Marketing Science Institute Working Paper Series 2025

Report No. 25-147

# When Connections Depress Contributions: The Hidden Cost of Social Filtering

Ziwei Cong, Jia Liu and Shijie Han

# When Connections Depress Contributions: The Hidden Cost of Social Filtering

Ziwei Cong, Jia Liu, Shijie Han*

**ABSTRACT**

Recommendation algorithms have reshaped how information and attention flow across digital platforms. Among them, social filtering—curating content through users' social networks—has become a defining feature of online engagement. Yet, little is known about its influence on user-generated content (UGC) contributions. Leveraging a quasi-experiment on a large Q&A platform that shifted from content-based to social filtering, this study uncovers a striking paradox: while social filtering fosters greater user consensus and satisfaction, it substantially undermines UGC contributions. After the algorithm change, answers per question fell by 44.7%, the likelihood of a question receiving any answer declined by 32.7%, first response times lengthened by 555.3%, and the quality of top answers dropped by 6.3%. Meanwhile, upvote ratios increased by 5.0%, reflecting fewer dissenting votes. Mechanism analyses reveal that social filtering narrows exposure within homophilic social clusters, reducing opportunities for diverse knowledge exchange. These findings highlight a critical trade-off in algorithmic design: recommendation systems that optimize social engagement may inadvertently erode the informational vitality of online communities.

*Ziwei Cong is an Assistant Professor of Marketing, Georgetown University. Email: zc260@georgetown.edu. Jia Liu is an Associate Professor of Marketing and an affiliated Associate Professor of Industrial Engineering and Decision Analytics, Hong Kong University of Science and Technology. Email: jialiu@ust.hk. Shijie Han is a PhD student in Industrial Engineering and Decision Analytics, Hong Kong University of Science and Technology. Email: kirk.han@connect.ust.hk.

**Keywords:** social media; algorithmic curations; user-generated content; social networks; user homophily; regression discontinuity in time

## INTRODUCTION

Algorithmic recommendation is one of the most intriguing digital technologies in the 21st century. Its wide adoption on digital platforms has transformed the ways that information is distributed, filtered, and discovered. A prominent implementation is social filtering, which curates content based on a user's online social connections. This algorithm is widely adopted across digital platforms, including social media (Facebook), Q&A community (Quora), and review sites (Yelp) (Rodriguez 2020, Yurieff 2020). Prior research shows that social filtering enhances content consumption, yet its implications for user-generated content (UGC) production remain unclear and underexplored. On the one hand, social filtering may encourage contributions by creating a sense of belonging and boosting user loyalty—a view commonly held by many content platforms. On the other hand, by confining exposure within homogeneous social networks, social filtering may limit the diversity of perspectives and discussions, potentially depressing the quantity and quality of UGC production. This can hurt platforms' ecosystems and long-term sustainability, given that UGC is a critical driver of user engagement.

This paper aims to fill this research gap by leveraging a large-scale quasi-experiment on Zhihu, the largest online knowledge-sharing platform in China. Zhihu launched in 2011 as a Q&A community where users share and seek information by posting questions and answers. Initially, Zhihu used a content-based filtering (CF) algorithm, distributing content (e.g., questions and answers) to the subscribers of the topic tags associated with the content. After nearly two years of operation, without any public notification, Zhihu switched to a social filtering (SF) algorithm, which distributes content to its creators' online social networks. This switch from topic-based to social network-based recommendations provides a unique opportunity for us to study the impact of social filtering on UGC contributions. In particular, the straightforward algorithm design allows us to uncover the mechanisms of possible intervention effects. Such transparency is rare in an era where most platforms' algorithms are complex and opaque. In addition, the sudden implementation of this platform-level change minimizes confounding factors that are challenging to address even in controlled

experimental settings, such as interference among connected users in online social networks (Su et al. 2016) and inherent user tendencies (Berman and Katona 2020).

We obtained comprehensive data from Zhihu that records the full online history of user content contributions, voting, subscriptions, and social networking. We apply a regression discontinuity in time (RDiT) framework (Hausman and Rapson 2018) to assess the short-run effects (up to 90 days) of the algorithm change on content contributions. Our analysis primarily focuses on user-contributed answers to newly posted questions over our study period, examining their quantity, quality, and viewer engagement through voting. The identification strategy relies on the assumption that, absent the algorithmic change, user behaviors would have evolved smoothly around the intervention date. We demonstrate the plausibility of this assumption using both institutional context and empirical evidence, and further reinforce validity with a series of robustness checks, including placebo tests that rule out unobservable confounding shocks.

Our main results show that the intervention led to substantial declines in both the quantity and quality of user-generated answers on Zhihu: the average number of answers per question fell by 44.7%, the odds of a new question receiving an answer dropped by 32.7%, the response time of the first answer increased by 555.3%, and the quality of the best answer per question declined by 6.3%. These effects occurred without materially altering the volume or characteristics of questions posted. Paradoxically, the decline in answer quantity and quality coincided with a 18.5% reduction in downvotes and a corresponding 5.0% increase in the share of upvotes per answer, suggesting more positive feedback and greater consensus from viewers.

We identify two parallel mechanisms driving the reduction in answer quantity and quality, both rooted in changes to the content distribution channel. First, social filtering constrains the question dissemination scale. The reduced visibility limits the likelihood that a question will be seen by users who are both willing and able to respond, potentially reducing the quantity and quality of answers it receives. Empirical evidence supports this mechanism: questions posted by low-follower askers—who thereby have lower visibility under social filtering—experience greater reductions in both quantity and quality of answers received. Second, social filtering changes the

exposure composition by increasing the likelihood of content dissemination within homophilic online social clusters. While this inflates the consensus among users who engage with the same content, it can also suppress the diversity of discussions and perspectives, potentially reducing answer volume and quality. Consistent with this mechanism, we find that the negative intervention effects on answer contributions are more pronounced for questions from askers with high social network homophily, where homophily is defined based on content interests and quantified using machine learning methods.

Our research findings have important implications for platforms, creators, and policymakers. First, platforms should exercise caution when deploying algorithms to group like-minded individuals. While exposing content to homophilic audiences may foster consumption and satisfaction, it may also suppress the diversity and richness of discussions, thereby reducing the quantity and quality of UGC, which are critical factors for long-term platform sustainability. Second, platform evaluations of recommendation algorithms should incorporate multiple engagement metrics to obtain a comprehensive assessment. In our case, the social filtering algorithm improves voting-based metrics; if the platform's evaluation relies solely on voting-based measures, it may erroneously conclude that the algorithm has positive effects. Third, while our findings highlight the potential drawbacks of social filtering, the presence of strong user homophily in online social networks suggests the value of leveraging social network data to uncover user content preferences. In sum, these implications advocate for a more balanced approach to algorithmic curation, acknowledging that optimizing for engagement/consumption alone may come at the cost of knowledge production, diversity, and long-term platform health.

The rest of this paper is organized as follows. We first review the relevant literature. Then, we introduce the research context and data, followed by the description of the empirical approach. We present the intervention effects on content contributions and an extensive exploration of the mechanisms in three separate sections. We conclude with a discussion.

## LITERATURE REVIEW

This paper relates to and contributes to two streams of literature: (1) the impact of recommendation algorithms on user behaviors, and (2) the role of user homophily in social networks. Below, we discuss each stream and highlight how this paper advances the existing literature. For ease of comparison, we summarize relevant empirical studies in Table 1 along a few dimensions.

Prior research has extensively examined how recommendation algorithms influence content consumption on digital content platforms (e.g., Chiou and Tucker 2017, Hosseinmardi et al. 2021, Athey et al. 2021, Liu and Cong 2023, Peukert et al. 2024). These studies have shown that (non-social) algorithms can shape user preferences, change engagement, and even polarize content consumption. However, the impact of algorithms on content contribution remains underexplored. While some theoretical work has explored how algorithms might affect creators' incentives through monetary rewards or traffic boosts (Berman and Katona 2020, Qian and Jain 2024), empirical evidence is scarce. This paper expands the research scope by providing one of the first empirical studies on how recommendation algorithms, particularly social filtering, affect content contributions. By leveraging an exogenous and transparent change in algorithm design, we isolate the effects of social filtering and demonstrate that it can affect both content quantity and quality through changing the scale and composition of the content distribution channel.

Note that Liu and Cong (2023) leverages the same intervention to pursue distinct research objectives, standing from a perspective that is fundamentally different from ours. Their analysis primarily focuses on social tie formation and content consumption (i.e., whom to follow and which topics to subscribe to), while offering only a platform-level overview of the intervention effects on answer quantity. They find that social filtering enhances social engagement and content consumption by motivating users to follow established contributors, thereby exposing them to valuable niche content that they might not have otherwise discovered. In contrast, our paper examines answer contributions in depth. Moving beyond a macro-level overview, we adopt a micro-level perspective, analyzing newly posted questions and their answers in terms of quantity,

Table 1: Studies Related to Recommendation Algorithms on Content Platforms

| Paper | Context | Algorithm | Study Method | Dependent Variables | | | Mechanisms |
|---|---|---|---|---|---|---|---|
| | | | | Consumption | Contribution | Homophily | |
| Chiou and Tucker (2017) | News websites | Non-personalized ranking | Event study | Search | – | – | †Ranking increases content salience and user attention |
| Athey et al. (2021) | News websites | Non-personalized ranking | Event study | Page view | – | – | †Ranking increases traffic and user attention |
| Peukert et al. (2024) | News websites | Content-based filtering | Field Experiment | Click | – | – | *Recommending content based on user preferences reduces content diversity |
| Hosanagar et al. (2014) | Music service | Collaborative filtering | Event study | Purchase | – | – | *Reinforcing popular content reduces exploration |
| Holtz et al. (2020) | Music service | Content-based filtering | Field experiment | Streaming | – | – | *Personalization increases short-term engagement but reduces long-term discovery |
| Hosseinmardi et al. (2021) | YouTube | Unknown | Descriptive | Page view | – | – | – |
| Bakshy et al. (2015) | Facebook | Unknown | Descriptive | Click | – | Yes | *Homophily in friend networks and users' choice to click limit exposure to cross-cutting content |
| Kitchens et al. (2020) | Three platforms | Unknown | Descriptive | Browse | – | – | – |
| Liu and Cong (2023) | Q&A platform | Social filtering | Event study | Subscription | Platform-level quantity | – | *Follower-followee asymmetry drives niche topic consumption |
| **This paper** | Q&A platform | Social filtering | Event study | Voting | Question-level quantity, quality, time efficiency, contributor identity | Yes | *Content distribution scale and composition affect content contributions |

Note: For studies that involve multiple social media sites, commonly studied ones are Facebook, Twitter, Instagram, and Reddit. †: Proposed mechanism only (not tested). * Empirically tested mechanism.

efficiency, quality, contributor identity, and viewer engagement. This micro-level approach enables us to document declines in both the quantity and quality of contributions and to uncover the mechanisms driving these adverse effects. Thus, our paper and Liu and Cong (2023) are distinct yet complementary, together offering a comprehensive view of the multidimensional effects of social filtering on platform ecosystems.

Our paper also relates to the literature on user homophily. User homophily, the tendency of individuals to associate with others who share similar traits, has been widely documented in social networks (McPherson et al. 2001). Prior research has explored how homophily is associated with individual and market-level outcomes, such as the diffusion of information (Centola 2011) and the formation of social ties (Lewis et al. 2012). In the context of content platforms, Goldenberg et al. (2023) shows that creators can leverage homophily to expand their reach by targeting influencers within their existing networks. To our knowledge, this paper offers the first causal evidence on how curation algorithms affect homophily in content distribution. We document that social filtering algorithms amplify exposure homophily, which is associated with increased consensus but reduced contributions. This finding sheds light on how platforms can leverage or circumvent user homophily to influence content generation and user discussions. In addition, we introduce an embedding-based machine learning framework to quantify user interest similarity concretely, which can be applied in academia and practice.

## EMPIRICAL APPROACH

### Research Context

Zhihu is the largest Q&A community in China, with key features being very similar to Quora. Users can seek knowledge, expertise, and experiences by asking questions. When posting questions, users usually attach relevant topic tags, which help categorize inquiries for easier navigation.


Figure 1: An Example of A Question and Its Answers

Questions typically receive multiple answers, which are ranked based on crowd-sourced voting behavior. See an example in Figure 1. The voting function allows users to express agreement or disagreement on answers through upvotes or downvotes.[1] In addition, users can subscribe to topic

[1]For upvotes, Zhihu displays the total counts and the identities of users who upvote an answer—clicking on the upvote count reveals a list of upvoters. For downvotes, both the total counts and voter identities are concealed from both the public and answer contributors (see Figure 1). Thus, we believe that the downvote can relatively better reflect user opinions than the upvote, as it is less influenced by social pressure or popularity bias.

tags and questions or follow other users. It has been reported that the majority of Zhihu users are interested in discussing topics of interest or in sharing their own experiences/opinions (Graziani 2018). Therefore, fostering authentic and engaging user discussions remains a key aspect of user experience on Zhihu, which is also true for many online social media platforms.

Each user's homepage displays a personalized feed of recent Q&A threads, accounting for over 40% of user activity on the platform during our study period. Initially, Zhihu utilized a CF algorithm, which shares content (e.g., questions and answers) with the subscribers of topic tags associated with the content. However, on August 16, 2012, Zhihu transitioned to an SF algorithm without any public announcement, distributing user-contributed questions and answers and user-initiated activities (such as voting and subscription) to the user's followers.[2] Zhihu also announced the algorithm change to users shortly after its actual implementation through a Q&A thread. This algorithm change was motivated by Zhihu's belief that leveraging user-created social networks could create a sense of belonging and increase user engagement.

To understand the effects of this algorithm change, it is important to clarify the nature of users' social networks on Zhihu around the intervention time. The platform did not provide separate, dedicated social spaces such as social groups at that time; instead, all social interactions were content-oriented. Therefore, following another user primarily reflected interest in their shared knowledge and expertise rather than offline friendship. Consequently, when Zhihu switched to a social filtering algorithm, users' feeds became organized around the activities of their followees, meaning the content each user saw was directly shaped by the contributions/activities of people they chose to follow. This suggests that the algorithm leveraged pre-existing content-driven social networks to determine exposure, rather than introducing a separate social layer.

[2]Zhihu was still in its early stage in 2012, and its recommendation algorithm was relatively simple, which makes our setting ideal for exploring the underlying mechanisms. A few key aspects of the social filtering algorithm are noteworthy. First, users who had not followed anyone before the intervention saw empty feeds afterward; they may have to rely on a non-personalized "Area" tab, which typically contained popular Q&A threads across a few broad topical areas. These users were typically inactive and had minimal impact on our empirical estimates. Second, the algorithm only leveraged first-degree connections, ignoring second-degree connections. Third, users who subscribed to specific questions received private message notifications whenever new answers were posted, but they were not exposed to other content from those answerers unless they also chose to follow them.

## Data Description

Zhihu provided comprehensive historical data on all registered users. This dataset includes a wide range of activities: posting questions and answers, voting on answers, subscribing to questions and topic tags, and following other users. Given our research objectives, we limit our study period to a 181-day window, including 90 days before and after the intervention date, to ensure a substantial sample of new postings while minimizing confounding variability such as seasonality. Within this window, we focus on newly created questions, their associated answers, and user engagement (i.e., voting) with these answers.

Table 2: Summary Statistics of Q&A

| Variables | Mean | Std. | Min | Max |
|---|---|---|---|---|
| Question Characteristics | | | | |
| Num. related topic tags | 1.99 | 1.87 | 0 | 20 |
| Num. answers | 2.19 | 4.77 | 0 | 252 |
| Num. subscribers | 7.17 | 21.77 | 0 | 2,099 |
| Answer Characteristics | | | | |
| Num. upvotes received | 2.05 | 10.80 | 0 | 1,008 |
| Num. downvotes received | 0.35 | 1.49 | 0 | 301 |

Note: This table is based on 116,741 questions (and their corresponding answers) created within our 181-day study window. All variables represent the cumulative counts within the first two weeks following the creation of a question or answer.

A total of 116,741 questions and 284,419 associated answers were created over our study period. Table 2 reports some summary statistics. On average, a question was linked to 1.99 topic tags, received 2.19 answers, and attracted 7.17 subscribers within the first two weeks of posting. Approximately 85% of answers and 82% of subscriptions occurred within the very first day of posting. An average answer received 2.05 upvotes and 0.35 downvotes within the first two weeks, of which approximately 95% happened within the first day. Given the short lifespan of Q&A threads, our analysis throughout the paper is based on user interactions with content within the first day of content creation.[3] This focus captures the most timely and valuable responses

[3]The only exception is the first-answer response time, defined as the time elapsed between a question's creation

Table 3: Summary Statistics of Users

| | Askers | | Answerers | | Voters | |
|---|---|---|---|---|---|---|
| | Mean | Std | Mean | Std | Mean | Std |
| Num. followers | 47.82 | 1,138.21 | 75.13 | 1,445.78 | 65.12 | 1347.77 |
| Num. followees | 22.37 | 245.04 | 28.19 | 263.27 | 28.77 | 246.31 |
| Num. questions contributed | 4.38 | 21.59 | 4.14 | 21.36 | 3.83 | 20.79 |
| Num. answers contributed | 13.39 | 64.74 | 20.50 | 73.59 | 17.51 | 70.39 |
| Num. questions subscribed | 39.02 | 252.14 | 53.96 | 281.87 | 54.48 | 273.79 |
| Num. topic tags subscribed | 18.23 | 51.74 | 22.83 | 58.99 | 22.98 | 59.24 |

Note: The left (middle) panel describes the 41,418 (35,506) users who posed at least one question in our study period (who answered at least one of these questions within the first day of question creation). The right panel describes the 38,324 users who voted on any of these answers within the first day of answer creation. All variables are the cumulative counts by the intervention date.

for content creators, ensures comparability across content posted at different times, and reflects interactions most likely shaped by the platform's content distribution mechanisms.

Table 3 provides summary statistics for users who engaged with the content summarized in Table 2. These users either asked a question (41,418 askers), answered a question (35,506 answerers), or voted on an answer (38,324 voters). Answerers tend to have a larger follower base, reflecting their roles as content experts and the reputational benefits of contributing answers on Zhihu. In contrast, askers and voters generally have fewer followers, consistent with their different roles in the content ecosystem.

**Measuring Answer Quality Using LLMs**

Measuring answer quality for large-scale user-generated Q&A platforms like Zhihu is very challenging due to the subjectivity of language, diversity of expression, the platform's wide topical coverage, and the specialized expertise required to evaluate content across domains. Traditional methods, including manual annotation and supervised machine learning, face significant limitations. Human labeling is costly, slow, and often lacks the domain expertise needed for nuanced or factual

and its first answer. We include cases in which the first answer arrives after the first day of question posting to fully capture variations in the efficiency of answer contributions.

evaluation (Snow et al. 2008, Pavlick and Kwiatkowski 2019), while machine learning models typically fail to generalize beyond their training data and cannot reliably assess deeper semantic qualities such as insightfulness or coherence (Taghipour and Ng 2016, Verma et al. 2021). To overcome these challenges, we leverage LLMs, which are pre-trained on vast and diverse corpora and thus possess broad factual knowledge, strong reasoning ability, and high consistency across topics. Recent computer science literature (Zheng et al. 2023, Liu et al. 2023) has also proved that LLMs enable scalable, domain-agnostic, and semantically nuanced assessments of content quality.

Web Appendix A provides all the details on our scoring framework and how to operationalize it using LLMs. We provide a brief summary here. According to literature in automated essay grading and Q&A evaluation (e.g., Agichtein et al. 2008, Anderson et al. 2012, Stab and Gurevych 2014, Roy et al. 2023), we designed a five-dimensional evaluation rubric: (Expected) Answer Length, Logic & Structure, Grammar & Diction, Relevance, and Accuracy / Insight. Each criterion is clearly defined on a 7-point scale to ensure consistency across diverse answer types. LLMs are prompted as professional evaluators, using a fixed system instruction and the embedded rubric, to assess each answer given the corresponding question title and details. Then, we compute a composite quality score for each answer as a weighted average across the five dimensions, assigning weights of 0.1 to (Expected) Answer Length, 0.3 to Accuracy / Insight, and 0.2 to each of the remaining three dimensions.[4] This structured and transparent prompting design ensures that the evaluation process is systematic, replicable, and interpretable across models.

We verify the reliability and validity of our prompt design through cross-model and behavioral validation. Specifically, we apply the same evaluation prompts to 18,142 answers associated with 5,000 randomly sampled questions, using six frontier LLMs (GPT-4o, GPT-5, Claude 3.7 Sonnet, Gemini 2.5 Pro, DeepSeek V3.1, and Qwen3-235B). As shown in Table 4, the Cronbach's alpha values across all five dimensions exceed 0.74, demonstrating strong inter-model consistency and confirming that our evaluation is not sensitive to any single LLM's idiosyncrasies. Moreover, the derived quality scores exhibit moderate and statistically significant correlations with user voting

[4] Our results remain consistent when equal weights are applied across all dimensions.

outcomes: Pearson $r = 0.254$ for upvote count and $r = 0.349$ for upvote ratio, both at $p < 0.01$. This finding reinforces the construct validity of our quality measures, while acknowledging that voting reflects additional social and contextual factors.

Table 4: Inter-Model Consistency of Quality Ratings

| Dimension | Cronbach's Alpha |
|---|---|
| (Expected) Answer Length | 0.748 |
| Grammar & Diction | 0.756 |
| Logic & Structure | 0.882 |
| Relevance | 0.864 |
| Accuracy / Insight | 0.867 |
| Total Score | 0.847 |

Note: The sample size is 18,142 answers. Quality scores are derived independently from six LLM models: GPT-4o, GPT-5, Claude 3.7 Sonnet, Gemini 2.5 Pro, DeepSeek V3.1, and Qwen3-235B.

We finally proceed to implement our evaluation framework at scale. For cost and computational efficiency, we employ a single model, GPT-4o, for evaluating all answers in the full dataset, using deterministic settings and automated safeguards to ensure robustness and data integrity. All 284,419 new answers generated during our study period were evaluated in a single batch in May 2025, thereby avoiding any variation that could arise from subsequent model updates. We find that the quality of an average answer is 3.17 (SD = 1.52); the mean and standard deviation of answer quality within each question are 3.15 and 1.27, respectively.

**Identification Strategy**

The key identification challenge is to construct a credible counterfactual outcome in the absence of the intervention against which we can compare the outcome after the intervention. We achieve such causal identification using the RDiT design (Hausman and Rapson 2018), which requires fewer assumptions than other non-experimental methods like difference-in-differences or instrumental variables, thus offering potentially more credible insights (Lee and Lemieux 2010). Because of

these advantages, RDiT has been increasingly applied to various marketing problems (e.g., Ozturk et al. 2019, Vana and Lambrecht 2021, He et al. 2021).

