around the world to consider regulating ranking algorithms (e.g. European Union's Digital Services Act). In addition, promoting low-quality content can be harmful to platforms if advertisers respond by reducing advertising spending due to brand safety reasons [Ahmad et al., 2023] or if there is a disconnect between short-term engagement metrics and long-term user welfare [Spence and Owen, 1977, Kleinberg et al., 2023, Allcott et al., 2022]. Therefore, managers must balance maximizing engagement with the quality of content they are promoting to satisfy both internal and external stakeholders.

Despite these competing narratives, the impact of personalized news feeds on the quality of content users engage with remains an important and largely unresolved question. The lack of evidence regarding these issues primarily stems from the substantial challenges to studying ranking algorithms on social media platforms, including a hesitance to share data and experiments with external researchers and the identification challenges of observational approaches [Eckles, 2022].

This paper analyzes how ranking algorithms on social media platforms impact the quality of news content with which users engage. To do so, I first estimate the causal effect of a news article's position in the feed on future engagement using a novel regression discontinuity. The treatment effects demonstrate the importance of the ranking algorithm in steering attention and engagement to promoted articles. Second, I estimate a model of user engagement where the ranking algorithm influences engagement through attention as I hypothesize that users are less likely to be exposed to articles ranked further down the page. The estimates of the treatment effect of rank on engagement identify how article rank impacts the likelihood of a user being exposed to a post in this model. Conditional on being exposed to an article, users choose whether or not to comment on the article depending on their preferences over article features. Users have heterogeneous preferences over article features and these heterogeneous preferences can be used by the platform to personalize content.

I then use this model to estimate engagement patterns under counterfactual ranking algorithms. First, I analyze how a personalized ranking algorithm that optimizes for engagement impacts the quality of content with which users engage. Second, I evaluate the cost to the platform in terms of foregone engagement of credibility-aware ranking algorithms that trade off optimizing for total engagement and engagement with high-quality content. Varying the weight placed on content quality allows me to trace the efficient frontier between engagement and content quality.

I explore this question in the context of political news on Reddit. In particular, I focus on the platform's largest politics community that centers on sharing and discussing news articles about US political news. In this community, users share news articles about US politics and then engage in discussion in comment threads alongside each article. I use the number of comments an article receives as a measure of engagement.

Reddit is an important platform to investigate the trade-off between engagement and content quality for a number of reasons. First, Reddit is among the ten most popular social media platforms in the world with over 70 million daily active users [SimilarWeb, 2024]. In addition, Reddit is a critical component of the digital ecosystem playing an influential role in both web search and a source of training data for large language models [Patel, 2024]. As a result, Reddit is an important setting to understand how algorithms influence the content with which users engage.

The Reddit politics community also provides an ideal laboratory to analyze the content promoted under alternative ranking algorithms. The community is important to the platform due its size and the strong preferences of advertisers to not appear alongside low-quality news content [Ahmad et al., 2023].[1] In addition, it is often challenging to evaluate the quality of content on social media and a benefit of analyzing the politics community, which focuses on discussing news articles, is that I am able to use established measures of publisher credibility as a neutral measure of quality [Lin et al., 2022]. Finally, many key platform stakeholders, including advertisers, employees, and policymakers, have demonstrated an interest in the credibility of news content that appears on platforms which motivates the focus on a community where users share and discuss news articles.

Throughout the analysis, a central challenge will be the endogeneity of an article's rank – its position in the feed. One should be concerned that an article's potential outcomes are correlated with its position in the feed, as I expect the existing feed to promote articles that are more 'commentable' relative to articles that are not promoted. Therefore, to identify position effects – the causal effect an article's position has on the number of comments it receives – I exploit a novel regression discontinuity revealed in an open-source mirror of the platform's code base. This open-source mirror allows me to inspect the ranking algorithm and recreate the numerical score that is used to rank articles. Consequently, this permits using a regression discontinuity design to identify the local average treatment effect of an article's rank on the number of comments it receives in the subsequent period. As the ranking score of a focal article passes the score of a competing article, there is a discontinuous jump in the probability the focal article is ranked lower on the page.[2] The treatment effect estimates suggest that the causal effect of an article being promoted from the second position in the feed to the first position results in a 42.5% increase in the number of comments the article receives immediately following the ranking. The effect of being promoted declines further down the feed, as the causal effect of moving one position higher on the feed is

[1] The politics community is consistently ranked as one of the most active communities on the platform.

[2] This identification strategy is most closely related to Narayanan and Kalyanam [2015], where data on the AdRank scores in Google auctions are used to estimate the position effects on Google advertisements, though to my knowledge this is the first application of such a strategy to a social media setting.

largest for the first position.

With an identification strategy for position effects, I turn to understanding the impact of optimizing for engagement via personalized rankings on the quality of content to which users are exposed and with which they engage. To do so, I estimate a micro-founded model of user comment decisions. I model engagement decisions based on two components: whether a user is exposed to an article and whether, conditional on exposure, their utility from commenting exceeds the utility of the outside option. Article rank impacts engagement in this model only through the exposure component, where the probability of being exposed to an article depends on the article's rank. Conditional on exposure, users then have heterogeneous preferences to comment on articles depending on the political slant and credibility rating of the publisher. This model is identified using the reduced form position effect estimates and individual engagement choices. I use the model to estimate engagement patterns under counterfactual ranking algorithms including both personalized and non-personalized engagement maximization. In addition, I evaluate alternative credibility-aware ranking algorithms that optimize for an objective function that balances total engagement and engagement with high-credibility publishers.

I find that a personalized engagement maximizing algorithm exacerbates inequality in the quality of user news diets – the share of a user's engagement with high-credibility publishers. The engagement maximizing algorithm tends to promote high-credibility publishers to users engaging with high-credibility publishers under Reddit's actual ranking algorithm and promotes lower-credibility publishers to users engaging with less-credible publishers under the actual ranking algorithm. Moreover, I find that personalized engagement maximization leads to engagement with publishers that are less politically diverse and more similar to publishers the user has engaged with previously.

I next analyze engagement patterns under credibility-aware counterfactual ranking algorithms that optimize alternative objective functions that explicitly trade-off total engagement and engagement with high-credibility publishers. At one extreme, this nests a credibility-maximizing algorithm that maximizes engagement with high-credibility publishers. This algorithm leads to a 5.0% decline in total user engagement. However, platforms can achieve over half of the increase in news diet quality from the credibility-maximizing algorithm for a more modest 1.9% decrease in engagement. This change in engagement is similar in magnitude to the difference between the personalized and non-personalized engagement-maximizing algorithms. However, the non-personalized algorithm does not meaningfully improve the quality of user news diets, while the credibility-aware algorithm increases the average user's share of engagement with high-credibility publishers by 5.9 percentage points. This result suggests there is room for managers to use personalized ranking algorithms to balance the competing objectives of maximizing total engagement while maintaining the quality of content on the platform.

Related Literature

This paper primarily contributes to two strands of the literature. A large literature has studied the impact of algorithmic recommendations on consumers. This literature considers algorithmic recommendations' impact on product sales [Fleder and Hosanagar, 2009, Oestreicher-Singer and Sundararajan, 2012, Hosanagar et al., 2014, Ghose et al., 2014, Lee and Hosanagar, 2019, Donnelly et al., 2023, Wang et al., 2023], content consumption [Peukert et al., 2023, Holtz et al., 2020, Aridor et al., 2022, Chen et al., 2023], and consumer welfare [Ghose et al., 2014, Donnelly et al., 2023]. In addition, this literature investigates how ranked feeds on social media platforms impact individual well-being [Kramer et al., 2014], media consumption [Bakshy et al., 2015, Levy, 2021, Dujeancourt and Garz, 2023, Guess et al., 2023], and exposure to content from politicians [Huszár et al., 2022]. Much of this literature explores the impact of algorithmic ranking on the diversity of consumption [Van Alstyne and Brynjolfsson, 2005, Fleder and Hosanagar, 2009, Peukert et al., 2023, Holtz et al., 2020, Chen et al., 2023] or product sales [Oestreicher-Singer and Sundararajan, 2012, Hosanagar et al., 2014, Lee and Hosanagar, 2019] and how likely users are to be exposed to cross-cutting news publishers [Bakshy et al., 2015, Levy, 2021].

Most related, Huszár et al. [2022] analyzes an experiment on Twitter and Guess et al. [2023] analyze an experiment on Facebook and Instagram that randomly assigns users to receive a reverse chronological ranking algorithm relative to the existing personalized algorithm. Huszár et al. [2022] find that Twitter's personalized algorithm amplified right-leaning publishers and Guess et al. [2023] find that the reverse chronological feed increases exposure to political and content from untrustworthy sources. This paper contributes to the literature by unpacking the heterogeneous effect of engagement maximizing algorithms on the credibility of content with which users engage and by investigating the engagement-credibility trade-off the platform faces. An additional benefit of the modeling based approach taken here is that it allows me to consider many counterfactual ranking algorithms without the costs of testing each ranking algorithm experimentally.

Second, this paper contributes to the large and growing literature studying interventions to improve the quality of information people consume online. This literature both documents the reach of misinformation on social media and how it spreads [Allcott and Gentzkow, 2017, Vosoughi et al., 2018, Grinberg et al., 2019, Guess et al., 2019, 2020] and evaluates interventions to curb the spread of misinformation (see Pennycook and Rand [2021] and Lazer et al. [2018] for a review). My findings are consistent with the literature showing that a minority of users account for a large share of consumption of low-credibility news content [Allcott and Gentzkow, 2017, Grinberg et al., 2019, Guess et al., 2019, 2020], and contribute to the literature by finding that personalized engagement-maximizing algorithms exacerbate this difference. In addition, this literature assesses many behavioral interventions through both lab and field experiments. That said, empirical evaluations of algorithmic interventions have been more difficult given limited access to platform data. I contribute to this literature by exploring the impact of algorithmic interventions that promote high-credibility publishers and estimate the cost of such interventions.

Implications for Managers and Policy Makers

These findings have important managerial implications. Given advertiser concerns over brand safety and a reluctance to appear alongside low-credibility news publishers [Ahmad et al., 2023], platform managers must balance total engagement with the credibility of content being promoted to users. Additional internal and external stakeholders, including policy makers and platform employees, have also demonstrated interest in reducing the spread of low-credibility content on digital platforms [Warner, 2023, Haugen, 2021]. The results presented here suggest codifying the trade-off explicitly in the objective function of the ranking algorithm is an effective method for limiting the spread of low-credibility news content for a modest cost in terms of foregone engagement.

The findings also have important implications for policy makers. Concerns regarding ranking algorithms promoting and incentivizing low-quality content have prompted policy makers around the world to consider regulation that can address these issues. A common regulatory approach is to require platforms to allow users to opt out of personalized recommendations. The European Union's Digital Services Act includes such provisions, as do proposed laws in the United States such as the Filter Bubble Transparency Act. The implications of this study are clear: ranking algorithms can be designed to improve the credibility of the news content users engage with, and personalization is a valuable tool to mitigate the cost of moving away from optimizing only for engagement. The results suggest that allowing users to opt out of personalized feeds would not have a material impact on the credibility of news content with which users engage.

2 Background and Data

Reddit is a large social news aggregator with over 70 million daily active users as of January 2024 and was valued at approximately $6.5 billion shortly after its initial public offering in 2024.[3] The platform is organized into over 100,000 virtual communities called subreddits that are focused on sharing and discussing content related to the community's topic. In this study, I focus on a subset of communities that are centered around sharing and discussing news articles. In these communities, users share news articles and then discuss the articles in comment threads. Reddit is structured such that users can submit two types of content, submissions and comments. In the communities studied, submissions must contain a link to a news article and I therefore use the terms submissions, articles, and posts interchangeably. Users then discuss articles by posting comments, and this commenting activity is the primary engagement measure I study.

2.1 Algorithmic Feeds on Reddit

Users interact with content on the platform via algorithmic feeds of a few different forms. Any user who visits a community page will see submissions from the community ranked by the platform's default ranking algorithm.[4] This algorithm sorts submissions according to the post's age and vote score – the net number of upvotes minus downvotes on a post – and is described in more detail

[3] `https://www.redditinc.com/`

[4] On the platform, this default algorithm is called the *hot* algorithm. I refer to the hot algorithm as the actual ranking algorithm throughout.

in Section 3. In addition to the default algorithm, users can choose to rank posts according to several alternative algorithms. The *new* algorithm implements a reverse chronological ranking; the *top* algorithm ranks posts according to the vote score in a given period; the *rising* algorithm favors recent posts; and the *controversial* algorithm promotes posts that have received more votes, either up or down, regardless of their direction. This paper focuses on the default algorithm's impact on engagement. All analyses presented here condition on the alternative algorithm rankings remaining unchanged. That is, when estimating the impact of post rank on engagement I estimate counterfactuals where post rank changes in the default feed but not in the alternative feeds.

Reddit users can also join communities. Posts from these communities are displayed on a user's Home feed, the default feed users encounter when visiting the platform. The Home feed sorts posts according to the same default algorithm used by individual communities but ranks posts from all communities that a user is a member of rather than only posts from a single community.[5] The Home feed has important implications for this analysis, as I estimate the effect a post's rank in the subreddit feed has on its future engagement and how alternative ranking algorithms on the subreddit feed impact engagement patterns. Importantly, this captures both the direct effect of changing a post's rank on the subreddit feed and the indirect effect of changing the rank on the Home feed, holding fixed posts from other communities. For example, if two posts from the politics feed (A and B) and one post from another community (C) are ranked A, C, B on the Home feed, then counterfactuals where post B is promoted on the politics community correspond to the counterfactual ranking B, C, A on the Home feed. Given the prominence of the Home feed, it is important that the position effect estimates and counterfactual analyses include the effect of post rank in the Home feed.

2.2 Data

2.2.1 Ranking and Engagement

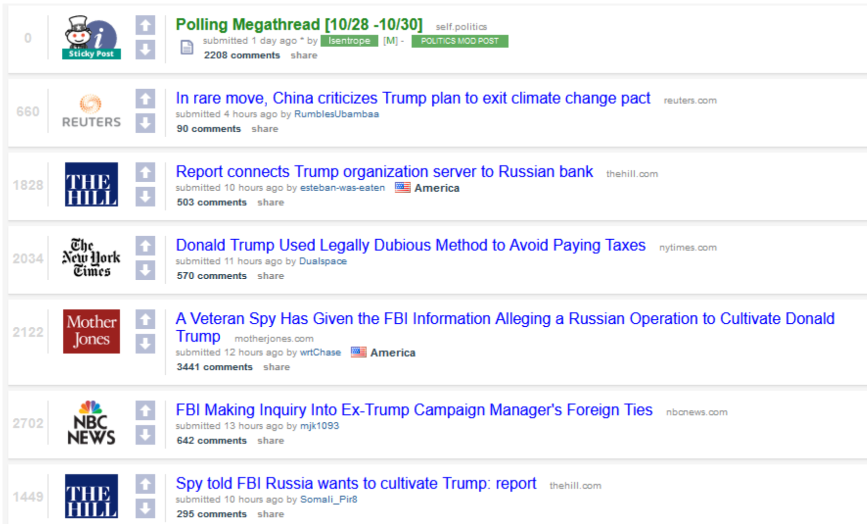
I merge data from several sources in this study. First, I scrape subreddit landing pages from the Internet Archive's Wayback Machine for each subreddit in the study. These data provides historical snapshots of subreddit feeds, allowing me to collect the top 25 ranked posts, their position in the feed, and post features. A snapshot from the politics community following the 2016 election is shown in Figure 1. Alongside the post position in the feed, parsing the Wayback Machine snapshots provides the age of a post, number of existing comments, vote score of each post (net number of upvotes minus downvotes), post title, and domain the post links to, if any. In addition, each snapshot reveals the number of subscribers each community has and the number of users online at the time of the snapshot.

Submissions on Reddit are either pinned to the top of the feed by community moderators or ranked organically.[6] I focus on organic posts displayed in blue in Figure 1. These posts are

[5] In 2018, after the period studied, Reddit changed the default algorithm used by the Home feed to the Best algorithm, as described in `https://www.reddit.com/r/changelog/comments/7spgg0/best_is_the_new_hotness/`.

[6] Posts pinned by moderators are shown in green and are typically threads created to discuss the major events of

Figure 1: Snapshot of Politics Community



Note: A Wayback Machine snapshot of the politics subreddit from November 2016. Posts pinned by moderators are shown in green and are typically threads created to discuss major events or frequent discussion topics such as polling. Posts in blue are algorithmically ranked organic posts that are the focus of this study.

submitted by users and ranked according to the algorithms described in Section 2.1. In the politics community considered here, posts are required to follow strict community guidelines: they must be on topic for the community, they must link to an article from a news publisher, and the post title must exactly match the headline of the article to which the post links. Any commentary on the article must be added in the comment sections, which I turn to next.

The primary engagement metric I consider is comments on articles. Reddit is a platform centered around sharing and discussing user-generated content. Comments themselves are a form of user-generated content that bring people to the platform, and encouraging additional comments is of direct interest to the platform [Burke et al., 2009]. In addition, experimental evidence suggests users who receive comments on their posts are more likely to generate content in the future, a finding that further supports the premise that encouraging more comments is desirable for Reddit [Eckles et al., 2016, Mummalaneni et al., 2022]. There is also correlational evidence that users who comment more also spend more time on social media platforms – a metric that more closely approximates the amount of advertisements the user sees [Wojcik and Hughes, 2019].

I merge data from Baumgartner et al. [2020] that contain a near-universe of submissions and comments to public Reddit communities to generate the engagement outcomes. These data contain user-level commenting behavior, where each comment includes a time-stamp, a user identifier of the comment author, the full text of the comment, the post the comment is responding to, and the vote-score the comment received, among other observables. This information allows me to reconstruct a post's full comment history, including the comments that occurred immediately following each of

the day. Importantly, these are not algorithmically ranked and I condition on these posts remaining in their position on the feed. That is, I only consider counterfactuals where the organic post positions change.

the Wayback Machine's snapshots.

These data on user comments serve several purposes. First, they allow me to construct the number of comments each post received in a window following each snapshot. This will be critical for estimating effects of position on engagement on the platform. Second, individual-level comment decisions are used to estimate a choice model of user engagement in Section 4. Here, the panel nature of the data allows me to identify rich user-level heterogeneity in comment preferences. Finally, studying comments allows me to analyze the text content to provide additional insight into user preferences and to understand how optimizing for engagement impacts the sentiment of comments submitted to the platform.

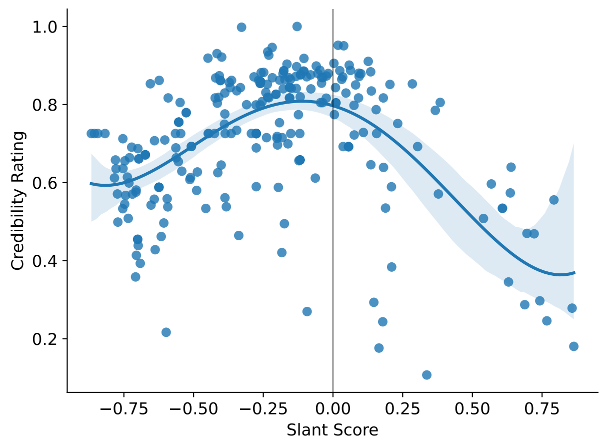
2.2.2 Publisher Slant and Credibility Ratings

I also collect two sets of publisher ratings that capture various aspects of an article's publisher. First are measures of a publisher's political slant [Robertson et al., 2018] that represent the relative propensity of a publisher domain being shared on Twitter by known Democratic party members relative to known Republican members, ranging from -1 to 1. A slant rating of -1 represents a domain that is only shared by Democrats while a slant rating of 1 represents a domain that is only shared by Republicans. Robertson et al. [2018] demonstrates this measure is consistent with a number of other expert, crowd-sourced, and audience-based ratings [Bakshy et al., 2015, Budak et al., 2016]. A primary benefit of the Robertson et al. [2018] scores compared to other measures of publisher slant is the high coverage, as the data set includes ratings for over 19,000 domains. This results in high coverage in our data, with over 90% of posts in the politics community containing a link to a publisher matching a domain in the Robertson et al. [2018] data.

I use credibility ratings, described in Lin et al. [2022], for over 11,520 news publishers. Lin et al. [2022] aggregate individual ratings from six rating organizations and demonstrate substantial agreement among individual sources. Importantly, the ratings released alongside Lin et al. [2022] show an extremely high correlation with NewsGuard ratings, a proprietary set of publisher ratings that employ extensive criteria including accuracy and balance of reporting, a process of publishing corrections, clear separation of opinion articles, and transparency of perceived conflicts. Figure 2 plots the joint distribution of publisher slant score and credibility rating for publishers that appear in at least 1% of the snapshots in the politics community. Table A.1 shows these ratings for six example domains. In evaluating user news diets, I discretize the credibility ratings into high- and low-credibility publishers for ease of interpretation. When doing so, I classify publishers as high credibility if their credibility rating is greater than 0.65 and I show robustness of key results to other thresholds in Appendix Section C.4.[7]

[7] The threshold of 0.65 is chosen as it is the median Lin et al. [2022] credibility rating within the Medium credibility category of Media Bias Fact Check, a professional rating organization.

Figure 2: Joint Distribution of Publisher Credibility and Slant Scores



Note: This figure plots the joint distribution of publisher slant score and credibility rating for the set of publishers that appear in at least 1% of the snapshots in the politics community. The dotted line displays the cutoff for high-credibility publishers. The regression line is a fourth-order polynomial fit. Confidence bands represent 95% confidence intervals.

Table 1: Summary Statistics

	Number	Share of Posts Missing			Number of Comments			
	Snapshots	Domain	Slant	Credibility	5 Min	10 Min	20 Min	60 Min
Politics	2105	0.01	0.07	0.11	2.98	5.97	11.93	35.76
US/World	5735	0.00	0.07	0.14	2.10	4.21	8.43	25.22
Sports	6390	0.22	0.66	0.85	0.78	1.53	3.01	8.51
Entertainment	3450	0.21	0.53	0.69	0.61	1.21	2.43	7.23
Gaming	2080	0.31	0.70	0.95	0.47	0.95	1.90	5.67
Technology	5912	0.06	0.43	0.66	0.28	0.57	1.13	3.41
Crypto	1267	0.38	0.77	0.87	0.23	0.45	0.90	2.60
Science	3561	0.10	0.37	0.49	0.10	0.20	0.41	1.20
Business	1632	0.08	0.27	0.38	0.05	0.09	0.19	0.58

Note: Summary statistics for the communities included in the study. Each row represents a category of news. The Number of Snapshots column contains the number of Wayback Machine snapshots for all communities in each category. The columns labeled Share of Posts Missing denote the share of submissions that lack information on publisher domain, slant score, and credibility rating. The columns labeled Number of Comments show the average number of comments a submission receives in the 5, 10, 20, and 60 minute periods following a snapshot. These columns average over both periods (i.e. snapshots) and positions in the feed.

2.3 Textual Analysis of Comments

A unique benefit of studying comments as the focal measure of engagement is that I can analyze the textual content in order to understand the types of comments users are submitting to the platform and how this varies depending on article features. This sentiment analysis is used in the model of user engagement decisions as I allow users to choose the sentiment of their comment conditional on article features. I use these estimates to evaluate the extent to which optimizing for engagement leads to deterioration in discussion quality.

