Peer Recognition and Content Provision on Online: An Empirical Study from a Question-and-Answer Platform*

Xintong Han† Yushen Li‡ Tong Wang§

Preliminary and Incomplete, Please Do Not Circulate

Abstract

This paper particularly identifies the incentive effect of peer recognition on the content provision online. Using unique data from the Chinese biggest Question-and-Answer Platform, we analyzed the time data of all the influencers with more than 10,000 followers on the platform for two years. We find that the OLS method leads to underestimating the incentive of "peer recognition" for online creation by 30%~40%, which mainly comes from two channels: reputation and privacy concern. In the meanwhile, our estimation results suggest that peer incentives affect individual and commercial users differently: commercial users tend to overreact for "marketing purposes." Such results indicate that platform policies may engender externalities on the marketing strategies of commercial content producers, thus changing the diversity of contents.

Keywords: Peer recognition, Content provision, Question-and-Answer platform

JEL Classification: C26, L82, M14

*Acknowledgements: We thank Yue Zhu as well as seminar and conference participants at Concordia, Edinburgh for helpful comments. This research has received financial support from the Economic and Social Research Council (ESRC) Impact Acceleration Funding. Authors gratefully thank Zhihu (Zhihu Inc.) for data provision. The conclusions drawn from the data are those of the researchers and do not reflect the views of Zhihu Inc. All remaining errors are our own.

†Concordia University and CIREQ, Department of Economics, 1455 Boulevard de Maisonneuve Ouest, Concordia University, Montreal, H3G 1M8, Canada. Email: xintong.han@concordia.ca

‡Concordia University, Department of Economics, 1455 Boulevard de Maisonneuve Ouest, Concordia University, Montreal, H3G 1M8, Canada. Email: yushen.li@concordia.ca

§University of Edinburgh, Business School, 29 Buccleuch Pl, Edinburgh EH8 9JS. United Kingdom. Email: tong.wang@ed.ac.uk

1 Introduction

Internet companies faced a binary choice: "free speech savior or shield for scoundrels?"

— Zuboff, 2018, The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power.

The influence of online platforms has penetrated into every aspect of the economy. On one hand, the proliferation of online content brings more attention and specific commercial value to these content providers (e.g., Sun and Zhu (2013), Wu and Zhu (2019), Xu, Nian, and Cabral (2019)). On the other hand, growing online footprints result in a higher chance of exposing content providers' information. This allows not only the platforms to abuse the identity information and open the door to mass surveillance (see, Tirole (2019)) but also other malicious users to intervene in providers' real life. As a result, the theory predicts that content providers have to balance the risk of exposing their private information on the platform with the potential profit of getting more attention with the credibility that certification brings. On the empirical side, how content providers are motivated to create content by non-monetary factors (e.g., peer recognition, social responsibility, etc.), and whether the platform's policy (e.g., real-name policy, badge policy, etc.) stimulates or limits the enthusiasm of creation are still poorly understood.

This paper particularly provides a pioneer empirical study in the following three questions: 1. How are a user's contributions on an online platform would be motivated by "peer recognition?"; 2. Are users affected by privacy or reputation concerns when providing content online? 3. How does the real-name policy affect the diversity of content on the platform?

To identify the above questions, we collect data from Zhihu,[1] a popular online

[1]The Chinese meaning of Zhihu is: did you know? Zhihu (https://www.zhihu.com/) is the

Question-and-Answer website in China. Our data contains all the activities of the influencers (e.g., answering questions, voting up to others, collecting articles, etc.) on the platform over two years. A key feature is that our data also includes all interactions between these influencers, which allow the building of a social network. This additional network information further allows us to construct instrumental variables to evaluate the impact of "peer recognition" from other influencers on a given influencer's content creation. Moreover, our data also includes important information about whether a user has submitted personal information for authentication and other characteristics about the quality of the user. Through econometric methods, we develop empirical strategies to identify the content provision incentives associated with "celebrity identity" and check for several possible factors that may affect influencers' online contributions: reputation, privacy concerns, knowledge spillover, etc.

The empirical results reveal the following interesting three findings. First, we find that the OLS method leads to underestimating the incentive of "peer recognition" for online creation by 30%~40%, which mainly comes from two channels: reputation and privacy concern. Second, on average, reputation concern is the most important factor that significantly affects the incentive effect of "peer recognition": "best answers" are less likely to be motivated by "peers" to answer questions because they valued their feathers more than other users. In addition, we find that users who are worried about their information being exposed may not expose any personal information to the platform at all, and they are least likely to be easily motivated to provide content by other influencers' votes. Last, peer incentives affect individual and commercial users differently: commercial users tend to overreact for "marketing purposes." Such results indicate that platform policies may engender externalities on the marketing strategies

Chineses largest online Question-and-Answer platform that is similar to Quora in the U.S. Unlike Quora, there is no multilingual version of Zhihu, and all content on the platform is provided in Chinese. From Alexa Traffic Rank on August 2018, website traffic of Zhihu ranked 112 among all the websites in the world.

of commercial content providers (e.g., Goh, Hui, and Png (2015)), thus changing the distribution of contents: there will be more marketing content and fewer individual opinion based answers/content on the platform.

Our paper contributes to the literature in the following three aspects.