Our approach aggregates all outcome variables at the platform-day level to assess the impacts of the algorithm change. Following the standard specification of the RDiT, we model the log-transformed outcome of interest $Y_t$ on day $t$ as

$$\ln(Y_t) = \alpha + \beta After_t + \delta_1 f(t) + \delta_2 f(t) \times After_t + \delta_3 X_t + \varepsilon_t, \tag{1}$$

where $After_t$ indicates the switch from zero to one at the intervention date.[5] The vector $f(t)$ contains a polynomial time trend to flexibly control for smooth changes that would persist in the absence of the intervention. The interaction term $f(t) \times After_t$ allows the trend to differ on either side of the intervention date. We consider lower polynomial orders ranging from one to three and determine the best order using the Bayesian information criterion (BIC) (Hausman and Rapson 2018). The control variable $X_t$ represents various temporal factors that may influence the outcome on day $t$: day-of-week fixed effects, month fixed effects, and the (log-transformed) daily number of new users and topic tags, respectively. The key coefficient, $\beta$, indicates the multiplicative effect of the intervention on the outcome of interest $Y_t$. If $\beta$ is smaller (larger) than zero, we interpret it as indicating the average level of $Y_t$ decreased (increased) by $100 \times (1 - e^{\beta})\%$ ($100 \times (e^{\beta} - 1)\%$) after the intervention.

An RDiT design is well-suited to our study context for several reasons. First, the intervention was implemented simultaneously for all users, creating an immediate and uniform shift in overall user activity. While it may have taken individual users some time to notice the changes in their feeds and adjust their online behaviors, the platform-wide impact was instantaneous. Second, there was no anticipatory behavior from users since the platform did not announce the intervention in advance. Finally, although user activity is influenced by various unobserved factors, the RDiT framework allows these factors to vary nonlinearly over time, as long as these factors were not

[5] As the intervention took place around noon on August 16, 2012, the observation for that day is removed when estimating the model, ensuring that the full treatment effect is captured by $\beta$.

discontinuous at the time of the intervention (Lee and Lemieux 2010).

Our key identification assumption is that in the absence of the intervention, there would be no sudden discontinuity in user behavior around the intervention date. We assume that user pre-intervention behavior forms a valid counterfactual for user behavior in the days immediately after the intervention, conditional on observed temporal shocks and a highly flexible and smooth time trend. We believe that this assumption holds in our context for several reasons. First, according to the platform, no other major changes were made to the Zhihu interface around the intervention date. Second, our RDiT specification controls for key observable and unobservable changes influencing user behavior, ensuring us to isolate the effect caused solely by the intervention (Hausman and Rapson 2018). Third, as our robustness checks will demonstrate, there is no evidence of significant discontinuities in observed control variables around the intervention date.

It is worth noting that RDiT primarily captures short-term effects due to its reliance on local temporal variation around the intervention date (Hausman and Rapson 2018, Ozturk et al. 2019). While extending the analysis window beyond 90 days might seem desirable for detecting long-term impacts, economic theory suggests this would not resolve the fundamental limitation: without a control group, post-intervention time trends become increasingly confounded by unobserved secular changes (Heckman et al. 1998). Thus, a longer time window will not necessarily enhance the robustness of the estimates because RDiT relies on the cutoff discontinuity.

## EFFECTS ON CONTENT CONTRIBUTIONS

This section examines the intervention effects on content contributions, focusing on the quantity, efficiency, and quality of user-generated answers. We show at the end of this section that the intervention had no significant influence on question contributions.

**Dependent Variables**

We measure answer quantity in three complementary dimensions. For each question, we measure the number of answers received by question $q$ ($AnsPerQ_q$) and the elapsed time in hours between question posting and the creation of its first answer ($FirstAnsResTime_q$). We also track the proportion of questions posted on day $t$ that receive at least one answer ($AnsOdds_t$). For the RDiT specification in Equation (1), across all questions posted on the same day $t$, we aggregate $AnsPerQ_q$ and $FirstAnsResTime_q$ by the daily average and median, respectively, denoted as $AnsPerQ_t$ and $FirstAnsResTime_t$.[6]

We measure answer quality using the LLM-based quality score. For each question, we calculate the average, highest, and lowest composite quality scores across answers, denoted as $AvgQuality_q$, $MaxQuality_q$, and $MinQuality_q$. These measures capture the overall, best, and worst quality of answers a question receives. Our primary focus is on $MaxQuality_q$, as the best answer is typically the most valuable to the asker. For the RDiT specification, we compute daily averages of these metrics across all questions posted on day $t$, denoted as $AvgQuality_t$, $MaxQuality_t$, and $MinQuality_t$, respectively.

We measure viewer engagement with answers through voting. For each question, we calculate the average number of upvotes and downvotes across its answers, and the proportion of upvotes in total votes, denoted as $Upvote_q$, $Downvote_q$, and $UpvoteRatio_q$. A higher $UpvoteRatio_q$ typically reflects greater consensus among viewers regarding an answer.[7] To align with the RDiT specification, we aggregate these measures to platform-day level by averaging across all questions posted on day $t$, denoted as $Upvote_t$, $Downvote_t$, and $UpvoteRatio_t$, respectively.

[6] For $FirstAnsResTime_q$, questions that remained unanswered by June 2017 (the end of our data availability window) are coded as missing. In the daily aggregation, we take the median across all questions posted on day $t$ to mitigate the influence of extreme outliers, such as questions receiving their first answer months later.

[7] Note that all question-level voting and quality metrics are computed for questions with at least one answer within the first day of posting.

## Graphical Evidence

We first present graphical evidence of the intervention effects using RDiT graphs, focusing on a 90-day window before and after the intervention date. Following the literature, we segment the raw data into several non-overlapping bins, and ensure that there are two separate bins before and after the intervention date. We then apply a non-parametric kernel regression, using a bandwidth of seven days, to estimate the trends. This graphical approach not only indicates the presence and magnitude of discontinuities in the outcomes of interest, but also aids in selecting the appropriate functional form for our regression models (Lee and Lemieux 2010, Hausman and Rapson 2018).

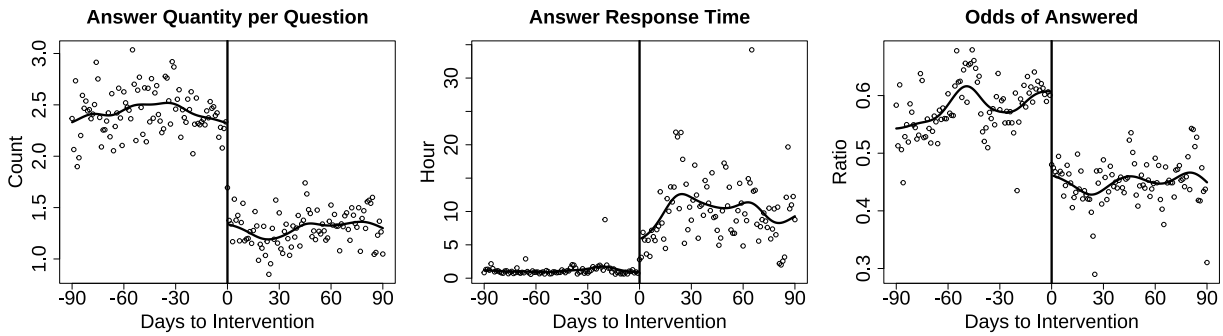



Figure 2: RDiT Graphs: Answer Quantity-Related Variables

Note: Each RDiT plot shows the raw dependent variable within a 90-day window on either side of the intervention date. The fitted lines are from kernel regressions with a bandwidth of seven on either side. Odds of Answered is the proportion of questions posted on day $t$ that receive at least one answer, among all answers posted that day.

Figure 2 displays the RDiT graphs for the quantity-related variables. Immediately following the intervention, we observe substantial decreases in the number of answers per question and the ratio of questions being answered, alongside a significant increase in the response time of the first answers. Figure 3 shows the RDiT graphs for the three quality-related variables. While the average answer quality did not change, the lowest-quality answers improved, and the quality of the best answers declined. In contrast, Figure 4 shows a sharp increase in the upvote ratio, driven primarily by a decrease in the number of downvotes. We interpret these patterns in greater detail in the next subsection.

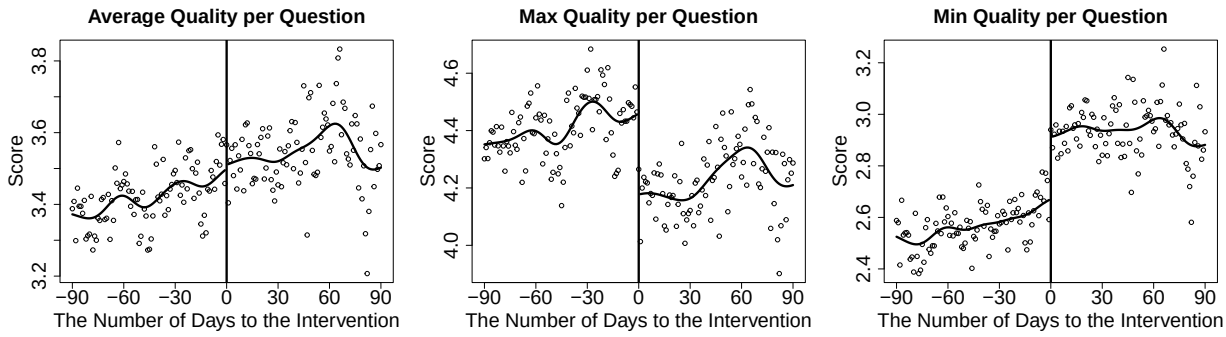



Figure 3: RDiT Graphs: Answer Quality-Related Variables

Note: Each RDiT plot shows the raw dependent variable within a 90-day window on either side of the intervention date. The fitted lines are from kernel regressions with a bandwidth of seven on either side.

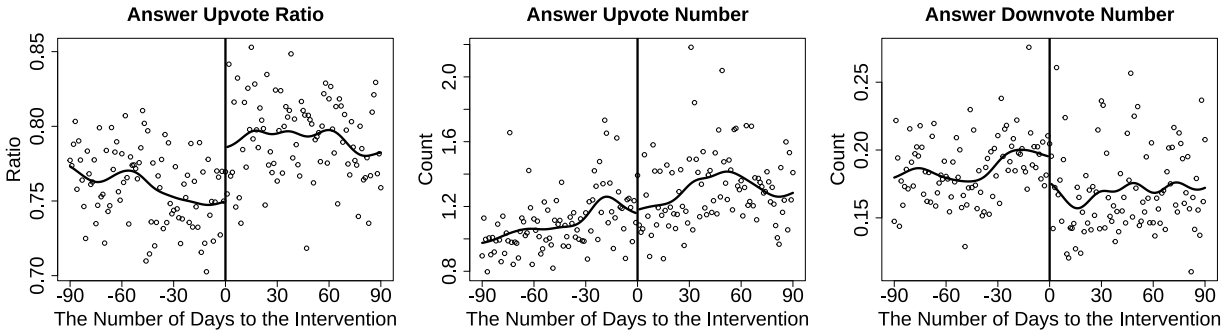



Figure 4: RDiT Graphs: Answer Voting-Related Variables

Note: Each RDiT plot shows the raw dependent variable within a 90-day window on either side of the intervention date. The fitted lines are from kernel regressions with a bandwidth of seven on either side.

## Main Results

We now turn to the regression framework to estimate the intervention effects on answer contributions. Tables 5 and 6 present the estimation results on answer quantity and quality, based on Equation (1) using a 90-day window before and after the intervention date. The findings align with the patterns in Figures 2 and 3. Specifically, the intervention significantly reduced the number of answers per question by $100 \times (1 - e^{-0.593}) = 44.7\%$ ($p < 0.01$) and the odds that a new question would be answered by $100 \times (1 - e^{-0.396}) = 32.7\%$ ($p < 0.01$).[8] The response time for the first answers also increased by $100 \times (e^{1.880} - 1) = 555.3\%$ ($p < 0.01$), where the large percentage change is

[8] For $AnsOdds_t$, we use a Binomial regression (instead of a linear model) in Equation (1).

partly attributable to the low pre-intervention baseline. In terms of quality, although the average answer quality per question did not change significantly ($p > 0.1$), the quality of the best answer decreased by $100 \times (1 - e^{-0.065}) = 6.3\%$ ($p < 0.01$) and that of the worst answer increased by $100 \times (e^{0.094} - 1) = 9.9\%$ ($p < 0.01$). These suggest that while quality dispersion narrowed after the intervention, the quality of the best answers—typically the most valuable to askers—declined.[9]

Table 5: Effects on Answer Volume.

| | *LnAnsPerQ* | *LnFirstAnsResTime* | *AnsOdds* |
|---|---|---|---|
| | (1) | (2) | (3) |
| *After* | $-0.593^{***}$ | $1.880^{***}$ | $-0.396^{***}$ |
| | (0.044) | (0.162) | (0.056) |
| *LnNewUserPlt* | $-0.027$ | 0.090 | $-0.152^{***}$ |
| | (0.028) | (0.103) | (0.021) |
| *LnNewTopicPlt* | $-0.060^{**}$ | 0.031 | $-0.015$ |
| | (0.030) | (0.111) | (0.022) |
| Fixed Effects | Yes | Yes | Yes |
| BIC-chosen order | 1 | 1 | 3 |
| (McFadden) R-sq. | 0.903 | 0.890 | 0.550 |

*Note*: The first two columns report the estimates from linear regressions of the log-transformed dependent variables using a 90-day window, while the last column is a binomial regression of the number of answered questions across all new questions per day. The coefficient on *After* captures the change after the intervention. All specifications include day-of-week and month fixed effects, the two continuous controls listed, and separate polynomial terms in the pre- and post-intervention periods, with order chosen by BIC. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

Such effects have important negative consequences for Zhihu, as both the quantity and quality of answers are critical to sustaining the platform's ecosystem. Quantity matters because Q&A platforms rely on a steady volume of answers to foster community vitality and downstream engagement, such as viewing and voting (Cao et al. 2024). The volume of answers also determines the likelihood that an asker receives feedback, which in turn promotes their retention, reciprocal participation, and social connections within the community (Cheng et al. 2014). Therefore, a decline in answer quantity can trigger a cascade effect of reduced user engagement. Quality, particularly that of

[9] In addition to the composite quality score, we also construct *AvgQuality*, *MaxQuality*, and *MinQuality* measures analogously across individual evaluation dimensions—(Expected) Answer Length, Logic & Structure, Grammar & Diction, Relevance, and Accuracy / Insight. The estimated intervention effects are robust across all dimensions (see Web Appendix B).

the best answer, is equally important, as users typically seek not just any answer but the most insightful, authoritative, or practically useful one (Li et al. 2020). These top-tier answers drive user satisfaction, earn the most engagement, and differentiate the platform from low-quality or generic sources (e.g., search results). When the best answers decline in quality, the platform loses its most valuable content that attracts returning users, garners high visibility, and reinforces trust in the platform's expertise (Deng et al. 2015). Given the importance of answer contributions, it is essential to uncover the underlying mechanisms through which social filtering diminishes both contribution quantity and quality.

Table 6: Effects on Answer Quality-related Metrics.

| | *LnAvgQuality* | *LnMaxQuality* | *LnMinQuality* |
|---|---|---|---|
| | (1) | (2) | (3) |
| *After* | 0.008 | $-0.065^{***}$ | $0.094^{***}$ |
| | (0.009) | (0.009) | (0.012) |
| *LnNewUserPlt* | $0.014^{**}$ | $0.011^{*}$ | $0.017^{**}$ |
| | (0.006) | (0.006) | (0.008) |
| *LnNewTopicPlt* | $-0.008$ | $-0.012^{*}$ | $-0.004$ |
| | (0.006) | (0.006) | (0.008) |
| Fixed Effects | Yes | Yes | Yes |
| BIC-chosen order | 1 | 1 | 1 |
| R-sq. | 0.513 | 0.586 | 0.844 |

*Note*: This table reports the estimates from linear regressions of the log-transformed dependent variables using a 90-day window. The coefficient on *After* captures the change after the intervention. All specifications include day-of-week and month fixed effects, the two continuous controls listed, and separate polynomial terms in the pre- and post-intervention periods, with order chosen by BIC. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

Voting-related results further underscore the importance of exploring underlying mechanisms. As shown in Table 7, the intervention led to a $100 \times (1 - e^{-0.204}) = 18.5\%$ drop in *Downvote* and a corresponding $100 \times (e^{0.049} - 1) = 5.0\%$ increase in $UpvoteRatio$. These effects indicate more favorable feedback and greater consensus from viewers, which stands in contrast to the observed declines in answer quality. This apparent tension motivates an investigation into mechanisms that can reconcile the multidimensional intervention effects.

Table 7: Effects on Voting-related Metrics.

| | *LnUpvoteRatio* | *LnUpvote* | *LnDownvote* |
|---|---|---|---|
| | (1) | (2) | (3) |
| *After* | $0.049^{***}$ | $-0.030$ | $-0.204^{***}$ |
| | (0.013) | (0.058) | (0.057) |
| *LnNewUserPlt* | 0.012 | $-0.084^{**}$ | $-0.131^{***}$ |
| | (0.008) | (0.037) | (0.037) |
| *LnNewTopicPlt* | $-0.001$ | $-0.022$ | 0.002 |
| | (0.009) | (0.040) | (0.039) |
| Fixed Effects | Yes | Yes | Yes |
| BIC-chosen order | 1 | 1 | 1 |
| R-sq. | 0.388 | 0.422 | 0.274 |

*Note*: This table reports the estimates from linear regressions of the log-transformed dependent variables using a 90-day window. The coefficient on *After* captures the change after the intervention. All specifications include day-of-week and month fixed effects, the two continuous controls listed, and separate polynomial terms in the pre- and post-intervention periods, with order chosen by BIC. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

## Robustness Checks

To affirm the validity of our RDiT design and the robustness of our main findings, we conduct a series of analyses, including considering alternative polynomial orders, performing donut RD, adding lagged dependent variables, adding lagged controls, using a shorter time window specification, placebo tests, and checking discontinuity in key control variables. Below is a summary of our approaches and key insights, with detailed results available in Web Appendix B.

*Alternative polynomial orders.* We replicate our analysis using all polynomial orders from one to three. Orders beyond three are not considered because high-order polynomials can lead to overfitting problems or noisy estimates (Gelman and Imbens 2019). Our main findings remain consistent across polynomial orders, albeit with slight differences in effect sizes.

*Donut RD.* A potential concern is that user behaviors around the intervention date may be distorted by short-run avoidance or anticipation, such as resistance to the algorithm change immediately after the intervention or behavioral adjustments in anticipation of the change. Such behaviors

may bias the estimation of the intervention effects. To mitigate this concern, we employ a donut RD design (Hausman and Rapson 2018, Barreca et al. 2011), excluding data from one to three days immediately before and after the intervention date and re-estimating the discontinuity on the remaining sample. The estimation results remain virtually unchanged, alleviating the concerns of short-run anticipation or avoidance.

*Potential serial correlations.* To address possible serial correlations in time-series data, we add lagged dependent variables over the past three days in Equation (1). This addition does not alter the magnitude or significance of our findings. The presence of a discontinuity in autoregressive models is strong evidence for the validity of our RDiT design (Lee and Lemieux 2010).

*Lagged control variables.* To control for potential lagged effects of temporal shocks (e.g., fluctuations in new users or topic tags), we extend our model to include these factors over the preceding two days, separately. These adjustments have minimal impact on the point estimates, affirming the robustness of our main findings.

*Shorter time window specifications.* While our main analysis uses a 90-day window to capture a sufficiently large number of newly posted content, we replicate the analysis using shorter windows of 30 and 60 days, respectively. We find that our key findings remain consistent, with slight differences in the magnitudes of effect sizes.

*Placebo tests.* We examine whether our estimated effects also appeared when they should not. Such evidence would cast doubt on our assumption that the intervention date is unrelated to discontinuous changes in the unobservable determinants of platform-level user activity. We consider two placebo periods: one using a 45-day window around July 02, 2012 (the midpoint of the pre-intervention period in our main specification), and another using a 90-day window around August 16, 2011 (a year prior to the actual intervention). These placebo tests did not reveal any significant effects, further strengthening the credibility of our estimation results.

*Discontinuities in control factors.* We also examine whether any discontinuities exist in control variables – the daily number of new users and topic tags – that could potentially invalidate our assumption of smoothness across the intervention date. We find no evidence of significant discontinuities

in observed control variables around the intervention date.

**Remarks on Question Contributions**

Given the nature of Q&A communities, the intervention may also cause changes in question contributions, which in turn lead to changes in answer contributions. To understand this, we conduct a series of analyses, with details provided in Web Appendix B. First, we examine whether the intervention affected question contributions, including the daily number of newly contributed questions on the platform (i.e., *NewQuesPlt*) and four key question characteristics: the number of associated tags, the popularity of associated tags, question sentiment (both positive and negative), and question type. Tag popularity is approximated using the cumulative number of questions associated with a tag on the platform up to the intervention date. The sentiment scores are derived from question titles and descriptions using lexicon-based text analysis approaches.[10] Each question is classified into either open- or closed-ended using an *XGBClassifier*, which is trained on the DuReader dataset (He et al. 2017) that contains labeled questions from Baidu Zhidao, a Q&A community similar to Zhihu. We compute the daily average of these question characteristics across all questions posted on the same day, and use the RDiT model in Equation (1) to estimate the intervention effects on these characteristics and question quantity. We find no significant discontinuities in either the volume or characteristics of question contributions.

In addition, we replicate our main analysis for answer contributions by including the daily number of new questions as a control variable in Equation (1). The parametric estimates for all dependent variables remain largely unchanged. Therefore, we conclude that the intervention did not cause systematic changes in question quantity and characteristics, and the previously reported changes in answer contributions were direct effects of the intervention, rather than through changing question contributions.

[10]We apply the Python package *cntext* that contains Chinese sentiment dictionaries. See https://github.com/hiDaDeng/cntext/.

## EFFECTS ON DISTRIBUTION CHANNEL

The key distinction between the SF and CF algorithms lies in their distribution channels. Under CF, content is delivered to subscribers of related tags, while under SF, content is delivered to creators' followers. If the intervention operated as designed, we would expect to see increased exposure among creators' followers. While we do not have access to actual exposure data (which the platform did not continuously store during our study period), we approximate distribution effects using actual engagement patterns. Our approach is reasonable in this context because the algorithm's design is simple and transparent, and the observed engagement patterns provide a reliable proxy that closely reflects the characteristics of exposure. This analysis serves as a manipulation check to assess whether the intervention functioned as intended, laying the foundation for our mechanism analysis, which rests on specific characteristics of the distribution channel.