I analyze the sentiment and emotional content of comments using a pre-trained neural network for sentiment analysis and emotion detection Pérez et al. [2021]. For each comment in the data, this model constructs a set of scores for predicted sentiment and emotion.

Comment sentiment is the primary quality measure of a comment's text that I use, which is highly correlated with predicted toxicity and the comment's emotional content (Appendix Figure A.1). Manual inspection reveals comments labeled as negative by the model are often extremely vulgar and unlikely to contribute productively to the discussion.

3 Effect of Rank on Engagement

In this section, I estimate the causal effect post rank has on the number of comments the post receives. Recall that I ultimately want to understand how engagement-maximizing ranking algorithms impact the type of content with which users engage and a key ingredient to this analysis is the causal effect of rank on engagement. Naive comparisons between posts with lower ranks (higher on the page) and posts with higher ranks (lower on the page) are unlikely to identify the causal effect of rank as I expect potential outcomes to be correlated with post rank [Narayanan and Kalyanam, 2015, Ursu, 2018]. Specifically, it is likely that posts with high potential outcomes, or latent commentability, are more likely to garner upvotes as the posts may be more inherently interesting. As I will describe, upvotes play a central role in the ranking algorithm and thus these posts are likely to be shown higher on the page.

This section introduces the identification strategy I use to overcome this challenge by estimating position effects – the causal effect of post rank on engagement. This serves two purposes. First, the treatment effect estimates provide important motivation for the analysis, given that I find post rank has a large causal effect on engagement, meaning the ranking algorithm plays an important role in shaping the posts with which users eventually engage. Second, the causal estimates from this section will be utilized directly in identifying the choice model of user engagement that is employed to analyze counterfactual ranking algorithms.

I exploit a regression discontinuity to identify the causal effect of rank on engagement. Until 2017, Reddit maintained an open-sourced mirror of its code base, which allows me to directly inspect the algorithm used to sort posts [reddit.com, 2017]. The algorithm assigns a ranking score to each post and ranks posts in descending order of these scores. Formally, a post's ranking score is defined as

$$s_{jt} = \text{sign}\left(u_{jt}\right)\log_{10}\left(\max\left\{\left|u_{jt}\right|, 1\right\}\right) - \frac{a_{jt}}{45,000} \tag{1}$$

where u_{jt} is the net number of upvotes minus downvotes that post j had at time t and a_{jt} is the age of the post in seconds.[8] This requires that for the ranking score of a post with a positive vote score to remain constant, every 12.5 hours the net number of upvotes minus downvotes must increase by a factor of 10 to offset the age penalty. Importantly, this defines a continuous score that determines post rank, creating a regression discontinuity that can be used to identify position effects. As the open source mirror was only maintained until early 2017, I only use data through 2016 for all analyses.

To give a concrete example of the regression discontinuity I exploit, consider two adjacent posts i, j with ranking scores s_i, s_j and observed ranks r_i, r_j. There is a discontinuous jump in the probability of post i being ranked lower than post j when the continuous forcing variable $s_i - s_j$ crosses zero. I take advantage of this discontinuity to identify the effect of rank on future engagement, under the assumption that potential outcomes (i.e. latent post quality) are continuous across the zero threshold of the forcing variable $(s_i - s_j)$.

3.1 Implementation Details of the Regression Discontinuity Design

I now discuss the implementation details of the regression discontinuity used to estimate the causal effect of post rank on future engagement. In particular, I focus on estimating the causal effect of moving up from position $r+1$ to position r on the feed. To simplify notation, let D_i be a treatment indicator (i.e. $D_i = 1[s_i > s_{-i}]$), and the forcing variable is denoted $\Delta s_i = s_i - s_{-i}$.

In this setting, the running variable is a composition of two scores, the ages and vote scores of the posts. This creates a cutoff frontier, shown in Appendix Figure B.6, analogous to geographic regression discontinuity designs. I take advantage of the multiple score nature of the problem and estimate the treatment effect at the origin, which ensures that posts are balanced on both post age and vote score, as described in Cattaneo et al. [2023].

The primary results use a local-linear approximation to the conditional expectation functions on either side of the discontinuity and a uniform kernel. I will show the estimates are similar under alternative specifications. I restrict to observations within a bandwidth λ of the cutoff chosen to minimize the mean squared error of the treatment effect estimator [Calonico et al., 2014, Cattaneo et al., 2020] and demonstrate the results are not sensitive to this choice (Appendix B.2).

For each of the 24 positions on the first page of the feed, I estimate the treatment effect of moving from position $r+1$ to position r as

$$\hat{\tau}_r = \hat{\mu}_r^+ - \hat{\mu}_r^- \tag{2}$$

where $\hat{\mu}_r^+$ is the estimated intercept from the local linear regression to the right of the discontinuity

[8] I normalize the post's timestamp by the period to interpret the second term as age. This is equivalent to adding a constant to all posts in a period and does not affect the ranking, but does make the ranking score more interpretable.

and $\hat{\mu}_r^-$ is the intercept to the left of the discontinuity [Cattaneo et al., 2020]. I estimate the treatment effect separately for each position in the feed, allowing the treatment effect of being promoted from position $r + 1$ to position r to vary by position. In Table B.2 and Appendix B.2, I show the results are not sensitive to the degree of polynomial approximation or the choice of kernel. Following best practices [Cattaneo et al., 2020], statistical inference uses robust bias-corrected standard errors that are clustered at the period level.

Measurement Error in the Running Variable

A challenge in this setting is that the running variable is constructed using data scraped from the Internet Archive's WayBack Machine and the reconstructed ranking scores do not completely determine the rank of a post. This is a result of several factors. First, Reddit explicitly adds noise to the vote scores shown to users to combat vote manipulation [Muchnik et al., 2013].[9] Second, Reddit caches votes and rankings for performance purposes due to the large amount of traffic the platform receives.[10] Caching means the ranking score and actual ranks are not continuously updated. This makes it possible for the observed ranks to differ from what is implied by the relative ranking scores as either the scores or observed rankings are a cached version.

Adding noise to the vote score introduces measurement error into the running variable and can bias traditional regression discontinuity estimates. To estimate the local average treatment effect of rank in the presence of measurement error in the running variable I follow Dong and Kolesár [2023] by excluding posts within a doughnut around the discontinuity. I manually select a doughnut width of 0.05 on either side of the cutoff and show in Appendix B.2 that the results are robust to the choice of doughnut width. Under the assumption that the doughnut excludes all periods where the posts are misclassified due to measurement error, Dong and Kolesár [2023] show that the usual regression discontinuity estimators identify a local average treatment effect.

After excluding posts within a doughnut of the discontinuity, I assume the remaining mismatch between post rank and the relative ranking scores is not due to measurement error. This assumption appears justified, as the probability of mismatch is constant as one moves away from the discontinuity (Figure 3a). If this were driven by the noise added to vote scores, the probability of mismatch would decline further away from the discontinuity as the probability the noise added is sufficiently large to misclassify the posts declines. Therefore, estimating the local average treatment effects using local linear regression results in conservative estimates of position effects.

3.2 First-Stage and Balance of the Regression Discontinuity Design

I now show evidence that Reddit ranks posts according to the algorithm I describe to establish a first stage in the regression discontinuity analysis. For each position on the feed $r \in \{1, \ldots, 24\}$, I consider the two posts ranked in position r and $r + 1$ and plot the probability that a post is in position r against the running variable (the difference in ranking scores of the two competing

[9] https://www.reddit.com/wiki/faq

[10] https://web.archive.org/web/20170121192832/https://redditblog.com/2017/1/17/caching-at-reddit/

posts). Figure 3a shows the discontinuity between position 1 and position 2; the plots for the remaining positions are shown in Appendix B.1. There is a clear discontinuity in the probability of being ranked lower when a post's ranking score surpasses the competing post's ranking score.

In addition, to test that post observables are balanced across the discontinuity, Figure B.11 plots the estimated treatment effect of rank on pre-treatment covariates including publisher slant, publisher credibility, post vote score, and post age. Nearly all estimates are insignificant at the 5% level, suggesting post observables are balanced across the discontinuity. While it is not possible to test the identifying assumption that potential outcomes are continuous through the discontinuity, this result is consistent with such an assumption holding. Appendix B.2.2 presents plots of the non-parametric conditional expectation functions of these covariates around the discontinuity.

3.3 Position Effect Estimates

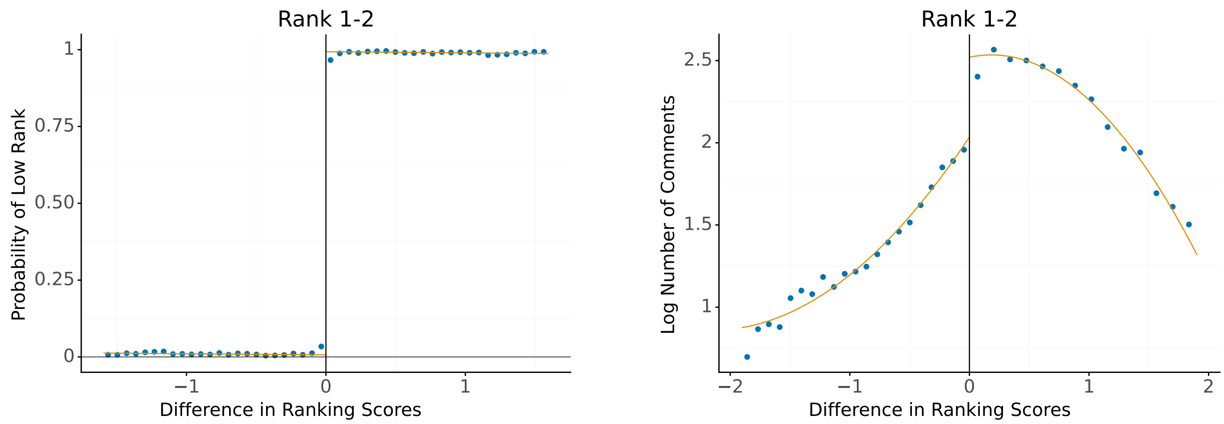
I now turn to estimating how post rank affects the engagement a post receives in the window following the snapshot. Figure 3b plots a binned scatter plot of the log of one plus the number of comments a post receives in the 20 minutes following each snapshot against the running variable (Δs_i) to visualize the discontinuity in the outcome variable. There is a clear discontinuity in engagement when a post is promoted to position 1 from position 2. Appendix B.1 shows the same plots for the remaining positions on the feed and the discontinuity in engagement quickly disappears further down the feed, suggesting treatment effects of rank are largest at the top of the feed.

I estimate the local average treatment effect using local linear regression, and present treatment effect estimates in Figure 4, which shows the effect of moving from rank $r+1$ to rank r on the log of one plus the number of comments a post receives in the 20 minutes following a snapshot. Being promoted to the first position has a large effect, with the treatment effect estimate suggesting a 42.5% increase in the number of comments received immediately following a snapshot relative to the second post. The importance of rank quickly dissipates further down the feed. Table B.2 also shows the results are robust to the polynomial choice and choice of kernel. Figure 4 includes naive OLS estimates of position effects. As expected, OLS substantially overestimates the effect of position on engagement, and this is particularly severe towards the top of the feed.

These treatment effect estimates demonstrate that the ranking algorithm has an important effect in determining the posts with which users engage. This in turn motivates further investigation of how the design of ranking algorithms impact the credibility of content users engage with as the platform has substantial power in determining what content users are exposed to and ultimately engage.

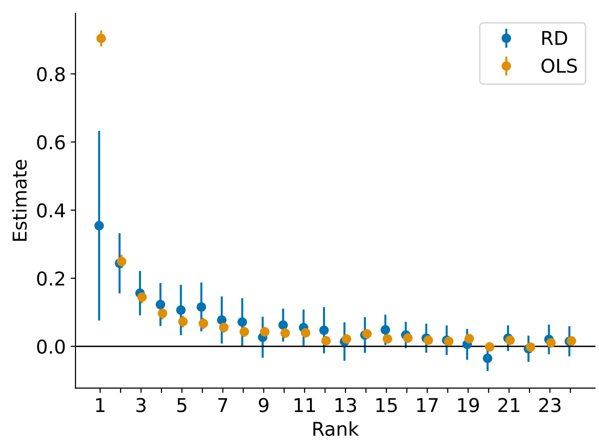
Figure 3: Regression Discontinuity Plots

(a) Effect on Post Rank (b) Effect on Engagement



Note: Regression discontinuity plots for the discontinuity around being promoted to the top position on the feed from the second position on the feed. Here, the x-axis is the running variable – the difference in the focal post's ranking score from that of the adjacent post – and the y-axis is (a) the probability a post is ranked lower on the page and (b) the log number of comments received in the 20 minutes following a snapshot plus one. This figure excludes posts within the doughnut which consists of posts where the absolute value of the running variable is less than 0.05. Fourth order polynomial fits are plotted alongside the binned mean values. The corresponding figures for the remaining positions on the feed are shown in Appendix B.1.

Figure 4: Position Effect Estimates



Note: Estimates of the local average treatment effect from a post moving from position $r+1$ to position r on the feed on the log of one plus the number of comments a post receives in the 20 minutes following a snapshot. The points labeled RD are from the regression discontinuity estimates and the points labeled OLS are from a naive comparison of articles ranked higher to those ranked lower. Error bars represent 95% confidence intervals. RD confidence intervals formed using robust bias-corrected standard errors that allow for misspecification of the conditional expectation function and that are clustered at the period level. OLS confidence intervals formed using cluster-robust standard errors clustered at the period level.

4 Model of Individual Engagement Decisions

I now estimate a model of user engagement that allows me to estimate engagement patterns under counterfactual ranking algorithms. Section 4.1 introduces the model of individual decisions, Section 4.2 describes identification and the estimation approach, and Section 4.3 summarizes the model estimates and fit.

4.1 Model

The model of user engagement decisions has several components. First, users visit the platform and are exposed to a subset of articles that is partially determined by the ranking algorithm. In this model, the ranking algorithm affects engagement by focusing user attention on articles promoted in the feed. Conditional on being exposed to an article, a user then comments if their utility from doing so exceeds the outside option. This model flexibly allows for users to have heterogeneous preferences over article features. This heterogeneity is important, as many of the counterfactual ranking algorithms I evaluate take advantage of personalization where preference heterogeneity may result in different rankings for different users.

Formally, users indexed by i visit the platform in periods indexed by t and are exposed to a ranked feed of posts indexed by j. In each period, users are exposed to a post in position r_{jt} if $v_{ijt} = 1$, which is an independent Bernoulli random variable equal to one with probability $p(r,t)$. I use a parsimonious parameterization of the exposure probability, $p(r,t) = p_t p_r$, where p_t is the probability of accessing the platform in period t and p_r is the probability of being exposed to a post in position r conditional on accessing the platform.[11] If exposed to a post, users receive utility

$$u_{ijt} = \delta_{ijt} + \varepsilon_{ijt} \tag{3}$$

if they comment on post j in period t, which I denote d_{ijt}. Users comment if they are exposed to the post and the utility from commenting exceeds the utility from the outside option (u_{i0t})[12]

$$d_{ijt} = 1\left[v_{ijt} = 1\right] 1\left[u_{ijt} \geq u_{i0t}\right]. \tag{4}$$

I model $\delta_{ijt} = x'_{jt}\left(\bar{\beta} + \beta_i\right) + \xi_{jt} = \delta_{jt} + x'_{jt}\beta_i$ where x_{jt} is a vector of observable article features, $\bar{\beta}$ represents average preferences, and β_i is a vector of the deviation of user i's preferences from the mean. Finally, ξ_{jt} is a article-period fixed effect that represents latent article commentability. This fixed effect flexibly captures anything users have vertical preferences over. For example, if

[11] Section 5.3 endogenizes the search process by allowing a user to decide whether to consider an article in position r by comparing the expected benefit of viewing articles in position r to a search cost of doing so.

[12] A key simplification is the stylized process by which users form consideration sets. A more flexible model that allows users to consider a subset of posts and comment on their most preferred post introduces substantial computational challenges as the number of potential consideration sets grows combinatorially. These models would likely yield similar results given users are highly unlikely to comment on more than one post in either the data or in the counterfactual simulations. Therefore, given the substantial computational benefits of assuming independence between comment decisions, I assume that users make engagement decisions on posts independently.

users prefer to comment on articles with many existing comments, this would be captured by the ξ_{jt} term.[13] Post rank is excluded from utility in this model, implying that rank does not impact choice conditional on exposure to a post.[14] Finally, normalize $\delta_{i0t} = 0$ without loss and assume ε_{ijt} is an independent and identically distributed Type 1 Extreme Value preference shock. This results in the mixed logit choice probabilities multiplied by the exposure parameter $p(\cdot)$.

$$P_{ijt} = P(v_{ijt} = 1, u_{ijt} \geq u_{i0t}) = p(r_{jt}, t) \frac{\exp \delta_{ijt}}{1 + \exp \delta_{ijt}} \tag{5}$$

Conditional on commenting on a post, users choose the sentiment of the comment to submit. Users can either submit a comment with negative sentiment or neutral sentiment. Users choose the probability with which their comment will be perceived negatively based on the user-specific and vertical components of comment utility. That is, conditional on commenting users choose the probability that comment ijt will be a negative comment as follows

$$\log \frac{b_{ijt}}{1 - b_{ijt}} = \beta^s_{i0} + \beta^s_{i1}(\delta_{ijt} - \xi_{jt}) + \beta^s_{i2}\xi_{jt} + \varepsilon^s_{ijt} \tag{6}$$

where b_{ijt} is the probability user i's comment on post jt is a negative comment, $\delta_{ijt} - \xi_{jt}$ is the user-specific component of comment utility user i receives when commenting on post jt, ξ_{jt} is the vertical commentability component of post jt, $\beta^s_i = \langle \beta^s_{i0}, \beta^s_{i1}, \beta^s_{i2} \rangle$ is a vector of individual i's sentiment preferences, and ε^s_{ijt} is an independent error term.

This model captures key elements of user engagement decisions that are relevant for the counterfactual ranking algorithms in a tractable manner. First, the model allows for ranking algorithms to impact engagement decisions by steering attention to promoted articles. Second, the model allows for rich heterogeneity in preferences over publisher features. In particular, the panel data allow me to estimate a separate β_i for each user. Given a leading counterfactual will consider personalized ranking algorithms, it is important to capture such preference heterogeneity in the model. In this counterfactual, I will assume that the platform has learned β_i and uses this to personalize the ranking of articles for each user.

4.2 Identification and Estimation

4.2.1 Identification of Model Parameters

The key identification challenge in this model is that observed post ranks are correlated with latent commentability, or $E[\xi_{jt} r_{jt}] \neq 0$. I now describe how this model is identified using the regression discontinuity from Section 3 and user-level engagement decisions.

[13] The latent commentability term ξ_{jt} is often referred to as latent quality in the literature estimating demand systems [Berry, 1994]. To avoid confusion with publisher credibility, I refer to ξ_{jt} as latent commentability, where this captures a vertical component making all users more likely to comment on the article.

[14] This assumption is motivated by the findings of Ursu [2018] that demonstrates empirically that rankings impact search probabilities but, conditional on search, do not affect purchase probabilities in an online travel platform. This is also consistent with recent work modeling personalized rankings in e-commerce [Donnelly et al., 2023]

I first describe how the exposure parameters (p_t, p_r) are identified. I assume that each user logs on to the platform with probability p_t, independent of article features or preference shocks. This probability is identified via the share of users who visit the platform in each period which I estimate using data on the share of users online at the start of each period. Conditional on accessing the platform, I assume all users are exposed to the top post on the feed implying that $p_1 = 1$.[15] The remaining exposure parameters p_r are identified by the reduced form treatment effects assuming constant treatment effects of rank on engagement. That is, I assume $E\left[Y_{jt}(r)\right] = e^{\tau_r} E\left[Y_{jt}(r+1)\right]$ where $Y_{jt}(r) = \sum_i d_{ijt}(r)$ is the potential outcome under rank r for the total number of comments post j in period t receives following a snapshot. This implies the following mapping between treatment effects and the ratio of exposure probabilities p_r

$$\tau_r = \log \frac{\sum_i E\left[d_{ijt}(r)\right]}{\sum_i E\left[d_{ijt}(r+1)\right]} = \log \frac{p_r}{p_{r+1}}. \tag{7}$$

The assumption of constant treatment effects could in principle be relaxed to allow for arbitrary individual heterogeneity and heterogeneity along observed article features, though the demands on the data grow substantially if this type of heterogeneity are included. For example, allowing for arbitrary individual heterogeneity would require estimating the reduced form treatment effects separately for each user.

Given exposure parameters, individual- and mean-preference parameters are identified from the assumption that article features are exogenous $E\left[x_{jt}\varepsilon_{ijt}\right] = 0$ and $E\left[x_{jt}\xi_{jt}\right] = 0$. Finally, the parameters in the sentiment model are identified by the assumption that ε^s_{ijt} is mean-independent of ε_{ijt}, ξ_{jt}, and x_{jt}.

4.2.2 Estimation

I estimate the choice model using data from the politics community given the relevance of this community to managers, policy makers, and users. I use individual-level comment decisions and restrict the sample to users who comment on at least 25 articles in the periods I study, where a period consists of the 60 minutes following a Wayback Machine snapshot.

I take this model to the data using a two step procedure that simplifies the computation given the large number of periods, users, and posts. In the first step, I estimate the exposure parameters p_t and p_r. I estimate p_t by combining data on the number of users online at the start of each period, which are observed in the Wayback Machine snapshots, with public statements by the platform on the average session duration to estimate the number of users who log on to the platform during each period using Little's law [Little, 1961]. I then use public usage statistics again to estimate the number of active community members and calculate the share of active community members who log on in each period. Finally, I smooth estimates of p_t by taking the fitted values of a regression of the raw values of p_t on quarter and day of week fixed effects. The full details of this process are described in Appendix C.3. I then estimate the remaining exposure parameters using an empirical

[15] The results are robust to other choices of p_1 as shown in Appendix Section C.4.

analog of Equation 7 ($p_r = \exp\left\{-\sum_{r'>1}^{r} \tau_{r'-1}\right\}$).

Given the estimates of p_t and p_r I estimate the individual parameters β_i using maximum likelihood

$$\mathcal{L}(\beta, \xi) = \sum_t \sum_i \sum_j d_{ijt} \log P_{ijt}(\beta_i, \xi_{jt}) + (1 - d_{ijt}) \log(1 - P_{ijt}(\beta_i, \xi_{jt})). \quad (8)$$

Finding the maximum likelihood estimate involves solving a high-dimensional optimization procedure due to the large number of individuals and posts. Therefore, I use the following iterative algorithm. First, I initialize a guess of ξ_{jt} and, conditional on these unobserved commentability parameters, estimate the individual-level preference parameters using maximum likelihood. I then invert observed engagement shares [Berry, 1994] using the Berry et al. [1995] contraction mapping to find the values of ξ_{jt} such that predicted market shares equal observed market shares.[16] I iterate between these two steps until convergence. Splitting the estimation algorithm into these two steps allows the maximum likelihood parameters to be estimated in parallel vastly reducing the required computation. Inference on the preference parameters uses cluster-robust standard errors.

The preference estimates of any individual user will contain sampling error. This implies that the distribution of preference estimates will be a convolution of the true distribution of preferences and sampling error, leading the distribution of estimates to be over-dispersed relative to the true distribution. To correct for this over dispersion, I shrink all preference estimates towards the grand-mean using the empirical Bayes procedure described in Appendix C.2.