First, it significantly contributes to the literature on the motivations of online content providers. Previous literature (Lerner and Tirole (2002), Fershtman and Gandal (2011), Xu, Nian, and Cabral (2019), Han and Xu (2018)) have both theoretically and empirically focused on why people are willing to contribute and collaborate online for free. Other articles illustrate the role of money in motivating online creativity (e.g., Sun and Zhu (2013), Kuang et al. (2019), Wu and Zhu (2019)). Our study is the first to quantify the impact of "peer recognition" on the provision of content for platform users. More specifically, the paper solves an essential, endogenous problem based on our data: An influencer may receive more votes (peer recognition) because he or she is actively creating answers, or the influencer may be actively creating because he or she has received, or expects to receive, more votes. In most cases, the cost of content creation and content quality are unobserved. As a consequence, it has been nearly impossible to identify the incentives of "peer recognition" for content creation in previous literature. In this project, for the first time, we use the network information to create instrumental variables and identify the impact of a variable on users' motivation to contribute in ways other than policy shock (e.g., Zhang and Zhu (2011), Wu and Zhu (2019)) which sometimes are not purely exogenous.[2]

Second, our paper also contributes to the literature of information asymmetry under the presence of a two-sided market. There is an extensive and growing literature

[2]For example, many online platform users may have anticipated the policy and responded to it in advance. To our knowledge, another recent example of using instrumental variables to identify peer effects under the platform is Bailey et al. (2019). Compared with our paper, authors use machine learning methods to analyze what each user posted on the platform and to determine whether a user changes the phone due to a malfunctioning device or simply to a peer effect. The instrumental variables that we introduce in the paper are more intuitive and easier to apply.

that discusses the impact of platform policies on eliminating information asymmetry (e.g., Roberts (2011), Saeedi (2014), Hui et al. (2016), Hui, Saeedi, and Sundaresan (2018)). A large amount of empirical evidence shows that for e-commerce platforms, badges (certification) issued by platforms to outstanding sellers will not only reduce information asymmetry but will allow sellers to monitor the product quality better to maintain "reputation and public praise" after being certified by the platform. However, the empirical study about the platform based on content remains largely unexplored. Using the data, we are the first to study such problems empirically. In our data, we can see multiple types of content providers: self-authenticated professionals, best answerers, business users (merchants who promote their products by answering questions), and uncertified users. Each user can be one or more of these types. The wide variety helps deepen our understanding of the conflicting effects of badge policies on influencers.

Finally, our paper also makes policy suggestions from the perspective of the growing concerns about the "chilling effect" caused by the excessive supervision of privacy by platforms (e.g., Penney, 2016 and Penney, 2019, Zuboff, 2018, Tirole (2019)). Most of the economics literature shows the existence of privacy concerns and the price people are willing to pay (e.g., Goldfarb and Tucker (2012), Goh, Hui, and Png, 2015, Athey, Catalini, and Tucker, 2017, Tang, 2019), but research on the contribution of content to this remains limited. Being one of the pioneer studies in economics (e.g., Chiou and Tucker (2017), Han and Zhao (2019)), our paper confirms that platforms' policies may cause the "chilling effect" due to reputation and privacy concerns, which ultimately makes content become less diversified in the long run.

The rest of the paper is organized as follows. Section 2 provides a brief description about the online Question-and-Answer platform we use. Section 3 gives a theoretical model to illustrate an influencer's decision of content provision. Section 4 describes

the data and provides basic statistics. Section 5 explores the identification strategies of the impact of "peer recognition." Section 6 shows the potential channels that makes influencers contribute less, and Section 7 concludes.

2 Question-and-Answer platform

Question-and-answer (Q&A) platforms have been growing fast recently and attracted a considerable amount of users. Different from web search engines, users on Q&A websites can ask specific questions. Amongst all the online Q&A platforms, Quora and Zhihu are two leading Q&A websites.[3] Compared with traditional Q&A websites (e.g., Yahoo Answers, WikiAnswers), these platforms have several improvements, for example:

- Platforms allow users to build social connections: Users can vote on each other, collect each other's content, and follow each other.
- Based on the votes that each answer receives, platforms use algorithms to analyze the quality of answers and rank them under each question.
- After creating an account, a corresponding personal homepage of the user will be generated. Each user's daily activities (e.g., voting, collecting answers, etc.) will be displayed on the timeline of the homepage and sorted by time. Followers will notified when an user makes a new action.

This paper uses data from Zhihu. Zhihu is the biggest Question-and-Answer community platform in China, which was launched in January 2011 and quickly became one of the most frequently visited websites by Chinese internet users. Users on Zhihu can ask and answer questions, write articles, make comments and vote on the answers

[3]We provide in Appendix A a simple comparison of the design of user interface in Zhihu and Quora.

and articles. Up to 2019, Zhihu had more than 200 million registered users, of whom 30 million were daily active users and ask hundreds of thousands of new questions or generate other content every day.

Figure 1: An example of a Zhihu user's homepage

Figure 1 shows an example of the homepage of a given influencer. By visiting an influencer's homepage, a user observes the influencer's personal information such as nicknames, place of residence, industry, and related personal profiles on the top of the homepage. These pieces of information are voluntarily disclosed by influencers and have not been verified by the platform.

On the right of the home page, the user can check the "badge" status as to whether he/she has a blue star (self-authenticated), yellow start (best-answerer) or nothing.

The authentication system and "best answerer" reward are two features unique to the platform. Users who own a blue star must submit relevant authentication materials to the platform, including but not limited to: personal real-name information, id card, work certificate, etc. It is worth mentioning that users with less than a Ph.D. degree and without a position in a science-related industry cannot obtain blue stars. The allocation of yellow stars is based on the platform algorithm, and the platform will only award yellow stars to a tiny number of users who are considered as the best answerers in their field. Most of the time the user does not know when he or she is going to get a yellow star because the algorithm is so complicated that even if he or she answers enough questions in the relevant field and gets many votes, it does not mean that a user is going to be awarded a yellow star. Below the authentication information, it shows how many votes the influencer has received, how many other users she/he follows, and how many followers she/he has.