Let $A(q)$ denote the set of answers to question $q$, and $F(q)$ indicate the set of users who had followed question $q$'s contributor before $q$ was posted. Let $A(q|F(q))$ indicate the subset of answers to question $q$ that were contributed by users in $F(q)$. Similarly, $V(a)$ denotes the set of users who voted on answer $a$, $F(a)$ denotes those who had followed the answer $a$'s contributor before $a$ was posted, and $V(a|F(a))$ denotes the subset of votes to answer $a$ that were cast by users in $F(a)$. We let $Q(t)$ denote the set of questions posted on day $t$. We consider the daily average number/ratio of *answers* from askers' followers,

$$FollowerAnsNum_t = \frac{1}{|Q(t)|} \sum_{q \in Q(t)} |A(q|F(q))| \tag{2}$$

$$FollowerAnsRatio_t = \frac{1}{|Q(t)|} \sum_{q \in Q(t)} \frac{|A(q|F(q))|}{|A(q)|} \tag{3}$$

and the daily average number/ratio of *votes* from answer contributors' followers,

$$FollowerVoteNum_t = \frac{1}{|Q(t)|}\sum_{q\in Q(t)}\frac{1}{|A(q)|}\sum_{a\in A(q)}|V(a|F(a))| \tag{4}$$

$$FollowerVoteRatio_t = \frac{1}{|Q(t)|}\sum_{q\in Q(t)}\frac{1}{|A(q)|}\sum_{a\in A(q)}\frac{|V(a|F(a))|}{|V(a)|} \tag{5}$$

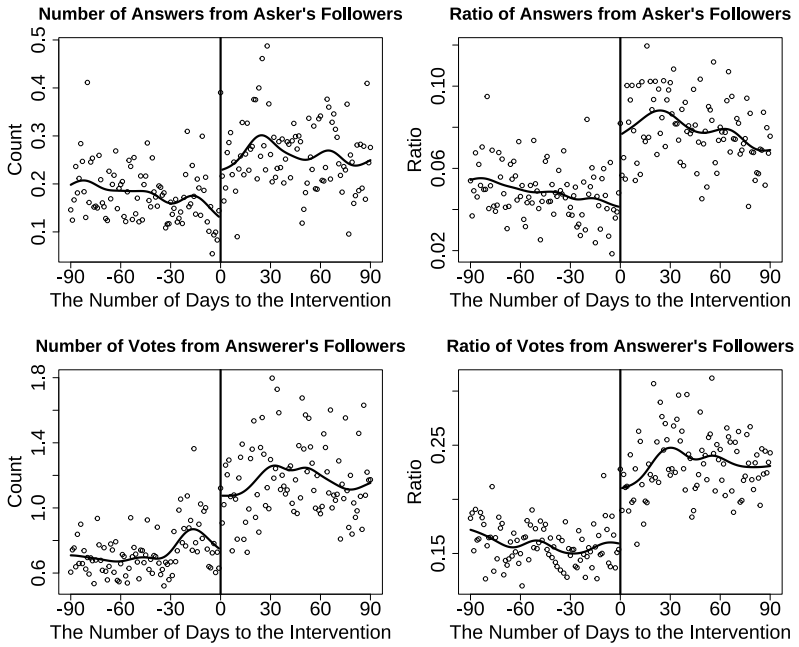



Figure 5: RDiT Graphs for Distribution Channel: Follower Engagement

Figure 5 displays the RDiT graphs for the four variables, each showing a pronounced increase immediately after the intervention. The RDiT estimates based on Equation (1) confirm these patterns: for an average question, the intervention increased the number (ratio) of answers from the asker's followers by $100\times(e^{0.652}-1)=91.9\%$ ($100\times(e^{0.892}-1)=144.0\%$) ($p<0.01$); for an average answer, it increased the number (ratio) of votes from the answerer's followers by $100\times(e^{0.303}-1)=35.4\%$ ($100\times(e^{0.358}-1)=43.0\%$) ($p<0.01$). Web Appendix C presents the detailed estimates and various robustness checks, which show consistent findings. Beyond first-degree followers, we also examine engagement from contributors' broader social circles, recognizing that first-degree followers may further distribute content to higher-degree connections.

The results are robust to this alternative measure. All details are reported in Web Appendix C.

Taken together, these findings confirm that the intervention operated as designed, distributing content among contributors' online social networks and thereby substantially increasing engagement from contributors' followers. These findings also suggest that some characteristics of the distribution channel likely drive the negative intervention effects on answer contributions. In the following two sections, we document the role of distribution scale (the extent to which content is distributed) and network homophily (the attributes of the audience), respectively.

## THE ROLE OF DISTRIBUTION SCALE

We begin with the distribution scale, as it is typically the most fundamental characteristic of a distribution channel. Intuitively, a question with broader visibility is more likely to be seen by users who are both willing and able to respond (Liu and Jansen 2013). As such, the scale of distribution is likely to influence both the quantity and quality of answers that a question receives.

### Comparing the Distribution Scale

We approximate and compare the distribution scale under the two algorithms. Consider a user $U(q)$ who posts a question $q$ tagged with a set of topic tags $T(q)$. Let $S(j)$ indicate the set of subscribers of tag $j \in T(q)$ and $F(u)$ be the set of followers of the user. Under the CF algorithm, the potential audience for question $q$ includes all users subscribed to any topic tag in $T(q)$; but it is confined to users in $F(u)$ under the SF algorithm. Therefore, we approximate the distribution scale of each question $q$, denoted by $DisScale_q$, as:

$$DisScale_q = \begin{cases} \sum_{j \in T(q)} |S(j)| & \text{if } q \text{ created before the intervention} \\ |F(u_q)| & \text{otherwise} \end{cases}$$

We examine all the questions created during our study period. We find that the average subscriber size (average $\sum_{j \in T(q)} |S(j)|$ across questions) is 7,770 (SD = 14,087), but the average follower size (average $|F(u)|$ across questions) is only 545 (SD = 5,784), significantly lower than the average subscriber size ($p < 0.01$). An alternative explanation is that each question may have multiple tags, naturally inflating $\sum_{j \in T(q)} |S(j)|$. To rule out this explanation, we compare the individual tags to users. Among the 353,140 users and 63,666 topic tags on Zhihu before the intervention, the average user had only 2 followers (SD = 482), whereas the average topic tag had 61 subscribers (SD = 1,246), with this difference being statistically significant ($p < 0.01$). These findings suggest that social networks serve as a narrower distribution channel than topic subscriber networks, likely due to differences in how attention is allocated between users and tags.

**Effects on Content Contributions by Network Scale**

We hypothesize that the reduction in content distribution scale, in part, explains the observed declines in answer quantity and quality. If this mechanism is at work, the reduction in answer quantity and quality should be more pronounced for questions raised by low-follower askers, whose content faces particularly constrained reach under the social filtering algorithm. We test this prediction by examining the heterogeneity in the intervention effects by asker follower size.

For each question, we focus on the key quantity and quality metrics: $AnsPerQ_q$ and $MaxQuality_q$. We classify questions into two groups based on whether the asker's follower count by the intervention date ($|F(u_q)|$) exceeds the median among 41,418 askers who posted at least one question during our study period. For each group, we aggregate question-level characteristics to the platform-day level and re-estimate the RDiT specification in Equation (1) for $AnsPerQ_t$ and $MaxQuality_t$. Table 8 reports the results. Consistent with our expectation, both quantity and quality decline more sharply for questions from low-follower askers. For quantity, the reduction is $100 \times (1 - e^{-0.384}) = 31.9\%$ ($p < 0.01$) for the low-follower group and $100 \times (1 - e^{-0.307}) = 26.4\%$ ($p < 0.01$) for the high-follower group, implying a 20.8% greater reduction among low-follower askers. For quality,

the decline is $100 \times (1 - e^{-0.061}) = 5.9\%$ ($p < 0.01$) in the low-follower group and $100 \times (1 - e^{-0.046}) = 4.5\%$ ($p < 0.01$) in the high-follower group, indicating a 31.1% larger reduction among low-follower askers.

Table 8: Heterogeneous Effects on Answer Contributions by Network Scale

| Dependent Variable | Low-Follower Askers | | High-Follower Askers | |
|---|---|---|---|---|
| | *LnAnsPerQ* | *LnMaxQuality* | *LnAnsPerQ* | *LnMaxQuality* |
| | (1) | (2) | (3) | (4) |
| $After$ | $-0.384^{***}$ | $-0.061^{***}$ | $-0.307^{***}$ | $-0.046^{***}$ |
| | (0.038) | (0.011) | (0.031) | (0.009) |
| *LnNewUserPlt* | $-0.048^{*}$ | $0.043^{***}$ | 0.014 | $0.012^{**}$ |
| | (0.025) | (0.094) | (0.020) | (0.006) |
| *LnNewTopicPlt* | $-0.017$ | $-1.043^{***}$ | $-0.039^{*}$ | $-0.010$ |
| | (0.026) | (0.094) | (0.022) | (0.006) |
| Day-of-week fixed effects | Yes | Yes | Yes | Yes |
| Month fixed effects | Yes | Yes | Yes | Yes |
| BIC-chosen order | 1 | 1 | 1 | 1 |
| R-sq | 0.678 | 0.356 | 0.864 | 0.432 |

Note: This table reports the estimates from linear regressions of the log-transformed dependent variables using a 90-day window. The coefficient on *After* captures the change after the intervention. All specifications include day-of-week and month fixed effects, the two continuous controls listed, and separate polynomial terms in the pre- and post-intervention periods, with order chosen by BIC. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

As an alternative identification strategy, we conduct question-level regression analysis to examine how the intervention effects vary across askers with different numbers of followers, which is captured by an interaction term between $After_t$ and $|F(u_q)|$. The estimation results are presented in Web Appendix D. As expected, we find that the negative intervention effect is mitigated when the asker has a larger follower base. In summary, results from both analyses confirm that changes in content distribution scale explain the observed effects on answer contributions, which is intuitive given the fundamental role of scale in content dissemination. However, distribution scale alone is unlikely to cause the paradoxical increase in the upvote ratio. This motivates a deeper examination of the attributes of the audience in the next section.

## THE ROLE OF NETWORK HOMOPHILY

In theory, individuals tend to connect with others who share similar traits (e.g., interests, background, etc.) (McPherson et al. 2001). If user homophily is prevalent within Zhihu's social networks, the social filtering algorithm would distribute content among users with similar interests. While greater alignment in content preferences can enhance consensus and satisfaction, it can also suppress the diversity of discussions and perspectives (Cinelli et al. 2021, Bakshy et al. 2015, Flaxman et al. 2016). Literature demonstrates that individuals from socially distinct groups embody diverse cognitive resources and perspectives that, when cooperatively combined, produce ideas, solutions, and designs that outperform those from homogeneous groups (Shi et al. 2019). Therefore, we hypothesize that network homophily is negatively associated with both the quantity and quality of answer contributions, particularly the latter. Under this mechanism, the observed increase in the upvote ratio is likely an "illusion of harmony" — a seemingly positive but potentially misleading signal that platforms should treat with caution.[11]

### Defining User Embeddings and Network Homophily

To investigate this mechanism, we measure network homophily in terms of content interests. Quantifying user content interests poses significant challenges in environments like Zhihu, marked by a vast array of Q&As, sparse user interactions, and the ephemeral nature of Q&A threads. To tackle these challenges, we employ a machine learning framework designed to effectively capture latent user characteristics. This framework derives user embeddings by leveraging the topic tags associated with questions users have interacted with, either by answering or subscribing. Consequently, we generate two types of embeddings for each user, reflecting content contribution and consumption interest, respectively.

[11] We acknowledge that the increase in upvote ratio may also stem from other SF–related factors, such as social desirability or conformity, rather than homophily alone. For instance, users may feel more reluctant to downvote content originating from their online social connections. We do not disentangle these factors in this current paper, as they are consistent with the interpretation of the "illusion of harmony."

Our procedure involves the following steps. Let $T(q)$ denote the set of tags associated with a question $q$ and $O^m(u)$ denote the set of questions with which user $u$ has engaged through activity type $m \in \{A, S\}$ prior to our study period, where $A$ indexes answer contribution and $S$ represents question subscription.[12] We define user $u$'s embedding for action type $m$ as the average topic embedding across tags linked to these questions in $O^m(u)$,

$$I_u^m = \frac{1}{|O^m(u)|} \sum_{q \in O^m(u)} I_q = \frac{1}{|O^m(u)|} \sum_{q \in O^m(u)} \frac{1}{|T(q)|} \sum_{j \in T(q)} h_j, \tag{6}$$

where $h_j$ is the latent representation of topic tag $j \in T(q)$, and $I_q$ denotes the latent representation of question $q \in O^m(u)$, defined as the average embedding of its associated tags. Thus, the key is to obtain the latent representation $\mathbf{h}$ for all topic tags. To derive $\mathbf{h}$, we leverage Zhihu's topic tree, which organizes topic tags in a "child-parent" hierarchy based on semantic similarities, and an existing dataset on pre-trained Zhihu topic embeddings (Li et al. 2018), which covers nearly half of the tags in our study. We input these datasets to a Graph Convolutional Network (GCN) designed to extract node features from graph-structured data (Kipf and Welling 2016). In our application, the GCN learns meaningful representations for each topic tag by considering both its own features (i.e., embeddings) and the features of its neighboring tags in the topic tree. We demonstrate the validity of the derived topic embeddings through visualization and link prediction. Technique details are in the Appendix.

For each focal user $u$' social network with follower set $F(u)$, we measure its network homophily by the intervention date as the average pairwise cosine similarity in embeddings among her followers:

$$SocHomo_u^m = \frac{2}{|F(u)|(|F(u)|-1)} \sum_{i,i' \in F(u)} Cos(I_i^m, I_{i'}^m), \tag{7}$$

where $Cos(.)$ denotes the cosine similarity between two users' interest embeddings for action type $m \in \{A, S\}$. Analogously, for each topic tag $j$ with subscriber set $S(j)$ at the time of the

[12] User embedding is set to missing if a user had no answer contribution or question subscription activity before our study period, or when the questions she interacted with had no tag.

intervention, we define this network homophily as the average pairwise cosine similarity among its subscribers:

$$SubHomo_j^m = \frac{2}{|S(j)|(|S(j)|-1)} \sum_{i,i' \in S(j)} Cos(I_i^m, I_{i'}^m). \tag{8}$$

We calculate these metrics for users and tags that, by the time of the intervention, have at least two followers or subscribers with valid embeddings.

Table 9 shows that, on average, $SocHomo_u^m$ is approximately 32% ($p < 0.01$) higher than $SubHomo_j^m$.[13] This finding indicates that social ties on Zhihu connect users with more similar content preferences than topical subscriptions, likely because topic subscriptions are less selective and more influenced by temporal trends. As a result, the social filtering algorithm increased the exposure within more homogeneous audiences than content-based filtering.

Table 9: Comparing Network Homophily

| | $SocHomo_u$ | | $SubHomo_j$ | | $p$-value |
|---|---|---|---|---|---|
| | Mean | Std | Mean | Std | |
| Answering Questions | 0.478 | 0.314 | 0.363 | 0.256 | 0.000 |
| Subscribing Questions | 0.561 | 0.317 | 0.427 | 0.263 | 0.000 |

Note: $SocHomo_u^A$ ($SocHomo_u^S$) are computed for 83,745 (88,410) users who, by the intervention date, have at least two followers with valid embeddings. $SubHomo_j^A$ ($SubHomo_j^S$) are computed for 14,621 (15,177) topic tags that, by the intervention date, have at least two subscribers with valid embeddings.

**Effects on Content Contributions by Network Homophily**

If the decline in answer contributions is partly driven by increased audience interest homophily, then we should observe larger effects on answer contributions to questions posted by askers whose

[13] Note that a single question may be linked to multiple topic tags. When aggregating all subscribers across these tags, the overall degree of subscriber homophily for a given question is likely even lower than the computed value of $SubHomo_j^m$. Furthermore, beyond the similarity among followers, we find that the average cosine similarity between a followee and her followers exceeds that between a tag and its subscribers (see Web Appendix E), reinforcing that social connections embody a stronger degree of network homophily.

social networks exhibit higher homophily. To test this conjecture, we classify questions into two groups based on whether the asker's social network homophily, $SocHomo^{S}_{u_q}$, exceeds the median value among all askers posting at least one question during our study period.[14] For each group, we aggregate question-level characteristics to the platform-day level and re-estimate the RDiT specification in Equation (1) for $AnsPerQ_t$ and $MaxQuality_t$. Table 10 reports the estimation results. We find that both quantity and quality decline more for questions posted by high-homophily askers. For quantity, the reduction is $100 \times (1 - e^{-0.264}) = 23.2\%$ ($p < 0.01$) in the low-homophily group and $100 \times (1 - e^{-0.379}) = 31.6\%$ ($p < 0.01$) in the high-homophily group, implying a 36.2% greater reduction among high-homophily askers. For quality, the reduction is $100 \times (1 - e^{-0.032}) = 3.1\%$ ($p < 0.01$) in the low-homophily group, only about half the magnitude observed in the high-homophily group. These provide suggestive evidence for the proposed mechanism.

Table 10: Heterogeneous Effects on Answer Contributions by Network Homophily

| Dependent Variable | Low Social Network Homophily | | High Social Network Homophily | |
|---|---|---|---|---|
| | *LnAnsPerQ* | *LnMaxQuality* | *LnAnsPerQ* | *LnMaxQuality* |
| | (1) | (2) | (3) | (4) |
| *After* | $-0.264^{***}$ | $-0.032^{***}$ | $-0.379^{***}$ | $-0.061^{***}$ |
| | (0.042) | (0.011) | (0.051) | (0.018) |
| *LnNewUserPlt* | 0.031 | $0.016^{**}$ | $-0.009$ | 0.006 |
| | (0.027) | (0.007) | (0.033) | (0.012) |
| *LnNewTopicPlt* | $-0.026$ | $-0.007^{***}$ | $-0.033$ | $-0.011$ |
| | (0.029) | (0.007) | (0.035) | (0.012) |
| Day-of-week fixed effects | Yes | Yes | Yes | Yes |
| Month fixed effects | Yes | Yes | Yes | Yes |
| BIC-chosen order | 1 | 1 | 1 | 1 |
| R-sq | 0.761 | 0.337 | 0.767 | 0.259 |

Note: This table reports the estimates from linear regressions of the log-transformed dependent variables using a 90-day window. The coefficient on *After* captures the change after the intervention. All specifications include day-of-week and month fixed effects, the two continuous controls listed, and separate polynomial terms in the pre- and post-intervention periods, with order chosen by BIC. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

In addition, we conduct a question-level regression analysis like before to examine how the intervention effects vary across askers with different levels of social network homophily, which is captured by an interaction term between $After_t$ and $SocHomo_{u_q}$. The estimation results are shown

[14] Results are qualitatively unchanged when using $SocHomo^{A}_{u_q}$.

in Web Appendix E. We find that the coefficient on the interaction term is significantly negative for both quantity ($p < 0.1$) and quality ($p < 0.01$), suggesting that the intervention effects are more pronounced when askers have higher network homophily.

A potential concern is that network homophily may depend on network scale. That is, social network homophily may inherently be higher among users with fewer followers. However, we find that the correlation between each question's social network homophily, $SocHomo_{u_q}$, and distribution scale, $|F(u_q)|$, is only around $-0.10$ ($p < 0.01$).[15] This low correlation suggests that the two variables capture distinct aspects of the distribution network. Furthermore, in the question-level analysis in Web Appendix E, the intervention term between $After_t$ and $SocHomo_{u_q}$ remains significantly negative for $MaxQuality_q$ ($p < 0.05$), even after controlling for distribution scale. For $AnsPerQ_q$, the intervention term becomes statistically insignificant once distribution scale is controlled for, but the coefficient remains negative. These results suggest that holding distribution scale constant, social network homophily continues to exert an important moderating effect, particularly on answer quality. This indicates that the network homophily operates through a mechanism different from the distribution scale.

**Effects on Engagement Homophily**

Our analysis thus far has focused on audience homophily, measured by the pre-intervention network homophily of askers, which provides a clean identification of the proposed mechanism. A natural question is whether homophily in exposure translates into homophily among the engaged users. Answers to this question are critical for deepening our understanding of the underlying mechanism and assessing its potential consequences for the platform.

We begin by measuring user homophily conditional on engagement. There are three roles associated with a Q&A thread: askers, answerers, and voters. We focus on examining four types of similarity metrics among these roles: asker-answerer, asker-voter, answerer-answerer, and

[15] Specifically, the correlation between $SocHomo^S_{u_q}$ and $|F(u_q)|$ is $-0.11$ ($p < 0.01$), and that between $SocHomo^A_{u_q}$ and $|F(u_q)|$ is $-0.09$ ($p < 0.01$).

answerer-voter. Let $u_q$ denote the user who asked question $q$, $U_q(a)$ the set of users who answered question $q$, and $U_{qa}(v)$ the set of users who voted on answer $a$ of question $q$. We construct the following engagement-based similarity metrics:

$$Homo(asker, answerer)_q^m = \frac{1}{|U_q(a)|} \sum_{i \in U_q(a)} Cos(I_{u_q}^m, I_i^m) \tag{9}$$

$$Homo(asker, voter)_q^m = \frac{1}{|A(q)|} \sum_{a \in A(q)} \frac{1}{|U_{qa}(v)|} \sum_{i \in U_{qa}(v)} Cos(I_{u_q}^m, I_i^m) \tag{10}$$

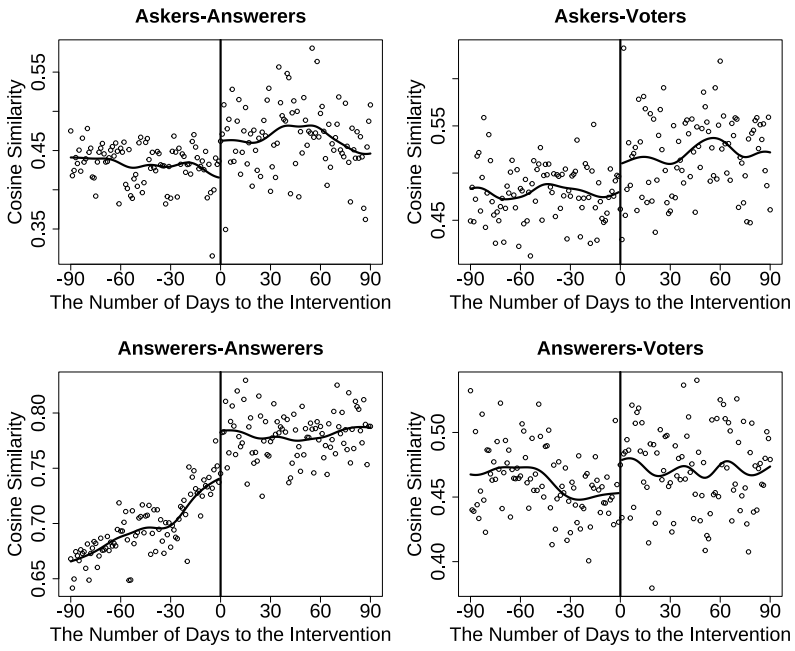
$$Homo(answerer, answerer)_q^m = \frac{2}{|U_q(a)|(|U_q(a)|-1)} \sum_{u,i \in U_q(a)} Cos(I_u^m, I_i^m) \tag{11}$$

$$Homo(answerer, voter)_q^m = \frac{1}{|U_q(a)|} \sum_{u \in U_q(a)} \frac{1}{|U_{qa}(v)|} \sum_{i \in U_{qa}(v)} Cos(I_u^m, I_i^m) \tag{12}$$

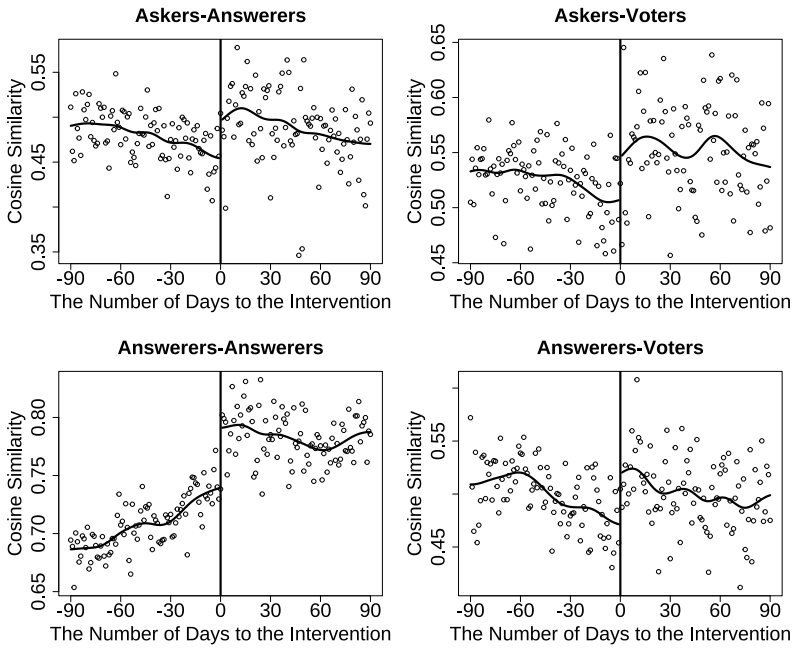
where $A(q)$ is the set of answers to question $q$ and $Cos(\cdot)$ denotes the cosine similarity between two users' interest embeddings.