4.3 Model Estimates

Table 3a summarizes the distribution of individual preference estimates $(\bar{\beta} + \beta_i)$. It is helpful to summarize preferences for publisher slant through each user's bliss point, which is defined as the slant the user most prefers

$$b_i^* = \begin{cases} \text{sign}(\beta_{is}) & \text{if } \beta_{is^2} \geq 0 \\ \min\left\{1, \max\left\{-1, -\frac{\beta_{is}}{2\beta_{is^2}}\right\}\right\} & \text{if } \beta_{is^2} < 0 \end{cases} \quad (9)$$

where β_{is} (β_{is^2}) is user i's taste parameter on post slant (slant squared). The marginal distribution of slant bliss points is shown in Figure 5. It is evident there is substantial preference heterogeneity in the politics community. Just over half of users prefer more credible publishers, while the remaining users prefer less credible publishers. Regarding political slant, there is also substantial heterogeneity, with a large mass of users preferring outlets slightly left of center. There are also mass points at each political extreme, with nearly 20% of users preferring outlets that are strongly left leaning and over 25% of users preferring outlets that are strongly right leaning.

Table 3b summarizes the estimates of individual-level preferences to submit a negative comment based on the vertical- and individual-specific components of comment utility (Equation 6). There is substantial heterogeneity in user preferences to submit a negative comment with 51.6% of users

[16]There are instances in the data where a post receives zero comments. I assume that ξ_{jt} is bounded below such that the minimum predicted market share is equal to 0.01% in these situations.

Table 2: Distribution of Individual Preference Estimates

	Mean	Std	1%	25%	50%	75%	99%
Constant	-4.55	0.75	-6.09	-5.06	-4.62	-4.11	-2.49
Slant Score	-0.20	0.64	-1.60	-0.65	-0.20	0.23	1.27
Slant Score^2	-0.30	0.82	-2.10	-0.88	-0.30	0.27	1.59
Credibility Rating	-0.02	0.86	-2.30	-0.53	0.03	0.56	1.86
ξ_{jt}	-0.00	0.93	-2.22	-0.61	0.02	0.61	2.12

(a) Individual Comment Preference Estimates

	Mean	Std	1%	25%	50%	75%	99%
Intercept	0.20	7.25	-18.46	-3.88	0.20	4.26	18.78
Heterogeneous component	0.04	1.53	-3.93	-0.82	0.06	0.90	3.94
Vertical component	0.03	0.08	-0.19	-0.01	0.03	0.08	0.23

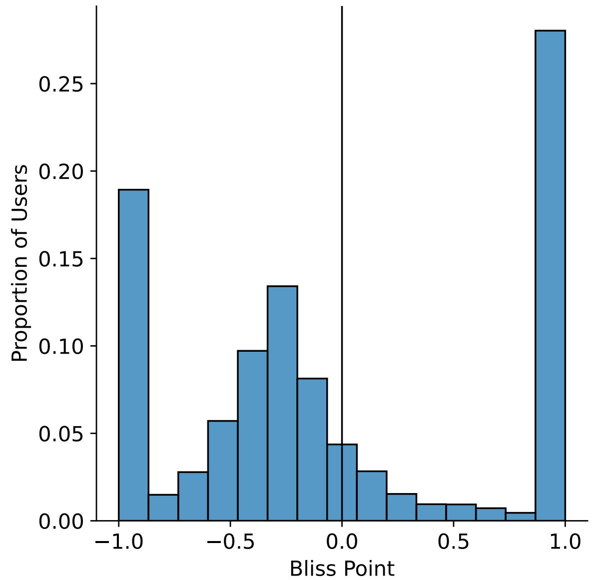
(b) Individual Sentiment Preference Estimates

Note: This table shows the user-level distribution of preference estimates. Panel (a) presents the distribution of comment preferences (Equation 4). The values for Constant, Slant Score, Slant Score^2, and Credibility Rating contain the user-level comment preferences. The values of ξ_{jt} are at the article level and show the distribution of latent article commentability. Panel (b) presents the distribution of sentiment preferences (Equation 6). The heterogeneous component captures how the likelihood of a user to submit a negative comment changes in response to changes in the user-specific component of post comment utility. The vertical component captures how the likelihood of a user to submit a negative comment changes in response to a change in the latent commentability term (ξ_{jt}). All preference parameters are shrunk to the grand mean using empirical Bayes.

more likely to comment negatively on posts in which they are more likely to comment. Figure C.18 reveals this is especially true for users more likely to comment on left-leaning posts (i.e. they have a negative bliss point) and users who prefer to comment on less credible publishers. Ranking algorithms that optimize solely for engagement will increase the share of negative posts for these users.

To assess the fit of the model, Table 3 presents summary statistics of actual engagement and engagement predicted by the model. The distribution of actual engagement with engagement predicted by the model is also shown graphically in Figure C.17. Figure C.17a demonstrates the high correlation between actual user engagement and predicted user engagement. Figure C.17b shows the correlation between actual and predicted user engagement by publisher credibility and the model again has a high correlation between actual and predicted engagement by group.

Figure 5: Distribution of Slant Bliss Points



Note: This figure plots the marginal distribution of user-level slant bliss points. The bliss point is the slant score for which a user is most likely to comment, all else being equal. A bliss point of -1 implies the user is most likely to comment on left-leaning articles and a bliss point of 1 implies the user is most likely to comment on right-leaning articles.

Table 3: Summary of Model Fit

		Actual		Model	
		Mean	Std	Mean	Std
Total		52.39	38.85	54.05	35.31
Credibility	High	41.55	30.73	42.72	27.88
	Low	10.84	9.62	11.33	8.26
Slant Partition	Strongly Left	13.38	11.86	14.04	10.42
	Left	7.99	6.75	8.14	5.31
	Middle	13.44	10.60	13.63	8.88
	Right	13.63	10.78	13.93	9.15
	Strongly Right	3.95	3.87	4.31	3.32

Note: Summary of the model fit. The Actual columns report the average and standard deviation of the total number of comments posted by each user and the number of comments by publisher rating. The Model columns report the model's predicted values for the same quantities under the existing ranking algorithm.

5 Counterfactual Ranking Algorithms

In this section, I use the model from Section 4 to analyze how counterfactual ranking algorithms impact user news diets. The goal of this analysis is to first understand how ranking content to maximize engagement impacts the credibility of news content with which users engage. Second, I evaluate the cost to the platform in terms of foregone engagement of alternative credibility-aware ranking algorithms that balance total engagement and engagement with high-credibility content in the ranking objective function. I describe the mechanics of the analysis of counterfactual ranking algorithms in Section 5.1, discuss the results in Section 5.2, and assess the robustness of the findings in Section 5.3.

5.1 Description of Counterfactual Rankings

I now describe the counterfactual ranking algorithms that I consider. In each period, the platform has an estimate of the probability that user i will comment on article j conditional on being exposed to that article ($\hat{P}_{ijt}$). The platform then rank content to maximize the following objective functions.

Personalized engagement maximizing: A leading counterfactual considered is personalized engagement maximization. The personalized engagement-maximizing ranking solves

$$r_{it}^{P} = \arg\max_{r \in \mathcal{R}} \sum_{j=1}^{J} p\left(r_j, t\right) \hat{P}_{ijt} \tag{10}$$

where $\mathcal{R}$ is the set of possible rankings and $r = \langle r_1, \ldots, r_J \rangle \in \mathcal{R}$ is a vector of possible article ranks. It is straightforward to show that when $p\left(r, t\right)$ is weakly decreasing in r, the optimal ranking sorts articles in descending order of $\hat{P}_{ijt}$.[17]

Credibility-aware algorithm: While short-term engagement is often used as a proxy for consumer welfare, a growing literature has emerged to study situations where these measures may differ. This disconnect can arise for rational economic agents [Spence and Owen, 1977] and in models with behavioral biases, including agents with present bias [Kleinberg et al., 2023], dual self models, [Kahneman, 2011], and digital addiction [Allcott et al., 2022]. Moreover, the platform may want to avoid promoting low-credibility publishers for brand-safety purposes or to prevent potential regulatory actions. These factors could lead the platform to consider publisher credibility in the ranking objective function. Therefore, I consider credibility-aware algorithm that maximizes an objective function that balances two competing objectives: total engagement and engagement with high-credibility publishers

$$\mathcal{S}_{ijt} = E\left[d_{ijt}\right]\left(\left(1-\lambda\right) + \lambda 1\left[c_{jt} \geq \underline{c}\right]\right) \tag{11}$$

[17] To show this, assume for contradiction there exists an optimal ranking with two posts j and j' such that $r_j < r_{j'}$ and $\hat{P}_{ijt} < \hat{P}_{ij't}$. Note that the objective under this ranking is less than the objective if the positions of the two posts are swapped

$$\left(\hat{P}_{ij't} - \hat{P}_{ijt}\right)\left(p\left(r_j, t\right) - p\left(r_{j'}, t\right)\right) \geq 0$$

because $p\left(\cdot\right)$ is weakly decreasing in r. Therefore, this ranking is not optimal, thus providing a contradiction.

where λ reflects the weight on engagement above a minimum credibility threshold $\underline{c}$. Note that this nests the personalized engagement-maximizing algorithm when $\lambda = 0$ and a credibility-maximizing algorithm when $\lambda = 1$. The credibility-aware algorithm solves

$$r_{it}^{O} = \arg\max_{r \in \mathcal{R}} \sum_{j=1}^{J} \hat{\mathcal{S}}_{ijt} = \arg\max_{r \in \mathcal{R}} \sum_{j=1}^{J} p\left(r_j, t\right) \left(\left(1-\lambda\right) + \lambda 1\left[c_{jt} \geq \underline{c}\right]\right) \hat{P}_{ijt} \tag{12}$$

and is solved by ranking articles in descending order of $\left(\left(1-\lambda\right) + \lambda 1\left[c_{jt} \geq \underline{c}\right]\right) \hat{P}_{ijt}$ for each user.

Non-personalized engagement maximizing: The non-personalized engagement maximizing algorithm solves the following maximization problem

$$r_{t}^{N} = \arg\max_{r \in \mathcal{R}} \sum_{j=1}^{J} p\left(r_j, t\right) E\left[\hat{P}_{ijt}\right] \tag{13}$$

which ranks articles in descending order of $E\left[\hat{P}_{ijt}\right]$ for each user.

Benchmarks: I compare the engagement patterns under the counterfactual algorithms described above to two benchmark algorithms, the ranking employed by the platform (Actual) and a random benchmark that randomly shuffles the articles shown on the page for each user (Random).

To form $\hat{P}_{ijt}$, I assume that the platform has high quality estimates of user preferences given their access to rich user-level behavioral data and therefore assume the platform observes β_i in the counterfactuals. I assume the platform does not, however, observe latent article commentability (ξ_{jt}) and must estimate this through observable article features. I model the platform's estimates of ξ_{jt} as a supervised learning problem where the platform forms estimates of the true latent article commentability ($\hat{\xi}_{jt}$) based on article observables. I operationalize this using a random forest that predicts ξ_{jt} using observable post features including the stock of total and top-level comments, vote score, post age, publisher slant, and publisher credibility rating. This model performs well in the prediction task as demonstrated in Appendix Figure C.19. With estimates of article commentability, observed post features, and observed user preferences, the platform can estimate engagement probabilities for each user and article conditional on exposure $\hat{P}_{ijt} = \frac{\exp \hat{\delta}_{ijt}}{1+\exp \hat{\delta}_{ijt}}$ where $\hat{\delta}_{ijt} = x_{jt}'\left(\bar{\beta} + \beta_i\right) + \hat{\xi}_{jt}$.

In this study, I focus on the implications of different objective functions in algorithmic ranking conditional on the candidates available. That is, the counterfactuals considered here only re-rank the top 25 posts in each period, which I treat as the set of candidate posts to be ranked. This decision is relatively innocuous for analyzing the impact of optimizing for engagement, as latent post commentability is highly correlated with post rank in the data (Figure C.15). Latent post commentability is an important factor in optimizing for engagement, meaning posts that are not in the top 25 posts would be less likely to be ranked high on the feed even if they were included in the candidate posts.[18] To calculate engagement under a counterfactual algorithm, I calculate

[18] This assumption does preclude analyzing simple proposed algorithms such as reverse chronological, as the restricted set of candidate posts excludes the high volume of low-quality posts that are often promoted under a reverse-

engagement probabilities for each post and user by multiplying the exposure probability for the post under the counterfactual ranking with the true estimated engagement probability conditional on exposure.

5.2 Counterfactual Ranking Algorithm Results

Summaries of engagement patterns under the different counterfactual ranking algorithms are shown in Table 4. I now describe the impact the various ranking algorithms have on user news diets.

5.2.1 Impact on Engagement Quantity

The counterfactual analysis suggests that the algorithm employed by the platform, which prioritizes simplicity and transparency, is far from engagement maximizing. That said, the actual algorithm does result in substantially higher engagement relative to the random benchmark. As expected, optimizing for engagement leads to a substantial increase in engagement quantity.

Much of the benefit comes from ranking articles according to expected engagement without personalization, which is evidenced by the 19.1% increase in engagement under the non-personalized engagement-maximizing algorithm. Personalizing user feeds increases engagement by 21.0% relative to the existing algorithm, providing a modest increase in engagement relative to the non-personalized engagement-maximizing algorithm. While modest in size, this lift does demonstrate the platform has an incentive to personalize rankings to drive engagement. Optimizing for engagement with high-credibility publishers also leads to a substantial increase in engagement relative to the actual algorithm employed (15.0%), but represents a substantial cost in terms of lost engagement relative to the engagement-maximizing algorithms.

5.2.2 Impact on Engagement with Publishers by Credibility

Reddit's algorithm does not materially impact the share of engagement with high-credibility publishers relative to a random ordering of posts, with both algorithms resulting in 79.0% of the average user's engagement being with high-credibility publishers. Optimizing for engagement also does not lead to a substantial change for the average user, with an average high-credibility engagement share of 79.9% for the non-personalized engagement-maximizing algorithm and 79.0% for the personalized engagement-maximizing algorithm. The credibility-maximizing algorithm does lead to a substantial increase in the share of engagement with high-credibility publishers, with the average user's high-credibility engagement share rising to 89.7%.

Focusing on average changes masks important heterogeneity. Figure 6a plots the empirical CDF of the change in high-credibility shares relative to the existing algorithm. The non-personalized algorithm has little impact on the quality of news diets as the share of engagement with high-credibility publishers does not change substantially for any user. The personalized engagement-maximizing algorithm, however, does have substantial impacts for many users despite the negligible

chronological ranking.

Table 4: Counterfactual Engagement Summaries

	Engagement	Dist. to Uniform	Credibility	Neg. Engagement
Intercept	54.048	0.280	0.790	0.512
	(0.387)	(0.001)	(0.001)	(0.002)
Random	-6.468	-0.004	-0.001	-0.001
	(0.043)	(0.000)	(0.000)	(0.000)
Non-Personalized	10.309	-0.008	0.008	0.001
	(0.062)	(0.000)	(0.000)	(0.000)
Personalized	11.357	0.030	-0.001	0.002
	(0.072)	(0.001)	(0.000)	(0.000)
Credibility Max.	8.118	0.008	0.107	0.001
	(0.054)	(0.000)	(0.000)	(0.000)
Observations	41675	41675	41675	41675
R-Squared	0.031	0.031	0.405	0.000

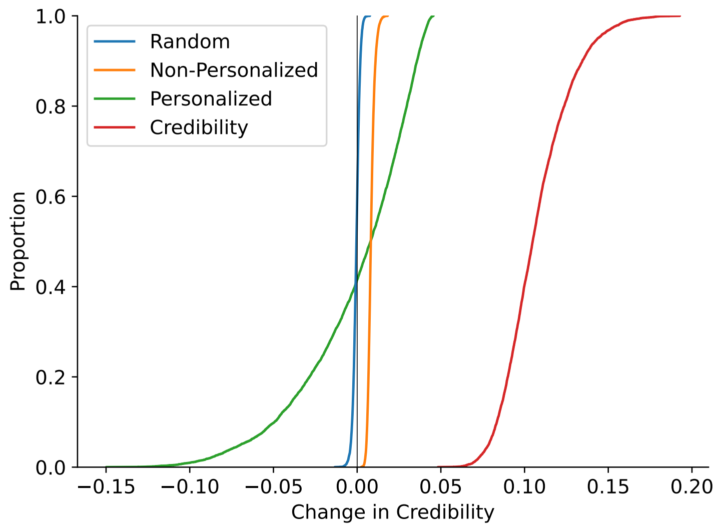
Note: This table reports estimates of a panel regression of each counterfactual outcome on counterfactual algorithm dummy variables. The intercept is the average quantity under the existing algorithm. (1) Engagement represents the total number of articles a user comments on. (2) Dist. to Uniform represents the first Wasserstein distance of engagement shares across publisher slant partitions from the uniform distribution. Recall distributions closer to uniform will have smaller distances, meaning they represent more diverse engagement. (3) Credibility represents the share of a users engagement with high-credibility publishers. (4) Neg. Engagement represents the share of comments that are negative sentiment. Standard errors are clustered at the user level.

average effect. The majority of users experience a modest improvement in the quality of their news diets, as a slightly larger share of their engagement is with high-credibility publishers. However, 41.5% of users experience a deterioration in the quality of their news diets, with a subset of these users seeing the share of their engagement with high-credibility publishers falling by over 10 percentage points. To better understand what users experience these declines, Figure 6b plots the relationship between news diet quality under the existing algorithm against news diet quality under the counterfactual algorithms. It is clear that users engaging with less credible publishers under Reddit's actual algorithm experience large declines in the quality of their news diets under the personalized engagement-maximizing algorithm. This suggests the engagement maximizing algorithm exacerbates differences in the quality of user news diets by promoting high-credibility publishers to the majority of users who typically engage with high-credibility publishers and promoting low-credibility publishers to users who have engaged with these publishers in the past. Moreover, the results are robust to the choice of threshold for high-credibility publishers (Appendix Section C.4) and I find personalization exacerbates differences in the quality of user news diets even for very low thresholds for high-credibility publishers.

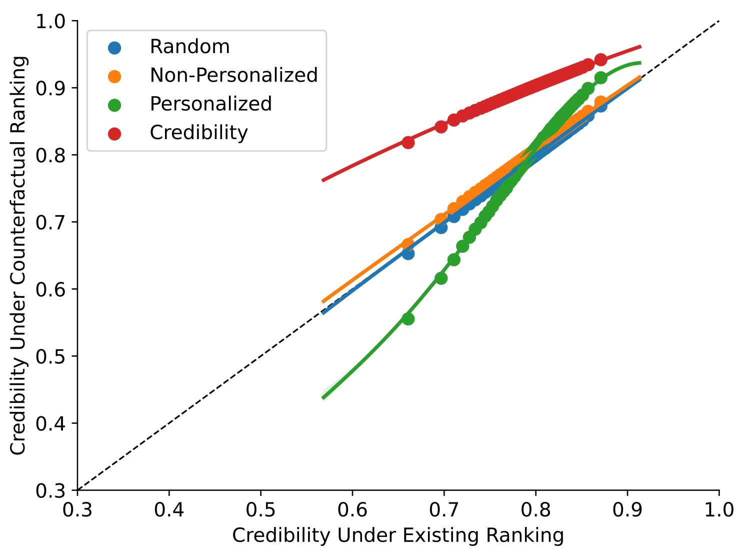
Turning to the credibility-maximizing algorithm, I find that optimizing for engagement with high-credibility publishers leads to substantial increases in the share of engagement with high-credibility publishers across all users. Importantly, though, Figure 6b shows that the users experiencing the largest increases are those who engage more with low-credibility publishers under Reddit's actual algorithm. This indicates that including publisher credibility in the objective function narrows the disparity between users with high- and low-quality news diets, a difference that was exacerbated when optimizing only for engagement.

Figure 6: Impact of Algorithm on Share of Engagement with High-Credibility Publishers

(a) Distribution of Change in High-Credibility Share



(b) High-Credibility Share by Baseline Credibility



Note: (a) Plots the empirical CDF of the change in the share of engagement with high credibility publishers under the counterfactual algorithms relative to the existing algorithm. (b) Plots binned mean credibility shares under the counterfactual algorithm against credibility shares under the existing algorithm. Regression line is a fourth-order polynomial fit.

5.2.3 Engagement-Credibility Trade-Off

The results thus far have compared engagement-maximizing algorithms with a credibility-maximizing algorithm. That said, platforms or society may balance these competing objectives in a more nuanced manner rather than preferring either extreme. I now describe the frontier of possible outcomes as λ, the weight placed on engagement with high-credibility publishers in the credibility-aware ranking algorithm, is varied. Figure 7 plots this trade-off along with points corresponding to the total engagement-maximizing algorithm, credibility-maximizing algorithm, non-personalized engagement-maximizing algorithm, and non-personalized credibility-maximizing algorithm. As can be seen, moving to the credibility maximizing algorithm reduces engagement by 5.0%. Nevertheless, platforms can achieve over half of the increase in news diet quality from the credibility-maximizing algorithm for a 1.9% decrease in engagement. This change in engagement is similar in magnitude to the difference between the non-personalized engagement-maximizing algorithm and the personalized engagement-maximizing algorithm. However, the non-personalized algorithm does not meaningfully improve the quality of users' news diets, while the credibility-aware algorithm increases the average share of engagement with high-credibility publishers by 5.9 percentage points for approximately the same total quantity of engagement.

The shape of this frontier is also important, as the gradient is relatively flat around the engagement maximizing algorithms. This suggests that, for small decreases in engagement, the platform can drastically increase the share of engagement with high-quality publishers. However, this also means that small differences in preferences between the platform and society can lead to large discrepancies in outcomes along the credibility dimension – again highlighting the importance of aligning the ranking algorithm's objective function.

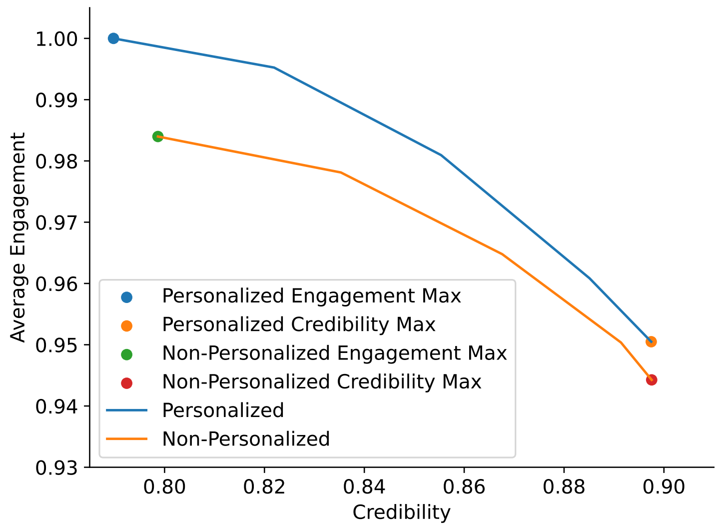
5.2.4 Additional Results

I now summarize additional impacts of the counterfactual ranking algorithms. First, I discuss the impact of the ranking algorithms on the quality of discussion which I operationalize through comment sentiment. Second, I discuss the heterogeneous impact on publisher market shares to understand how the various ranking algorithms impact the incentives of publishers. Finally, I discuss how the ranking algorithms impact the diversity of engagement with publishers across the political spectrum.