In the center of the home page, we can see the timeline data. This includes the user's historical answers, historical questions, and total number of articles. In the center of the home page, we can see the timeline data. Relevant information includes the influencer's historical answers, historical questions, as well as the total number of articles. We can also find below in detail what questions the influencer did answer, what articles she/he did create, what answers/articles she/he did vote for, and when did he do these activities etc.

3 Theoretical model

Following the previous work of Dasgupta and Prat (2008), Dasgupta and Prat (2008), Guerrieri and Kondor (2012), we develop a simple model with reputation concern to further motivate our empirical analysis. Considering a continuum of influencers who write answers and care about their reputation, i.e. they not only care about votes

obtained in the current period but those in future as well. The model has two periods, $t = 0, 1, 2$. There is a continuum groups of risk neutral platform users, with a total measure of n, and n is sufficiently large. Every user is endowed with a voting power per period, and they could choose to vote for the answer that they views. At the beginning of $t = 1$, users choose whether to follow the influencer that the platform pushes. We use π_1 to denote the user's gross utility at $t = 1$. That is, if a user follows an influencer, the expected utility that the user gets from viewing the influencer's answer is π_1; otherwise if the user does not follow, $\pi_1 = q_r$, where q_r is the expected quality of content when the user randomly browse. In period 2, the user will decide again whether to follow the same influencer for another period or exit the following relationship. We use π_2 to denote the user's gross utility at $t = 2$. A users' choice can be summarized by D_t for $t = \{1, 2\}$. If the user decides to follow an influencer in period t, $D_t = 1$, and 0 otherwise. Therefore, the objective function of a user is to choose D_t to maximize his expected utility $max_{D_1, D_2} \pi_1 + \pi_2$.

An influencer may be either a good type g or a bad type b and the type information is observable only to himself. An influencer can choose from two answer strategies: either a broad-coverage strategy or a focus strategy. Assuming that the platform always pushes a random sample from a variety of potential questions to an influencer in each period, the influencer will then choose to ignore some of them and respond to others. If the influencer adopts a focus strategy, she will only respond to those questions that he is really good at; and in this case, the quality of the answers could be relatively high while the possibility that the influencer could find very fitted questions in a certain period is lower. Otherwise, if the influencer chooses a broad-cover strategy, she will respond to a broad range of questions pushed by the platform, in which case the quality of the answers is comparatively low. In the following analysis, we denote a broad-coverage strategy as strategy I, and focus strategy as strategy II.

We assume that the quality of answers are exogenous and it is equal to the number of votes obtained from the user groups. The influencer cares about peer recognition, thus she would like to maximize the sum of votes in both periods. If a type-θ influencer adopts strategy$i$, the expected quality/votes has a binary distribution:

$$q_i = \begin{cases} q_i^+ & p = p_i^\theta, \\ q_i^- & p = 1 - p_i^\theta, \end{cases} \quad \theta \in \{g, b\}, i \in \{I, II\}, q_i^+ > q_r > q_i^-.$$

To motivate our empirical analysis, $q_i$could be proxied by the average votes from the matching group of users during a certain length of time intervals. Naturally, a good influencer has higher possibility to produce higher quality outcomes. Therefore $p_i^g > p_i^b$, $q_I^+ < q_{II}^+$ and $q_I^- < q_{II}^-$. Also, we assume $p_I^\theta > p_{II}^\theta, \theta \in \{g, b\}$. This condition make the broad-coverage strategy to be a "safer" strategy, by responding to a variety of questions, the influencer has a larger possibility to obtain a moderate number of votes. On the contrary, the focus strategy is more likely to produce extremely highly voted answers, while the possibility of such an event taking place is comparatively low.

An influencer's reputation is defined as the likelihood that the influencer is of type g. When an influencer enters the matching market, her type is drawn independently from a distribution with probability ρ_0 of being type g. Therefore, the initial reputation is ρ_0. After an influencer's performance $q_i$is realized at $t = 1$, her reputation is updated following the Bayes' rule: $\rho = \frac{Pr(q_i|g) \times \rho_0}{Pr(q_i|g) \times \rho_0 + Pr(q_i|b) \times (1-\rho_0)}$.

In the beginning of $t = 2$, a bunch of new influencers arrive at the market, drawn from the same distribution as the $t = 1$. So, the pool of available influencers includes both those with track records from 1 and the new arrivals. The matching market for influencers and groups of users are organized as follows. At time t, for $t \in \{1, 2\}$, the influencer with the highest reputation first randomly matches with a group of users. If

the matched group decides to follow the influencer, the pair leaves the market and the next round of matching starts for the influencer with the second highest reputation. If the group of users decides not to follow the influencer, they will initiate a random browsing with a utility q_r and the influencer is randomly matched with another group of users. This process is iterated until all groups of users have made their decisions, or all influencers get a follower. If multiple influencers have the same reputation level, seasoned influencers (who arrived at $t = 0$) match first. If multiple influencers have the same reputation level and starting time, they match in random order. For the case where a group of users are indifferent between follow a particular influencer or random browsing, we assume that they will follow the influencer. The timing is summarized in Figure 2:

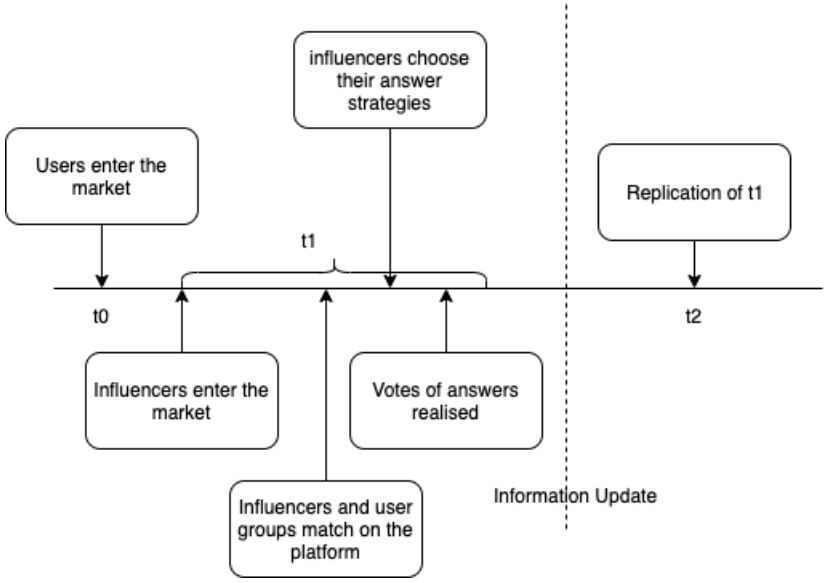

Figure 2: Timing of the model

To rule out the trivial case when all the influencers are followed or not followed, we assume that if all the information is public, the users are only willing to follow type-$g$influencers, that is, the average quality $\bar{q}_i^{\theta} = p_i^{\theta} q_i^{+} + (1 - p_i^{\theta}) q_i^{-}$ satisfies $\bar{q}_i^{g} > q_r > \bar{q}_i^{b}$. A user will choose to follow an average influencer with reputation ρ_0, as long as

$\rho_0 \bar{q}_i^g + (1-\rho_0)\bar{q}_i^b > q_r$. Correspondingly, if an influencer is followed in period 1, her objective is to choose strategies to maximize $E[q_i + D_1 q_j|\theta]$,where $i, j \in \{1, 2\}$ and $\theta \in \{g, b\}$. Alternatively, if an influencer is followed in period 2, the objective function is simply $E[q_i|\theta]$.

Obviously, in the absence of reputation concerns, an influencer always takes the strategy with a higher expected quality, i.e. the strategy that generates more votes. If the highest expected quality strategy is different for good and bad influencers, influencers' types will be immediately revealed. In the rest of this section, we restrict our attention to the reputation concern case.

The reputation concern rises from the fact that, if a rookie influencer fails to produce a q_i^+ votes in period 1, her reputation will be lower than ρ_0. As a consequence, all group of users will be reluctant to follow her in period 2 because there are plenty of new rookies with an average reputation ρ_0 entering this game. Therefore, the type-b influencers have the incentive to mimic the behaviour pattern of type-g influencers. For simplicity, we assume out the first-order difference between strategies for type-b influencers, i.e. the board-coverage and focus strategies generate the same expected votes for type-b influencers, $\bar{q}_I^b = \bar{q}_{II}^b$.

Given the above model settings, The strategy of users is simple: since the rookie level ρ_0 is preferred than randomly browsing, any influencer whose reputation is equal or higher than ρ_0 will be followed by the user; moreover, since in the beginning of period 2, there is plenty of rookies in the market, so any influencer with a reputation lower than ρ_0 will be not be followed by any group of users, since users can always match a rookie influencer with reputation ρ_0. The decision variable thus is:

$$D_t = \begin{cases} 0 & \rho < \rho_0, \\ 1 & \rho \geq \rho_0. \end{cases}$$

Given the equilibrium behaviour of users, we can immediately infer that type-g influencers will definitely prefer the broad-coverage strategy if $\bar{q}_I^g = \bar{q}_{II}^g$. The reason is quite straightforward: If the expected votes are the same, broad-coverage strategy have higher possibility to produce a positive outcome, because $p_I^g > p_{II}^g$. A positive outcome will lead another group of followers in period 2. Therefore, there exists a reputation premium R, so that type-$g$influencers will choose focus strategy if and only if the difference between votes gains from the focus strategy and those from the broad-coverage strategy is equal or higher than R. Obviously, type-b influencers will mimic the strategy of type-$g$influencer at no cost in period 1 and truthfully reveal their types in period 2, when the reputation concern no more exists. The equilibrium strategy when all the influencers' reputation is ρ_0 thus is:

$$S_1^\theta(\rho) = \begin{cases} I & \bar{q}_{II}^g - \bar{q}_I^g < R, \\ II & \bar{q}_{II}^g - \bar{q}_I^g \geq R, \end{cases} S_2^\theta(\rho) = \begin{cases} I & \bar{q}_I^\theta > \bar{q}_{II}^\theta, \\ II & \bar{q}_I^\theta \leq \bar{q}_{II}^\theta. \end{cases} \theta \in \{g, b\}.$$

And R has to be higher enough to cover lost of votes caused by the difference of probability of being unfollowed in period 2. Therefore, $R = (p_1^g - p_2^g) max\{\bar{q}_I^\theta, \bar{q}_{II}^\theta\}$.

Now, considering that a small proportion of influencers in period 0 has a reputation $\rho_h > \rho_0$. Obviously, if ρ_h is high enough to accommodate any possible reputation loss in period 1 (i.e. after any possible outcome in period 1, the reputation is constantly higher than ρ_0), reputation concern will not lead to any behavioural distortion. Therefore, to make our model interesting, we assume that ρ_h is not that high[4]. Intuitively, when influencers' initial reputation is high enough, a small reputation loss or gain makes no substantial difference on the attractiveness of influencers, unless she produces a really low-quality article q_I^-. Therefore, the reputation premium of strategy I drastically changes from positive to negative for these high-reputation influencers:

[4]It implies that the maximal $\rho_h = \frac{(1-p_1^b)\rho_0}{(1-p_1^b)\rho_0 + (1-p_1^g)(1-\rho_0)}$.