To apply the RDiT design in Equation (1), we use the daily averages of the above metrics across all questions posted each day $t$, denoted as $Homo(asker, answerer)_t^m$, $Homo(asker, voter)_t^m$, $Homo(answerer, answerer)_t^m$, and $Homo(answerer, voter)_t^m$. We present the RDiT graph and point estimates for all eight dependent variables in Figure 6 and Table 11. We observe a significant and consistent increase immediately following the intervention across all similarity metrics. Specifically, the similarity between askers and answerers/voters increased by around 10% ($p < 0.001$); the similarity across answerers to the same question increased by 5% – 7% ($p < 0.001$); and the similarity between answerers and their voters increased by 5% – 11% ($p < 0.05$). In addition, we perform a series of robustness checks similar to those in our previous analysis, yielding mostly consistent findings. Details are available in Web Appendix E.

In summary, these findings suggest that audience homophily, conditional on mere exposure, translates into homophily among engaged users. This further supports the network homophily mechanism and underscores its significant implications for user engagement and platform ecosystems.




(a) Based on Interests in Answering Questions




(b) Based on Interests in Subscribing Questions

Figure 6: RDiT Graphs: Engagement Homophily

Table 11: Effects on Engagement Homophily

| | Panel A: Interests in Answering Questions | | | |
|---|---|---|---|---|
| | Askers-Answerers | Askers-Voters | Answerers-Answerers | Answerers-Voters |
| $After$ | $0.094^{***}$ | $0.091^{***}$ | $0.055^{***}$ | $0.055^{**}$ |
| | (0.030) | (0.028) | (0.010) | (0.025) |
| $LnNewUserPlt$ | $-0.006$ | 0.014 | 0.005 | $-0.002$ |
| | (0.019) | (0.018) | (0.006) | (0.016) |
| $LnNewTopicPlt$ | $0.038^{*}$ | $-0.029$ | 0.004 | 0.001 |
| | (0.021) | (0.019) | (0.007) | (0.017) |
| Fixed Effects | Yes | Yes | Yes | Yes |
| BIC-chosen order | 1 | 1 | 1 | 1 |
| R-sq. | 0.331 | 0.335 | 0.871 | 0.157 |
| | Panel B: Interests in Subscribing Questions | | | |
| | Askers-Answerers | Askers-Voters | Answerers-Answerers | Answerers-Voters |
| $After$ | $0.098^{***}$ | $0.117^{***}$ | $0.068^{***}$ | $0.102^{***}$ |
| | (0.027) | (0.026) | (0.009) | (0.023) |
| $LnNewUserPlt$ | 0.016 | $0.032^{*}$ | $-0.006$ | $-0.006$ |
| | (0.017) | (0.017) | (0.006) | (0.015) |
| $LnNewTopicPlt$ | $0.032^{*}$ | $-0.006$ | 0.003 | $-0.011$ |
| | (0.019) | (0.018) | (0.006) | (0.016) |
| Fixed Effects | Yes | Yes | Yes | Yes |
| BIC-chosen order | 1 | 1 | 1 | 1 |
| R-sq. | 0.243 | 0.255 | 0.849 | 0.246 |

Note: This table reports the estimates from several linear regression models based on the log-transformed interest similarity metrics using a 90-day time window. The coefficient of $After$ captures the change in a dependent variable after the intervention date. All of the specifications include day-of-week and month fixed effects, the two continuous control variables listed in the table, and separate polynomial terms in the pre- and post-intervention periods, respectively. The polynomial order is selected using the BIC. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

## CONCLUSION

This paper investigates the impact of this social filtering function on platforms' content ecosystem using a large-scale quasi-experiment on a knowledge-sharing platform. This platform initially used a content-based algorithm, which distributes content to subscribers of topic tags associated with the content, and later completely switched to a social filtering algorithm, which distributes content to the creator's online social networks. We find that this intervention substantially decreased the quantity and quality of answer contributions, while slightly increasing the consensus across users

who voted for the same answer. We identify two parallel mechanisms driving the observed effects. First, social filtering limits the dissemination scale of questions. The reduced visibility limits the likelihood that a question will be seen by users who are both willing and able to respond, potentially reducing the quantity and quality of answers it receives. Second, social filtering increases the likelihood of content dissemination within homophilic online social clusters. While this inflates the consensus among users who engage with the same content, it decreases exposure to potentially diverse perspectives that may foster answer volume and quality, creating a harmful "illusion of harmony." For illustration, Figure 7 visualizes the differential mechanisms of the SF and CF algorithms.

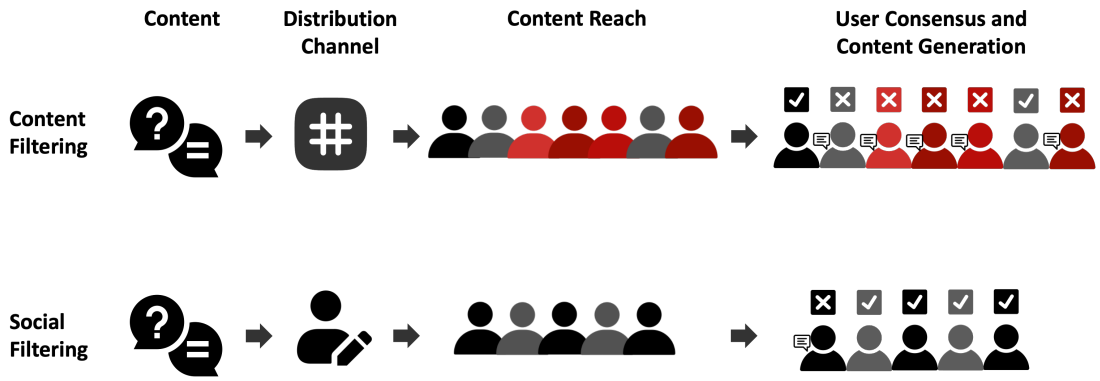



Figure 7: Overview of the Intervention Effects and Mechanisms

Note: In the final two steps, different colors represent distinct interests. The box with a checkmark indicates upvotes, while the box with a cross indicates downvotes. The chatbox symbol represents content contributions.

While the actual effects of social filtering algorithms may vary across platforms, the mechanisms uncovered in our study are general in nature and apply across diverse platforms. First, user homophily is a well-documented characteristic of social networks, and the relatively limited scale of user follower size is also evident on other platforms (Wasko and Faraj 2005). For example, anecdotal evidence from Instagram suggests the limited reach of individual social networks: only 6% of Instagram users have more than 50,000 followers (Dean 2025, Iyanu 2024), yet 50,000 is merely a modest figure compared to the reach of many Instagram hashtags (Forsey 2023). Second, many large-scale platforms (e.g., Quora, Reddit, Wikipedia) rely on algorithmic user–content matching. Although the algorithms differ in design and are not always based on social filtering, the

core principle—connecting similar users with similar content—remains consistent. In this sense, our mechanisms, particularly the role of user homophily, extend beyond social filtering to various homophily-driven algorithms. Third, while answer contribution is a distinctive behavior in Q&A contexts, it represents a broader category of UGC behaviors, such as reviewing or commenting, that can be shaped by similar recommendation dynamics. Therefore, our findings yield actionable insights for platform design well beyond the Q&A domain, with relevance for platforms such as Reddit, Instagram, as well as niche community forums and review sites.

Our findings imply that while social filtering algorithms may enhance content consumption by promoting consensus and reducing downvotes, they can significantly harm content contributions. Platforms should carefully evaluate these trade-offs, as an overemphasis on boosting consumption metrics could suppress the volume, efficiency, quality, and diversity of user content contributions, which are vital for sustaining vibrant platform ecosystems. More broadly, this highlights the importance of evaluating algorithms based on their impact on multiple user behaviors, as focusing on a single metric may lead to ineffective policy decisions.[16]

Furthermore, our research findings underscore the need for platforms to carefully balance homophily-driven recommendations with mechanisms that promote diverse interactions. When deploying homophily-driven algorithms, managers should consider integrating mechanisms that expose users to diverse and opinion-challenging viewpoints or occasionally surface content from outside a user's "comfort zone." These insights are particularly relevant and timely in light of the rise of generative AI tools, which are putting significant pressure on content platforms to increase traffic and foster genuine discussions and interactions. Lastly, by documenting the negative effects of social filtering on the platforms themselves, this paper provides valuable insights and levers for policymakers who seek to regulate platform behavior and mitigate echo chamber phenomena.

[16]Based on our discussion with the platform, Zhihu did not immediately track the effects of the intervention after its implementation. The platform may not have been aware of the significant drop in answer contributions, particularly because the intervention coincided with an increase in social interactions (as documented in Liu and Cong (2023)), which could have masked early signs of reduced content creation. Nevertheless, approximately one year after the intervention, Zhihu implemented a hybrid algorithm combining social filtering with content-based filtering, which remains in use today. This suggests that the platform may eventually learn about the trade-offs between algorithm designs.

Our study has several limitations that open fruitful directions for future research. First, we explore the comparative effects of the SF algorithm against the CF algorithm. Further studies could expand this to include other types of algorithms. Such research could offer a more comprehensive understanding of how different types of algorithms influence content creation. Second, the effectiveness of SF algorithms may vary substantially based on platform-specific characteristics and network structures. While our findings provide insights into these mechanisms, their empirical magnitude should be interpreted with caution when extrapolating to other contexts. Future investigations could explore these heterogeneous effects to better tailor algorithms to specific environments. Third, our study primarily focuses on the short-term (up to 90 days) impacts of algorithm change, due to the limitations in our research context and identification strategy. Future work could investigate the long-term effects of recommender systems on user behavior and community culture if data and context allow. Fourth, we acknowledge that the increase in upvote ratio may also stem from other SF-related factors, such as social desirability or conformity, rather than homophily alone. Future research could delve deeper into the nuanced consumer psychology under social filtering and examine how these social dynamics shape user evaluations and participation.

# APPENDIX

## DERIVING TOPIC EMBEDDINGS

### The GCN

To construct the GCN model, we start by transforming the topic tree as a directed graph object $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with $N$ nodes $v_i \in \mathcal{V}$, edges $(v_i, v_j) \in \mathcal{E}$, and an adjacency matrix $A \in \mathbb{R}^{N \times N}$. We consider a one-layer GCN model of size 128 with the following propagation rule: $H = \delta(\hat{D}^{-1/2}\hat{A}\hat{D}^{-1/2}X\Theta)$, where $\Theta$ is a learnable linear transformation applied to every node; $\hat{A} =$

$A + I_N$ is the adjacency matrix of the graph $\mathcal{G}$ with added self-connections, with $I_N$ representing the identity matrix; and $\hat{D}$ is the corresponding degree matrix, i.e., $\hat{D}_{ii} = \sum_j \hat{A}_{ij}$. The matrix $X = \{\vec{x}_1, \vec{x}_2, ..., \vec{x}_n\}$ is a matrix of node feature vectors. We apply ReLU for the nonlinearity parameter $\delta$.

To obtain node features, we use pre-trained Zhihu topic embedding from the Chinese-Word-Vectors repository by Li et al. (2018), which contains 9,612 topic tags present in our data.[17] We train the GCN model using the Deep Graph Infomax (DGI) algorithm (Veličković et al. 2018) which introduces a contrastive learning objective, where the model aims to distinguish between a node's neighborhood (i.e., positive examples) and other random nodes in the graph (i.e., negative examples). We use the trained GCN model to generate 128-dimensional embeddings for the 20,919 topic tags that were created before the intervention and had at least one subscriber.

**Validation with Visualization**

We demonstrate the validity of the derived topic embeddings through visualization and link prediction. In preparation for visualization, we classify all 20,919 topic tags into 17 categories, including Design, Arts, Business, Careers, Economics & Finance, Education, Food, Healthcare, Internet, Law, Lifestyle, Music & Games & Movies, Psychology, Reading & Writing, Science & Technology, Sports, and Travel.[18] Each category corresponds to one or more tags on Zhihu's topic tree. For example, the category "Law" maps directly to the tag "Law", while "Science & Technology" corresponds to both "Science" and "Technology". Therefore, we create a many-to-one tag-category mapping that pairs each category with its associated topic tags (hereafter, "anchor" tags). Note that these anchor tags are typically non-terminal nodes in the topic tree, meaning that they have child tags and represent broad and important disciplines.

[17] This Chinese word embedding repository provides word embeddings trained on corpora of different domains, including Chinese Wikipedia, Baidu Encyclopedia, Zhihu, and Weibo. Its Zhihu word embeddings are trained on a mass of textual data comprising 32,137 answers and 3,239,114 questions on Zhihu. These embeddings exhibit good performances on sentiment classification and other downstream tasks (Qiu et al. 2018).

[18] This category list is used by Zhihu for its knowledge market launched in 2017.

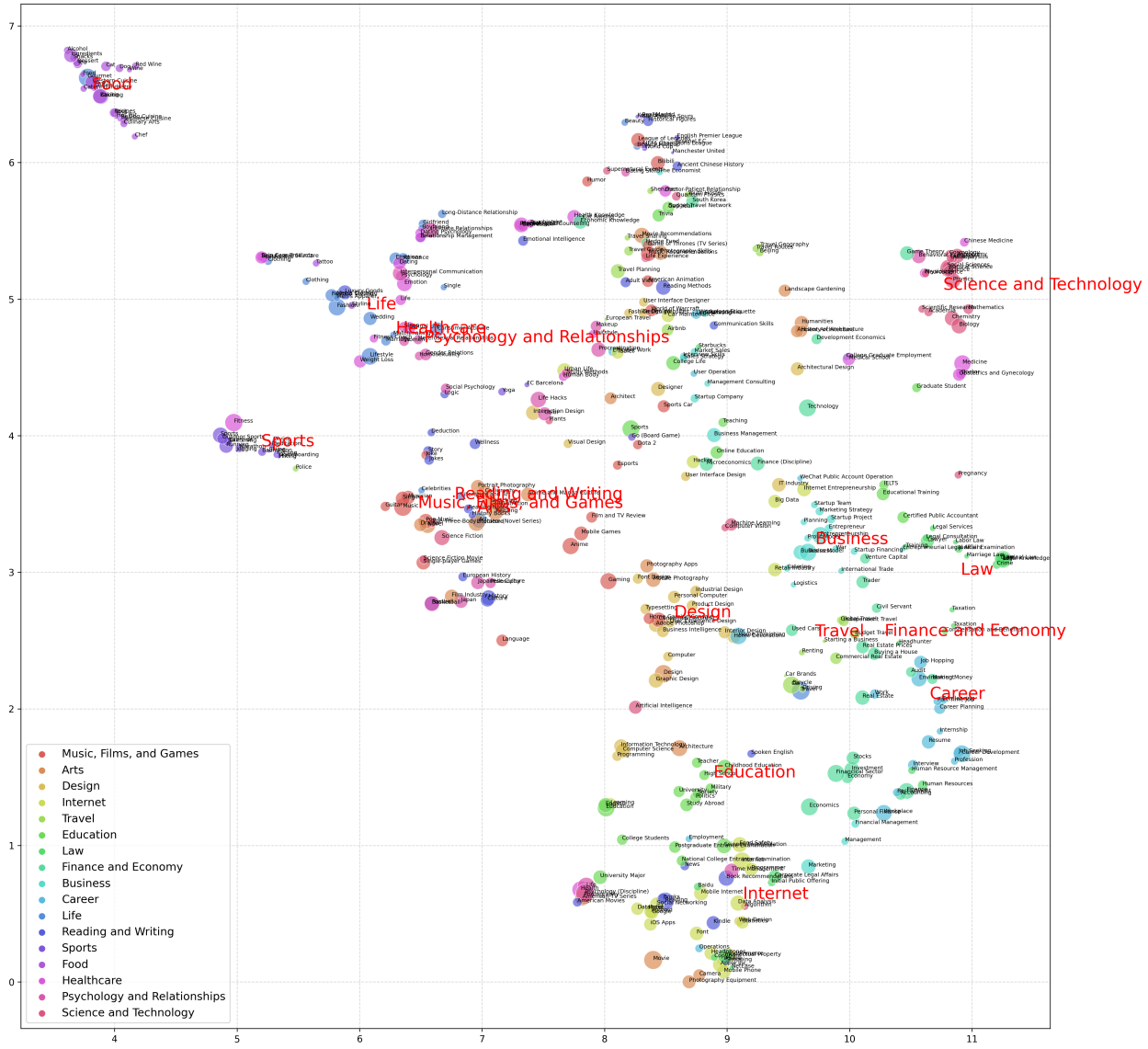



Figure 8: Visualization of Tag Embeddings

Note: The text in red are 17 representative parent tags on Zhihu's topic tree (referred to as anchor category). Each tag is then classified to its closest anchor category and colored accordingly. Details of the tag classification process are reported in the Appendix.

Our classification contains the following steps. First, we adopt the Breadth First Search (BFS) algorithm to traverse every tag's parent tags on the topic tree, until we find the first parent topic tag that appears in the tag-category mapping. Due to the nature of BFS, this parent tag has the shortest distance to the target tag relative to other anchor tags in the mapping. Consequently, we classify the target tag to the category this parent tag maps to. This step successfully classifies

5,672 topics. For the remaining tags, we use the GCN-based embeddings to find their most similar categories. For every uncategorized tag, we calculate its cosine similarity with every anchor tag in the tag-category mapping and find the anchor tag with the highest similarity. The target tag is then classified into the category corresponding to that anchor tag.[19]

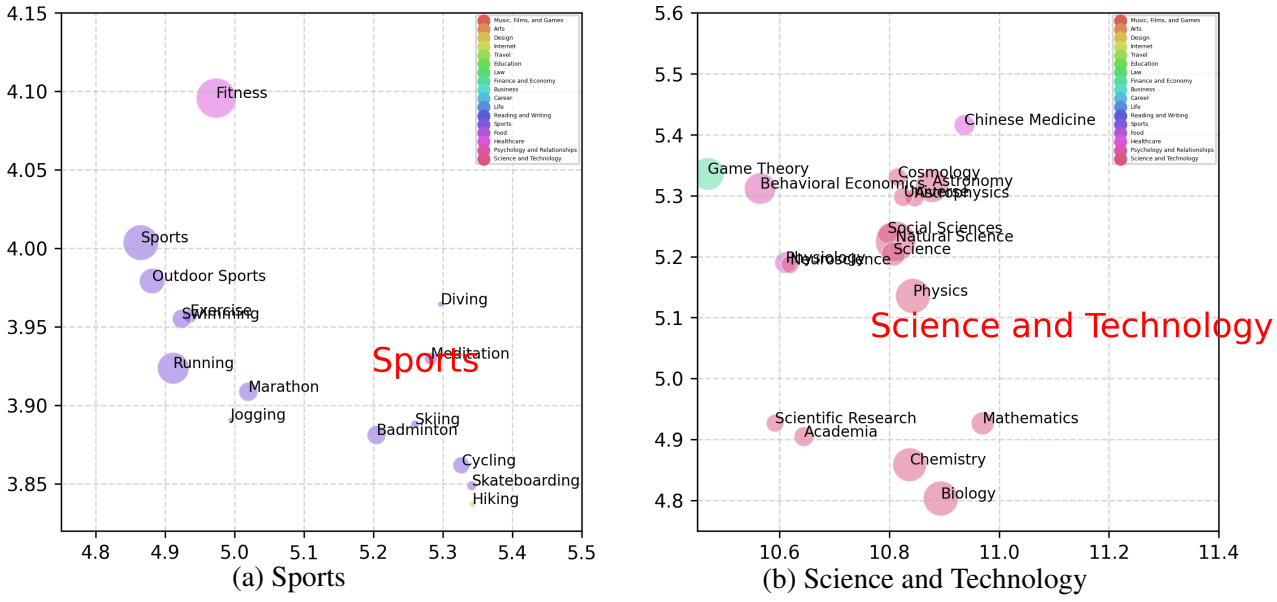
To facilitate visualization, we select the top 25 most-subscribed tags within each category, resulting in a subset of $17 \times 25 = 425$ selected tags. We use the Uniform Manifold Approximation and Projection (UMAP) (McInnes et al. 2018), a nonlinear dimension reduction method, to project these tags along with the 17 categories (referred to as anchor category) on a two-dimensional plane. For categories associated with multiple topic tags, we compute the category's embedding by averaging the embeddings of the relevant tags. For example, the embedding for "Science & Technology" is the average of the "Science" and "Technology" embeddings.