First, I report findings on the impact optimizing for engagement has on discussion quality, as measured through the sentiment of comments submitted to the platform. Table 4 demonstrates that both the non-personalized and personalized algorithms slightly elevate the share of negative-sentiment comments submitted by the average user relative to the existing algorithm. Recall a negative-sentiment comment is significantly more likely to contain strongly negative emotions such as anger and disgust and more likely to be classified as toxic. Moreover, inspecting negative comments reveals they are often extremely vulgar and unlikely to contribute to the discussion in a meaningful way.

While the effect on the sentiment of the average user is small, personalization increases the

Figure 7: Engagement-Quality Frontier



Note: This figure plots the frontier of possible outcomes when varying λ in the credibility aware algorithm. The y axis is average total engagement and the x axis is the average share of engagement with high credibility publishers. The y-axis is normalized to 1 at its maximal value. Points indicate outcomes under the counterfactual algorithms described in Section 5.1.

variance in sentiment leading to some users commenting more positively while others are shown content that makes them respond negatively (Figure C.20a). Figure C.20b plots the relationship between the change in the negative-sentiment share of users against user preferences for publisher credibility and I find that users who prefer less-credible publishers have a larger increase in their negative-sentiment share. The same is true of users who prefer left-leaning outlets, consistent with the sentiment preference estimates in Figure C.18.

Next, I change the unit of analysis to the publisher and summarize how the counterfactual ranking algorithms impact different types of publishers. Figure C.21 plots the change in publisher market share by publisher slant (Figure C.21a) and publisher credibility (Figure C.21b).[19] Optimizing solely for engagement leads to a reallocation of market share from left-leaning publishers to right-leaning publishers and a slight increase in the market shares of low-credibility publishers. Optimizing for engagement with high-credibility publishers leads to a reallocation of engagement from politically slanted publishers to more neutral publishers and a reallocation from low- to high-credibility publishers.

Finally, I estimate the impact the counterfactual algorithms have on the diversity of engagement across the political spectrum. To do so, I discretize publisher slant into quintiles and calculate the first-order Wasserstein distance of engagement or promotion shares across these five bins of publisher slant relative to a uniform distribution. This distance metric is better suited to this setting relative

[19]Here, publisher market share is defined as a publisher's share of total engagement in the counterfactuals. This differs from how publisher market share would traditionally be defined, and one should think of market share in this context as the share of traffic from the platform.

to other common measures of diversity used in the literature, including the Herfindahl-Hirschman Index and Shannon Entropy, due to the ordered nature of slant partitions. For example, the Wasserstein distance between a user's engagement and the uniform distribution is larger (i.e. less diverse) for a user who engages only with publishers from a politically slanted partition versus a user only engaging with moderate publishers. The distance is minimized when users engage equally with publishers from all slant partitions and is largest when only engaging with publishers from a politically extreme partition.

The counterfactuals suggest that the random and non-personalized engagement-maximizing algorithms lead to slight increases in engagement diversity relative to the actual algorithm. That said, the personalized algorithm results in a decline in engagement diversity. This decline occurs for a large majority of users, with 71.6% of users experiencing a decline in their engagement diversity in the personalized engagement-maximizing counterfactual. Turning to the credibility-maximizing algorithm, I also find a decrease in the diversity of engagement as the average distance to uniform engagement shares rose by 3.0%.

5.3 Robustness of Findings

5.3.1 Robustness to Alternative Method of Personalization and External Validity

The results reported thus far are based on a micro-founded model of user engagement decisions. This approach requires many assumptions and here I seek to show the results are not sensitive to the assumptions made. To do so, I summarize the findings of an analysis reported in Appendix D that analyzes the type of publishers that are promoted when personalizing the ranking algorithm to maximize engagement using a reduced form collaborative-filtering based recommender system. The recommender system then recommends publishers on which a user is most likely to comment in a period. I validate this recommender system by estimating heterogeneous treatment effects when the regression discontinuity experiments align with the recommender system's predictions. I find that the recommender system effectively predicts treatment effects, a result that suggests the model has learned important aspects of user preferences. I then study the types of content that gets promoted under this simple recommender system to understand the extent to which personalized engagement maximization impacts individual news diets.

This approach has two advantages over the discrete choice model and counterfactual analysis I study in Section 4 and Section 5. First, this model is trained using comment decisions from over 500,000 users and is evaluated on comment decisions of over 180,000 users. This is a much larger sample than that used in the choice model approach, as I can use comment decisions on articles during periods not captured in Wayback Machine snapshots during the training process. Second, this approach relies on a different set of assumptions than the model of engagement decisions and can be validated by predicting treatment effects of which the model is predictive. I find consistent results across both approaches when comparable, which gives confidence that the findings of the choice model approach can be generalized to a broader set of users. That said, I emphasize the model-based approach as the main findings given it allows me to study in more detail the implications of various algorithm designs, including the credibility-aware ranking algorithms, on engagement rather than simply studying what content the algorithm would have recommended.

5.3.2 Robustness to Endogenous Search

A potential limitation of the analysis of counterfactual ranking algorithms presented above is the partial equilibrium nature of the analysis. That is, in equilibrium many things might adjust including user attention, the supply of articles, and the slate of existing comments on each article. Here I demonstrate the findings are not sensitive to endogenous attention allocation by allowing for endogenous search.

Recall in Section 4.1, a user is exposed to an article if $v_{ijt} = 1$ where v_{ijt} is an independent Bernoulli draw with probability $p(r,t)$. Here, I allow for a more flexible model of exposure where I model v_{ijt} as the composite of two random variables

$$v_{ijt} = v_{it} 1 \left[\bar{u}_{ir_j} > c_{ir_j} + \eta_{ijt} \right]$$

where v_{it} is an independent Bernoulli draw equal to one if user i is active in period t, $\bar{u}_{ir} = E\left[\max\left\{u_{ijt}, u_{i0t}\right\} | r_j = r\right]$ is the expected utility from viewing an article in position r, c_{ir} is the mean search cost of viewing an article in position r, and η_{ijt} is an idiosyncratic search cost. I assume that $\eta_{ijt} \sim N(0,1)$ and $\eta_{ijt} \perp v_{it}, x_{jt}, \xi_{jt}, \varepsilon_{ijt}$. I further assume for simplicity that search costs are such that there is no heterogeneity in user exposure probabilities (i.e. $c_{ir} = \bar{u}_{ir} - \Phi^{-1}(p_r)$). In equilibrium, this model of exposure is equivalent to the model in Section 4.1 but it allows for users to reoptimize p_r – the probability of being exposed to an article in position r conditional on being active – in response to the alternative ranking algorithms.

Under this model of exposure, the treatment effect of rank on engagement could come from two channels. First, articles in different positions have different search costs (i.e. $c_{ir} \neq c_{ir+1}$). For example, one may expect it to be more costly for users to view articles ranked lower on the page. Second, an article being promoted from position $r+1$ to position r would induce an update in user beliefs about the expected utility from viewing that article (i..e $\bar{u}_{ir} \neq \bar{u}_{ir+1}$). Under a counterfactual ranking algorithm, only this latter channel would change in equilibrium as users rationally update their beliefs about the expected utility they would receive from viewing articles in different positions. Therefore, understanding which channel drives the treatment effects I observe is important for understanding how attention may adjust in equilibrium.

Figure C.23 plots the components of the search model with panel C.23a showing the average $\bar{u}_{ir}$ by article position and counterfactual ranking algorithm and panel C.23b showing the average c_{ir} by article position. As expected, both channels contribute to the estimated treatment effects with the average $\bar{u}_{ir}$ declining with rank and the average search cost increasing with rank. Moreover, for the various counterfactual ranking algorithms I also observe changes to $\bar{u}_{ir}$ as one would expect. The personalized engagement maximizing algorithm by construction has the sharpest decline in $\bar{u}_{ir}$ given the algorithm is explicitly promoting articles with high utility and therefore users update their beliefs about the quality of article they encounter. Also of note is that the credibility maximizing algorithm is non-monotonic. This is because a subset of users prefer low-credibility publishers and these are moved to the bottom of the page by this algorithm. Therefore, the expected quality of articles in these positions is higher for these users.

However, the magnitudes of the changes in $\bar{u}_{ir}$ are small relative to the search costs (Figure C.23b). Therefore, the search costs are the primary driver of the treatment effects and allowing users to update reoptimize their attention across the feed has little impact on the counterfactual results. Figure C.24 plots the average $p_{ir} = P\left(\bar{u}_{ir_j} > c_{ir_j} + \eta_{ijt}\right)$ by rank and counterfactual algorithm and it is clear there is little change in the exposure probabilities and thus the main findings are robust to endogenous attention allocation within the feed.

6 Discussion and Conclusion

In this study, I evaluate the impact of optimizing for engagement in social media news feed algorithms on user news diets. To address this question, I exploit a regression discontinuity revealed in the platform's code to identify the causal effect of rank on engagement and use these causal estimates to identify a model of user engagement. Using this model, I estimate engagement patterns under counterfactual ranking algorithms including personalized and non-personalized engagement-maximizing and a credibility-aware algorithm that explicitly trades-off total engagement and engagement with high-credibility publishers.

The counterfactual analysis demonstrates that social media platforms have a strong incentive to optimize their ranking algorithms for engagement. Optimizing for engagement leads to a dramatic increase in the quantity of engagement and much of this results from promoting posts with which all users are likely to engage. The marginal benefit of personalizing feeds is modest in terms of engagement quantity but has substantial impacts on the credibility and diversity of publishers with which users engage. In particular, personalized engagement maximization exacerbates differences in the quality of user news diets. That is, the personalized engagement-maximizing ranking increases the inequality in the share of high-credibility engagement between users engaging with lower-credibility publishers and those engaging with higher-credibility publishers under the existing algorithm.

Advertiser concerns about brand safety give platform managers a direct motive to promote credible publishers. Many advertisers seek to avoid advertising on platforms that promote content that is inconsistent with their values or that would create backlash from their consumers. For example, the #StopHateForProfit movement led over 1,000 large advertisers to halt or reduce advertising on Facebook to pressure the platform to expand its efforts to combat hate speech and misinformation [Hsu and Friedman, 2020]. There is also evidence suggesting that firms advertising on platforms alongside misinformation can experience customer backlash [Ahmad et al., 2023]. The credibility-aware algorithm demonstrates one method managers can use to improve the credibility of news content that is promoted on their platforms. The gradient of the engagement-credibility frontier indicates that moving away from the engagement-maximizing algorithm and towards the credibility-maximizing algorithm incurs a relatively small cost in terms of lost engagement. However, this also implies that small differences in the preferences of the platform and society can generate large changes in the amount of engagement with low-credibility publishers despite reasonably small changes to total engagement.

In addition, these results are also relevant for managers of publishers and the incentives they face when advertising revenue on traffic originating from social media referrals comprises an important component of their income. I find that personalized engagement maximization benefits publishers with a strong conservative slant and those producing low-credibility journalism. This introduces an incentive for publishers to change their coverage to match the increased demand for politically slanted and low-credibility journalism.

Finally, these results have implications for regulating digital platforms. A growing regulatory trend is to require or incentivize platforms to allow users to opt-out of personalized recommendations or feeds. Examples include the European Union's Digital Services Act or proposed legislation (such as the Filter Bubble Transparency Act, Justice Against Malicious Algorithms Act, and the Protecting Americans from Dangerous Algorithms Act) in the United States. The findings presented here suggest the emphasis on allowing users to opt out of personalization may be misguided. Rather, as the results show, personalization has substantial benefits when the objective function aligns with society's preferences. Recall that for approximately the same level of engagement as the non-personalized engagement-maximizing algorithm, the credibility-aware algorithm can increase the share of the average user's engagement with high-credibility publishers by 5.9 percentage points. A solution that takes advantage of the benefits of personalization, while protecting individual autonomy, would be to allow users to adjust the weights on different components within the ranking algorithm objective function, including the weight placed on publisher credibility. What weights users would choose and the results of such a design remain open questions and merit future work. Alternatively, regulators could incentivize platforms to align their ranking objective function with the preferences of society to take advantage of personalized ranking algorithms' substantial benefits while mitigating potential negative effects, though implementing such a policy would face substantial legal and ethical challenges.

References

Wajeeha Ahmad, Ananya Sen, Charles Eesley, and Erik Brynjolfsson. The role of advertisers and platforms in monetizing misinformation: Descriptive and experimental evidence. Technical report, 2023. 1, 1, 6

Hunt Allcott and Matthew Gentzkow. Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, 31(2):211–236, 2017. 1

Hunt Allcott, Matthew Gentzkow, and Lena Song. Digital addiction. *American Economic Review*, 112(7):2424–63, 2022. 1, 5.1

Guy Aridor, Duarte Gonçalves, Daniel Kluver, Ruoyan Kong, and Joseph Konstan. The economics of recommender systems: Evidence from a field experiment on movielens. *arXiv preprint arXiv:2211.14219*, 2022. 1

Eytan Bakshy, Solomon Messing, and Lada A Adamic. Exposure to ideologically diverse news and opinion on facebook. *Science*, 348(6239):1130–1132, 2015. 1, 2.2.2

Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. The pushshift reddit dataset. In *Proceedings of the international AAAI conference on web and social media*, volume 14, pages 830–839, 2020. 2.2.1

Steven Berry, James Levinsohn, and Ariel Pakes. Automobile prices in market equilibrium. *Econometrica: Journal of the Econometric Society*, pages 841–890, 1995. 4.2.2

Steven T Berry. Estimating discrete-choice models of product differentiation. *The RAND Journal of Economics*, pages 242–262, 1994. 13, 4.2.2

Ceren Budak, Sharad Goel, and Justin M Rao. Fair and balanced? quantifying media bias through crowdsourced content analysis. *Public Opinion Quarterly*, 80(S1):250–271, 2016. 2.2.2

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. 2.2.1

Sebastian Calonico, Matias D Cattaneo, and Rocio Titiunik. Robust nonparametric confidence intervals for regression-discontinuity designs. *Econometrica*, 82(6):2295–2326, 2014. 3.1

Matias D. Cattaneo, Nicolás Idrobo, and Rocío Titiunik. *A Practical Introduction to Regression Discontinuity Designs: Foundations*. Elements in Quantitative and Computational Methods for the Social Sciences. Cambridge University Press, 2020. doi: 10.1017/9781108684606. 3.1, 3.1

Matias D. Cattaneo, Nicolas Idrobo, and Rocio Titiunik. A practical introduction to regression discontinuity designs: Extensions, 2023. 3.1

Guangying Chen, Tat Chan, Dennis Zhang, Senmao Liu, and Yuxiang Wu. The effects of diversity in algorithmic recommendations on digital content consumption: A field experiment. *Available at SSRN 4365121*, 2023. 1

Yingying Dong and Michal Kolesár. When can we ignore measurement error in the running variable? *Journal of Applied Econometrics*, 38(5):735–750, 2023. 3.1

Robert Donnelly, Ayush Kanodia, and Ilya Morozov. Welfare effects of personalized rankings. *Marketing Science*, 2023. 1, 14

Jack Dorsey. They simply try to put the tweets that you're *most likely* to engage with at the top... `https://twitter.com/jack/status/1525662031440924672`, May 2022. Tweet. 1

Erwan Dujeancourt and Marcel Garz. The effects of algorithmic content selection on user engagement with news on twitter. *The Information Society*, 39(5):263–281, 2023. 1

Dean Eckles. Algorithmic transparency and assessing effects of algorithmic ranking. 2022. 1

Dean Eckles, René F Kizilcec, and Eytan Bakshy. Estimating peer effects in networks with peer encouragement designs. *Proceedings of the National Academy of Sciences*, 113(27):7316–7322, 2016. 2.2.1

Daniel Fleder and Kartik Hosanagar. Blockbuster culture's next rise or fall: The impact of recommender systems on sales diversity. *Management science*, 55(5):697–712, 2009. 1

Anindya Ghose, Panagiotis G Ipeirotis, and Beibei Li. Examining the impact of ranking on consumer behavior and search engine revenue. *Management Science*, 60(7):1632–1654, 2014. 1

Nir Grinberg, Kenneth Joseph, Lisa Friedland, Briony Swire-Thompson, and David Lazer. Fake news on twitter during the 2016 us presidential election. *Science*, 363(6425):374–378, 2019. 1

Andrew Guess, Jonathan Nagler, and Joshua Tucker. Less than you think: Prevalence and predictors of fake news dissemination on facebook. *Science Advances*, 5(1):eaau4586, 2019. 1

Andrew M Guess, Brendan Nyhan, and Jason Reifler. Exposure to untrustworthy websites in the 2016 us election. *Nature Human Behaviour*, 4(5):472–480, 2020. 1

Andrew M Guess, Neil Malhotra, Jennifer Pan, Pablo Barberá, Hunt Allcott, Taylor Brown, Adriana Crespo-Tenorio, Drew Dimmery, Deen Freelon, Matthew Gentzkow, et al. How do social media feed algorithms affect attitudes and behavior in an election campaign? *Science*, 381(6656): 398–404, 2023. 1

Frances Haugen. Statement of frances haugen. *United States Senate Committee on Commerce, Science and Transportation*, 2021. 1

David Holtz, Ben Carterette, Praveen Chandar, Zahra Nazari, Henriette Cramer, and Sinan Aral. The engagement-diversity connection: Evidence from a field experiment on spotify. In *Proceedings of the 21st ACM Conference on Economics and Computation*, pages 75–76, 2020. 1

Kartik Hosanagar, Daniel Fleder, Dokyun Lee, and Andreas Buja. Will the global village fracture into tribes? recommender systems and their effects on consumer fragmentation. *Management Science*, 60(4):805–823, 2014. 1

Tiffany Hsu and Gillian Friedman. Facebook boycott: Starbucks and diageo to pull ads. *The New York Times*, 2020. URL `https://www.nytimes.com/2020/06/26/business/media/Facebook-advertising-boycott.html`. 6

Yifan Hu, Yehuda Koren, and Chris Volinsky. Collaborative filtering for implicit feedback datasets. In *2008 Eighth IEEE international conference on data mining*, pages 263–272. Ieee, 2008. D.1

Ferenc Huszár, Sofia Ira Ktena, Conor O'Brien, Luca Belli, Andrew Schlaikjer, and Moritz Hardt. Algorithmic amplification of politics on twitter. *Proceedings of the National Academy of Sciences*, 119(1):e2025334119, 2022. 1

Daniel Kahneman. *Thinking, fast and slow*. macmillan, 2011. 5.1

Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. The challenge of understanding what users want: Inconsistent preferences and engagement optimization. *Management Science*, 2023. 1, 5.1

Adam DI Kramer, Jamie E Guillory, and Jeffrey T Hancock. Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences*, 111(24):8788–8790, 2014. 1

David MJ Lazer, Matthew A Baum, Yochai Benkler, Adam J Berinsky, Kelly M Greenhill, Filippo Menczer, Miriam J Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, et al. The science of fake news. *Science*, 359(6380):1094–1096, 2018. 1

Dokyun Lee and Kartik Hosanagar. How do recommender systems affect sales diversity? a cross-category investigation via randomized field experiment. *Information Systems Research*, 30(1): 239–259, 2019. 1

Ro'ee Levy. Social media, news consumption, and polarization: Evidence from a field experiment. *American Economic Review*, 111(3):831–70, 2021. 1

Hause Lin, Jana Lasser, Stephan Lewandowsky, Rocky Cole, Andrew Gully, David Rand, and Gordon Pennycook. High level of agreement across different news domain quality ratings. 2022. 1, 2.2.2, 7

John DC Little. A proof for the queuing formula: L= λ w. *Operations research*, 9(3):383–387, 1961. 4.2.2, C.3

Lev Muchnik, Sinan Aral, and Sean J Taylor. Social influence bias: A randomized experiment. *Science*, 341(6146):647–651, 2013. 3.1

Simha Mummalaneni, Hema Yoganarasimhan, and Varad V Pathak. Producer and consumer engagement on social media platforms. *Available at SSRN 4173537*, 2022. 2.2.1

Arvind Narayanan. Understanding social media recommendation algorithms. *Knight First Amendment Institute*, 2023. URL `https://knightcolumbia.org/content/understanding-social-media-recommendation-algorithms`. 1

Sridhar Narayanan and Kirthi Kalyanam. Position effects in search advertising and their moderators: A regression discontinuity approach. *Marketing Science*, 34(3):388–407, 2015. 2, 3

Gal Oestreicher-Singer and Arun Sundararajan. Recommendation networks and the long tail of electronic commerce. *MIS Quarterly*, pages 65–83, 2012. 1

Jeff Orlowski. The social dilemma. Netflix, 2020. URL `https://www.thesocialdilemma.com/`. 1

Eli Pariser. *The filter bubble: How the new personalized web is changing what we read and how we think*. Penguin, 2011. 1

Rajan Patel. Google expands partnership with reddit, 2024. URL https://blog.google/inside-google/company-announcements/expanded-reddit-partnership/. 1

Gordon Pennycook and David G Rand. The psychology of fake news. *Trends in cognitive sciences*, 25(5):388–402, 2021. 1

Juan Manuel Pérez, Juan Carlos Giudici, and Franco Luque. pysentimiento: A python toolkit for sentiment analysis and socialnlp tasks. *arXiv preprint arXiv:2106.09462*, 2021. 2.3

Christian Peukert, Ananya Sen, and Jörg Claussen. The editor and the algorithm: Recommendation technology in online news. *Management science*, 2023. 1

reddit.com. Reddit, 11 2017. URL https://github.com/reddit-archive/reddit. 3

Ronald E Robertson, Shan Jiang, Kenneth Joseph, Lisa Friedland, David Lazer, and Christo Wilson. Auditing partisan audience bias within google search. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):1–22, 2018. 2.2.2

SimilarWeb. Top social media networks websites ranking in march 2024, 2024. URL https://www.similarweb.com/top-websites/computers-electronics-and-technology/social-networks-and-online-communities/. 1

Michael Spence and Bruce Owen. Television programming, monopolistic competition, and welfare. *The Quarterly Journal of Economics*, 91(1):103–126, 1977. 1, 5.1

Luke Thorburn, Priyanjana Bengani, and Jonathan Stray. How platform recommenders work. *Medium*, 2022. URL https://medium.com/understanding-recommenders/how-platform-recommenders-work-15e260d9a15a. 1

Raluca M Ursu. The power of rankings: Quantifying the effect of rankings on online consumer search and purchase decisions. *Marketing Science*, 37(4):530–552, 2018. 3, 14

Marshall Van Alstyne and Erik Brynjolfsson. Global village or cyber-balkans? modeling and measuring the integration of electronic communities. *Management Science*, 51(6):851–868, 2005. 1

Soroush Vosoughi, Deb Roy, and Sinan Aral. The spread of true and false news online. *Science*, 359(6380):1146–1151, 2018. 1

Yuyan Wang, Long Tao, and Xing Zhang. Recommending for a three-sided food delivery marketplace: A multi-objective hierarchical approach. Technical report, Working Paper, 2023. 1

Mark R. Warner. Warner presses meta on facebook's role in inciting violence and spreading misinformation around the world. Press Release, 2023. URL https://www.warner.senate.gov/public/index.cfm/2023/2/warner-presses-meta-on-facebook-s-role-in-inciting-violence-and-spreading-misinformation-ar 1

Stefan Wojcik and Adam Hughes. Sizing up twitter users, April 2019. URL https://www.pewresearch.org/internet/2019/04/24/sizing-up-twitter-users/. 2.2.1

Shuai Zhang, Lina Yao, Aixin Sun, and Yi Tay. Deep learning based recommender system: A survey and new perspectives. *ACM computing surveys (CSUR)*, 52(1):1–38, 2019. D.2

Online Appendix

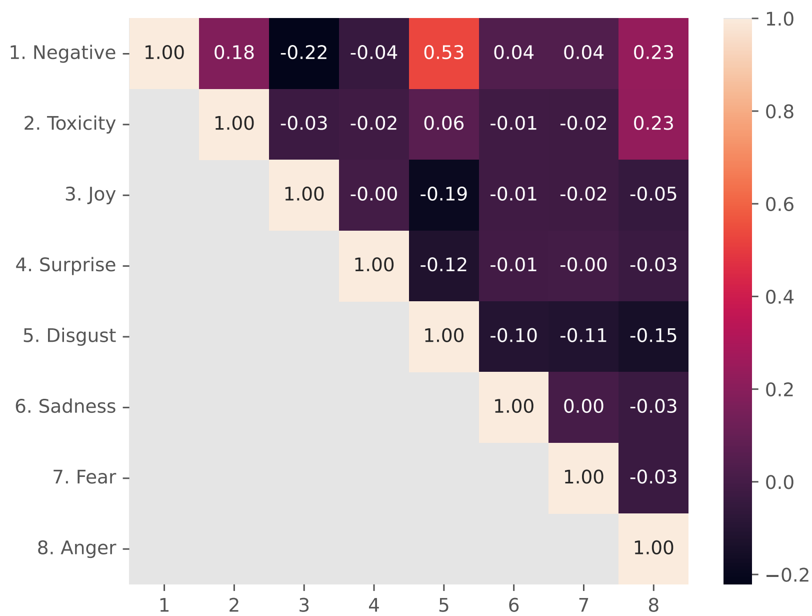
A Data Appendix

Table A.1: Summary of Publisher Ratings

	Slant Score	Credibility
msnbc.com	-0.62	0.59
huffpost.com	-0.31	0.57
nytimes.com	-0.26	0.86
wsj.com	0.01	0.80
foxnews.com	0.61	0.53
breitbart.com	0.74	0.30

Note: Publisher slant and credibility ratings for six widely known publishers.