$R = (p_1^g - 1)max\{\bar{q}_I^\theta, \bar{q}_{II}^\theta\} < 0$. i.e. Unless adopting a broad-cover strategy leads to sufficiently more votes, high-reputation influencers prefer adopting the focus strategy and only responding to fewer questions, rather than adopting the broad-coverage strategy and responding to a variety of questions.

This result is striking and counter-intuitive, since most of the high-reputation influencers have impressive tracking records of the past outcomes, and no obvious reason stop them continuing being productive. However, our theoretical model predicts that, despite of the incentives from peer recognition, these influencers are likely to produce fewer answers compared with those with an average reputation.

4 Data description

We collect the raw data provided by Zhihu.com, which contains information of all influencers (users with more than 10,000 followers) from January 2016 to August 2017. At that time, the platform had about 17 million registered users. The initial data contains 3686 users where some of them do not have complete information, have been kicked out (because of violating platform's regulations), or are no longer active. In addition to these users' daily activity information, we also captured their follower changes twice in early 2017 and around August 2017, and obtained some additional variables as the total number of received votes during these six months, as well as relevant information of some other variables. Figure 3 illustrate the structure of our dataset.

Therefore, we select the sample data based on the following criteria:

- Available information of badge received or not; (3437 remaining users)
- Available information on users' activities from January 2016 to August 2017; (3003 remaining users)
- Available information on the number of followers on March and August 2017; (1888 remaining users)

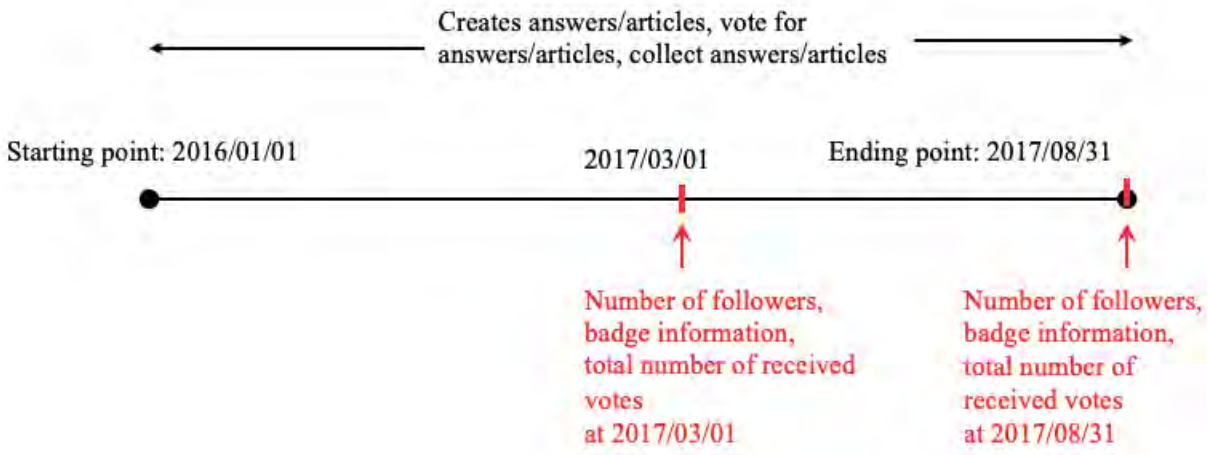

Figure 3: An illustration of the Data structure

- Total number of followers (78,404,520) of all selected users represent more than 80% of the raw sample; (1500 remaining users)
- Frequency of answer creation $\geqslant 1$ per month; (1500 remaining users)

After matching variable information and removing users who had been branded by the platform, we are left with 1888 influencers in our sample. We further select with the number of followers and frequency of activities and finally get a sample with 1500 header influencers: they answered questions at least once a month on average, and we could see their full profile on the platform, as well as the changing status of their followers over the last six months.

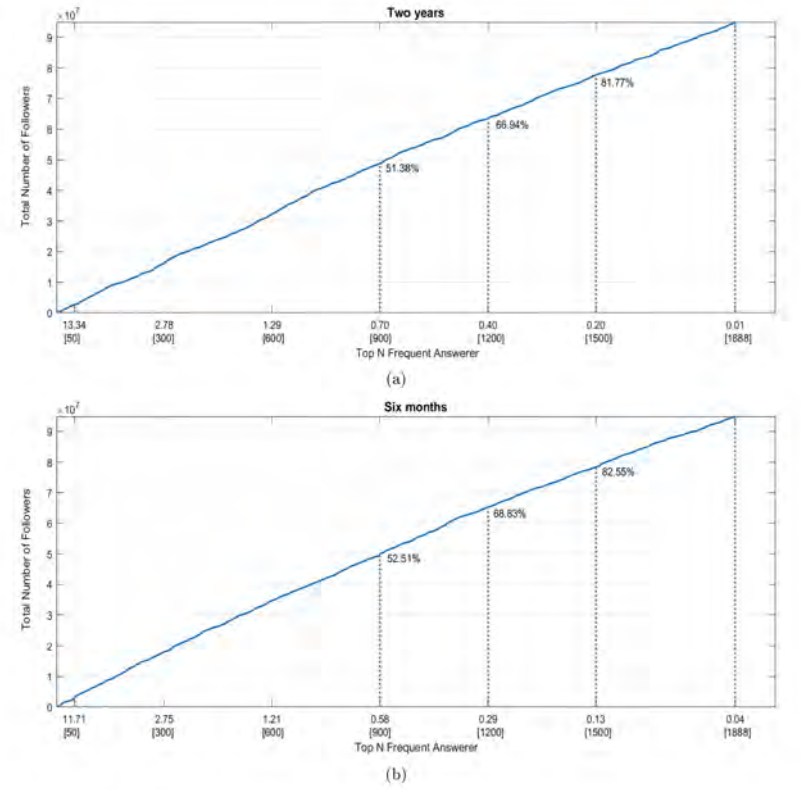

Figure 4: Top N frequent influencers (brackets show the number of remaining influencers) v.s. Accumulated total number of followers