The visualization results are presented in Figure 8, where categories are in text and tags in the same category are color-coded. We find that tags with similar meanings are clustered closely, while those with different themes are positioned apart. For example, tags related to "Law", "Business", and "Finance and Economy" are clustered in the bottom-right corner, while food-related tags are placed distinctly in the top-left corner. To provide further details, we magnify two regions of this visualization. Figure 9a zooms in on the sports-related cluster in the upper-left region. Notably, "Jogging" is positioned closer to "Marathon" than to "Meditation," reflecting the stronger semantic alignment between "Jogging" and "Marathon." Figure 9b features the science and technology-related cluster in the upper-right quadrant. We observe that "Biology" is positioned closer to "Chemistry" than to "Behavioral Economics," consistent with our common knowledge.

[19] Although every tag has an embedding generated after the GCN training, tags that have no edges on the topic tree and no pre-trained embeddings only gain nonsense embeddings. These tags are classified as the "Others" category. In addition, tags that do not have pre-trained embeddings and only connect to the "Uncategorized" topic in the topic tree are also classified as "Others." In total, 4,388 topics are classified as "Others." We exclude the "Others" category throughout our analysis.

### Validation with Link Prediction

The link prediction task predicts the existence of an edge between two arbitrary tags on Zhihu's topic tree. For each candidate node pair, we compute an edge score that represents the likelihood of an edge existing between them as the dot product of their embeddings. We then compute the Area Under the Receiver Operating Characteristics Curve (AUC-ROC) based on the edge scores and the true labels (whether an edge exists or not). We find a good predictive performance with an AUC-ROC score of 0.84. Taken together, these validation steps suggest that the GCN-based tag embeddings effectively capture the semantic relations among tags.




(a) Sports (b) Science and Technology

Figure 9: Selected Regions of the Tag Embedding Space

Note: The text in red are 17 representative parent tags on Zhihu's topic tree (referred to as anchor category). Each tag is classified to its closest anchor category and colored accordingly.

# REFERENCES


Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Eugene Agichtein, Carlos Castillo, Debora Donato, Aristides Gionis, and Gilad Mishne. Finding high-quality content in social media. In *Proceedings of the 2008 international conference on web search and data mining*, pages 183–194, 2008.

Ashton Anderson, Daniel Huttenlocher, Jon Kleinberg, and Jure Leskovec. Discovering value from community activity on focused question answering sites: a case study of stack overflow. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 850–858, 2012.

Susan Athey, Markus Mobius, and Jeno Pal. The impact of aggregators on internet news consumption. Technical report, National Bureau of Economic Research, 2021.

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

Alan I Barreca, Melanie Guldi, Jason M Lindo, and Glen R Waddell. Saving babies? revisiting the effect of very low birth weight classification. *The quarterly journal of economics*, 126(4):2117–2123, 2011.

Ron Berman and Zsolt Katona. Curation algorithms and filter bubbles in social networks. *Marketing Science*, 39(2):296–316, 2020.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Zike Cao, Yingpeng Zhu, Gen Li, and Liangfei Qiu. Consequences of information feed integration on user engagement and contribution: A natural experiment in an online knowledge-sharing community. *Information Systems Research*, 35(3):1114–1136, 2024.

Damon Centola. An experimental study of homophily in the adoption of health behavior. *Science*, 334 (6060):1269–1272, 2011.

Justin Cheng, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. How community feedback shapes user behavior. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 8, pages 41–50, 2014.

Lesley Chiou and Catherine Tucker. Content aggregation by platforms: The case of the news media. *Journal of Economics & Management Strategy*, 26(4):782–805, 2017.

Matteo Cinelli, Gianmarco De Francisci Morales, Alessandro Galeazzi, Walter Quattrociocchi, and Michele Starnini. The echo chamber effect on social media. *Proceedings of the National Academy of Sciences*, 118(9):e2023301118, 2021.

Emily Dean. 80+ instagram statistics for 2025, January 2025.

Shengli Deng, Yuling Fang, Yong Liu, and Hongxiu Li. Understanding the factors influencing user experience of social question and answer services. *Information Research: An International Electronic Journal*, 20(4):n4, 2015.

Seth Flaxman, Sharad Goel, and Justin M Rao. Filter bubbles, echo chambers, and online news consumption. *Public Opinion Quarterly*, 80(S1):298–320, 2016.

Caroline Forsey. 601 most popular instagram hashtags in 2023. https://blog.hubspot.com/marketing/instagram-hashtags, October 2023.

Andrew Gelman and Guido Imbens. Why high-order polynomials should not be used in regression discontinuity designs. *Journal of Business & Economic Statistics*, 37(3):447–456, 2019.

Jacob Goldenberg, Andreas Lanz, Daniel Shapira, and Florian Stahl. Express: Targeting nearby influencers: the acceleration of natural triadic closure by leveraging interconnectors. *Journal of Marketing*, page 00222429231223420, 2023.

Thomas Graziani. Zhihu: China's largest q&a platform is a content marketer's dream. https://walkthechat.com/zhihu-chinas-largest-qa-platform-content-marketers-dream/, June 2018.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, et al. A survey on llm-as-a-judge. *arXiv preprint arXiv:2411.15594*, 2024.

Sanda Harabagiu, Dan Moldovan, Marius Pasca, Rada Mihalcea, Mihai Surdeanu, Razvan Bunsecu, Roxana Girju, Vasile Rus, and Paul Morarescu. The role of lexico-semantic feedback in open-domain

textual question-answering. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, pages 282–289, 2001.

Catherine Hausman and David S Rapson. Regression discontinuity in time: Considerations for empirical applications. *Annual Review of Resource Economics*, 10:533–552, 2018.

Cheng He, O Cem Ozturk, Chris Gu, and Jorge Mario Silva-Risso. The end of the express road for hybrid vehicles: Can governments' green product incentives backfire? *Marketing Science*, 40(1):80–100, 2021.

Wei He, Kai Liu, Jing Liu, Yajuan Lyu, Shiqi Zhao, Xinyan Xiao, Yuan Liu, Yizhong Wang, Hua Wu, Qiaoqiao She, et al. Dureader: a chinese machine reading comprehension dataset from real-world applications. *arXiv preprint arXiv:1711.05073*, 2017.

James J Heckman, Hidehiko Ichimura, and Petra Todd. Matching as an econometric evaluation estimator. *The review of economic studies*, 65(2):261–294, 1998.

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.

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.

Homa Hosseinmardi, Amir Ghasemian, Aaron Clauset, Markus Mobius, David M Rothschild, and Duncan J Watts. Examining the consumption of radical content on youtube. *Proceedings of the National Academy of Sciences*, 118(32):e2101967118, 2021.

Taiwo Iyanu. How many instagram followers does the average person have? (2024), March 2024.

Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

Brent Kitchens, Steven L Johnson, and Peter Gray. Understanding echo chambers and filter bubbles: The impact of social media on diversification and partisan shifts in news consumption. *MIS quarterly*, 44(4), 2020.

David S Lee and Thomas Lemieux. Regression discontinuity designs in economics. *Journal of economic literature*, 48(2):281–355, 2010.

Kevin Lewis, Marco Gonzalez, and Jason Kaufman. Social selection and peer influence in an online social network. *Proceedings of the National Academy of Sciences*, 109(1):68–72, 2012.

Lei Li, Chengzhi Zhang, Daqing He, and Jia Tina Du. Researchers' judgment criteria of high-quality answers on academic social q&a platforms. *Online Information Review*, 44(3):603–623, 2020.

Shen Li, Zhe Zhao, Renfen Hu, Wensi Li, Tao Liu, and Xiaoyong Du. Analogical reasoning on chinese morphological and semantic relations. *arXiv preprint arXiv:1805.06504*, 2018.

Jia Liu and Ziwei Cong. The daily me versus the daily others: How do recommendation algorithms change user interests? evidence from a knowledge-sharing platform. *Journal of Marketing Research*, page 00222437221134237, 2023.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*, 2023.

Zhe Liu and Bernard J Jansen. Factors influencing the response rate in social question and answering behavior. In *Proceedings of the 2013 conference on Computer supported cooperative work*, pages 1263–1274, 2013.

Jane Lockwood. Handbook of automated essay evaluation current applications and new directions mark d. shermis and jill burstein (eds.)(2013). *Writing & Pedagogy*, 6(2):437–442, 2014.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, page 1906. Association for Computational Linguistics, 2020.

Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

Miller McPherson, Lynn Smith-Lovin, and James M Cook. Birds of a feather: Homophily in social networks. *Annual review of sociology*, 27(1):415–444, 2001.

Grégoire Mialon, Roberto Dessi, Maria Lomeli, Christoforos Nalmpantis, Ramakanth Pasunuru, Roberta Raileanu, Baptiste Roziere, Timo Schick, Jane Dwivedi-Yu, Asli Celikyilmaz, et al. Augmented language models: a survey. *Transactions on Machine Learning Research*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

O Cem Ozturk, Pradeep K Chintagunta, and Sriram Venkataraman. Consumer response to chapter 11 bankruptcy: Negative demand spillover to competitors. *Marketing Science*, 38(2):296–316, 2019.

Bo Pang, Erik Nijkamp, Wenjuan Han, Linqi Zhou, Yixian Liu, Kewei Tu, et al. Towards holistic and automatic evaluation of open-domain dialogue generation. Association for Computational Linguistics (ACL), 2020.

Ellie Pavlick and Tom Kwiatkowski. Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694, 2019.

Christian Peukert, Ananya Sen, and Jörg Claussen. The editor and the algorithm: Recommendation technology in online news. *Management science*, 70(9):5816–5831, 2024.

Kun Qian and Sanjay Jain. Digital content creation: An analysis of the impact of recommendation systems. *Management Science*, 2024.

Yuanyuan Qiu, Hongzheng Li, Shen Li, Yingdi Jiang, Renfen Hu, and Lijiao Yang. Revisiting correlations between intrinsic and extrinsic evaluations of word embeddings. In *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data: 17th China National Conference, CCL 2018, and 6th International Symposium, NLP-NABD 2018, Changsha, China, October 19–21, 2018, Proceedings 17*, pages 209–221. Springer, 2018.

Salvador Rodriguez. Mark zuckerberg shifted facebook's focus to groups after the 2016 election, and it's changed how people use the site. https://www.cnbc.com/2020/02/16/zuckerbergs-focus-on-facebook-groups-increases-facebook-engagement.html, February 2020.

Pradeep Kumar Roy, Sunil Saumya, Jyoti Prakash Singh, Snehasish Banerjee, and Adnan Gutub. Analysis of community question-answering issues via machine learning and deep learning: State-of-the-art review. *CAAI Transactions on Intelligence Technology*, 8(1):95–117, 2023.

Feng Shi, Misha Teplitskiy, Eamon Duede, and James A Evans. The wisdom of polarized crowds. *Nature human behaviour*, 3(4):329–336, 2019.

Rion Snow, Brendan O'connor, Dan Jurafsky, and Andrew Y Ng. Cheap and fast–but is it good? evaluating

non-expert annotations for natural language tasks. In *Proceedings of the 2008 conference on empirical methods in natural language processing*, pages 254–263, 2008.

Christian Stab and Iryna Gurevych. Annotating argument components and relations in persuasive essays. In *Proceedings of COLING 2014, the 25th international conference on computational linguistics: Technical papers*, pages 1501–1510, 2014.

Jessica Su, Aneesh Sharma, and Sharad Goel. The effect of recommendations on network structure. In *Proceedings of the 25th international conference on World Wide Web*, pages 1157–1167, 2016.

Kaveh Taghipour and Hwee Tou Ng. A neural approach to automated essay scoring. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 1882–1891, 2016.

Prasad Vana and Anja Lambrecht. The effect of individual online reviews on purchase likelihood. *Marketing Science*, 40(4):708–730, 2021.

Petar Veličković, William Fedus, William L Hamilton, Pietro Liò, Yoshua Bengio, and R Devon Hjelm. Deep graph infomax. *arXiv preprint arXiv:1809.10341*, 2018.

Vikas Verma, Thang Luong, Kenji Kawaguchi, Hieu Pham, and Quoc Le. Towards domain-agnostic contrastive learning. In *International Conference on Machine Learning*, pages 10530–10541. PMLR, 2021.

Molly McLure Wasko and Samer Faraj. Why should i share? examining social capital and knowledge contribution in electronic networks of practice. *MIS quarterly*, pages 35–57, 2005.

Douglas Brent West et al. *Introduction to graph theory*, volume 2. Prentice hall Upper Saddle River, 2001.

Kaya Yurieff. Facebook is pushing groups to your news feed. https://www.cnn.com/2020/10/01/tech/facebook-groups-news-feed/index.html, October 2020.

Jun Zhang, Mark S Ackerman, and Lada Adamic. Expertise networks in online communities: structure and algorithms. In *Proceedings of the 16th international conference on World Wide Web*, pages 221–230, 2007.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623, 2023.

Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy

Ba. Large language models are human-level prompt engineers. In *The Eleventh International Conference on Learning Representations*, 2022.

# Web Appendix

## When Connections Depress Contributions: The Hidden Cost of Social Filtering

## A Measuring Answer Quality Using LLMs

Evaluating user-generated text remains a challenging task due to the inherent subjectivity of language and the diversity of valid expression. Traditional approaches of scoring answer quality, including manual evaluation and trained machine learning models with labeled data, present significant limitations for large-scale, complex Q&A datasets such as Zhihu. Specifically, manual annotation is generally costly, time-consuming, and difficult to scale. Because answers span diverse topics and vary in formality on Zhihu, it is unrealistic to assume that crowd workers or annotators possess sufficient domain knowledge to evaluate factual correctness or insightfulness, particularly in nuanced or ambiguous cases (Snow et al. 2008, Pavlick and Kwiatkowski 2019). Similarly, the performance of traditional supervised machine learning models is tightly bound to the quality and scope of the annotated training dataset (Taghipour and Ng 2016, Verma et al. 2021). Such models often fail to generalize to out-of-domain questions or to assess deeper semantic quality, such as originality and logical coherence. They also lack external knowledge access, and hence tend to misjudge fabricated content (Maynez et al. 2020).

We address these challenges by leveraging the strengths of LLMs for evaluating user-generated answers. First, LLMs are pre-trained on massive, diverse corpora, giving them broad knowledge coverage and the ability to assess both surface fluency and content-level correctness across a wide range of topics (Brown et al. 2020, Achiam et al. 2023). This makes them especially well-suited for evaluating fact-based answers without needing extensive task-specific fine-tuning. Second, LLMs demonstrate strong generalization and reasoning ability, enabling them to judge open-ended responses for qualities such as coherence, relevance, and insight. Prior research has shown that LLMs can replicate human judgment in evaluation tasks with high consistency (Zheng et al. 2023),

and even outperform human annotators in cases requiring domain-agnostic reasoning. Third, LLMs offer high scalability and consistency. Unlike human raters, they do not suffer from fatigue, bias drift, or inter-rater variability. When paired with well-engineered prompts, LLMs can deliver structured, explainable, and reproducible assessments across large datasets (Liu et al. 2023).

In conclusion, the platform's broad topical coverage, user-generated nature, and mixture of fact-based and open-ended questions present a unique challenge that aligns well with LLMs' capabilities. Rather than relying on domain-specific classifiers or expensive human labor, LLM-based evaluation enables scalable, flexible, and semantically-aware scoring that aligns with both current best practices and emerging trends in NLP-based assessment (Pang et al. 2020).

### A.1 The Quality Evaluation Framework

To systematically evaluate the quality of answers on Zhihu, we designed a multi-dimensional evaluation framework informed by prior literature on answer quality assessment and automated essay scoring. The final rubric comprises five key dimensions: (Expected) Answer Length, Logic & Structure, Grammar & Diction, Relevance, and Accuracy / Insight. The rubric aims to balance comprehensiveness with practical applicability, ensuring that each dimension is clearly defined and can be applied consistently across diverse responses. These criteria reflect a synthesis of linguistic quality, argumentative soundness, and content fidelity—core elements in assessing answer quality in user-generated contexts.

- **(Expected) Answer Length** serves as a proxy for information richness and user engagement. Rather than assessing absolute length, we consider whether a response aligns with the expected depth given the question's complexity. Prior studies show that elaborated answers are often perceived as more helpful (Agichtein et al. 2008, Anderson et al. 2012), though excessive verbosity may undermine clarity. We therefore evaluate whether an answer falls within a reasonable range of length relative to the question itself, as both overly brief and unnecessarily long responses may signal inadequate or unfocused content.

- **Logic & Structure** captures the internal coherence of the response, including the presence of clear argumentation, transitions, and hierarchical organization. This criterion draws upon frameworks in argument mining and discourse analysis (Stab and Gurevych 2014), which emphasize the importance of well-structured reasoning in educational and QA contexts.

- **Grammar & Diction** evaluates the fluency and clarity of expression, including syntactic correctness and lexical choice. Prior work in automated essay scoring (Lockwood 2014) and peer-review quality detection (Roy et al. 2023) consistently show that linguistic fluency contributes significantly to perceived credibility and professionalism of an answer.

- **Relevance** assesses the degree to which a response directly addresses the posed question. This aligns with foundational QA evaluation metrics (Harabagiu et al. 2001), which distinguish between general elaboration and context-sensitive, question-specific information.

- **Accuracy / Insight** is arguably the most substantive dimension, encompassing factual correctness in closed questions and depth of perspective in open-ended ones. This criterion is informed by research on factual consistency in natural language generation (Maynez et al. 2020) and on the value of insightful contributions in online communities (Zhang et al. 2007). For fact-based questions, we prioritize factual alignment and penalize fabricated or misleading content. For open-ended questions, subjective viewpoints are acceptable, but we emphasize the importance of originality, coherence, and critical thinking.

### A.2 Prompt Engineering for Quality Evaluation

To operationalize the evaluation framework within an LLM, we adopt a structured prompt engineering approach that integrates all key components of the scoring task. Effective prompt design is critical to ensuring that LLM-generated judgments remain consistent, interpretable, and aligned with human-defined rubrics (Zhou et al. 2022, Gu et al. 2024). Our prompt is composed of three functional components: (1) system instruction, which defines the evaluator's role and

constrains the output format; (2) scoring rubric embedding, which provides detailed guidelines for assessing each dimension; and (3) contextual input, which delivers the question-and-answer pair to be evaluated.

*System instruction.* The prompt begins with a system-level instruction that defines the model's role and expected behavior. Specifically, the model is positioned as *a professional evaluator for a Q&A platform*, which anchors its output style in an expert, task-oriented context. This form of persona priming has been shown to improve the performance of LLMs in alignment-sensitive tasks such as grading, reviewing, and content moderation (Ouyang et al. 2022). Equally important is the explicit instruction to only output five integer scores, one per line, without any explanations. This ensures that outputs are clean, structured, and machine-readable, facilitating downstream statistical processing. It also mitigates the risk of hallucinated justifications or inconsistent reasoning, which can undermine evaluation validity (Liu et al. 2023).

*Embedded scoring rubrics.* Each of the five evaluation dimensions is explicitly defined using a 7-point scale, with thresholds and descriptions calibrated according to our rubric design (in Web Appendix A.1). This part of the prompt not only helps standardize the interpretation of each criterion but also enables fine-grained scoring that reflects the complexity of open-ended text quality.

*Contextual input: question and answer information.* To support accurate and context-aware scoring, the prompt includes structured input fields containing the question title, detailed question description, and the user-provided answer. This format ensures that the model receives sufficient background to understand the scope and intent of the question. Such structured context injection has been shown to improve LLM task performance, especially in cases where nuanced understanding of user intent and response alignment is required (Liu et al. 2023, Mialon et al.).

The complete prompt is shown below, and its structure allows the LLM to act as a context-aware evaluator, leveraging both its linguistic understanding and factual knowledge to assign interpretable and structured quality scores.

**Prompt for LLM-based Answer Evaluation**

**You are a professional evaluator for a Q&A platform.**

Please rate the following answer according to five criteria, based on the provided question. Each score must be an integer from 1 to 7. Output only 5 numbers, one per line, in the following order: Answer Length, Logic & Structure, Grammar & Diction, Relevance, Accuracy or Insight.

**Scoring Rubrics:**

1. Answer Length
    - 1–3 = Too short or too long, lacking appropriate density
    - 4–5 = Reasonable length, acceptable density
    - 6–7 = Optimal length and rich, well-balanced content
2. Logic & Structure
    - 1–3 = Disorganized, difficult to follow
    - 4–5 = Basic structure, some logical flow
    - 6–7 = Clear structure and robust argument flow
3. Grammar & Diction
    - 1–3 = Frequent grammatical or wording errors
    - 4–5 = Occasional minor issues
    - 6–7 = Flawless grammar and precise terminology
4. Relevance
    - 1–3 = Off-topic or mostly irrelevant
    - 4–5 = Mostly on point, but misses nuances or sub-questions
    - 6–7 = Fully aligned and addresses all aspects of the question
5. Accuracy/Insight
    - Fact-based Questions:
        - 1–3 = Inaccurate or misleading

- 4–5 = Mostly accurate with minor issues
- 6–7 = Comprehensive, precise, and authoritative

- Open-ended Questions:
  - 1–3 = No insight or shallow response
  - 4–5 = Some insight or useful opinion
  - 6–7 = Deep, original insights with thoughtful perspectives

**Question Title:** `{title}`

**Question Detail:** `{question_detail}`

**Answer:** `{answer_content}`

Only output the 5 numbers, one per line. No explanations.

### A.3 Implementation Details

The evaluation is conducted using OpenAI's GPT-4o model via the ChatCompletion API, with deterministic sampling (temperature=0) to enhance scoring consistency. The input data consists of annotated question–answer pairs from Zhihu, each comprising a title, a question detail, and an answer content. For each instance, a structured prompt is constructed using the full scoring rubric (as described in Web Appendix A.2) and populated with the corresponding Q&A content.

To ensure robustness, the implementation included the following technical features:

- Retry mechanism: Each request is allowed up to 5 attempts in case of transient API failures or invalid response formats. Between attempts, a brief delay (5 seconds) is introduced to comply with rate limits and mitigate instability.

- Score validation: Returned outputs are parsed line-by-line. Only responses with exactly five valid integers in the expected range are accepted.

- Checkpointing: Intermediate results are periodically saved after every 1,000 successfully

scored answers. This allows safe recovery in case of interruption and avoids reprocessing previously completed records.

- Failure logging: Responses that failed all retry attempts are logged into a separate file for later review or manual scoring.

After successful scoring, the script has produced a final merged dataset including all original fields and five individual score dimensions. A composite score *total score* is then computed as a weighted average using the following weighting scheme: 0.1 for (Expected) Answer Length, 0.3 for Accuracy / Insight, and 0.2 for each of the remaining three dimensions. Note that our research findings remain consistent if uniform weights are applied across the five dimensions. The resulting dataset is exported to a CSV file for downstream statistical analysis and modeling. This design ensures that LLM-based scoring remains scalable, reproducible, and resilient to output variability, thereby enabling consistent evaluation over tens of thousands of Q&A instances.