Figure A.1: Correlation Matrix of Text Features

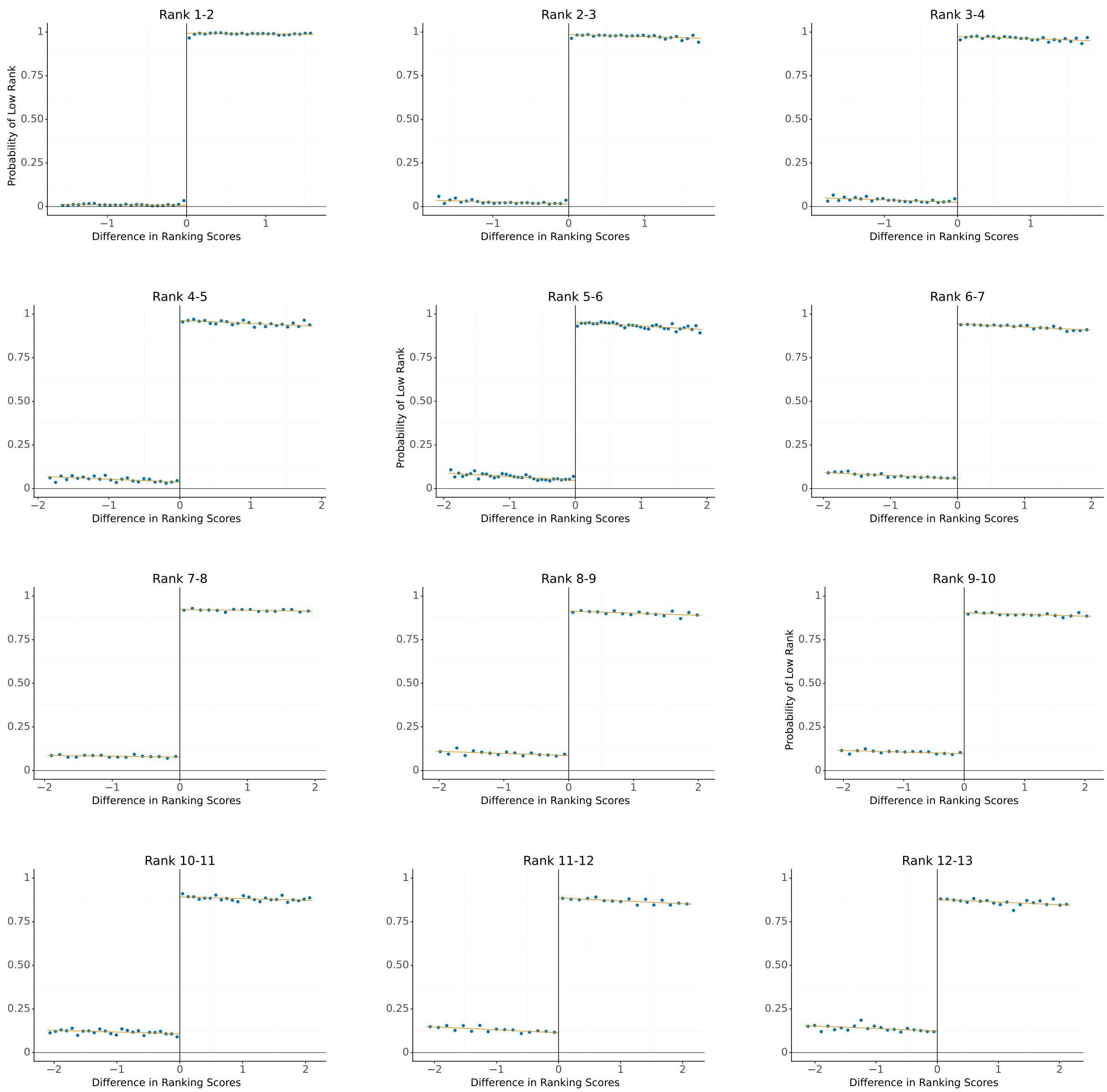


Note: This figure plots the correlation matrix of comment text features. Negative corresponds to negative sentiment, Toxicity corresponds to the predicted toxicity of the comment, while the remaining 6 features correspond to the emotional content of the post.

B Reduced Form Appendix

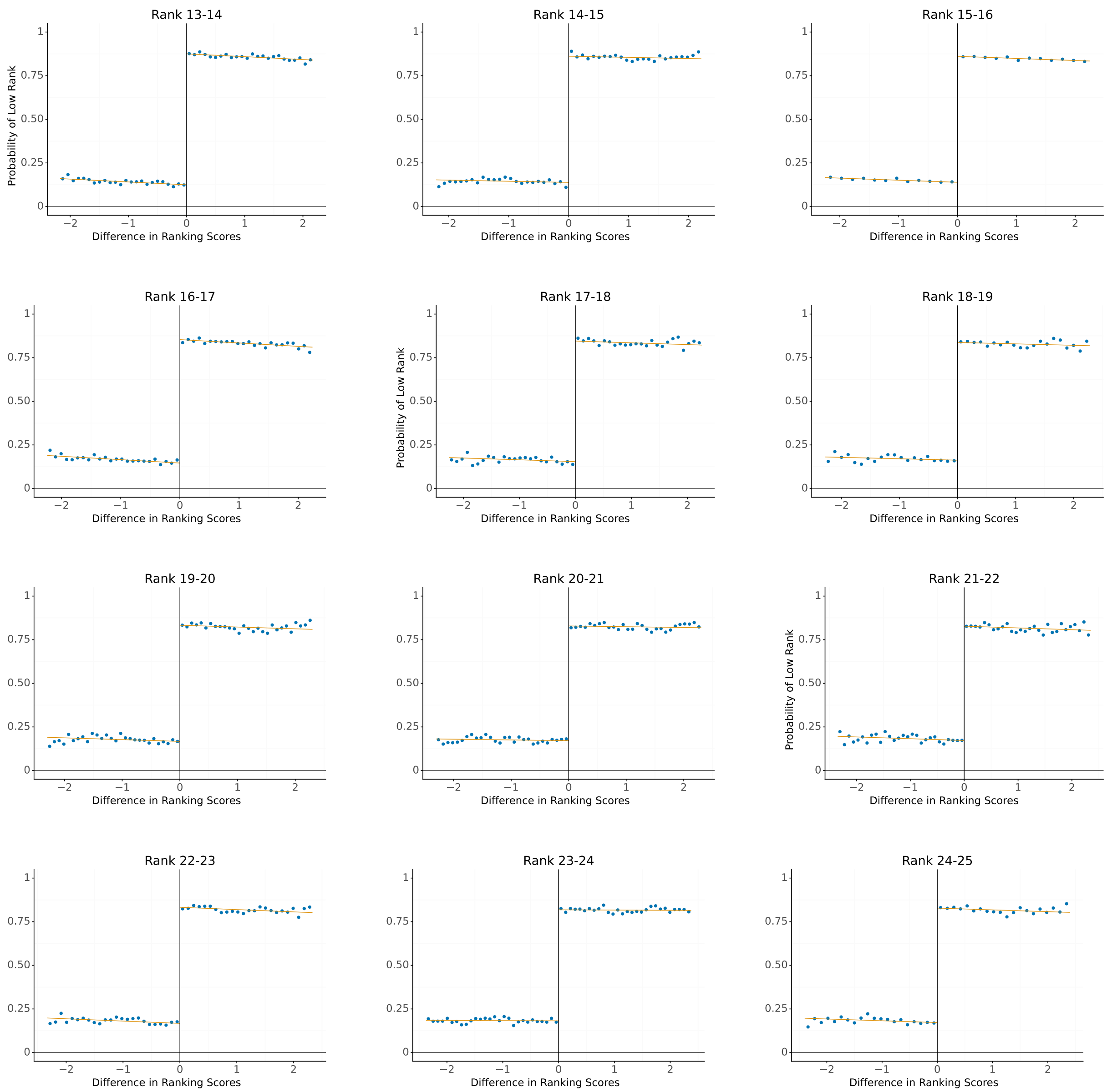
B.1 Additional Figures and Tables

Figure B.2: Regression Discontinuity Plots: First Stage



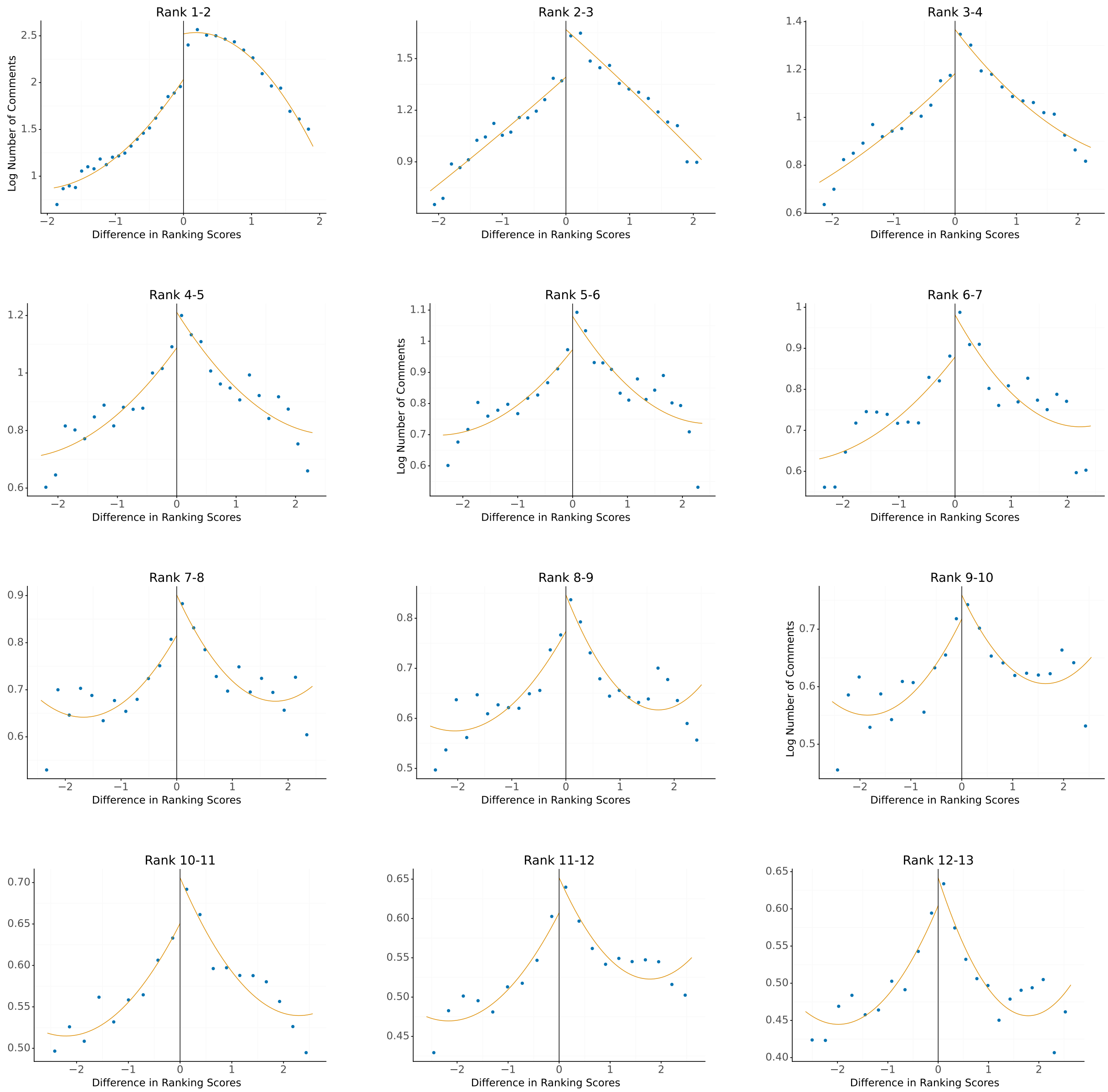
Note: Regression discontinuity first stage plots of the probability a post is ranked lower on the feed against the running variable. This figure excludes posts within the doughnut which includes posts where the absolute value of the running variable is less than 0.05. Fourth order polynomial is plotted alongside the binned mean values.

Figure B.3: Regression Discontinuity Plots: First Stage



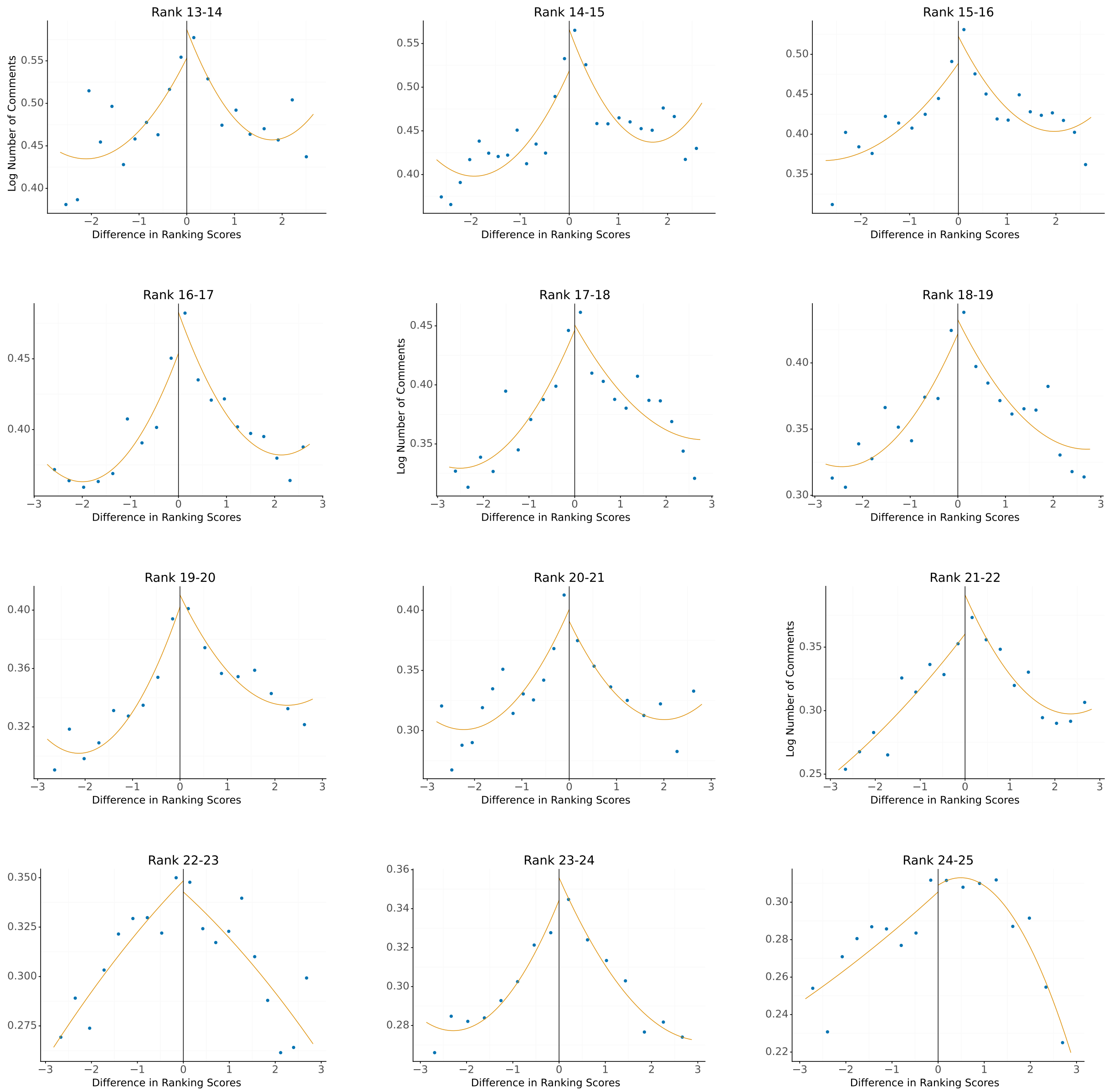
Note: Regression discontinuity first stage plots of the probability a post is ranked lower on the feed against the running variable. This figure excludes posts within the doughnut which includes posts where the absolute value of the running variable is less than 0.05. Fourth order polynomial is plotted alongside the binned mean values.

Figure B.4: Regression Discontinuity Plots: Engagement



Note: Regression discontinuity outcome plots of the log number of comments received in the 20 minutes following a snapshot against the running variable. This figure excludes posts within the doughnut which includes posts where the absolute value of the running variable is less than 0.05. Fourth order polynomial is plotted alongside the binned mean values.

Figure B.5: Regression Discontinuity Plots: Engagement



Note: Regression discontinuity outcome plots of the log number of comments received in the 20 minutes following a snapshot against the running variable. This figure excludes posts within the doughnut which includes posts where the absolute value of the running variable is less than 0.05. Fourth order polynomial is plotted alongside the binned mean values.

Table B.2: Position Effect Estimates

		Regression Discontinuity		
Rank	OLS	Local Linear	Local Constant	Triangular Kernel
1	0.904	0.354	0.208	0.365
	(0.012)	(0.142)	(0.318)	(0.164)
2	0.249	0.244	0.204	0.250
	(0.010)	(0.045)	(0.093)	(0.050)
3	0.145	0.156	0.120	0.160
	(0.009)	(0.033)	(0.068)	(0.037)
4	0.097	0.123	0.128	0.113
	(0.009)	(0.032)	(0.066)	(0.036)
5	0.073	0.107	0.088	0.111
	(0.009)	(0.038)	(0.079)	(0.042)
6	0.067	0.116	0.140	0.106
	(0.008)	(0.037)	(0.077)	(0.040)
7	0.056	0.077	0.062	0.067
	(0.008)	(0.035)	(0.074)	(0.039)
8	0.043	0.071	0.077	0.052
	(0.008)	(0.036)	(0.075)	(0.040)
9	0.043	0.027	0.004	0.028
	(0.007)	(0.031)	(0.064)	(0.034)
10	0.039	0.063	0.075	0.065
	(0.007)	(0.025)	(0.051)	(0.027)
11	0.040	0.055	0.079	0.054
	(0.007)	(0.027)	(0.056)	(0.030)
12	0.016	0.047	0.054	0.054
	(0.007)	(0.035)	(0.074)	(0.038)

Note: Estimates of the local average treatment effect from a post moving from position $r+1$ to position r on the feed on the log of one plus the number of comments a post receives in the 20 minutes following a snapshot. Robust bias-corrected standard errors that allow for misspecification of the conditional expectation function and that are clustered at the period level are shown in parentheses. Estimates exclude posts within the doughnut which includes posts where the absolute value of the running variable is less than 0.05. The bandwidths for each rank are not varied across the various regression discontinuity specifications to isolate the difference due to the different specifications.

Table B.3: Position Effect Estimates

		Regression Discontinuity		
Rank	OLS	Local Linear	Local Constant	Triangular Kernel
13	0.022	0.014	-0.007	0.019
	(0.007)	(0.029)	(0.061)	(0.032)
14	0.037	0.033	0.015	0.026
	(0.006)	(0.027)	(0.057)	(0.029)
15	0.022	0.049	0.082	0.045
	(0.006)	(0.023)	(0.047)	(0.025)
16	0.025	0.033	0.042	0.032
	(0.006)	(0.020)	(0.040)	(0.022)
17	0.018	0.024	0.054	0.022
	(0.006)	(0.022)	(0.045)	(0.024)
18	0.015	0.018	0.029	0.021
	(0.006)	(0.022)	(0.047)	(0.024)
19	0.023	0.006	0.001	0.017
	(0.005)	(0.023)	(0.049)	(0.026)
20	-0.001	-0.035	-0.081	-0.037
	(0.005)	(0.019)	(0.040)	(0.021)
21	0.018	0.024	0.018	0.019
	(0.005)	(0.019)	(0.041)	(0.021)
22	-0.002	-0.007	-0.000	-0.010
	(0.005)	(0.020)	(0.042)	(0.022)
23	0.010	0.020	0.040	0.007
	(0.005)	(0.022)	(0.047)	(0.025)
24	0.016	0.015	0.032	0.024
	(0.005)	(0.023)	(0.049)	(0.025)

Note: Estimates of the local average treatment effect from a post moving from position $r + 1$ to position r on the feed on the log of one plus the number of comments a post receives in the 20 minutes following a snapshot. Robust bias-corrected standard errors that allow for misspecification of the conditional expectation function and that are clustered at the period level are shown in parentheses. Estimates exclude posts within the doughnut which includes posts where the absolute value of the running variable is less than 0.05. The bandwidths for each rank are not varied across the various regression discontinuity specifications to isolate the difference due to the different specifications.

B.2 Robustness of Regression Discontinuity

B.2.1 Regression Discontinuity with Two-Dimensional Score

Recall the running variable in the regression discontinuity analysis is a composition of two continuous scores, the difference in vote scores and the difference in post age. Figure B.6 plots the joint distribution of these two scores along the discontinuity frontier.