Figure 4a) shows the cumulative distribution function of followers, and we sort influencers by their weekly frequency of answering questions. The influencer who is ranking at 1500 creates around one answer per month, and these 1500 most active people probably account for more than 83 percent of followers. We reported both results from the sample based on six months (where we can see more variables) and the sample based on two years. We find a high degree of overlap in the results, and the groups of selected influencers based on our selection criteria did not change over time. The only difference was that the overall frequency of answering questions in the last six months of the sample declined compared to the overall sample. We will consider and analyze this factor in the subsequent regression model.[5]

[5]In Appendix B, we provide a more detailed discussion about the representativeness of the selected sample and compare the statistics based on the 6-month sample with the full sample.

Type	Type.1	Type.2	Type.3	Type.4	Type.5
Self-authenticated (Blue Star)	√		√		
Best-answerer (Yellow Star)		√	√		
Commercial				√	
Number of users	49	301	61	55	1034
Proportion	3.267%	20.067%	4.067%	3.667%	68.933%
Increasing rate of followers	0.377	0.217	0.320	0.488	0.239
(min, max)	(0.003, 2.223)	(-0.001, 1.447)	(0.006, 1.805)	(0.019, 2.268)	(-0.009, 2.057)
Avg. # of answers created	2.591	1.691	1.967	2.370	2.525
(min, max)	(0.125, 23.333)	(0.125, 31.167)	(0.125, 7.958)	(0.125, 11.708)	(0.125, 71.375)
Avg. # of votes received	6.102	3.348	5.181	3.523	3.386
(min, max)	(0.042, 62.417)	(0.042, 50.917)	(0.083, 27.875)	(0.042, 45.958)	(0.042, 68.917)
Avg. # of articles created	0.829	0.290	0.536	2.292	0.425
(min, max)	(0, 8.291)	(0, 6.125)	(0, 14.125)	(0, 9.541)	(0, 17.292)
Avg. # of answers collected	0.346	0.696	0.622	0	0.758
(min, max)	(0, 6.375)	(0, 19.292)	(0, 8.875)	(0, 0)	(0, 93.875)
Avg. # of articles collected	0.059	0.183	0.216	0	0.152
(min, max)	(0, 0.833)	(0, 12.667)	(0, 3.125)	(0, 0)	(0, 7.542)

Note: 1. Commercial users in Type. 4 can be either self-reported or best-answerer or both.

2. Increasing rate of followers is measured in last six months data. It is calculated approximately by $\Delta log(follower) = log(follower_{t+1}) - log(follower_t)$;

3. Avg. # of answers created is measured by week, Avg. # of votes received only contains the votes from other influencers.

Table 1: Descriptive Statistics (based on six-month selected sample)

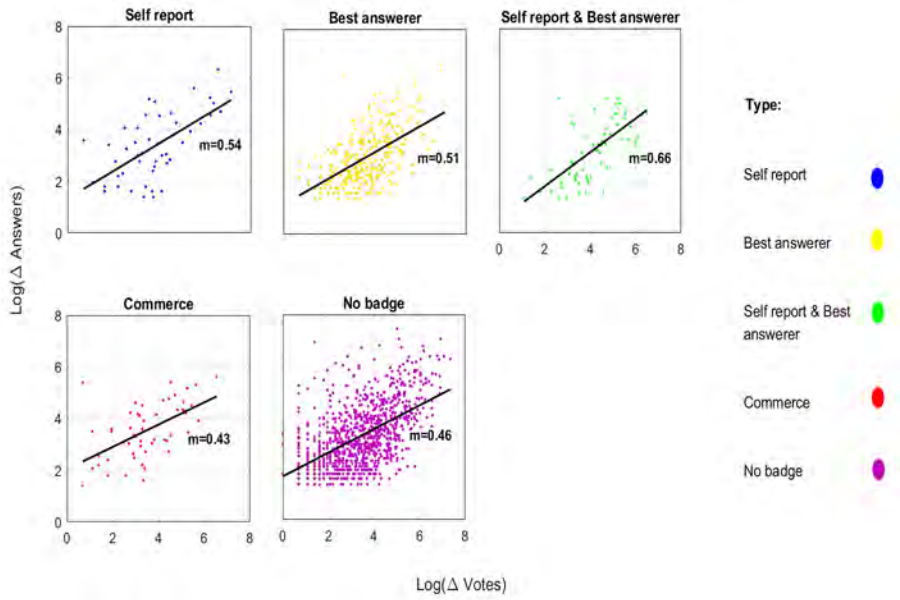

Figure 5: Log(ΔReceived Votes) versus Log(ΔAnswers) (the slop is calculated by m)

An important feature of the data is that the platform assigns different types to the users. In the data, the platform provides three kinds of tags to distinguish an ordinary user from "a special user": self-authenticated professionals (blue star users), best answerers (yellow star users), business users (merchants who promote their products by answering questions), and uncertified users.[6] According to the badge information provided by the platform, we divide the users into five types and report the relevant description statistics of each type of user in Table 1. We find that platform authenticated users and business users account for a smaller share of

[6]Blue star users are experts in their field, for example, lawyers, engineers, and accountants. Users need to provide licenses to the platform in order to authenticate themselves and get the blue star. The minimum requirement for users in academia is to be at least a Ph.D. student in progress; a student ID is accepted for the real-name system, while professors need to provide certification of employment. The platform only requires the professional level and employment status online should be truthful, but authenticated users could show either their real names or net names on the site.