## B Effects on Content Contributions

Table W1: Effects on Separate Dimensions of Answer Quality-related Metrics

| Dimensions | *LnAvgQuality* | *LnMaxQuality* | *LnMinQuality* |
|---|---|---|---|
| (Expected) Answer Length | $-0.007$ | $-0.112^{***}$ | $0.091^{***}$ |
| | (0.013) | (0.014) | (0.017) |
| Logic & Structure | $-0.004$ | $-0.094^{***}$ | $0.106^{***}$ |
| | (0.012) | (0.011) | (0.016) |
| Grammar & Diction | 0.006 | $-0.020^{***}$ | $0.061^{***}$ |
| | (0.007) | (0.007) | (0.009) |
| Relevance | 0.022 | $-0.055^{***}$ | $0.141^{***}$ |
| | (0.014) | (0.009) | (0.020) |
| Accuracy / Insight | $-0.001$ | $-0.089^{***}$ | $0.114^{***}$ |
| | (0.011) | (0.010) | (0.016) |

Note: This table reports the estimates from linear regressions of the log-transformed dependent variables using a 90-day window. The coefficient on all dimensions capture the change after the intervention. All specifications include day-of-week and month fixed effects, the two continuous controls, and separate polynomial terms in the pre- and post-intervention periods, with order chosen by BIC. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

Table W2: Effects on Answer Quantity-Related Metrics: Robustness to Polynomial Orders

| Order | *LnAnsPerQ* | *LnFirstAnsResTime* | *AnsOdds* |
|---|---|---|---|
| 1 | $-0.593^{***}$ | $1.880^{***}$ | $-0.643^{***}$ |
| | (0.044) | (0.162) | (0.032) |
| 2 | $-0.562^{***}$ | $1.770^{***}$ | $-0.624^{***}$ |
| | (0.056) | (0.206) | (0.040) |
| 3 | $-0.435^{***}$ | $1.342^{***}$ | $-0.396^{***}$ |
| | (0.077) | (0.284) | (0.056) |

*Note*: The first two columns report the estimates from linear regressions of the log-transformed dependent variables using a 90-day window, while the last column is a binomial regression of the number of answered questions across all new questions per day. The coefficient of $After$ captures the change in the dependent variable after the intervention. All specifications include day-of-week and month fixed effects, the two continuous control variables (daily new users and topics), and separate polynomial terms in the pre- and post-intervention periods, respectively. Polynomial orders from one to three are compared. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

Table W3: Effects on Answer Quality- and Voting-Related Metrics: Robustness to Polynomial Orders

| Order | *LnAvgQuality* | *LnMaxQuality* | *LnMinQuality* | *LnUpvoteRatio* | *LnUpvote* | *LnDownvote* |
|---|---|---|---|---|---|---|
| 1 | 0.008 | $-0.065^{***}$ | $0.094^{***}$ | $0.049^{***}$ | $-0.030$ | $-0.204^{***}$ |
| | (0.009) | (0.009) | (0.012) | (0.013) | (0.058) | (0.057) |
| 2 | $-0.003$ | $-0.077^{***}$ | $0.087^{***}$ | $0.045^{***}$ | $-0.101$ | $-0.155^{**}$ |
| | (0.012) | (0.011) | (0.016) | (0.016) | (0.074) | (0.072) |
| 3 | $-0.014$ | $-0.083^{***}$ | $0.073^{***}$ | $0.047^{**}$ | $-0.153$ | $-0.180^{*}$ |
| | (0.016) | (0.016) | (0.022) | (0.023) | (0.105) | (0.102) |

Note: This table reports the estimates from several linear regression models of the log-transformed dependent variables using a 90-day time window. The coefficient of $After$ captures the change in a dependent variable after the intervention date. All of the specifications include day-of-week and month fixed effects, the two continuous control variables (i.e., the daily number of new users and topics), and separate polynomial terms in the pre- and post-intervention periods, respectively. We compare orders from one to three. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

Table W4: Effects on Answer Quantity-Related Metrics: Robustness to "Donut" RD

| Dropping Window | *LnAnsPerQ* | *LnFirstAnsResTime* | *AnsOdds* |
|---|---|---|---|
| 1-Day | $-0.602^{***}$ | $1.939^{***}$ | $-0.412^{***}$ |
| | (0.045) | (0.164) | (0.057) |
| 2-Day | $-0.611^{***}$ | $1.961^{***}$ | $-0.424^{***}$ |
| | (0.046) | (0.170) | (0.056) |
| 3-Day | $-0.612^{***}$ | $1.978^{***}$ | $-0.446^{***}$ |
| | (0.048) | (0.176) | (0.055) |

*Note*: The first two columns report estimates from linear regressions of the log-transformed dependent variables using a 90-day window, while the last column is a binomial regression of the number of answered questions across all new questions per day, dropping all observations within one, two, or three days before and after the intervention date. The coefficient of $After$ captures the change in the dependent variable after the intervention. All specifications include day-of-week and month fixed effects, the two continuous control variables (daily new users and topics), and separate polynomial terms in the pre- and post-intervention periods. The polynomial order is selected using the BIC. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

Table W5: Effects on Answer Quality- and Voting-Related Metrics: Robustness to "Donut" RD

| Dropping Window | *LnAvgQuality* | *LnMaxQuality* | *LnMinQuality* | *LnUpvoteRatio* | *LnUpvote* | *LnDownvote* |
|---|---|---|---|---|---|---|
| 1-Day | 0.011 | $-0.060^{***}$ | $0.096^{***}$ | $0.054^{***}$ | $-0.028$ | $-0.214^{***}$ |
| | (0.009) | (0.009) | (0.013) | (0.013) | (0.060) | (0.058) |
| 2-Day | 0.014 | $-0.058^{***}$ | $0.100^{***}$ | $0.051^{***}$ | $-0.006$ | $-0.202^{***}$ |
| | (0.010) | (0.010) | (0.013) | (0.013) | (0.062) | (0.060) |
| 3-Day | $0.018^{*}$ | $-0.056^{***}$ | $0.104^{***}$ | $0.051^{***}$ | 0.001 | $-0.213^{***}$ |
| | (0.010) | (0.010) | (0.013) | (0.014) | (0.064) | (0.062) |

Note: This table reports the estimates from several linear regression models of the log-transformed dependent variables using a 90-day window specification while dropping all of the observations within one/two/three days right before and after the intervention date. The coefficient of $After$ captures the change in a dependent variable after the intervention date. All of the specifications include day-of-week and month fixed effects, the two continuous control variables (i.e., the daily number of new users and topics), and separate polynomial terms in the pre- and post-intervention periods, respectively. The polynomial order is selected using the BIC. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

Table W6: Effects on Answer-Quantity Related Metrics: Robustness to Serial Correlations in Outcome Variables

| | *LnAnsPerQ* | *LnFirstAnsResTime* | *AnsOdds* |
|---|---|---|---|
| $After$ | $-0.565^{***}$ | $1.370^{***}$ | $-0.411^{***}$ |
| | (0.053) | (0.202) | (0.059) |
| $LnY_{t-1}$ | 0.052 | $0.368^{***}$ | 0.122 |
| | (0.091) | (0.100) | (0.188) |
| $LnY_{t-2}$ | 0.103 | 0.125 | $0.309^{*}$ |
| | (0.085) | (0.102) | (0.187) |
| $LnY_{t-3}$ | $-0.082$ | $-0.073$ | $0.586^{***}$ |
| | (0.073) | (0.096) | (0.165) |
| $LnNewUserPlt$ | $-0.022$ | 0.037 | $-0.141^{***}$ |
| | (0.029) | (0.099) | (0.021) |
| $LnNewTopicPlt$ | $-0.061^{**}$ | 0.049 | $-0.010$ |
| | (0.030) | (0.106) | (0.022) |
| Fixed Effects | Yes | Yes | Yes |
| BIC-chosen order | 1 | 1 | 3 |
| (McFadden) R-sq. | 0.905 | 0.902 | 0.553 |

*Note*: The first two columns report the estimates from linear regressions of the log-transformed dependent variables using a 90-day window, while the last column is a binomial regression of the number of answered questions across all new questions per day. The coefficient of $After$ captures the change in the dependent variable after the intervention. All specifications include day-of-week and month fixed effects, the two continuous control variables (daily new users and topics), the lagged dependent variables over the previous three days (the last column uses answered rate), and separate polynomial terms in the pre- and post-intervention periods. The polynomial order is selected using the BIC. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

Table W7: Effects on Answer Quality- and Voting-Related Metrics: Robustness to Serial Correlations in Outcome Variables

| | *LnAvgQuality* | *LnMaxQuality* | *LnMinQuality* | *LnUpvoteRatio* | *LnUpvote* | *LnDownvote* |
|---|---|---|---|---|---|---|
| $After$ | 0.007 | $-0.064^{***}$ | $0.095^{***}$ | $0.050^{***}$ | $-0.028$ | $-0.187^{***}$ |
| | (0.009) | (0.009) | (0.012) | (0.013) | (0.059) | (0.058) |
| $LnY_{t-1}$ | 0.027 | 0.022 | 0.029 | $-0.049$ | $-0.013$ | $-0.247$ |
| | (0.022) | (0.020) | (0.034) | (0.073) | (0.122) | (0.462) |
| $LnY_{t-2}$ | $-0.033$ | $-0.020$ | $-0.052$ | 0.039 | $-0.120$ | 0.598 |
| | (0.022) | (0.020) | (0.033) | (0.071) | (0.117) | (0.428) |
| $LnY_{t-3}$ | 0.002 | 0.006 | 0.002 | $-0.051$ | 0.098 | 0.487 |
| | (0.017) | (0.015) | (0.025) | (0.055) | (0.109) | (0.423) |
| *LnNewUserPlt* | $0.016^{***}$ | $0.012^{**}$ | $0.020^{**}$ | 0.012 | $-0.090^{**}$ | $-0.137^{***}$ |
| | (0.006) | (0.006) | (0.008) | (0.008) | (0.038) | (0.037) |
| *LnNewTopicPlt* | $-0.008$ | $-0.012^{*}$ | $-0.005$ | $-0.001$ | $-0.015$ | 0.000 |
| | (0.006) | (0.006) | (0.008) | (0.009) | (0.040) | (0.039) |
| Fixed Effects | Yes | Yes | Yes | Yes | Yes | Yes |
| BIC-chosen order | 1 | 1 | 1 | 1 | 1 | 1 |
| R-sq. | 0.524 | 0.591 | 0.848 | 0.395 | 0.427 | 0.295 |

Note: This table reports the estimates from several linear regression models of the log-transformed dependent variables using a 90-day time window. The coefficient of $After$ captures the change in a dependent variable after the intervention date. All of the specifications include day-of-week and month fixed effects, the two continuous control variables (i.e., the daily number of new users and topics), the lagged dependent variables over the previous three days, and separate polynomial terms in the pre- and post-intervention periods, respectively. The polynomial order is selected using the BIC. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

Table W8: Effects on Answer Quantity-Related Metrics: Robustness to Lagged Control Variables

| | *LnAnsPerQ* | *LnFirstAnsResTime* | *AnsOdds* |
|---|---|---|---|
| $After$ | $-0.594^{***}$ | $1.885^{***}$ | $-0.419^{***}$ |
| | (0.044) | (0.164) | (0.057) |
| $LnNewUserPlt$ | $-0.035$ | 0.075 | $-0.127^{***}$ |
| | (0.030) | (0.113) | (0.022) |
| $LnNewUserPlt_{t-1}$ | $-0.029$ | 0.088 | $-0.108^{***}$ |
| | (0.027) | (0.101) | (0.020) |
| $LnNewUserPlt_{t-2}$ | 0.037 | $-0.046$ | $0.044^{**}$ |
| | (0.027) | (0.100) | (0.020) |
| $LnNewTopicPlt$ | $-0.053^{*}$ | 0.012 | 0.005 |
| | (0.031) | (0.116) | (0.023) |
| $LnNewTopicPlt_{t-1}$ | 0.009 | 0.012 | $0.050^{**}$ |
| | (0.031) | (0.116) | (0.022) |
| $LnNewTopicPlt_{t-2}$ | $-0.017$ | $-0.001$ | $-0.051^{**}$ |
| | (0.031) | (0.117) | (0.023) |
| Fixed Effects | Yes | Yes | Yes |
| BIC-chosen order | 1 | 1 | 3 |
| (McFadden) R-sq. | 0.905 | 0.891 | 0.562 |

*Note*: The first two columns report the estimates from linear regressions of the log-transformed dependent variables using a 90-day window, while the last column is a binomial regression of the number of answered questions across all new questions per day. The coefficient of $After$ captures the change in the dependent variable after the intervention date. All specifications include day-of-week and month fixed effects, the six continuous control variables listed in the table, and separate polynomial terms in the pre- and post-intervention periods, respectively. The polynomial order is selected using the BIC. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

Table W9: Effects on Answer Quality- and Voting-Related Metrics: Robustness to Lagged Control Variables

| | *LnAvgQuality* | *LnMaxQuality* | *LnMinQuality* | *LnUpvoteRatio* | *LnUpvote* | *LnDownvote* |
|---|---|---|---|---|---|---|
| $After$ | 0.008 | $-0.064^{***}$ | $0.094^{***}$ | $0.049^{***}$ | $-0.029$ | $-0.203^{***}$ |
| | (0.009) | (0.009) | (0.012) | (0.013) | (0.059) | (0.058) |
| *LnNewUserPlt* | $0.013^{**}$ | 0.009 | $0.019^{**}$ | 0.009 | $-0.079^{*}$ | $-0.126^{***}$ |
| | (0.006) | (0.006) | (0.009) | (0.009) | (0.041) | (0.040) |
| $LnNewUserPlt_{t-1}$ | 0.005 | 0.005 | 0.004 | 0.002 | 0.027 | 0.006 |
| | (0.006) | (0.006) | (0.008) | (0.008) | (0.036) | (0.036) |
| $LnNewUserPlt_{t-2}$ | 0.001 | 0.002 | 0.000 | 0.010 | $-0.030$ | $-0.035$ |
| | (0.006) | (0.006) | (0.008) | (0.008) | (0.036) | (0.036) |
| *LnNewTopicPlt* | $-0.007$ | $-0.011^{*}$ | $-0.003$ | 0.002 | $-0.028$ | $-0.013$ |
| | (0.007) | (0.007) | (0.009) | (0.009) | (0.042) | (0.041) |
| $LnNewTopicPlt_{t-1}$ | $-0.002$ | $-0.003$ | $-0.004$ | $-0.005$ | $-0.010$ | 0.026 |
| | (0.007) | (0.007) | (0.009) | (0.009) | (0.042) | (0.041) |
| $LnNewTopicPlt_{t-2}$ | $-0.007$ | $-0.003$ | $-0.010$ | $-0.014$ | 0.014 | 0.035 |
| | (0.007) | (0.007) | (0.009) | (0.009) | (0.042) | (0.041) |
| Fixed Effects | Yes | Yes | Yes | Yes | Yes | Yes |
| BIC-chosen order | 1 | 1 | 1 | 1 | 1 | 1 |
| R-sq. | 0.522 | 0.590 | 0.848 | 0.399 | 0.427 | 0.284 |

Note: This table reports the estimates from several linear regression models of the log-transformed dependent variables using a 90-day time window. The coefficient of $After$ captures the change in a dependent variable after the intervention date. All of the specifications include day-of-week and month fixed effects, the six continuous control variables listed in the table, and separate polynomial terms in the pre- and post-intervention periods, respectively. The polynomial order is selected using the BIC. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

Table W10: Effects on Answer Quantity-Related Metrics: Robustness to Shorter Window Specifications

| | *LnAnsPerQ* | *LnFirstAnsResTime* | *AnsOdds* |
|---|---|---|---|
| 60-Day Window Specification | | | |
| *After* | $-0.583^{***}$ | $1.858^{***}$ | $-0.603^{***}$ |
| | (0.045) | (0.156) | (0.067) |
| R-sq. | 0.918 | 0.913 | 0.617 |
| 30-Day Window Specification | | | |
| *After* | $-0.548^{***}$ | $1.814^{***}$ | $-0.666^{***}$ |
| | (0.065) | (0.255) | (0.099) |
| (McFadden) R-sq. | 0.924 | 0.893 | 0.641 |

*Note*: The first two columns report the estimates from linear regressions of the log-transformed dependent variables under different window specifications, while the last column is a binomial regression of the number of answered questions across all new questions per day. The coefficient of *After* captures the change in the dependent variable after the intervention date. All specifications include day-of-week and month fixed effects, the two continuous control variables (daily new users and topics), and separate polynomial terms in the pre- and post-intervention periods. The polynomial order is selected using the BIC. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

Table W11: Effects on Answer Quality- and Voting-Related Metrics: Robustness to Shorter Window Specifications

| | *LnAvgQuality* | *LnMaxQuality* | *LnMinQuality* | *LnUpvoteRatio* | *LnUpvote* | *LnDownvote* |
|---|---|---|---|---|---|---|
| 60-Day Window Specification | | | | | | |
| *After* | 0.005 | $-0.069^{***}$ | $0.094^{***}$ | $0.046^{***}$ | $-0.070$ | $-0.190^{***}$ |
| | (0.009) | (0.009) | (0.012) | (0.014) | (0.062) | (0.061) |
| R-sq. | 0.450 | 0.673 | 0.856 | 0.436 | 0.422 | 0.359 |
| 30-Day Window Specification | | | | | | |
| *After* | 0.011 | $-0.055^{***}$ | $0.089^{***}$ | $0.043^{*}$ | $-0.083$ | $-0.162^{*}$ |
| | (0.011) | (0.011) | (0.014) | (0.022) | (0.087) | (0.095) |
| R-sq. | 0.412 | 0.839 | 0.892 | 0.483 | 0.378 | 0.443 |

Note: This table reports the estimates from several linear regression models of the log-transformed dependent variables under different window specifications. The coefficient of *After* captures the change in a dependent variable after the intervention date. All of the specifications include day-of-week and month fixed effects, the two continuous control variables (i.e., the daily number of new users and topics), and separate polynomial terms in the pre- and post-intervention periods, respectively. The polynomial order is selected using the BIC. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

Table W12: Effects on Answer Quantity-Related Metrics: Placebo Tests

| | *LnAnsPerQ* | *LnFirstAnsResTime* | *AnsOdds* |
|---|---|---|---|
| Placebo intervention date 2012-07-02 | | | |
| $After$ | 0.097 | −0.055 | −0.025 |
| | (0.110) | (0.446) | (0.122) |
| (McFadden) R-sq. | 0.255 | 0.300 | 0.228 |
| Placebo intervention date 2011-08-16 | | | |
| $After$ | 0.052 | −0.190 | -0.041 |
| | (0.057) | (0.245) | (0.039) |
| (McFadden) R-sq. | 0.348 | 0.531 | 0.108 |

*Note*: The first two columns report estimates from linear regressions of the log-transformed dependent variables, while the last column is a binomial regression of the number of answered questions across all new questions per day. The coefficient of $After$ captures the change in the dependent variable after each placebo intervention date. All specifications include day-of-week and month fixed effects, the two continuous control variables (daily new users and topics), and separate polynomial terms in the pre- and post-intervention periods. The polynomial order is selected using the BIC. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

Table W13: Effects on Answer Quality- and Voting-Related Metrics: Placebo Tests

| | *LnAvgQuality* | *LnMaxQuality* | *LnMinQuality* | *LnUpvoteRatio* | *LnUpvote* | *LnDownvote* |
|---|---|---|---|---|---|---|
| Placebo intervention date 2012-07-02 | | | | | | |
| $After$ | 0.021 | 0.024 | 0.050 | 0.019 | 0.098 | −0.051 |
| | (0.022) | (0.022) | (0.032) | (0.037) | (0.155) | (0.128) |
| R-sq. | 0.583 | 0.449 | 0.624 | 0.302 | 0.396 | 0.391 |
| Placebo intervention date 2011-08-16 | | | | | | |
| $After$ | 0.009 | 0.014 | 0.020 | 0.007 | −0.134 | −0.097 |
| | (0.014) | (0.013) | (0.040) | (0.018) | (0.096) | (0.083) |
| R-sq. | 0.153 | 0.082 | 0.090 | 0.209 | 0.638 | 0.625 |

Note: This table reports the estimates from several linear regression models of the log-transformed dependent variables from 45 (or 90) days before to 45 (or 90) days after each placebo intervention date. The coefficient of $After$ captures the change in a dependent variable after the intervention date. All of the specifications include day-of-week and month fixed effects, the two continuous control variables (i.e., the daily number of new users and topics), and separate polynomial terms in the pre- and post-intervention periods, respectively. The polynomial order is selected using the BIC. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

Table W14: Effects on Controls

| | *LnNewUserPlt* | *LnNewTopicPlt* |
|---|---|---|
| $After$ | −0.083 | 0.069 |
| | (0.143) | (0.135) |
| Fixed Effects | Yes | Yes |
| BIC-chosen order | 1 | 1 |
| R-sq. | 0.368 | 0.353 |

Note: This table reports the estimated change in the two control variables used in the main specification using a 90-day time window. All of the specifications include day-of-week and month fixed effects and separate polynomial terms in the pre- and post-intervention periods, respectively. The polynomial order is selected using the BIC. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

Table W15: Effects on Question Quantity and Characteristics

| | *LnNewQuesPlt* | *LnTagNum* | *LnTagPopularity* | *LnPosScore* | *LnNegScore* | *Type* |
|---|---|---|---|---|---|---|
| $After$ | −0.071 | 0.005 | −0.029 | −0.017 | 0.003 | 0.020 |
| | (0.049) | (0.043) | (0.029) | (0.030) | (0.041) | (0.038) |
| *LnNewUserPlt* | 0.132*** | −0.065*** | −0.003 | 0.020 | 0.048* | 0.038 |
| | (0.025) | (0.016) | (0.018) | (0.019) | (0.026) | (0.024) |
| *LnNewTopicPlt* | 0.102*** | −0.009 | −0.045** | −0.030 | −0.037 | −0.024 |
| | (0.027) | (0.017) | (0.020) | (0.021) | (0.028) | (0.026) |
| Fixed Effects | Yes | Yes | Yes | Yes | Yes | Yes |
| BIC-chosen order | 2 | 3 | 1 | 1 | 1 | 1 |
| R-sq. | 0.759 | 0.538 | 0.599 | 0.218 | 0.172 | 0.105 |

Note: This table reports the estimates from several linear regression models of the question quantity and characteristics using a 90-day time window. All dependent variables are log-transformed continuous variables, except for $Type$, which is one if the question is classified as closed-ended and zero if open-ended. The coefficient of $After$ captures the change in a dependent variable after the intervention date. All of the specifications include day-of-week and month fixed effects, the two continuous control variables (i.e., the daily number of new users and topics), and separate polynomial terms in the pre- and post-intervention periods, respectively. The polynomial order is selected using the BIC. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