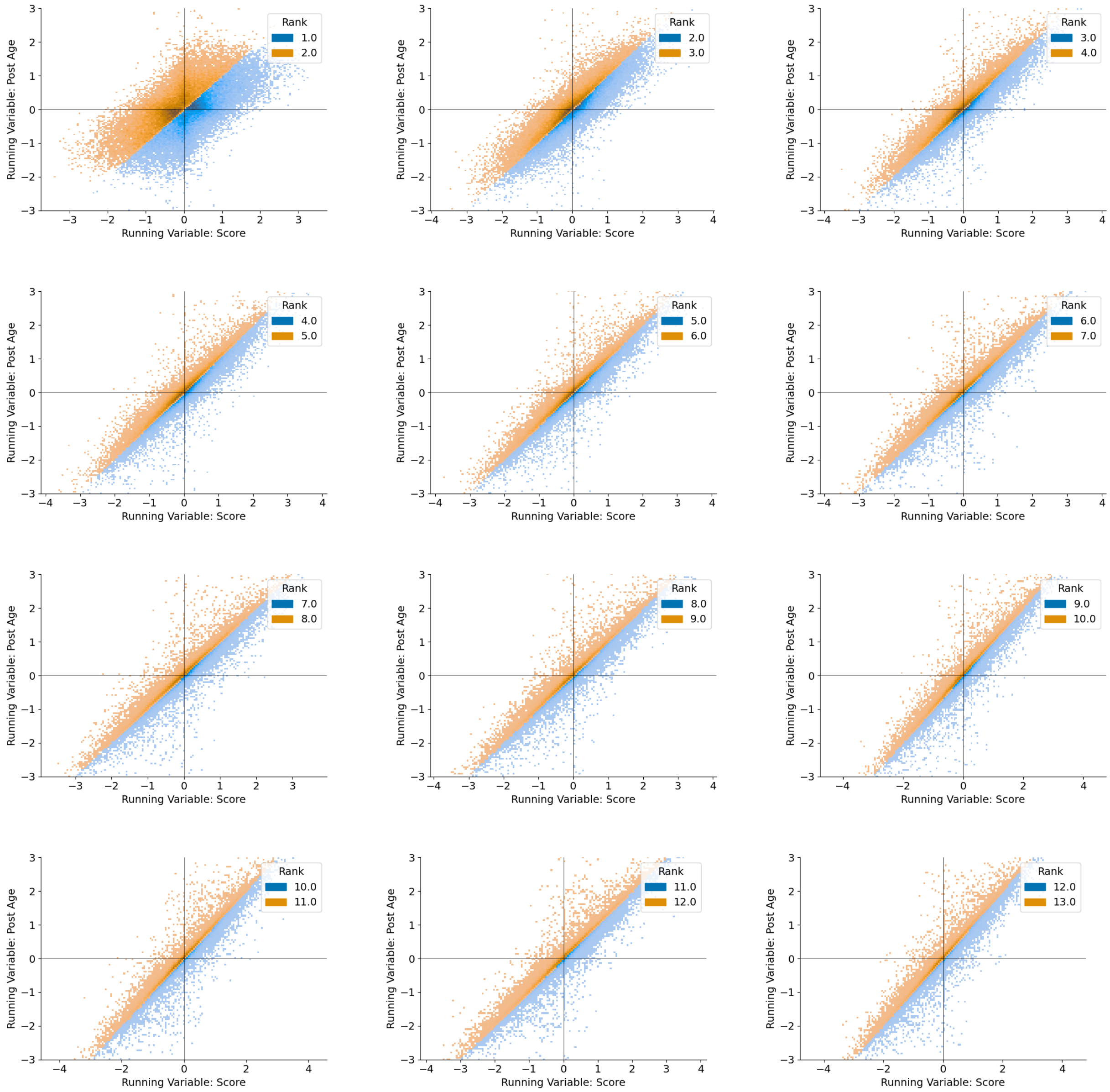
B.2.2 Balance of Covariates

Here, I show evidence that pre-treatment observable post features are continuous through the discontinuity. I show the full regression discontinuity plots for the top 12 positions on the feed for post vote score (Figure B.7), post age (Figure B.8), publisher slant (Figure B.9), and publisher credibility (Figure B.10). Estimates of the local average treatment effect on each of these covariates using local linear regression are displayed in Figure B.11.

B.2.3 Robustness of Bandwidth, Donut, and Comment Window

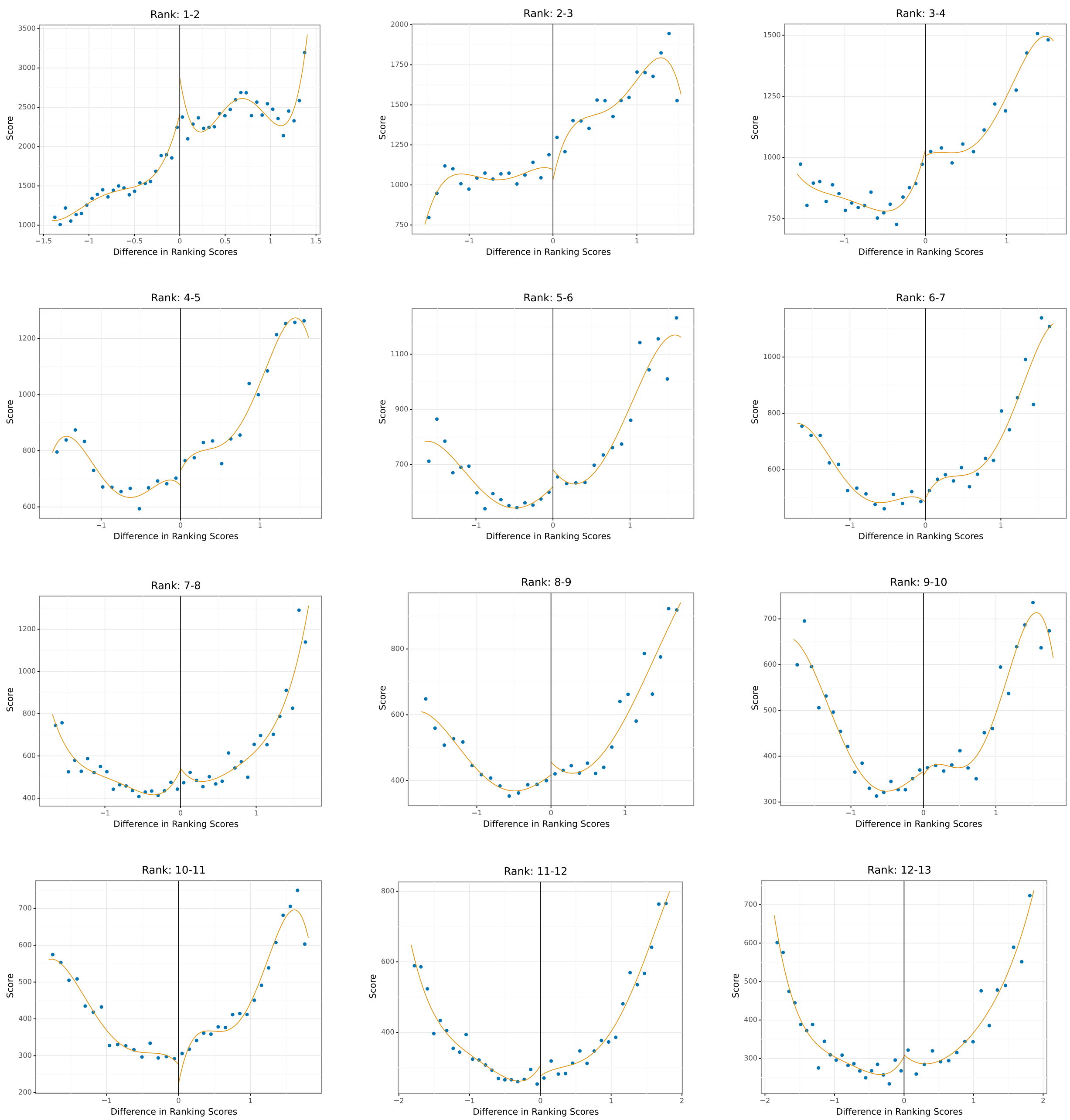
Here, I show the position effect estimates are robust to researcher choices regarding the regression discontinuity bandwidth (Figure B.12), the donut of data excluded around the discontinuity (Figure B.13), and the window of comments included after a post snapshot (Figure B.14).

Figure B.6: Regression Discontinuity with Multiple Scores



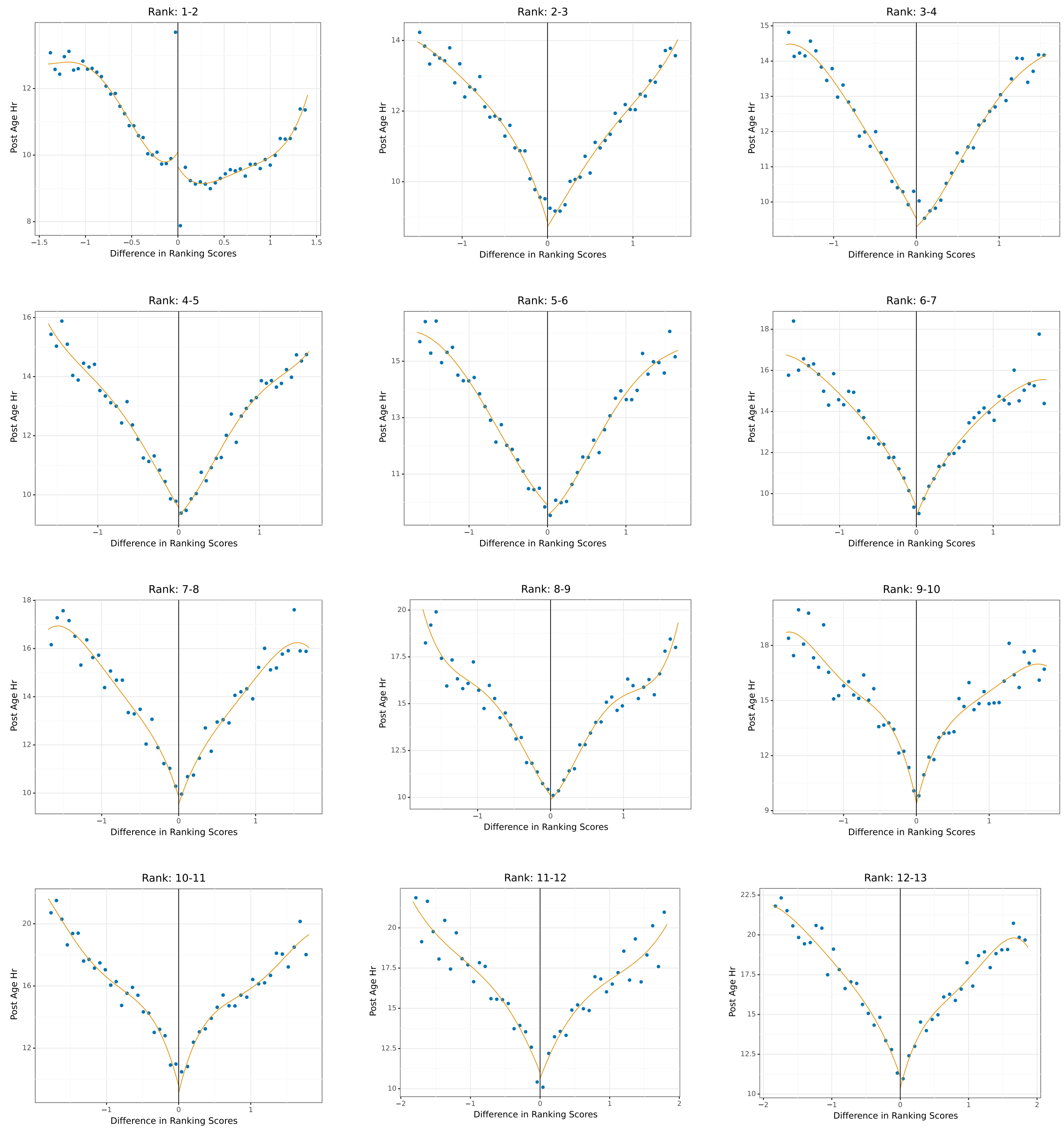
Note: This plot shows the regression discontinuity in two dimensions. The. x axis plots the difference in the normalized post vote scores and the y axis plots the difference in the normalized post ages. The discontinuity frontier corresponds to the 45 degree line. To make the charts easier to view, I only plot posts that are correctly classified by the running variable.

Figure B.7: Balance of Vote Score



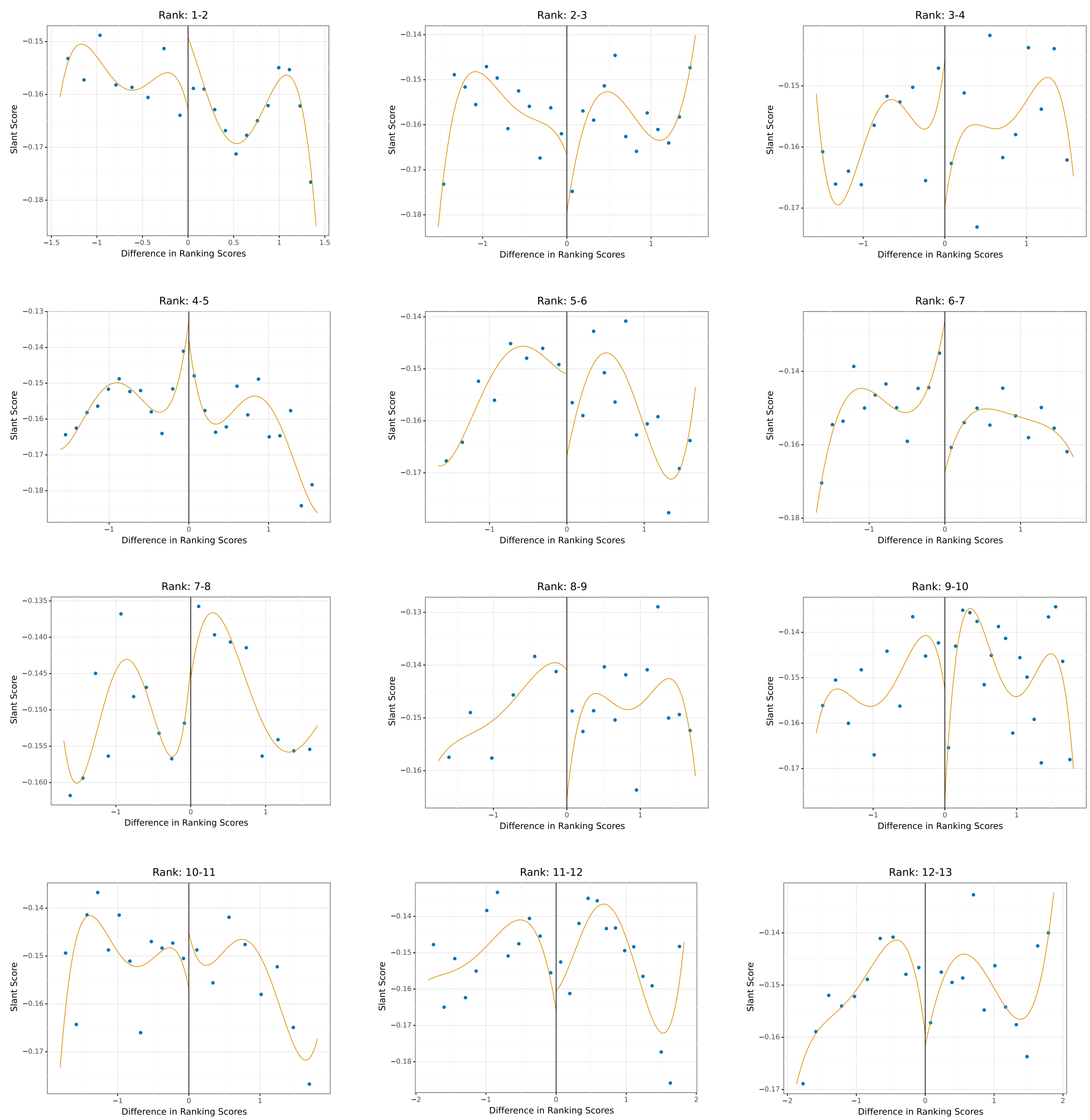
Note: This plot shows the binned means of post vote score against the running variable on the top 12 positions on the feed. This figure excludes posts within the doughnut which includes posts where the absolute value of the running variable is less than 0.05. The line represents the fourth order polynomial fit.

Figure B.8: Balance of Post Age



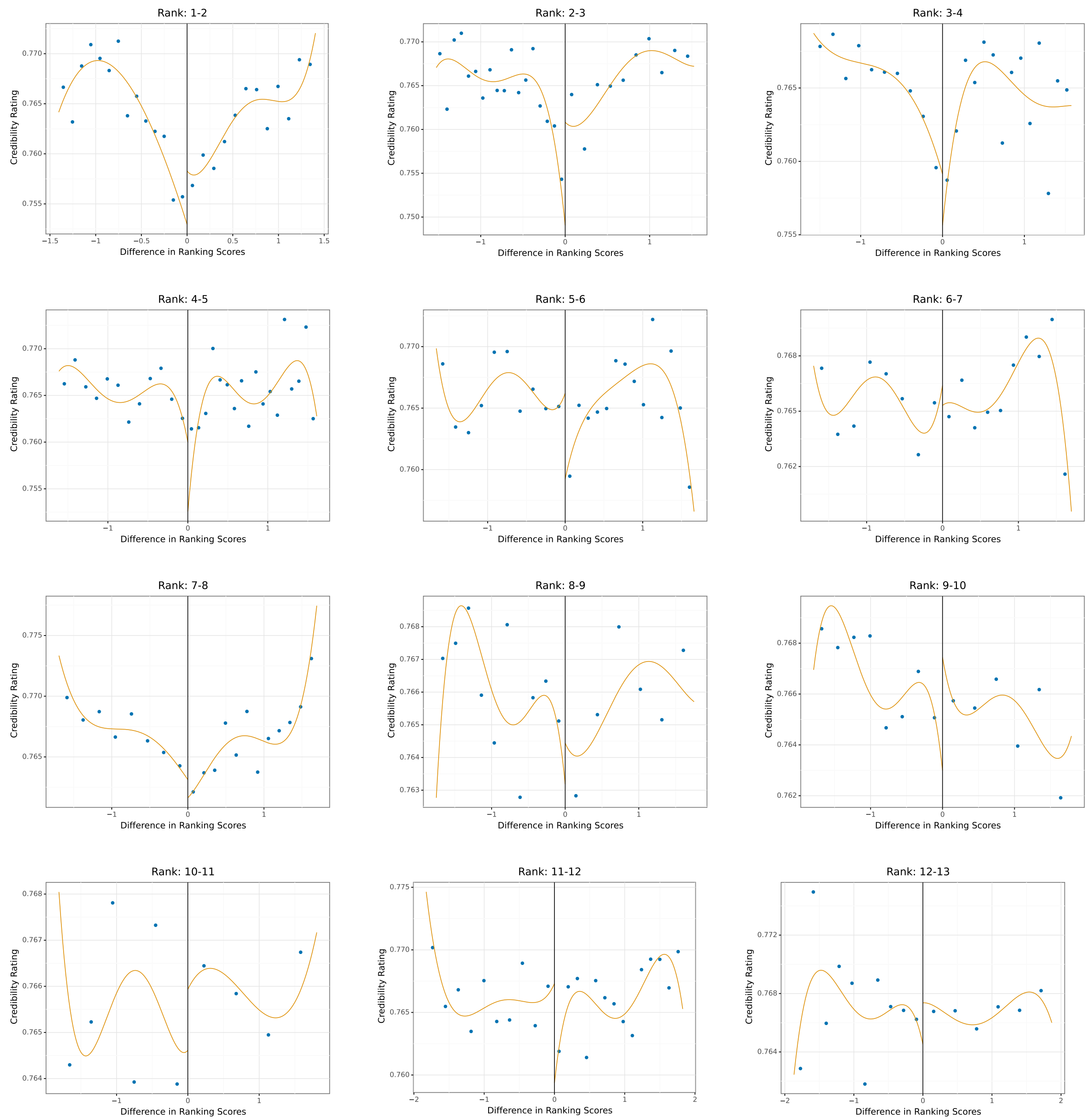
Note: This plot shows the binned means of post age against the running variable on the top 12 positions on the feed. This figure excludes posts within the doughnut which includes posts where the absolute value of the running variable is less than 0.05. The line represents the fourth order polynomial fit.

Figure B.9: Balance of Slant Score



Note: This plot shows the binned means of publisher slant score against the running variable on the top 12 positions on the feed. This figure excludes posts within the doughnut which includes posts where the absolute value of the running variable is less than 0.05. The line represents the fourth order polynomial fit.

Figure B.10: Balance of Credibility Rating

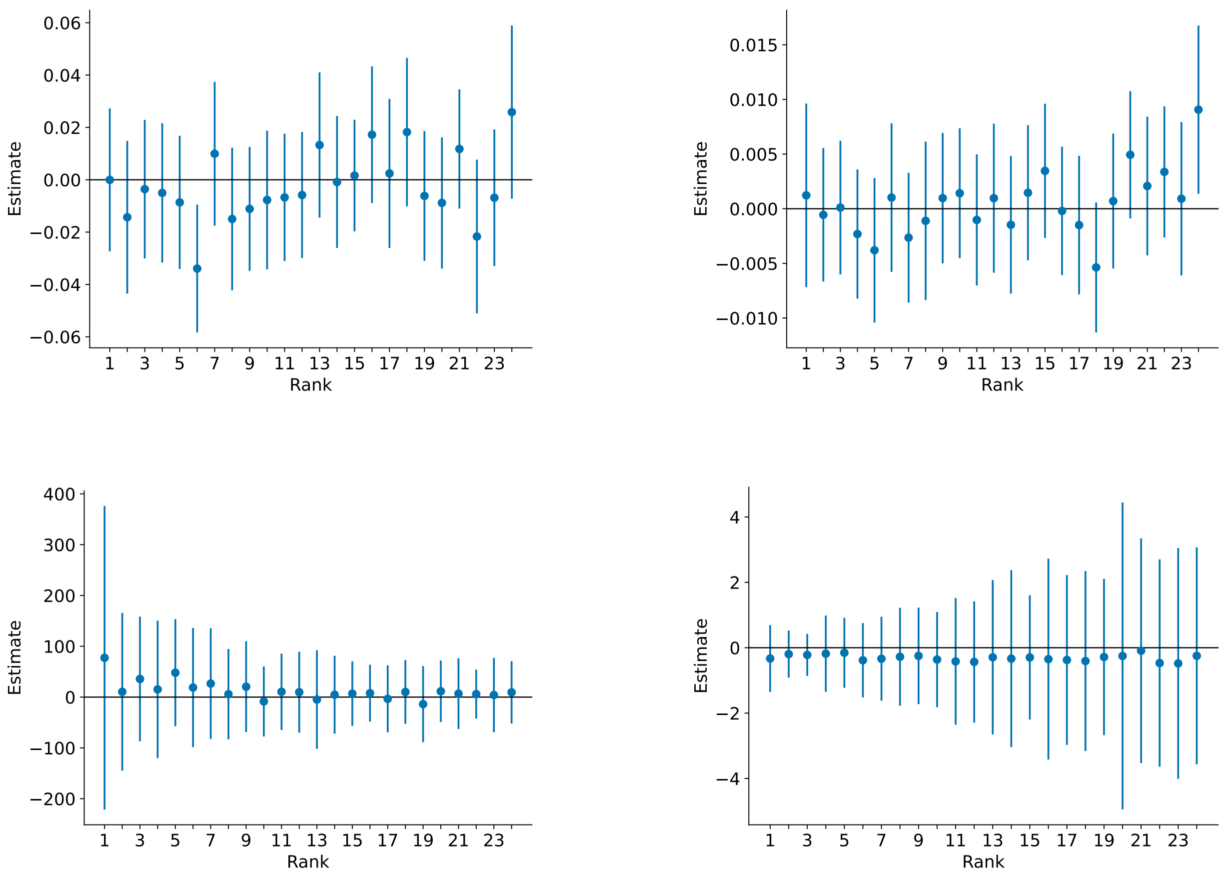


Note: This plot shows the binned means of publisher credibility rating against the running variable on the top 12 positions on the feed. This figure excludes posts within the doughnut which includes posts where the absolute value of the running variable is less than 0.05. The line represents the fourth order polynomial fit.