Yellow star users are labelled as "best answerer" by the platform. They are set up by the platform in order to enable readers to find valuable answers more quickly and accurately and to motivate content providers to output professional answers continuously. The platform automatically identifies the best respondents in each field according to the algorithm, and the algorithm calculates the topic weight of the reference user in a specific field. Excellent answerers can only be provided by the system and do not support applications or self-recommendation.

influencers, but have a higher follower growth rate than other types of users and tend to garner more votes. On average, yellow star influencers answer fewer questions than other types of users.

The results of descriptive statistics imply that different types of users may receive different levels of incentives, which in turn leads to differences in the enthusiasm of providing content to the platform. We show this empirical evidence again in Figure 5, where the horizontal axis is the number of votes each user receives from other users, and the vertical axis is the total number of creations made by a user. We find that different types of users are motivated by different degrees of "peer recognition" (the slope of each graph). In order to understand the impact of "peer recognition," we look for using more sophisticated regression models to identify.

5 Identifying the impact of "peer recognition"

5.1 Basic regression model

Our first objective is to identify the impact of being recognized as an incentive to share content. We use the number of votes an influencer receives from other influential users as a proxy of "the incentive of recognition." Since one user's answer is endorsed by another means that the answer appears on the vote-up person's timeline and will be seen by all of his followers, voting from other influencers helps the content provider spread the content and increases her/his influence on the voter's network,[7] which is also discussed regarding the impact of users' choices in online communities on their followers. We constructed the output variable from the number of questions each influencer answered per week and used the number of votes received by other

[7] Sun, Zhang, and Zhu (2019) also discussed the impact of users' choices in online communities on their followers. More recently, Bailey et al. (2019) used data from facebook to explore the impact of peer effects on phone purchase decisions in the U.S. market.

influencers in the previous week as the key explanatory variable to construct the regression model:

$$\log\left(Anwers_{i,t}+1\right)=\beta_0+\log\left(Votes_{i,t-1}+1\right)\beta_v+x'_{i,t}\beta_x+\delta_i+\delta_t+\eta_{i,t}, \quad (1)$$

where $x_{i,t}$ is a vector of control variables; δ_i and δ_t are two fixed components captures individual and time fixed effects on the answer creation and $\eta_{i,t}$ is an unobserved error term, where δ_i captures the effect of personal specific unobserved characteristics on the number of questions answered, and δ_t captures the presence of underlying specific time effects (for example, during national holidays, because most people choose to travel, the frequency of answering questions may decrease. The emergence of hot social topics in a certain period will also increase the frequency of answering the overall question). β_v captures the effect of "peer recognition". If there is no concern about endogeneity β_v is simply the statistical linear correlation between the number of created new answers and received votes that an influencer i receives during the week t. We use the lagged variable of received votes at t to avoid the potential endogeneity problem: the higher number of received votes during the period t may be due to the higher number of questions answered.

Table 2 reports Ordinary Least Square (OLS) regression results using through using weekly panel data from March 2017 to August 2017.[8] The direct estimation results form the first four columns show that an additional 10% votes received in the given week correspond to an increase in around 1% more created answer in the following week even after controlling the fixed effects. Receiving votes from other influencers (the act of approving the answer) has a statistically significant positive effect on the content creation. In the last five columns, we gradually add time trends, follower base and badge information as control variables. We note that since the

[8]The regression model estimation is based on the six months selected sample.

	Dependent variable: $\log(Anwers_t + 1)$								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	OLS	OLS	OLS	OLS	OLS	OLS	OLS	OLS	OLS
$\log(Votes_{t-1} + 1)$	0.109***	0.113***	0.083***	0.088***	0.113***	0.115***	0.113***	0.113***	0.113***
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)
$Week$					-0.130***	-0.112***	-0.130***	-0.127***	-0.128***
					(0.019)	(0.024)	(0.020)	(0.020)	(0.020)
$Week^2$					0.007***	0.006***	0.007***	0.007***	0.007***
					(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
$\log(Follower_0)$						-0.019			
						(0.016)			
$Self\ Authenticated$							0.058		0.088**
							(0.044)		(0.044)
$Best\ Answerer$								-0.113***	-0.124***
								(0.030)	(0.031)
Individual Fixed Effects			√	√					
Time Fixed Effects		√		√	√	√	√	√	√
N	34500	34500	34500	34500	34500	34500	34500	34500	34500
R^2	0.012	0.036	0.012	0.037	0.036	0.036	0.036	0.036	0.036

Note: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$, standard errors in parentheses.

$\log(Follower_0)$ represents the log number of follower on March 2017.

Table 2: OLS regression results of the impact of "peer recognition"

number of initial followers is time invariant, we cannot control both individual fixed effects and the number of followers simultaneously.

After adding control variables, we find our regression result is in line with the intuition. The amount of influencers' content contribution gradually decreases as time goes by. We suspect this is due to the drying of the knowledge (every user's knowledge is limited, so it is very challenging to provide content continuously), or the reason multihoming (some influencers may provide content on several platforms at once after they become famous, thus reducing their content contributions on the original platform). We also find that the authenticated influencers provide answers more frequently. This means that, on average, these influencers authenticated by the platform's real name submit more content because the answers they provide are more reliable and tend to rank higher among similar answers, which is consistent with Tang, Gu, and Whinston (2012). In contrast, the number of content contributions by excellent respondents is generally smaller than that of other influencers. We suspect there are two possible reasons: reputation concern (they tend to spend more time improving the quality of their content than answering more questions) or multihoming.