Table W16: Effects on Answer Quantity-Related Metrics: Robustness to Controlling for Daily Number of New Questions

| | *LnAnsPerQ* | *LnFirstAnsResTime* | *AnsOdds* |
|---|---|---|---|
| $After$ | $-0.628^{***}$ | $1.905^{***}$ | $-0.363^{***}$ |
| | (0.044) | (0.168) | (0.057) |
| *LnNewUserPlt* | 0.006 | 0.066 | $-0.080^{***}$ |
| | (0.029) | (0.112) | (0.023) |
| *LnNewTopicPlt* | $-0.030$ | 0.009 | 0.028 |
| | (0.031) | (0.117) | (0.023) |
| *LnNewQuesPlt* | $-0.250^{***}$ | 0.180 | $-0.551^{***}$ |
| | (0.084) | (0.321) | (0.071) |
| Fixed Effects | Yes | Yes | Yes |
| BIC-chosen order | 1 | 1 | 3 |
| (McFadden) R-sq. | 0.908 | 0.890 | 0.565 |

*Note*: The first two columns report the estimates from linear regressions of the log-transformed dependent variables using a 90-day window, while the last column is a binomial regression of the number of answered questions across all new questions per day. The coefficient of $After$ captures the change in the dependent variable after the intervention. All specifications include day-of-week and month fixed effects, the three continuous control variables listed in the table, and separate polynomial terms in the pre- and post-intervention periods. The polynomial order is selected using the BIC. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

Table W17: Effects on Answer Quality- and Voting-Related Metrics: Robustness to Controlling for Daily Number of New Questions

| | *LnAvgQuality* | *LnMaxQuality* | *LnMinQuality* | *LnUpvoteRatio* | *LnUpvote* | *LnDownvote* |
|---|---|---|---|---|---|---|
| $After$ | 0.008 | $-0.063^{***}$ | $0.092^{***}$ | $0.046^{***}$ | $-0.038$ | $-0.212^{***}$ |
| | (0.010) | (0.009) | (0.013) | (0.013) | (0.061) | (0.060) |
| *LnNewUserPlt* | $0.014^{**}$ | 0.010 | $0.019^{**}$ | 0.014 | $-0.076^{*}$ | $-0.123^{***}$ |
| | (0.006) | (0.006) | (0.009) | (0.009) | (0.040) | (0.040) |
| *LnNewTopicPlt* | $-0.008$ | $-0.013^{**}$ | $-0.003$ | 0.001 | $-0.015$ | 0.009 |
| | (0.007) | (0.007) | (0.009) | (0.009) | (0.042) | (0.042) |
| *LnNewQuesPlt* | 0.000 | 0.012 | $-0.015$ | $-0.020$ | $-0.058$ | $-0.061$ |
| | (0.018) | (0.018) | (0.025) | (0.025) | (0.116) | (0.114) |
| Fixed Effects | Yes | Yes | Yes | Yes | Yes | Yes |
| BIC-chosen order | 1 | 1 | 1 | 1 | 1 | 1 |
| R-sq. | 0.513 | 0.588 | 0.844 | 0.390 | 0.423 | 0.275 |

Note: This table reports the estimates from several linear regression models of the log-transformed dependent variables using a 90-day time window. The coefficient of $After$ captures the change in a dependent variable after the intervention date. All of the specifications include day-of-week and month fixed effects, the three continuous control variables listed in the table, and separate polynomial terms in the pre- and post-intervention periods, respectively. The polynomial order is selected using the BIC. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

# C Effect on Distribution Channel

## C.1 Follower Engagement

Table W18: Effects on Content Distribution Channel: Follower Engagement

| | *LnFollowerAnsNum* | *LnFollowerAnsRatio* | *LnFollowerVoteNum* | *LnFollowerVoteRatio* |
|---|---|---|---|---|
| $After$ | $0.652^{***}$ | $0.892^{***}$ | $0.303^{***}$ | $0.358^{***}$ |
| | (0.146) | (0.115) | (0.078) | (0.058) |
| *LnNewUserPlt* | $-0.081$ | $-0.027$ | $-0.119^{**}$ | 0.015 |
| | (0.093) | (0.073) | (0.050) | (0.037) |
| *LnNewTopicPlt* | 0.138 | 0.082 | $0.089^{*}$ | 0.012 |
| | (0.100) | (0.079) | (0.053) | (0.040) |
| Fixed Effects | Yes | Yes | Yes | Yes |
| BIC-chosen order | 1 | 1 | 1 | 1 |
| R-sq. | 0.513 | 0.672 | 0.669 | 0.755 |

Note: This table reports the estimates from several linear regression models on log-transformed follower engagement variables using a 90-day time window. The coefficient of $After$ captures the change in a dependent variable after the intervention date. All of the specifications include day-of-week and month fixed effects, the two continuous control variables listed in the table, and separate polynomial terms in the pre- and post-intervention periods, respectively. The polynomial order is selected using the BIC. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

## C.2 Alternative Measure: Social Distance

### C.2.1 Defining Social Distance

As an empirical extension, we extend the measure of online social networks by calculating social distances among users. The social distance metrics potentially take into account all interconnected users within Zhihu's social networks and thus provide a more comprehensive view of the social connections. Utilizing concepts from graph theory (West et al. 2001), we determine the social distance between two users as the minimum number of intermediary users required to connect one user to another within Zhihu's social network prior to our study period. Given our interest in the relative closeness between users, we consider Zhihu's social network as undirected when calculating the social distance. If no direct path exists between user pairs, we assign a distance

value of 10 to signify a substantial degree of separation.

There are three roles associated with each Q&A thread: askers, answerers, and voters. We focus on examining four types of social distances among these roles: asker-answerer, asker-voter, answerer-answerer, and answerer-voter. Let $u_q$ denote the user who asked question $q$, $U_q(a)$ the set of users who answered question $q$, and $U_{qa}(v)$ the set of users who voted on answer $a$ of question $q$. For any two users $(i, j)$, we let $d(i, j)$ denote their shortest social distance. For each relationship type, we calculate the mean social distance for all relevant user pairs per question. The actual formulas for quantifying these four types of social distance are listed below. To align with the RDiT specification, we aggregate these measures to platform-day level by averaging across all questions posted on the same day.

$$D(asker, answerer)_q = \frac{1}{|U_q(a)|} \sum_{i \in U_q(a)} d(u_q, i) \tag{W1}$$

$$D(asker, voter)_q = \frac{1}{|A(q)|} \sum_{a \in A(q)} \frac{1}{|U_{qa}(v)|} \sum_{i \in U_{qa}(v)} d(u_q, i) \tag{W2}$$

$$D(answerer, answerer)_q = \frac{2}{|U_q(a)|(|U_q(a)| - 1)} \sum_{u,i \in U_q(a)} d(u, i) \tag{W3}$$

$$D(answerer, voter)_q = \frac{1}{|U_q(a)|} \sum_{u \in U_q(a)} \frac{1}{|U_{qa}(v)|} \sum_{i \in U_{qa}(v)} d(u, i) \tag{W4}$$

### C.2.2 The RDiT Results

Figure W1 displays the RDiT graphs for the four social distance variables. We observe dramatic decreases in these variables immediately after the intervention. Table W19 reports the estimation results from the RDiT model, using the 90-day time window. The effects are consistent with the visual evidence in Figure W1. Specifically, the intervention significantly reduced the social distance among users who engage with the same Q&A threads by 8 to 12% ($p < 0.01$). These suggest that the SF algorithm is more likely to engage "nearby" users.

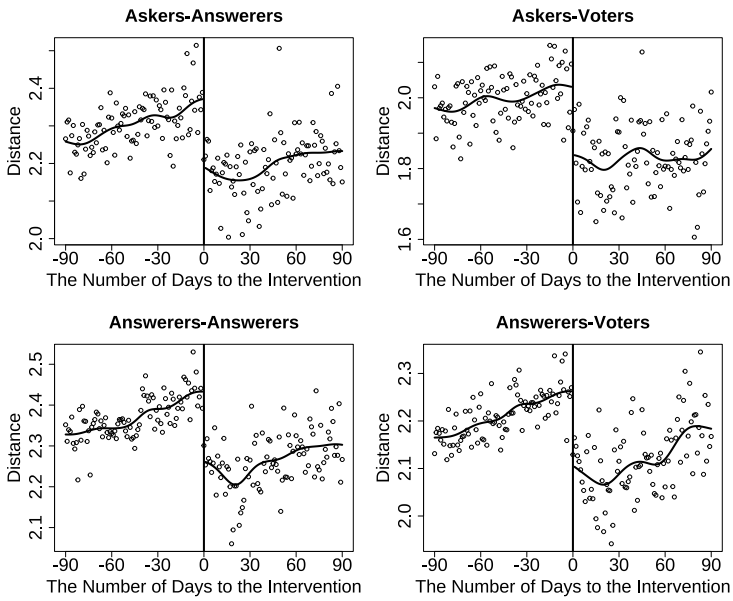



Figure W1: RDiT Graphs for Content Distribution: Social Distance

Table W19: Effects on Content Distribution Channel: Social Distance

| | Askers-Answerers | Askers-Voters | Answerers-Answerers | Answerers-Voters |
|---|---|---|---|---|
| $After$ | $-0.083^{***}$ | $-0.124^{***}$ | $-0.086^{***}$ | $-0.090^{***}$ |
| | (0.012) | (0.017) | (0.009) | (0.010) |
| $LnNewUserPlt$ | 0.003 | $-0.020^{*}$ | $0.014^{***}$ | 0.009 |
| | (0.007) | (0.011) | (0.005) | (0.006) |
| $LnNewTopicPlt$ | $-0.016^{**}$ | $-0.002$ | $-0.004$ | $-0.009$ |
| | (0.008) | (0.012) | (0.006) | (0.007) |
| Fixed Effects | Yes | Yes | Yes | Yes |
| BIC-chosen order | 1 | 1 | 1 | 1 |
| R-sq. | 0.534 | 0.560 | 0.653 | 0.574 |

Note: This table reports the estimates from several linear regression models on log-transformed social distance variables using a 90-day time window. The coefficient of $After$ captures the change in a dependent variable after the intervention date. All of the specifications include day-of-week and month fixed effects, the two continuous control variables listed in the table, and separate polynomial terms in the pre- and post-intervention periods, respectively. The polynomial order is selected using the BIC. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

### C.2.3 Robustness Checks

Table W20: Effects on Follower Engagement and Social Distance: Robustness to Polynomial Orders

| | Panel A: Follower Engagement | | | |
|---|---|---|---|---|
| Order | *LnFollowerAnsNum* | *LnFollowerAnsRatio* | *LnFollowerVoteNum* | *LnFollowerVoteRatio* |
| 1 | $0.652^{***}$ | $0.892^{***}$ | $0.303^{***}$ | $0.358^{***}$ |
| | (0.146) | (0.115) | (0.078) | (0.058) |
| 2 | $0.980^{***}$ | $1.019^{***}$ | $0.232^{**}$ | $0.357^{***}$ |
| | (0.183) | (0.147) | (0.100) | (0.074) |
| 3 | $0.868^{***}$ | $0.761^{***}$ | $0.253^{*}$ | $0.310^{***}$ |
| | (0.259) | (0.206) | (0.139) | (0.104) |
| | Panel B: Social Distance | | | |
| Order | Askers-Answerers | Askers-Voters | Answerers-Answerers | Answerers-Voters |
| 1 | $-0.083^{***}$ | $-0.124^{***}$ | $-0.086^{***}$ | $-0.090^{***}$ |
| | (0.012) | (0.017) | (0.009) | (0.010) |
| 2 | $-0.079^{***}$ | $-0.115^{***}$ | $-0.092^{***}$ | $-0.077^{***}$ |
| | (0.015) | (0.022) | (0.011) | (0.013) |
| 3 | $-0.068^{***}$ | $-0.104^{***}$ | $-0.073^{***}$ | $-0.051^{***}$ |
| | (0.021) | (0.032) | (0.015) | (0.018) |

Note: This table reports the estimates from several linear regression models of the log-transformed follower engagement and social distance variables using a 90-day time window. The coefficient of $After$ captures the change in a dependent variable after the intervention date. All of the specifications include day-of-week and month fixed effects, the two continuous control variables (i.e., the daily number of new users and topics), and separate polynomial terms in the pre- and post-intervention periods, respectively. We compare orders from one to three. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

Table W21: Effects on Follower Engagement and Social Distance: Robustness to "Donut" RD

| | Panel A: Follower Engagement | | | |
|---|---|---|---|---|
| Dropping Window | *LnFollowerAnsNum* | *LnFollowerAnsRatio* | *LnFollowerVoteNum* | *LnFollowerVoteRatio* |
| 1-Day | $0.682^{***}$ | $0.947^{***}$ | $0.293^{***}$ | $0.370^{***}$ |
| | (0.149) | (0.117) | (0.079) | (0.059) |
| 2-Day | $0.638^{***}$ | $0.930^{***}$ | $0.290^{***}$ | $0.374^{***}$ |
| | (0.152) | (0.120) | (0.081) | (0.061) |
| 3-Day | $0.585^{***}$ | $0.919^{***}$ | $0.250^{***}$ | $0.345^{***}$ |
| | (0.157) | (0.125) | (0.083) | (0.061) |
| | Panel B: Social Distance | | | |
| Dropping Window | Askers-Answerers | Askers-Voters | Answerers-Answerers | Answerers-Voters |
| 1-Day | $-0.084^{***}$ | $-0.127^{***}$ | $-0.087^{***}$ | $-0.093^{***}$ |
| | (0.012) | (0.018) | (0.009) | (0.010) |
| 2-Day | $-0.088^{***}$ | $-0.130^{***}$ | $-0.088^{***}$ | $-0.095^{***}$ |
| | (0.012) | (0.018) | (0.009) | (0.010) |
| 3-Day | $-0.089^{***}$ | $-0.128^{***}$ | $-0.088^{***}$ | $-0.096^{***}$ |
| | (0.013) | (0.019) | (0.009) | (0.011) |

Note: This table reports the estimates from several linear regression models of the log-transformed follower engagement and social distance variables using a 90-day window specification while dropping all of the observations within one/two/three days right before and after the intervention date. The coefficient of $After$ captures the change in a dependent variable after the intervention date. All of the specifications include day-of-week and month fixed effects, the two continuous control variables (i.e., the daily number of new users and topics), and separate polynomial terms in the pre- and post-intervention periods, respectively. The polynomial order is selected using the BIC. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

Table W22: Effects on Follower Engagement and Social Distance: Robustness to Serial Correlations in Outcome Variables

| | Panel A: Follower Engagement | | | |
|---|---|---|---|---|
| | *LnFollowerAnsNum* | *LnFollowerAnsRatio* | *LnFollowerVoteNum* | *LnFollowerVoteRatio* |
| $After$ | $0.563^{***}$ | $0.916^{***}$ | $0.276^{***}$ | $0.372^{***}$ |
| | (0.160) | (0.150) | (0.082) | (0.066) |
| $LnY_{t-1}$ | 0.296 | $-2.644$ | 0.141 | $-0.181$ |
| | (0.698) | (2.167) | (0.193) | (0.620) |
| $LnY_{t-2}$ | 0.765 | 1.246 | 0.193 | 0.309 |
| | (0.683) | (2.078) | (0.184) | (0.578) |
| $LnY_{t-3}$ | 0.507 | 1.033 | $-0.055$ | $-0.672$ |
| | (0.676) | (2.013) | (0.179) | (0.561) |
| *LnNewUserPlt* | $-0.074$ | $-0.033$ | $-0.112^{**}$ | 0.017 |
| | (0.094) | (0.074) | (0.051) | (0.037) |
| *LnNewTopicPlt* | 0.134 | 0.081 | 0.077 | 0.014 |
| | (0.100) | (0.079) | (0.054) | (0.040) |
| Fixed Effects | Yes | Yes | Yes | Yes |
| BIC-chosen order | 1 | 1 | 1 | 1 |
| R-sq. | 0.520 | 0.676 | 0.673 | 0.758 |
| | Panel B: Social Distance | | | |
| | Askers-Answerers | Askers-Voters | Answerers-Answerers | Answerers-Voters |
| $After$ | $-0.085^{***}$ | $-0.126^{***}$ | $-0.086^{***}$ | $-0.089^{***}$ |
| | (0.012) | (0.018) | (0.009) | (0.010) |
| $LnY_{t-1}$ | 0.003 | $-0.101^{*}$ | $-0.003$ | 0.033 |
| | (0.034) | (0.051) | (0.025) | (0.031) |
| $LnY_{t-2}$ | $-0.009$ | 0.065 | $-0.006$ | $-0.015$ |
| | (0.033) | (0.048) | (0.025) | (0.031) |
| $LnY_{t-3}$ | $-0.014$ | 0.001 | $-0.003$ | $-0.005$ |
| | (0.025) | (0.038) | (0.019) | (0.023) |
| *LnNewUserPlt* | 0.005 | $-0.023^{**}$ | $0.015^{***}$ | 0.010 |
| | (0.008) | (0.011) | (0.006) | (0.007) |
| *LnNewTopicPlt* | $-0.016^{**}$ | $-0.003$ | $-0.004$ | $-0.009$ |
| | (0.008) | (0.012) | (0.006) | (0.007) |
| Fixed Effects | Yes | Yes | Yes | Yes |
| BIC-chosen order | 1 | 1 | 1 | 1 |
| R-sq. | 0.538 | 0.572 | 0.654 | 0.577 |

Note: This table reports the estimates from several linear regression models based on the log-transformed follower engagement and social distance variables using a 90-day time window. The coefficient of $After$ captures the change in a dependent variable after the intervention date. All of the specifications include day-of-week and month fixed effects, the two continuous control variables (i.e., the daily number of new users and topics), the lagged dependent variables over the previous three days, and separate polynomial terms in the pre- and post-intervention periods, respectively. The polynomial order is selected using the BIC. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

Table W23: Effects on Follower Engagement and Social Distance: Robustness to Lagged Control Variables

| | Panel A: Follower Engagement | | | |
|---|---|---|---|---|
| | *LnFollowerAnsNum* | *LnFollowerAnsRatio* | *LnFollowerVoteNum* | *LnFollowerVoteRatio* |
| $After$ | $0.650^{***}$ | $0.887^{***}$ | $0.304^{***}$ | $0.357^{***}$ |
| | (0.146) | (0.114) | (0.078) | (0.058) |
| $LnNewUserPlt$ | $-0.076$ | $-0.004$ | $-0.097^{*}$ | 0.034 |
| | (0.101) | (0.079) | (0.054) | (0.040) |
| $LnNewUserPlt_{t-1}$ | $-0.106$ | $-0.134^{*}$ | 0.024 | $-0.007$ |
| | (0.090) | (0.070) | (0.048) | (0.036) |
| $LnNewUserPlt_{t-2}$ | 0.007 | $-0.012$ | $-0.084^{*}$ | $-0.033$ |
| | (0.090) | (0.070) | (0.048) | (0.036) |
| $LnNewTopicPlt$ | 0.122 | 0.058 | 0.070 | 0.009 |
| | (0.103) | (0.081) | (0.056) | (0.041) |
| $LnNewTopicPlt_{t-1}$ | 0.081 | 0.119 | $-0.021$ | $-0.018$ |
| | (0.104) | (0.081) | (0.056) | (0.042) |
| $LnNewTopicPlt_{t-2}$ | 0.091 | 0.099 | 0.091 | 0.040 |
| | (0.104) | (0.082) | (0.056) | (0.042) |
| Fixed Effects | Yes | Yes | Yes | Yes |
| BIC-chosen order | 1 | 1 | 1 | 1 |
| R-sq. | 0.525 | 0.686 | 0.676 | 0.759 |
| | Panel B: Social Distance | | | |
| | Askers-Answerers | Askers-Voters | Answerers-Answerers | Answerers-Voters |
| $After$ | $-0.084^{***}$ | $-0.124^{***}$ | $-0.086^{***}$ | $-0.090^{***}$ |
| | (0.012) | (0.018) | (0.009) | (0.010) |
| $LnNewUserPlt$ | 0.004 | $-0.018$ | $0.016^{**}$ | 0.009 |
| | (0.008) | (0.012) | (0.006) | (0.007) |
| $LnNewUserPlt_{t-1}$ | $-0.004$ | $-0.003$ | $-0.002$ | 0.000 |
| | (0.007) | (0.011) | (0.005) | (0.006) |
| $LnNewUserPlt_{t-2}$ | 0.001 | 0.000 | $-0.002$ | $-0.001$ |
| | (0.007) | (0.011) | (0.005) | (0.006) |
| $LnNewTopicPlt$ | $-0.017^{**}$ | 0.000 | $-0.005$ | $-0.009$ |
| | (0.008) | (0.012) | (0.006) | (0.007) |
| $LnNewTopicPlt_{t-1}$ | 0.010 | $-0.010$ | 0.001 | 0.000 |
| | (0.008) | (0.012) | (0.006) | (0.007) |
| $LnNewTopicPlt_{t-2}$ | $-0.006$ | 0.006 | 0.002 | 0.001 |
| | (0.008) | (0.013) | (0.006) | (0.007) |
| Fixed Effects | Yes | Yes | Yes | Yes |
| BIC-chosen order | 1 | 1 | 1 | 1 |
| R-sq. | 0.540 | 0.566 | 0.654 | 0.575 |

Note: This table reports the estimates from several linear regression models based on the log-transformed follower engagement and social distance variables using a 90-day time window. The coefficient of $After$ captures the change in a dependent variable after the intervention date. All of the specifications include day-of-week and month fixed effects, the six continuous control variables listed in the table, and separate polynomial terms in the pre- and post-intervention periods, respectively. The polynomial order is selected using the BIC. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

Table W24: Effects on Follower Engagement and Social Distance: Robustness to Shorter Window Specifications

| | Panel A: Follower Engagement | | | |
|---|---|---|---|---|
| | *LnFollowerAnsNum* | *LnFollowerAnsRatio* | *LnFollowerVoteNum* | *LnFollowerVoteRatio* |
| 60-Day Window Specification | | | | |
| $After$ | $0.736^{***}$ | $0.943^{***}$ | $0.263^{***}$ | $0.362^{***}$ |
| | (0.151) | (0.127) | (0.085) | (0.064) |
| R-sq. | 0.579 | 0.702 | 0.647 | 0.750 |
| 30-Day Window Specification | | | | |
| $After$ | $0.664^{***}$ | $0.749^{***}$ | $0.265^{*}$ | $0.304^{***}$ |
| | (0.222) | (0.199) | (0.135) | (0.095) |
| R-sq. | 0.692 | 0.741 | 0.516 | 0.770 |
| | Panel B: Social Distance | | | |
| | Askers-Answerers | Askers-Voters | Answerers-Answerers | Answerers-Voters |
| 60-Day Window Specification | | | | |
| $After$ | $-0.081^{***}$ | $-0.119^{***}$ | $-0.090^{***}$ | $-0.083^{***}$ |
| | (0.013) | (0.018) | (0.009) | (0.010) |
| R-sq. | 0.596 | 0.639 | 0.735 | 0.682 |
| 30-Day Window Specification | | | | |
| $After$ | $-0.081^{***}$ | $-0.098^{***}$ | $-0.083^{***}$ | $-0.073^{***}$ |
| | (0.019) | (0.027) | (0.013) | (0.017) |
| R-sq. | 0.718 | 0.703 | 0.839 | 0.740 |