Figure B.11: Regression Discontinuity Placebo Tests

(a) Publisher Slant

(b) Hour Posted

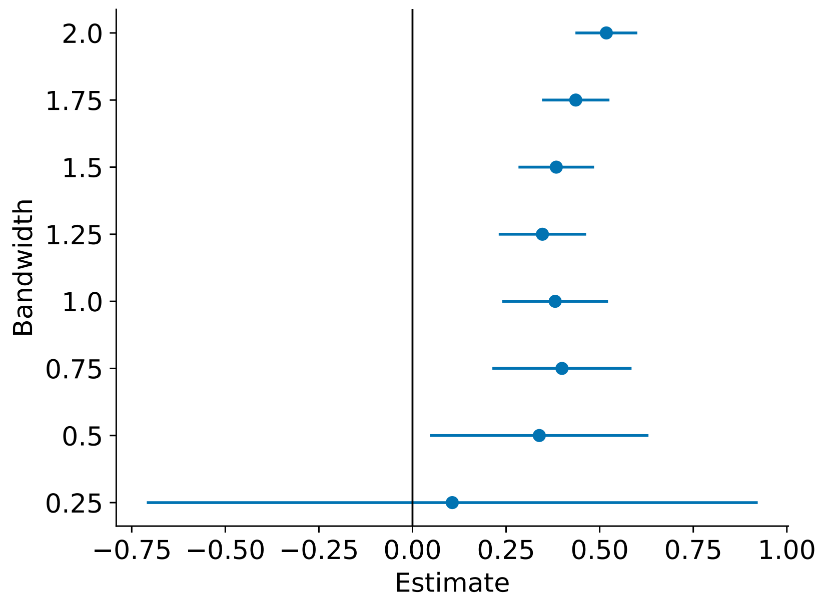


(c) Vote Score

(d) Post Age (hours)

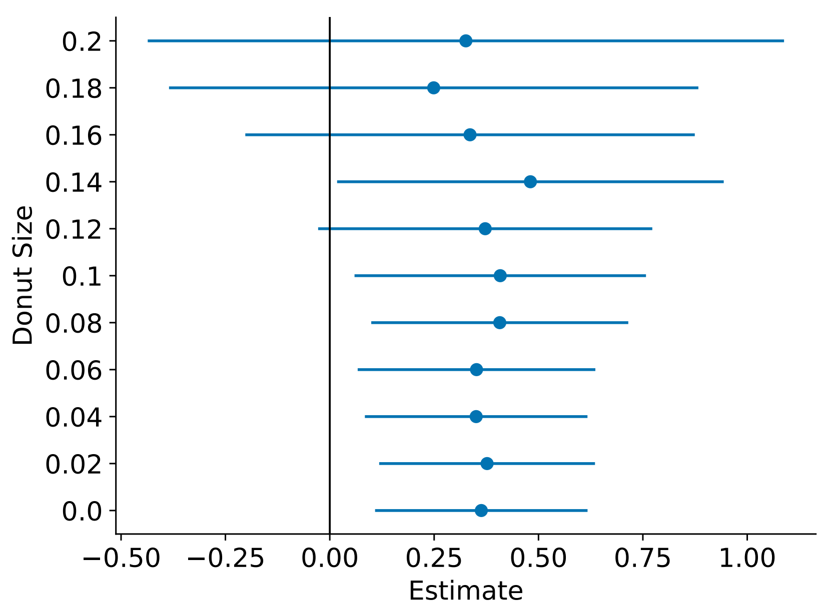
Note: Placebo test for discontinuity of observable pre-treatment covariates. Each figure plots local average treatment effect estimates of moving from rank $r + 1$ to rank r using a local linear regression for publisher slant score, hour posted, vote score, and post age.

Figure B.12: Robustness of Position Effect Estimates to Bandwidth



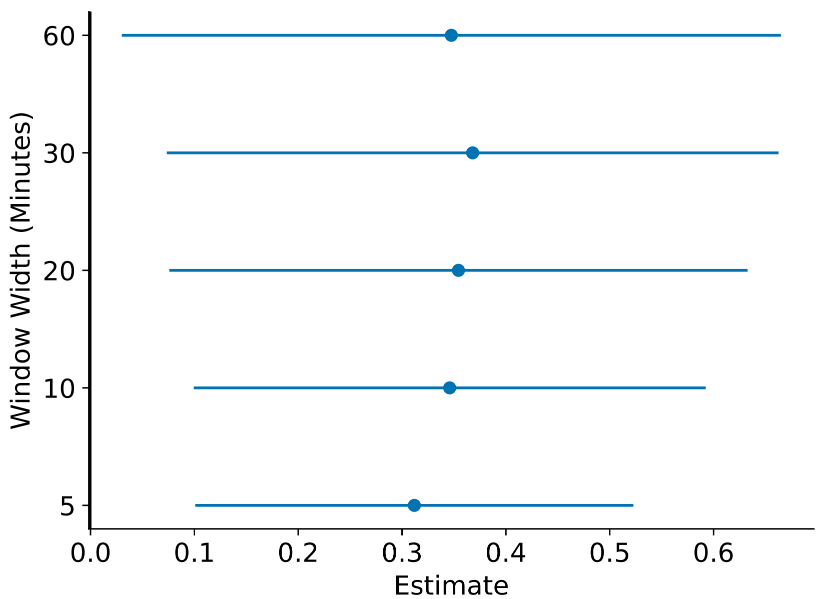
Note: This plot shows the robustness of the position effect estimate to bandwidth size. Each point represents the treatment effect estimate of being promoted to the top position on the feed relative to the second position on the log number of comments a post receives. Error bars represent 95% confidence intervals using robust bias-corrected standard errors.

Figure B.13: Robustness of Position Effect Estimates to Donut Width



Note: This plot shows the robustness of the position effect estimate to donut size. Each point represents the treatment effect estimate of being promoted to the top position on the feed relative to the second position on the log number of comments a post receives. Error bars represent 95% confidence intervals using robust bias-corrected standard errors.

Figure B.14: Robustness of Position Effect Estimates to Comment Window

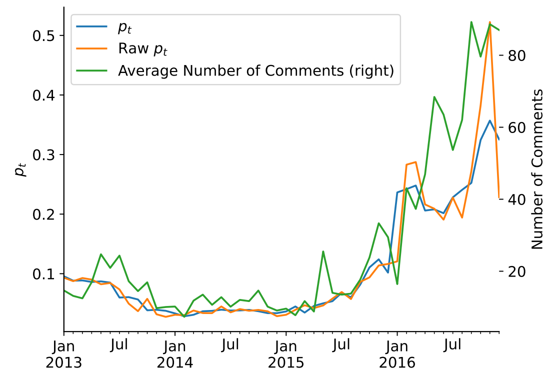


Note: This plot shows the robustness of the position effect estimate to window length. Each point represents the treatment effect estimate of being promoted to the top position on the feed relative to the second position on the log number of comments a post receives in the window following a snapshot. Error bars represent 95% confidence intervals using robust bias-corrected standard errors.

C Choice Model Appendix

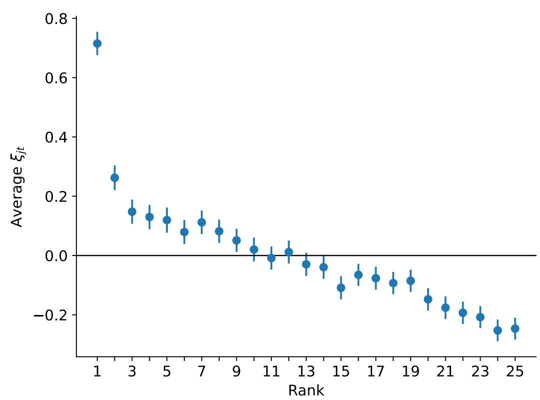
C.1 Additional Figures and Tables

Figure C.15: p_t and Average Top-Post Engagement



Note: This plot shows the time series of the raw and smoothed p_t estimates (left axis). The right axis shows the average number of comments the top post on the feed receives from the users in the sample.

Figure C.16: Average ξ_{jt} by Rank

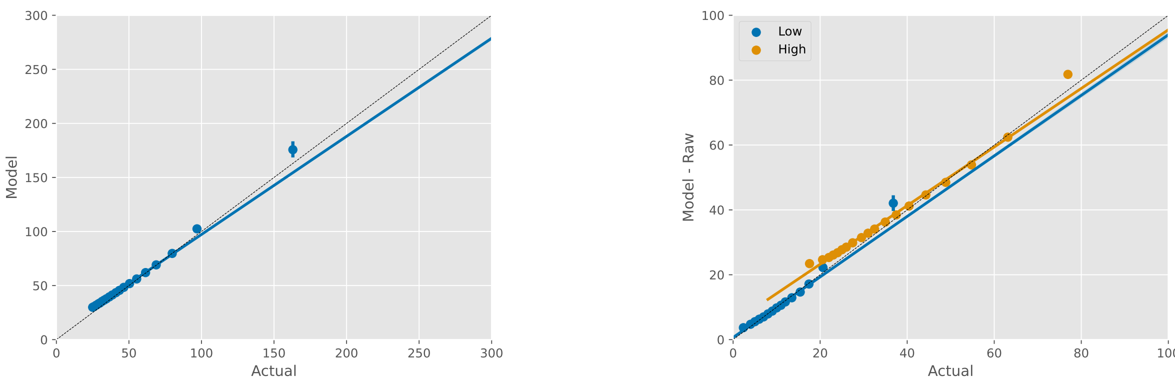


Note: This plot shows the average ξ_{jt} value by post rank for the posts in the sample. Bars represent 95% confidence intervals.

Figure C.17: Summary of Model Fit

(a) Total Number of Comments

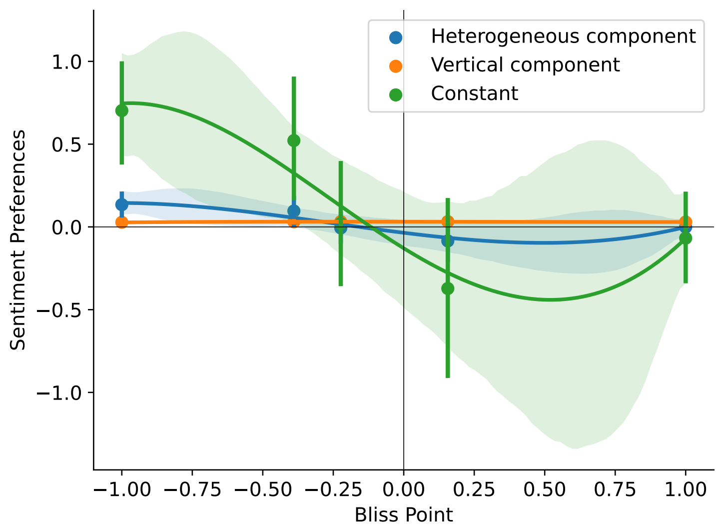
(b) Engagement by Credibility



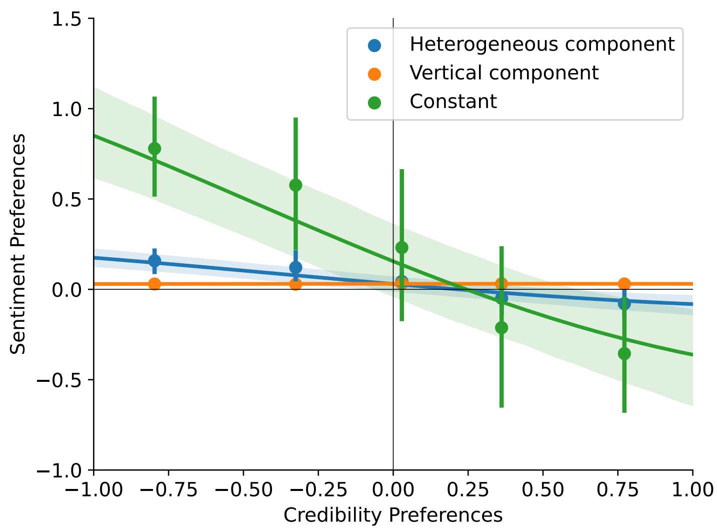
Note: This plot shows additional summaries of the model fit. (a) The relationship between the actual number of comments a user posts against the model fitted number of comments the user submitted. (b) The same relationship, broken out by publisher credibility.

Figure C.18: Correlation of Sentiment Preferences with Comment Preferences

(a) Slant Bliss Point



(b) Credibility Preferences

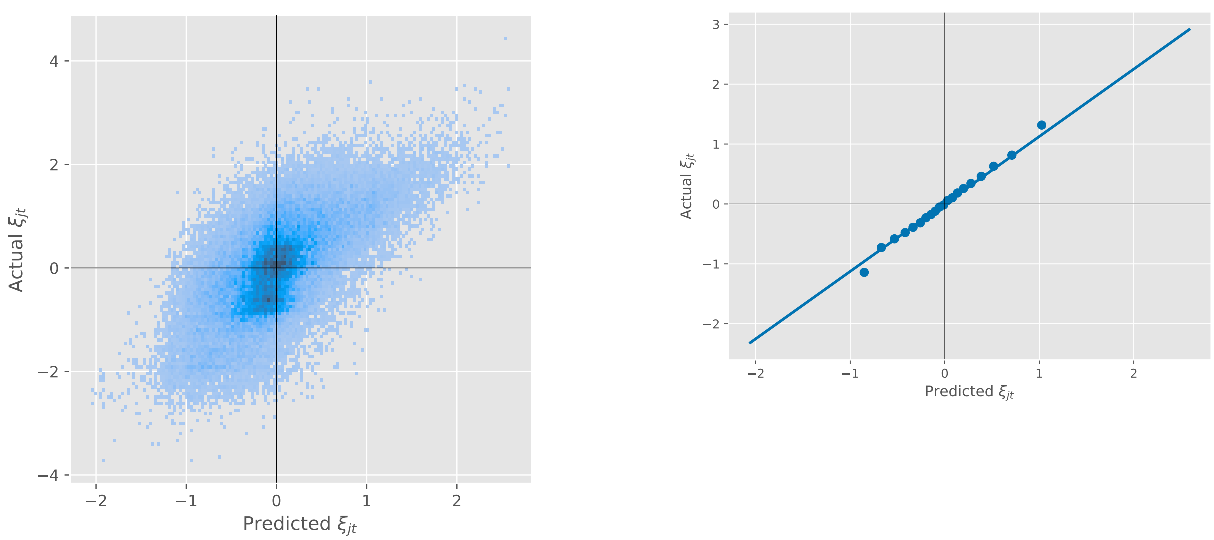


Note: This figure plots binned mean sentiment preferences against (a) user bliss points and (b) credibility preferences. The heterogeneous component captures how the likelihood of a user to submit a negative comment changes in response to changes in the user-specific component of post comment utility. The vertical component captures how the likelihood of a user to submit a negative comment changes in response to a change in the latent commentability term (ξ_{jt}). Positive values mean the user is more likely to submit a negative comment on articles they are likely to comment on. Regression line is a fourth-order polynomial fit. Confidence bands represent 95% confidence intervals.

Figure C.19: Summary of ξ_{jt} Model

(a) Joint Distribution

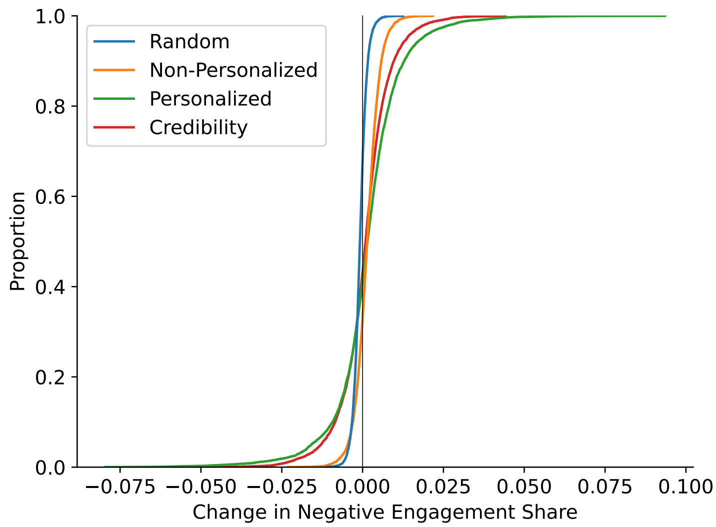
(b) Regression Plot



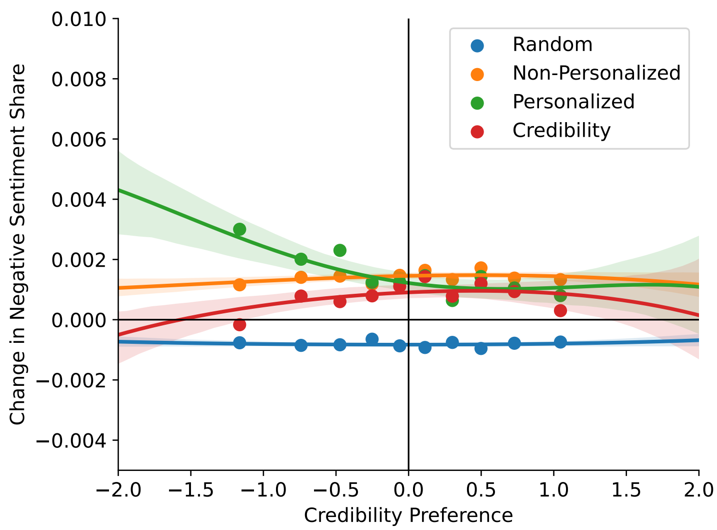
Note: This plot shows a summary of the random forest model used in the counterfactuals to estimate ξ_{jt} in each period. (a) shows the joint distribution between ξ_{jt} and $\hat{\xi}_{jt}$ and (b) shows binned means of this relationship along with a linear fit.

Figure C.20: Impact of Algorithm on Negative-Sentiment Share

(a) Distribution in Change in Negative-Sentiment Share



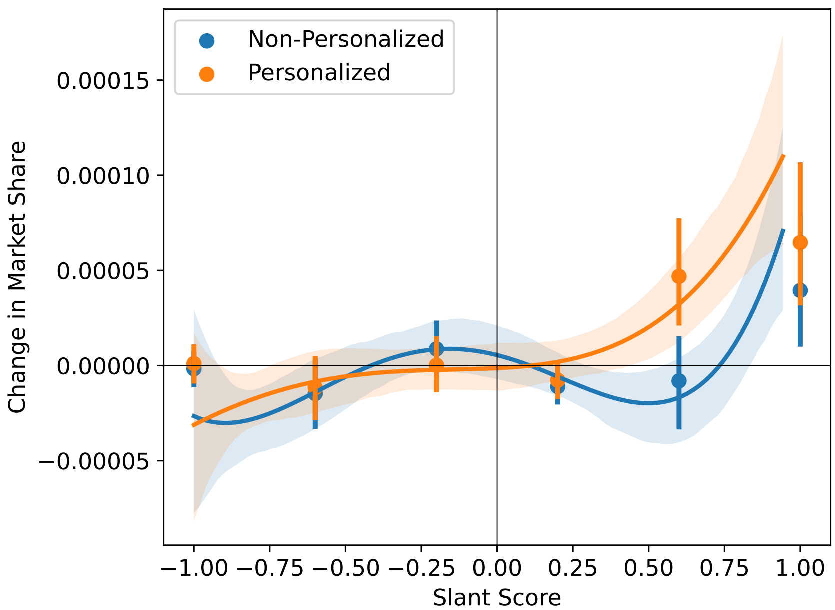
(b) Change in Negative-Sentiment Share by Credibility Preference



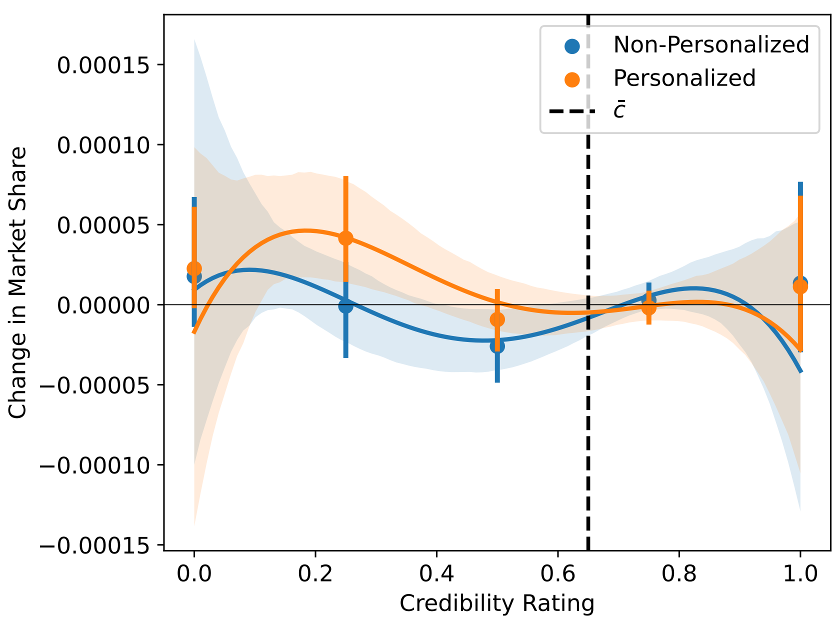
Note: (a) Plots the empirical CDF of the change in the share of users' comments that are negative sentiment under the counterfactual algorithms relative to the existing algorithm. (b) Plots the binned mean in users' change in negative sentiment score against their preferences for publisher credibility. Regression line is a fourth-order polynomial fit. Confidence bands represent 95% confidence intervals.

Figure C.21: Change in Publisher Market Shares

(a) Publisher Slant



(b) Publisher Credibility



Note: Figure C.21a plots the binned mean change in publisher market share by publisher slant and Figure C.21b plots the binned mean change in publisher market share by publisher credibility rating. In both figures, the regression lines are fourth-order polynomial fits. Confidence bands represent 95% confidence intervals.