Note: This table reports the estimates from several linear regression models based on the log-transformed follower engagement and social distance variables under different window specifications. The coefficient of $After$ captures the change in a dependent variable after the intervention date. All of the specifications include day-of-week and month fixed effects, the two continuous control variables (i.e., the daily number of new users and topics), and separate polynomial terms in the pre- and post-intervention periods, respectively. The polynomial order is selected using the BIC. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

Table W25: Effects on Follower Engagement and Social Distance: Placebo Tests

| | Panel A: Follower Engagement | | | |
|---|---|---|---|---|
| | *LnFollowerAnsNum* | *LnFollowerAnsRatio* | *LnFollowerVoteNum* | *LnFollowerVoteRatio* |
| Placebo intervention date 2012-07-02 | | | | |
| $After$ | −0.244 | −0.497 | 0.164 | −0.245 |
| | (0.498) | (0.406) | (0.210) | (0.187) |
| R-sq. | 0.215 | 0.247 | 0.324 | 0.215 |
| Placebo intervention date 2011-08-16 | | | | |
| $After$ | 0.304** | 0.191 | 0.054 | 0.076 |
| | (0.122) | (0.116) | (0.112) | (0.094) |
| R-sq. | 0.423 | 0.509 | 0.208 | 0.295 |
| | Panel B: Social Distance | | | |
| | Askers-Answerers | Askers-Voters | Answerers-Answerers | Answerers-Voters |
| Placebo intervention date 2012-07-02 | | | | |
| $After$ | 0.032 | −0.007 | 0.006 | 0.025 |
| | (0.029) | (0.040) | (0.019) | (0.021) |
| R-sq. | 0.386 | 0.194 | 0.535 | 0.461 |
| Placebo intervention date 2011-08-16 | | | | |
| $After$ | −0.002 | 0.020 | 0.009 | −0.001 |
| | (0.019) | (0.037) | (0.008) | (0.013) |
| R-sq. | 0.152 | 0.082 | 0.146 | 0.140 |

Note: This table reports the estimates from several linear regression models based on the log-transformed follower engagement and social distance variables from 45 (or 90) days before to 45 (or 90) days after each placebo intervention date. The coefficient of $After$ captures the change in a dependent variable after the intervention date. All of the specifications include day-of-week and month fixed effects, the two continuous control variables (i.e., the daily number of new users and topics), and separate polynomial terms in the pre- and post-intervention periods, respectively. The polynomial order is selected using the BIC. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

## D The Role of Distribution Scale

Table W26: Heterogeneous Effects on Content Contributions by Network Scale (Question-Level)

| Dependent Variable | *LnAnsPerQ* | *LnMaxQuality* |
|---|---|---|
| | (1) | (2) |
| $After$ | $-0.270^{***}$ | $-0.059^{***}$ |
| | (0.012) | (0.007) |
| $\lvert F(u_q)\rvert$ | $0.059^{***}$ | $0.008^{***}$ |
| | (0.004) | (0.002) |
| $After \times \lvert F(u_q)\rvert$ | $0.033^{***}$ | $0.010^{***}$ |
| | (0.005) | (0.002) |
| *LnNewUserPlt* | $-0.043^{***}$ | $0.012^{***}$ |
| | (0.008) | (0.004) |
| *LnNewTopicPlt* | $-0.026^{***}$ | $-0.014^{***}$ |
| | (0.008) | (0.005) |
| Day-of-week fixed effects | Yes | Yes |
| Month fixed effects | Yes | Yes |
| BIC-selected order | 1 | 1 |
| $R^2$ | 0.036 | 0.009 |
| Num. obs. | 116,741 | 60,046 |

Note: $\lvert F(u_q)\rvert$ is standardized prior to estimation. Column (2) excludes questions with no answers. All specifications include question-creation day-of-week and month fixed effects, the two continuous controls, and separate polynomial time trends for the pre- and post-intervention periods (order chosen by BIC). $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

# E The Role of Network Homophily

## E.1 Defining Network Homophily

Table W27: Comparing Followee-Follower and Tag-Subscriber Similarity

| | Followee-Follower | | Tag-Subscriber | | $p$-value |
|---|---|---|---|---|---|
| | Mean | Std | Mean | Std | |
| Answering Questions | 0.387 | 0.351 | 0.089 | 0.390 | 0.000 |
| Subscribing Questions | 0.429 | 0.388 | 0.085 | 0.404 | 0.000 |

Note: This table reports the average cosine similarity between a focal user and each of her followers and that between a focal topic tag and each of its subscribers. Followee-follower similarity in answer contribution (question subscription) is computed for 61,122 (79,804) users who have valid embeddings and at least one follower with a valid embedding by the intervention. Tag-subscriber similarity in answer contribution (question subscription) is computed for 18,786 (19,302) topic tags with valid embeddings and at least one subscriber with a valid embedding by the intervention.

## E.2 Effects on Content Contributions by Network Homophily

Table W28: Heterogeneous Effects on Content Contributions by Network Homophily (Question-Level)

| Dependent Variable | $LnAnsPerQ$ | $LnMaxQuality$ | $LnAnsPerQ$ | $LnMaxQuality$ |
|---|---|---|---|---|
| | (1) | (2) | (3) | (4) |
| $After$ | $-0.220^{***}$ | $-0.022^{*}$ | $-0.291^{***}$ | $-0.033^{***}$ |
| | (0.027) | (0.011) | (0.022) | (0.011) |
| $LnSocHomo(u_q)^S$ | $-0.213^{***}$ | $-0.059^{***}$ | $-0.166^{***}$ | $-0.046^{***}$ |
| | (0.024) | (0.012) | (0.023) | (0.012) |
| $After \times LnSocHomo(u_q)^S$ | $-0.064^{*}$ | $-0.066^{***}$ | $-0.019$ | $-0.050^{**}$ |
| | (0.038) | (0.020) | (0.038) | (0.020) |
| $LnNewUserPlt$ | $-0.001$ | $0.016^{***}$ | 0.004 | $0.016^{***}$ |
| | (0.011) | (0.005) | (0.011) | (0.005) |
| $LnNewTopicPlt$ | $-0.016$ | $-0.014^{**}$ | $-0.023^{*}$ | $-0.014^{**}$ |
| | (0.012) | (0.006) | (0.012) | (0.006) |
| $Ln\|F(u_q)\|$ | | | $0.078^{***}$ | $0.015^{***}$ |
| | | | (0.002) | (0.001) |
| Day-of-week fixed effects | Yes | Yes | Yes | Yes |
| Month fixed effects | Yes | Yes | Yes | Yes |
| BIC-selected order | 2 | 1 | 1 | 1 |
| $R^2$ | 0.039 | 0.007 | 0.072 | 0.017 |
| Num. obs. | 57,424 | 35,237 | 57,424 | 35,237 |

Note: All specifications include day-of-week and month fixed effects, the two continuous controls, and separate polynomial time trends for the pre- and post-intervention periods (order chosen by BIC). Column (3)–(4) additionally control for distribution scale, $|F(u_q)|$. Column (1) and (3) exclude questions whose askers have no followers or whose followers lack valid embeddings needed to compute $SocHomo(u_q)^S$. Column (2) and (4) further exclude questions with no answers. Results are qualitatively unchanged when using $SocHomo(u_q)^A$. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

## E.3 Effects on Engagement Homophily

Table W29: Effects on Engagement Homophily: Robustness to Polynomial Orders

| | Panel A: Interests in Answering Questions | | | |
|---|---|---|---|---|
| Order | Askers-Answerers | Askers-Voters | Answerers-Answerers | Answerers-Voters |
| 1 | $0.094^{***}$ | $0.091^{***}$ | $0.055^{***}$ | $0.055^{**}$ |
| | (0.030) | (0.028) | (0.010) | (0.025) |
| 2 | 0.052 | 0.050 | $0.042^{***}$ | 0.049 |
| | (0.038) | (0.035) | (0.013) | (0.032) |
| 3 | 0.052 | $-0.002$ | 0.029 | 0.031 |
| | (0.054) | (0.049) | (0.018) | (0.045) |
| | Panel B: Interests in Subscribing Questions | | | |
| Order | Askers-Answerers | Askers-Voters | Answerers-Answerers | Answerers-Voters |
| 1 | $0.098^{***}$ | $0.117^{***}$ | $0.068^{***}$ | $0.102^{***}$ |
| | (0.027) | (0.026) | (0.009) | (0.023) |
| 2 | $0.089^{**}$ | $0.087^{***}$ | $0.062^{***}$ | $0.095^{***}$ |
| | (0.035) | (0.033) | (0.012) | (0.030) |
| 3 | 0.055 | 0.024 | $0.047^{***}$ | $0.096^{**}$ |
| | (0.049) | (0.046) | (0.017) | (0.042) |

Note: This table reports the estimates from several linear regression models of the log-transformed interest similarity metrics using a 90-day time window. The coefficient of $After$ captures the change in a dependent variable after the intervention date. All of the specifications include day-of-week and month fixed effects, the two continuous control variables (i.e., the daily number of new users and topics), and separate polynomial terms in the pre- and post-intervention periods, respectively. We compare orders from one to three.

Table W30: Effects on Engagement Homophily: Robustness to "Donut" RD

| | Panel A: Interests in Answering Questions | | | |
|---|---|---|---|---|
| Dropping Window | Askers-Answerers | Askers-Voters | Answerers-Answerers | Answerers-Voters |
| 1-day | $0.097^{***}$ | $0.109^{***}$ | $0.056^{***}$ | $0.059^{**}$ |
| | (0.031) | (0.028) | (0.010) | (0.025) |
| 2-day | $0.089^{***}$ | $0.095^{***}$ | $0.060^{***}$ | $0.060^{**}$ |
| | (0.032) | (0.028) | (0.010) | (0.026) |
| 3-day | $0.115^{***}$ | $0.106^{***}$ | $0.063^{***}$ | $0.068^{**}$ |
| | (0.032) | (0.029) | (0.011) | (0.027) |
| | Panel B: Interests in Subscribing Questions | | | |
| Order | Askers-Answerers | Askers-Voters | Answerers-Answerers | Answerers-Voters |
| 1-day | $0.101^{***}$ | $0.128^{***}$ | $0.067^{***}$ | $0.105^{***}$ |
| | (0.028) | (0.026) | (0.010) | (0.024) |
| 2-day | $0.110^{***}$ | $0.121^{***}$ | $0.068^{***}$ | $0.106^{***}$ |
| | (0.029) | (0.026) | (0.010) | (0.024) |
| 3-day | $0.127^{***}$ | $0.136^{***}$ | $0.073^{***}$ | $0.109^{***}$ |
| | (0.029) | (0.027) | (0.010) | (0.025) |

Note: This table reports the estimates from several linear regression models of the log-transformed interest similarity metrics using a 90-day window specification while dropping all of the observations within one/two/three days right before and after the intervention date. The coefficient of $After$ captures the change in a dependent variable after the intervention date. All of the specifications include day-of-week and month fixed effects, the two continuous control variables (i.e., the daily number of new users and topics), and separate polynomial terms in the pre- and post-intervention periods, respectively. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

Table W31: Effects on Engagement Homophily: Robustness to Serial Correlations in Outcome Variables

| | Panel A: Interests in Answering Questions | | | |
|---|---|---|---|---|
| | Askers-Answerers | Askers-Voters | Answerers-Answerers | Answerers-Voters |
| $After$ | $0.096^{***}$ | $0.088^{***}$ | $0.056^{***}$ | $0.059^{**}$ |
| | (0.031) | (0.028) | (0.010) | (0.025) |
| $LnY_{t-1}$ | $-0.053$ | $-0.014$ | $-0.084$ | $-0.316^{**}$ |
| | (0.185) | (0.170) | (0.065) | (0.157) |
| $LnY_{t-2}$ | $-0.150$ | $-0.071$ | $-0.009$ | $-0.059$ |
| | (0.174) | (0.162) | (0.062) | (0.143) |
| $LnY_{t-3}$ | 0.132 | $0.236^{*}$ | 0.057 | 0.069 |
| | (0.148) | (0.137) | (0.049) | (0.123) |
| $LnNewUserPlt$ | $-0.007$ | 0.011 | 0.003 | $-0.004$ |
| | (0.020) | (0.018) | (0.006) | (0.016) |
| $LnNewTopicPlt$ | $0.039^{*}$ | $-0.029$ | 0.005 | $-0.001$ |
| | (0.021) | (0.019) | (0.007) | (0.017) |
| Fixed Effects | Yes | Yes | Yes | Yes |
| BIC-chosen order | 1 | 1 | 1 | 1 |
| R-sq. | 0.336 | 0.349 | 0.873 | 0.184 |
| | Panel B: Interests in Subscribing Questions | | | |
| | Askers-Answerers | Askers-Voters | Answerers-Answerers | Answerers-Voters |
| $After$ | $0.100^{***}$ | $0.114^{***}$ | $0.069^{***}$ | $0.106^{***}$ |
| | (0.028) | (0.027) | (0.009) | (0.023) |
| $LnY_{t-1}$ | $-0.221$ | 0.077 | $-0.009$ | $-0.049$ |
| | (0.167) | (0.157) | (0.061) | (0.146) |
| $LnY_{t-2}$ | 0.175 | $-0.063$ | $-0.124^{**}$ | $-0.239^{*}$ |
| | (0.151) | (0.151) | (0.060) | (0.138) |
| $LnY_{t-3}$ | $-0.011$ | 0.060 | $0.096^{**}$ | 0.114 |
| | (0.133) | (0.125) | (0.046) | (0.113) |
| $LnNewUserPlt$ | 0.013 | $0.030^{*}$ | $-0.007$ | $-0.006$ |
| | (0.018) | (0.017) | (0.006) | (0.015) |
| $LnNewTopicPlt$ | $0.033^{*}$ | $-0.006$ | 0.003 | $-0.012$ |
| | (0.019) | (0.018) | (0.006) | (0.016) |
| Fixed Effects | Yes | Yes | Yes | Yes |
| BIC-chosen order | 1 | 1 | 1 | 1 |
| R-sq. | 0.254 | 0.257 | 0.855 | 0.265 |

Note: This table reports the estimates from several linear regression models based on the log-transformed interest similarity metrics using a 90-day time window. The coefficient of $After$ captures the change in a dependent variable after the intervention date. All of the specifications include day-of-week and month fixed effects, the two continuous control variables (i.e., the daily number of new users and topics), the lagged dependent variables over the previous three days, and separate polynomial terms in the pre- and post-intervention periods, respectively. The polynomial order is selected using the BIC. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

Table W32: Effects on Engagement Homophily: Robustness to Lagged Control Variables

| | Panel A: Interests in Answering Questions | | | |
|---|---|---|---|---|
| | Askers-Answerers | Askers-Voters | Answerers-Answerers | Answerers-Voters |
| $After$ | $0.091^{***}$ | $0.091^{***}$ | $0.055^{***}$ | $0.053^{**}$ |
| | (0.030) | (0.028) | (0.010) | (0.024) |
| $LnNewUserPlt$ | 0.006 | 0.008 | 0.007 | 0.005 |
| | (0.021) | (0.020) | (0.007) | (0.017) |
| $LnNewUserPlt_{t-1}$ | $-0.027$ | 0.007 | 0.000 | $-0.012$ |
| | (0.019) | (0.017) | (0.006) | (0.015) |
| $LnNewUserPlt_{t-2}$ | $-0.009$ | 0.008 | $-0.002$ | 0.004 |
| | (0.019) | (0.017) | (0.006) | (0.015) |
| $LnNewTopic$ | 0.034 | $-0.030$ | 0.005 | 0.007 |
| | (0.021) | (0.020) | (0.007) | (0.017) |
| $LnNewTopic_{t-1}$ | 0.020 | 0.009 | $-0.003$ | $-0.019$ |
| | (0.022) | (0.020) | (0.007) | (0.017) |
| $LnNewTopic_{t-2}$ | 0.013 | $-0.016$ | 0.001 | $-0.001$ |
| | (0.022) | (0.020) | (0.007) | (0.017) |
| Fixed Effects | Yes | Yes | Yes | Yes |
| BIC-chosen order | 1 | 1 | 1 | 1 |
| R-sq. | 0.343 | 0.343 | 0.872 | 0.196 |
| | Panel B: Interests in Subscribing Questions | | | |
| | Askers-Answerers | Askers-Voters | Answerers-Answerers | Answerers-Voters |
| $After$ | $0.097^{***}$ | $0.117^{***}$ | $0.067^{***}$ | $0.101^{***}$ |
| | (0.028) | (0.026) | (0.009) | (0.023) |
| $LnNewUserPlt$ | 0.022 | 0.021 | $-0.004$ | 0.001 |
| | (0.019) | (0.018) | (0.007) | (0.016) |
| $LnNewUserPlt_{t-1}$ | $-0.017$ | 0.006 | 0.001 | 0.001 |
| | (0.017) | (0.016) | (0.006) | (0.014) |
| $LnNewUserPlt_{t-2}$ | $-0.008$ | 0.026 | 0.000 | $-0.006$ |
| | (0.017) | (0.016) | (0.006) | (0.014) |
| $LnNewTopicPlt$ | 0.029 | $-0.002$ | 0.003 | $-0.008$ |
| | (0.020) | (0.018) | (0.007) | (0.016) |
| $LnNewTopicPlt_{t-1}$ | 0.010 | 0.010 | 0.002 | $-0.016$ |
| | (0.020) | (0.018) | (0.007) | (0.017) |
| $LnNewTopicPlt_{t-2}$ | 0.018 | $-0.041^{**}$ | $-0.008$ | 0.002 |
| | (0.020) | (0.018) | (0.007) | (0.017) |
| Fixed Effects | Yes | Yes | Yes | Yes |
| BIC-chosen order | 1 | 1 | 1 | 1 |
| R-sq. | 0.252 | 0.283 | 0.854 | 0.266 |

Note: This table reports the estimates from several linear regression models based on the log-transformed interest similarity metrics using a 90-day time window. The coefficient of $After$ captures the change in a dependent variable after the intervention date. All of the specifications include day-of-week and month fixed effects, the six continuous control variables listed in the table, and separate polynomial terms in the pre- and post-intervention periods, respectively. The polynomial order is selected using the BIC. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

Table W33: Effects on Engagement Homophily: Robustness to Shorter Window Specifications

| | Panel A: Interests in Answering Questions | | | |
|---|---|---|---|---|
| | Askers-Answerers | Askers-Voters | Answerers-Answerers | Answerers-Voters |
| 60-Day Window Specification | | | | |
| $After$ | $0.068^{***}$ | $0.073^{**}$ | $0.051^{***}$ | $0.058^{**}$ |
| | (0.033) | (0.030) | (0.011) | (0.026) |
| R-sq. | 0.404 | 0.351 | 0.814 | 0.229 |
| 30-Day Window Specification | | | | |
| $After$ | 0.060 | 0.046 | 0.022 | 0.029 |
| | (0.049) | (0.052) | (0.016) | (0.036) |
| R-sq. | 0.386 | 0.253 | 0.776 | 0.370 |
| | Panel B: Interests in Subscribing Questions | | | |
| | Askers-Answerers | Askers-Voters | Answerers-Answerers | Answerers-Voters |
| 60-Day Window Specification | | | | |
| $After$ | $0.093^{***}$ | $0.102^{***}$ | $0.068^{***}$ | $0.101^{***}$ |
| | (0.032) | (0.028) | (0.011) | (0.024) |
| R-sq. | 0.257 | 0.353 | 0.802 | 0.315 |
| 30-Day Window Specification | | | | |
| $After$ | $0.088^{***}$ | 0.067 | $0.049^{***}$ | $0.113^{***}$ |
| | (0.042) | (0.045) | (0.017) | (0.038) |
| R-sq. | 0.374 | 0.399 | 0.746 | 0.383 |

Note: This table reports the estimates from several linear regression models based on the log-transformed interest similarity metrics under different window specifications. The coefficient of $After$ captures the change in a dependent variable after the intervention date. All of the specifications include day-of-week and month fixed effects, the two continuous control variables (i.e., the daily number of new users and topics), and separate polynomial terms in the pre- and post-intervention periods, respectively. The polynomial order is selected using the BIC. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

Table W34: Effects on Engagement Homophily: Placebo Tests

| | Panel A: Interests in Answering Questions | | | |
|---|---|---|---|---|
| | Askers-Answerers | Askers-Voters | Answerers-Answerers | Answerers-Voters |
| Placebo intervention date 2012-07-02 | | | | |
| $After$ | 0.053 | −0.082 | −0.032 | 0.013 |
| | (0.065) | (0.068) | (0.026) | (0.061) |
| R-sq. | 0.363 | 0.256 | 0.730 | 0.270 |
| Placebo intervention date 2011-08-16 | | | | |
| $After$ | −0.019 | −0.010 | −0.006 | 0.009 |
| | (0.062) | (0.051) | (0.014) | (0.052) |
| R-sq. | 0.138 | 0.092 | 0.519 | 0.160 |
| | Panel B: Interests in Subscribing Questions | | | |
| | Askers-Answerers | Askers-Voters | Answerers-Answerers | Answerers-Voters |
| Placebo intervention date 2012-07-02 | | | | |
| $After$ | −0.040 | −0.056 | −0.020 | 0.018 |
| | (0.056) | (0.060) | (0.026) | (0.061) |
| R-sq. | 0.400 | 0.291 | 0.613 | 0.370 |
| Placebo intervention date 2011-08-16 | | | | |
| $After$ | −0.067 | −0.060 | 0.000 | 0.031 |
| | (0.063) | (0.056) | (0.015) | (0.048) |
| R-sq. | 0.095 | 0.101 | 0.386 | 0.243 |

Note: This table reports the estimates from several linear regression models based on the log-transformed interest similarity metrics from 45 (or 90) days before to 45 (or 90) days after each placebo intervention date. The coefficient of $After$ captures the change in a dependent variable after the intervention date. All of the specifications include day-of-week and month fixed effects, the two continuous control variables (i.e., the daily number of new users and topics), and separate polynomial terms in the pre- and post-intervention periods, respectively. The polynomial order is selected using the BIC. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.