C.2 Empirical Bayes Shrinkage

To shrink individual preference estimates towards the grand mean and adjust for the over-dispersion due to sampling error I use the following empirical Bayes procedure. I assume that the true individual preference parameters are drawn independently and identically distributed from a multivariate normal distribution

$$\beta_i \sim N(\mu, \Sigma)$$

and we observe noisy estimates of these parameters $\hat{\beta}_i = \beta_i + \nu_i$ where $\nu_i \sim N(0, \Sigma_i)$ is independent sampling error and Σ_i are estimated covariance matrices of preferences for each user. I form estimates of the grand-mean and covariance matrix using empirical analogs of the following expectations.[20]

$$\mu = E\left[\hat{\beta}_i\right]$$

$$\Sigma = E\left[\left(\hat{\beta}_i - \mu\right)\left(\hat{\beta}_i - \mu\right)'\right] - E\left[\Sigma_i\right]$$

and then form estimates of the posterior mean for each β_i as

$$E\left[\beta_i | \hat{\beta}_i, \Sigma_i, \mu, \Sigma\right] = \left(\Sigma^{-1} + \Sigma_i^{-1}\right)^{-1}\left(\Sigma^{-1}\mu + \Sigma_r^{-1}\hat{\beta}_i\right).$$

This shrinks each estimated preference parameter towards the grand mean and corrects for the over-dispersion created by sampling error.

C.3 Estimating the Share of Users Accessing the Platform

Little's law shows that in a stationary system, the average number of users on the platform can be expressed as

$$L_t = \lambda_t W \tag{14}$$

where L_t is the average number of users on the platform at any point during period t, λ_t is the arrival rate of customers during period t, and W is the average session length [Little, 1961]. I assume $W = 10.82$ given that the average session length on Reddit in 2016 lasted 10 minutes and 49 seconds.[21] I assume that the number of users $A_t^0 = L_t$: the number of users online at the start of each period is equal to the average over the period. I can re-arrange equation 14 to show that $A_t = \frac{l}{W} A_t^0$ which says the total number of users to visit the platform during period t (A_t) equals the length of the period in minutes (l) divided by the session length (W) multiplied by the number of users online at any given time. To calibrate the number of active community members in a subreddit, I use two snapshots of the politics community's usage statistics from 2015 and 2016 to calculate the average number of unique users per day.[22] When combined with the number of

[20] When estimating the grand mean, I use inverse variance weights to improve precision of the estimated mean.

[21] `https://web.archive.org/web/20161203082123/https://www.similarweb.com/website/reddit.com/`

[22] `https://web.archive.org/web/20160905095430/https://www.reddit.com/r/politics/about/traffic`
`https://web.archive.org/web/20150513102644/http://www.reddit.com/r/politics/about/traffic/`

subscribers a community has, I can estimate the share of subscribers that are active in a given day which averages 0.071 over the months covered in the two snapshots. I then calculate $N_t = 0.071 \times S_t$ where S_t is the number of subscribers the community has at period t. Robustness to the scaling factor is shown in Appendix Section C.4. Finally, I smooth estimates of p_t by taking the fitted values of the following regression model

$$\frac{A_t}{N_t} = \gamma_0 + \gamma_{quarter} + \gamma_{day} + \eta_t \tag{15}$$

where $\gamma_{quarter}$ and γ_{day} are quarter and day of week fixed effects, respectively.

C.4 Choice Model and Counterfactual Robustness

C.4.1 Robustness to Scaling $p(\cdot)$

First, I show that the counterfactual results are robust to scaling the exposure probability $p(\cdot)$ in the choice model. This shows the results are robust to the decision to scale the number of active users in Section C.3 and to the assumption that all users are exposed to the first post in the feed ($p_1 = 1$) as both simply multiply either p_t or p_r by a constant, so showing $p(\cdot)$ is robust to being multiplied by a constant demonstrates robustness to both.

Table C.4: Counterfactual Engagement Summaries Robustness: $p'(\cdot) = 0.5p(\cdot)$

	Engagement	Diversity	Max Partition Share	Credibility	Negative Engagement Share
Intercept	53.537	1.519	0.290	0.791	0.512
	(0.376)	(0.000)	(0.000)	(0.000)	(0.002)
Random	-6.144	0.005	-0.001	-0.000	-0.001
	(0.037)	(0.000)	(0.000)	(0.000)	(0.000)
Non-personalized	11.211	0.003	-0.005	0.010	0.002
	(0.057)	(0.000)	(0.000)	(0.000)	(0.000)
Personalized	12.270	-0.017	0.021	0.000	0.002
	(0.066)	(0.000)	(0.000)	(0.000)	(0.000)
Observations	33340	33340	33340	33340	33340
R-Squared	0.043	0.048	0.074	0.006	0.000

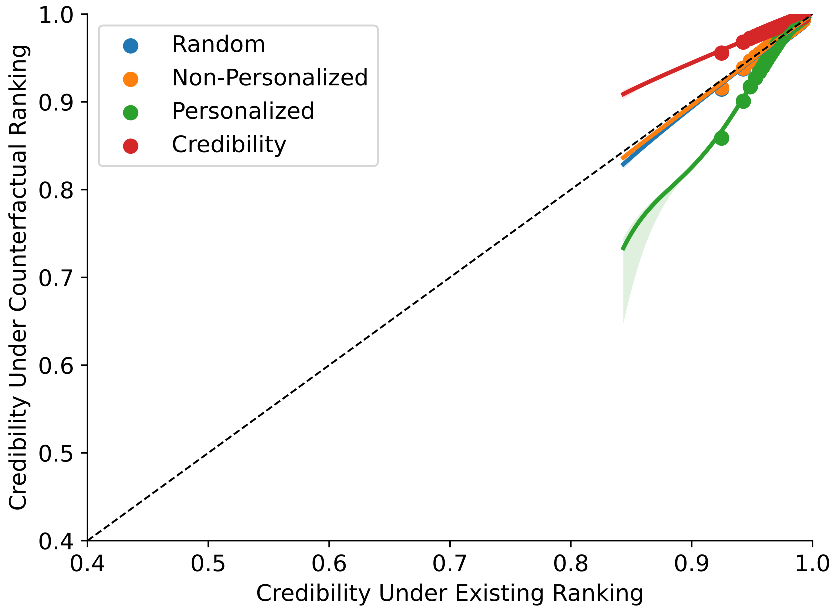
Note: This table reports estimates of a panel regression of each counterfactual outcome on counterfactual algorithm dummy variables in the robustness exercise where $p(\cdot)$ is multiplied by a factor of 0.5. The intercept is the average quantity under the existing algorithm. Standard errors are clustered at the user level.

C.4.2 Robustness to Choice of $\underline{c}$

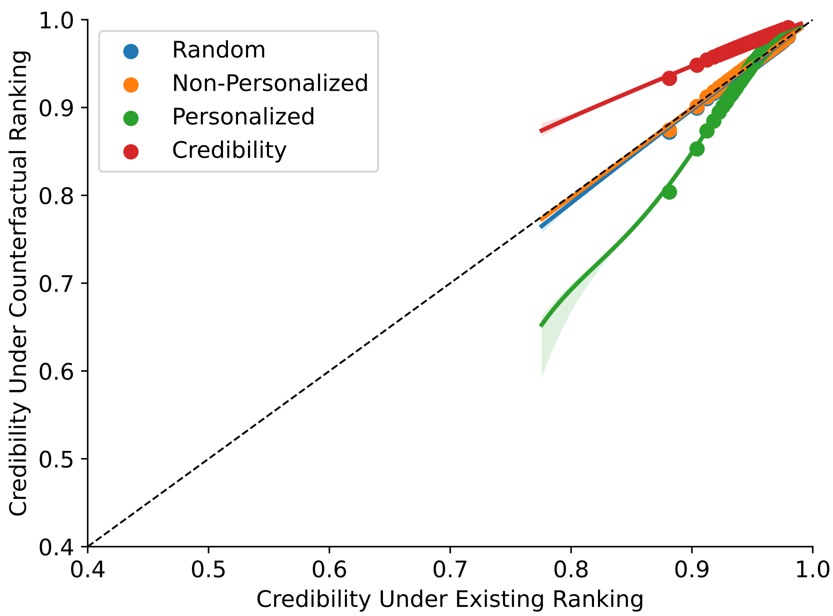
I replicate the quality analysis for various choices of $\underline{c}$ and find the results are qualitatively similar (Figure C.22). For values of $\underline{c}$ as low as 0.4, I find the personalized engagement maximizing exacerbates differences in users along the credibility dimension. As the threshold for credibility is lowered, by definition the share of high quality engagement rises as the threshold does not impact the counterfactuals directly, only how the counterfactuals are evaluated.

Figure C.22: High-Credibility Share by Baseline Credibility: Robustness

(a) $\underline{c} = 0.4$



(b) $\underline{c} = 0.5$

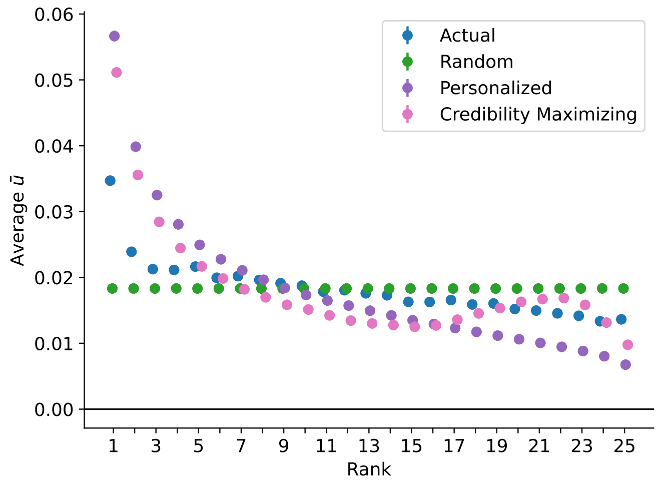


Note: This figure plots binned mean credibility shares under the counterfactual algorithm against credibility shares under the existing algorithm for various thresholds of high-quality publishers. Regression line is a fourth-order polynomial fit. Confidence bands represent 95% confidence intervals.

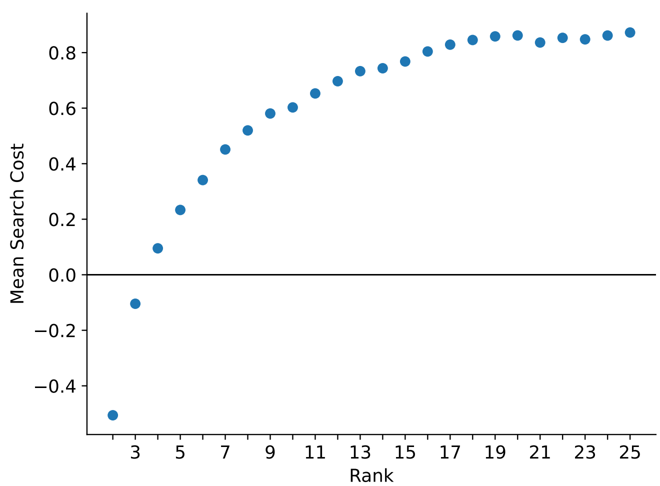
C.4.3 Robustness to Endogenous Search

Figure C.23: Decomposing Treatment Effects in Endogenous Search Model

(a) Average $\bar{u}_{ir}$ by Article Rank

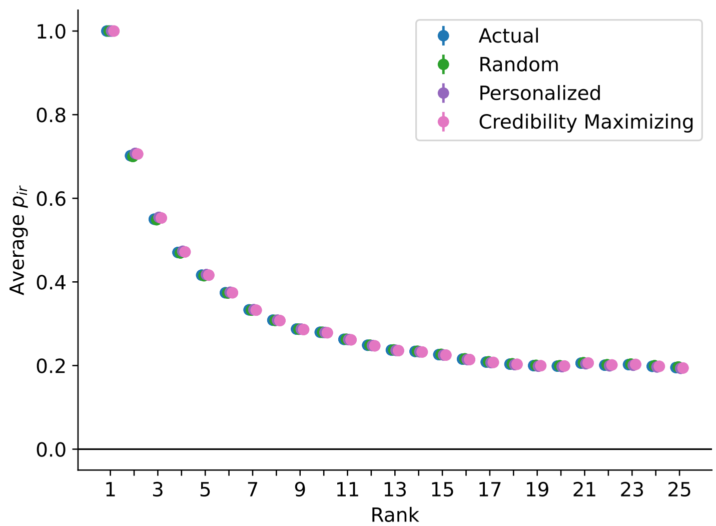


(b) Average c_{ir} by Article Rank



Note: (a) Average expected utility from viewing an article by position ($\bar{u}_{ir}$) and counterfactual algorithm. (b) Average search cost by article position (c_{ir}).

Figure C.24: Reoptimized Exposure Probabilities



Note: Average exposure probabilities ($E\left[P\left(\bar{u}_{ir_r} > c_{ir_j} + \eta_{ijt}\right)\right]$) in the endogenous search model by position and counterfactual algorithm.

D A Recommender System Approach to Personalization

In this section, I study the types of content that are promoted when personalizing the ranking algorithm to maximizing engagement using a reduced form approach. To explore this, I train a collaborative-filtering based recommender system using the matrix of user-level comment counts by publishers. The recommender system then recommends publishers on which a user is most likely to comment in a period. I validate this recommender system by estimating heterogeneous treatment effects when the regression discontinuity experiments align with the recommender system's predictions. I find that the recommender system effectively predicts treatment effects, a result that suggests the model has learned important aspects of user preferences. I then study the types of content that gets promoted under this simple recommender system to understand the extent to which personalized engagement maximization impacts individual news diets.

The primary purpose of this section is to provide reduced-form evidence that personalizing content to maximize engagement promotes low-credibility content to a subset of users and lowers the diversity of publishers that are promoted. This approach has two advantages over the discrete choice model and counterfactual analysis I study in Section 4 and Section 5. First, this model is trained using comment decisions from over 500,000 users and is evaluated on comment decisions of over 180,000 users. This is a much larger sample than that used in the choice model approach, as I can use comment decisions on articles during periods not captured in Wayback Machine snapshots during the training process. I find consistent results across both approaches, which gives confidence that the findings of the choice model approach can be generalized to a broader set of users. Second, I can evaluate this simple approach by predicting treatment effects to give confidence that the model has learned important aspects of preferences.

D.1 Training and Validating the Recommender System

To train the recommender system, I split user-level comment data into training and test sets. The test set consists of comments on articles that appear in Wayback Machine snapshots and the training set consists of comments on articles that do not appear in Wayback Machine snapshots. The test set is used to evaluate the recommendations through heterogeneous treatment effects. I focus this analysis on the politics community because of its importance to managers, policy makers, and users. In the training set, I generate a matrix of user comment counts by publisher domain, where each row represents a user and each column a publisher. I use this matrix to train a collaborative filtering model for implicit data, following Hu et al. [2008]. This simple model assumes that user preferences for a publisher can be represented by the dot product of low-rank vectors of latent user and publisher features. Appendix D.3 shows the publisher embeddings learned by the model are correlated with observable features. More specifically, publisher popularity and slant are the observable publisher features most correlated with the latent embeddings. Given the publisher and user features, the recommender system then recommends publishers that a user is more likely to prefer.

To evaluate the recommender system, I estimate heterogeneous treatment effects comparing periods when the model predicts a user's preferred publisher was promoted to the top of the feed in the regression discontinuity experiments relative to when the non-preferred publisher was promoted. For each period and user, I determine if the preferred post of a user is promoted, the preferred post is demoted, or the user is indifferent. A user is indifferent in a period if the two publishers are within 1 percentile of one another in the model's recommendations for that user. The preferred publisher is promoted for a user in a period if the publisher of the first post is at least 1 percentile higher than the publisher of the second post in the model's recommendations for the user. Likewise, the preferred publisher is demoted if the publisher of the first post is at least 1 percentile lower than the publisher of the second post. I then sum the total number of comments for each post and period across users based on whether the user-period is classified as the preferred post being promoted, demoted, or indifferent. Finally, I estimate the regression discontinuity heterogeneous treatment effects through local linear regression. Given the reduced power in identifying heterogeneous treatment effects, I inflate the bandwidth used in Section 3.3 by a factor of two and use cluster robust standard errors rather than standard errors that are robust to misspecification of the conditional expectation function.[23]

Heterogeneous treatment effect estimates are shown in Table D.5 and suggest that the recommender system effectively predicts treatment effects for the top position in the feed. The treatment effect is substantially – 13 percentage points – larger when a user's preferred publisher is promoted versus when the user's preferred publisher is demoted. That the recommender system is able to predict treatment effects confirms that the recommender system has learned important aspects of user preferences.

D.2 Recommender System Results

I now turn to summarizing the properties of the recommender system to understand the types of content promoted when personalizing rankings and to motivate the choice model presented in Section 4. For each user-period, I determine the most preferred publisher according to the recommender system and calculate the share of promoted publishers that are classified as highly credible. I also calculate the primary measure of slant diversity, which is the first Wasserstein distance between the share of publishers promoted in each slant partition and the uniform distribution. The distributions of these summaries are shown in Figure D.25 alongside the quantity under the existing ranking. The distributions indicate that the majority of users experience improved news diet quality in terms of publisher credibility, though an important minority of users experience a material deterioration in the quality of their news diets. In terms of diversity, a large majority of users are recommended a less diverse set of publishers.

While these results suggest that optimizing for engagement using personalized rankings has a heterogeneous impact on the credibility of publishers that are promoted and a near-uniform decrease

[23] This assumes that the conditional expectation function is linear within the bandwidth and does not account for misspecification.

Table D.5: Validating the Recommender System

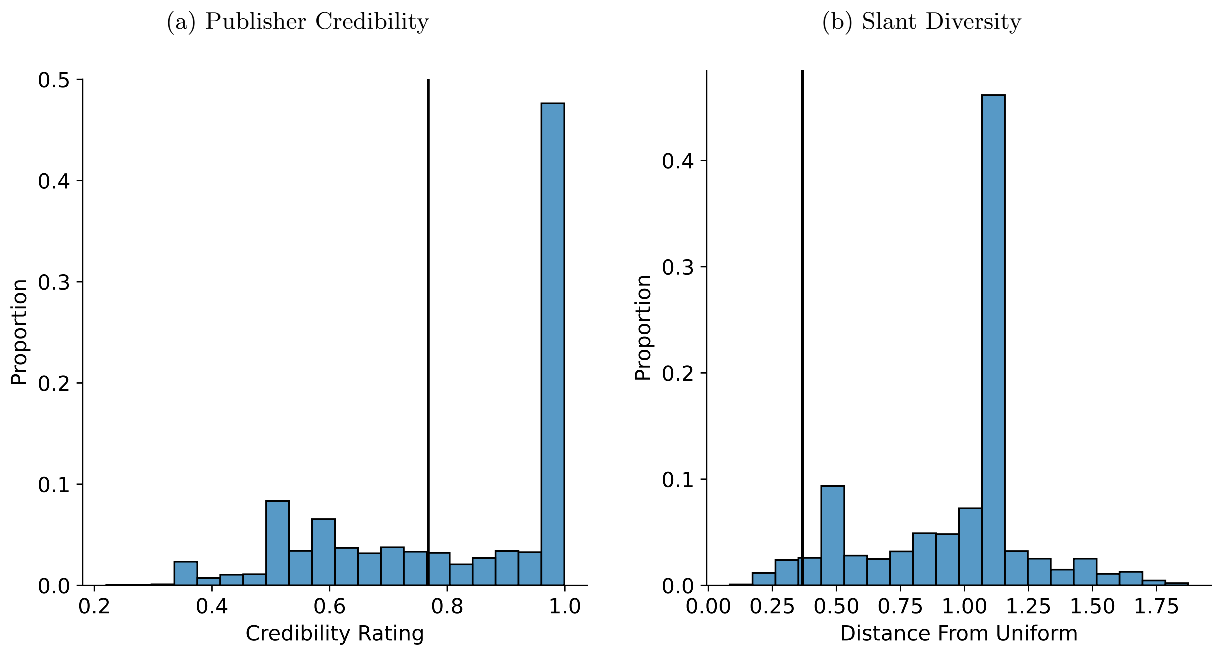
	Preferred Promoted	Preferred Demoted	Indifferent
D	0.80	0.67	0.62
	(0.04)	(0.04)	(0.04)
Δs_{jt}	0.62	0.49	0.72
	(0.12)	(0.12)	(0.13)
$\Delta s_{jt} \times D$	-0.97	-0.57	-1.44
	(0.25)	(0.25)	(0.28)
Intercept	2.14	2.14	1.95
	(0.06)	(0.06)	(0.07)
Obs	3538	3538	3538
R^2	0.09	0.08	0.04

Note: This table shows regression discontinuity heterogeneous treatment effect estimates using a local linear regression. Each column presents estimates of the local linear regression of the outcome (log of one plus the number of comments of each type) on an intercept, treatment indicator, running variable, and running variable interacted with the treatment indicator. The first row contains the coefficient on treatment, which is the local average treatment effect of being promoted to the first position on the feed from the second position. The treatment effect is estimated separately depending on whether a user's preferred publisher is promoted (first column), demoted (second column), or the user is indifferent between the publishers (third column) in the given period.

in slant diversity, this approach has important limitations. First, the recommender system is trained on observational data without accounting for endogenous post rank. The simple collaborative filtering model trained here also differs substantially from the more advanced models – which often employ deep learning – used in practice (see Zhang et al. [2019] for an overview of current deep learning based approaches to recommendation systems). In addition, this approach does not allow for within-publisher article heterogeneity, wherein certain articles are likely to garner more attention irrespective of the publisher. Finally, this approach is limited to analyzing the type of content promoted rather than modeling the content users eventually engage with under counterfactual rankings. Because it allows me to quantify the counterfactual ranking algorithms' impact on engagement – an outcome that serves as a closer proxy to advertising revenue – modeling engagement is critical to understanding the implications for the platform. The choice model and counterfactual analysis presented in Section 4 and Section 5 address these limitations directly.

Despite these limitations, that this model can accurately predict treatment effects indicates the model has learned useful information about user preferences. Moreover, this model can be estimated using data from a larger set of users since engagement on posts not included in the Wayback Machine snapshots can be included in training. This allows the recommender system approach to encompass over 180,000 users while the choice model is estimated on a smaller set of highly active users. As I will argue, the results from both analyses are similar and give confidence that the results generalize to a broader set of users.

Figure D.25: Summary of Promoted Publishers in Recommender System Approach



Note: This figure summarizes the user-level distribution of promoted publishers in the recommender system approach to personalization. In each user-period, I find the publisher out of the top 25 posts that the recommender system would promote first. These figures plot the user-level distribution of the high-credibility publisher share and the first Wasserstein distance between the share of promoted publishers from each slant partition and the uniform distribution. The distance is zero when a user is equally likely to be promoted a publisher from each partition of publisher political slant. Higher values of the Wasserstein distance indicate the user is being promoted a less diverse set of publishers. The maximum distance is 2, which would only occurs when a user is promoted entirely publishers from either the extreme left or extreme right publisher partitions.

D.3 Recommender System Features

Table D.6 shows the projection of the first 3 principal components of the publisher features learned in the collaborative filtering model onto the vector of publisher ratings. These regressions demonstrate that the publisher ratings do explain some of the variation in the publisher features learned by the recommender system.

Table D.6: Recommender System Publisher Factors

	(1)	(2)	(3)
Slant Score	-0.19***	-0.01	-0.07**
	(0.03)	(0.03)	(0.03)
Credibility Rating	-0.00	-0.00	-0.01
	(0.04)	(0.03)	(0.03)
Average Rank	0.00	-0.01	-0.02
	(0.02)	(0.01)	(0.02)
Quantity	0.05***	0.58***	0.17**
	(0.02)	(0.20)	(0.08)
Intercept	-0.07**	-0.04	0.20***
	(0.03)	(0.02)	(0.03)
Obs	1378	1378	1378
R^2	0.04	0.34	0.04

Note: This table shows estimates from a regression of the first 3 principal components of publisher features on publisher observables.