Another concern is that we focus on the influence of votes from other influencers on the content creation. However, "peer recognition" may not only be reflected in the votes of other influencers, but in all the votes received from all users which may dominate the effect. In Appendix C we check and compare the impact from both "all votes" and "votes from influencers", the results indicate that the votes from other influencers are good enough to capture the "peer recognition" effect.

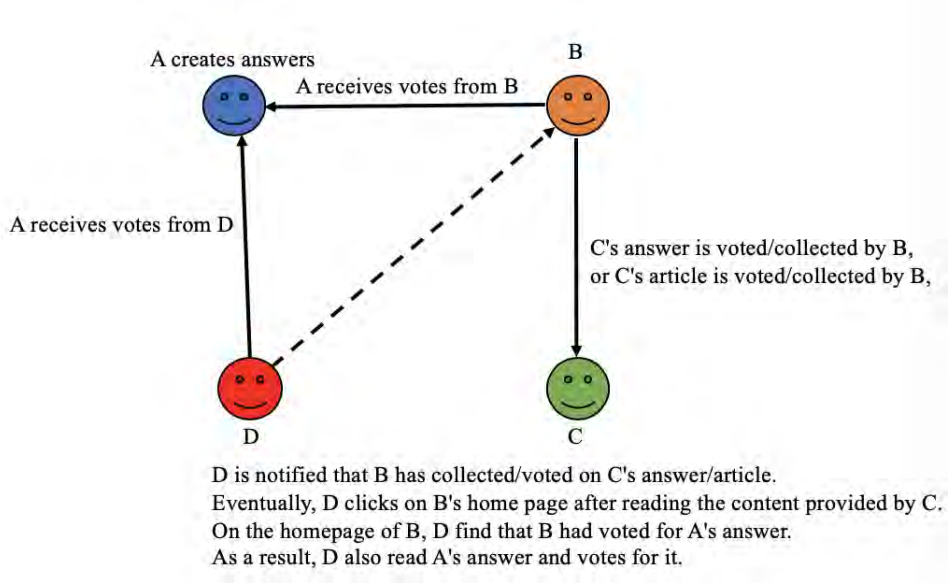

Figure 6: Illustration of instrumental variables

5.2 From the correlation to the causality

Although we use lag variables in the above formula to avoid possible endogeneity problems, the regression results are still limited to showing the statistical correlation between "the number of received votes" and "the number of created answers," rather than its incentive effect on "answer creation." The endogeneity problem resists because of the omission of important variables. For example, we do not see the cost of providing answers, and it may take more time to produce good answers. At the same time, a topic in a particular field may be unusually hot at some point in time, so that some content providers in this field provide more answers than others. To address these underlying considerations, we need better methods to identify the causal relationship between "the number of received votes" and "the number of creations."

In this paper, we provide an instrumental variable method to tackle the endogeneity issue. The construction of the instrumental variable is based on user interactions on social networks to identify such causal relationships. Figure 6 illustrates the construction and validation of instrumental variables. We assume there are four influ-

encers A, B, C, and D. A creates answers. B read and voted for one of A's answers. Meanwhile, B also read one answer/article from C and chose to vote for/collect it. B's behaviour makes the contents of A and C appear on B's home page at the same time. D is pushed by the platform when B votes for or collects C's content. D comes across A's answer after reading C's contents on B's homepage. So, D also reads A's answer and sends out the vote. In general, our instrumental variables assume that in a social network, votes from one user to other influencers have a positive spillover effect on content sharing (see, Fershtman and Gandal (2011)).

Since A's content creation quantity is only motivated by the number of votes it obtains, B's behaviours toward C can be used as valid instruments. On the one hand, they do not directly affect the creation enthusiasm of A (exclusion restriction); on the other hand, B's behaviour toward C will bring additional votes to A's content through D (reveal condition). In data, one influencer may have four actions on another: answer voting, answer collection, article voting, and article collection. Each of these four actions may cause additional votes by other influencers on the answer for a given influencer, we use them as potential instrumental variables to identify the effect of "peer recognition." Our identification strategy comes from the unique nature of the online content platform, where users' communication with each other allows us to correctly identify the impact of a variable on users' motivations to contribute in ways other than policy shock (e.g., Zhang and Zhu (2011), Wu and Zhu (2019)) for the first time. We consider a system of equations below:

$$\log\left(Anwers_{i,t}+1\right)=\beta_0+\log\left(Votes_{i,t-1}+1\right)\beta_v+x'_{i,t}\beta_x+\delta_i+\delta_t+\eta_{i,t}, \tag{2}$$

$$\log\left(Votes_{i,t-1}+1\right)=\gamma_0+z_{i,t}\gamma_z+x'_{i,t}\gamma_x+\delta_i+\delta_t+v_{i,t}. \tag{3}$$

where $z_{i,t}$ is a set of potential instrumental variables that we have discussed above.

Table 3 reports the first-stage results of IV estimation by controlling potential

	Dependent var: $\log(Votes_{t-1}+1)$		
Instrumental var	(1)	(2)	(3)
$\log\left(Article\ votes_{t-1}+1\right)$	0.049***	0.168***	
	(0.016)	(0.019)	
$\log(Article\ collections_{t-1}+1)$	-0.115***	0.605***	
	(0.016)	(0.016)	
$\log(Answer\ votes_{t-1})$	0.311***		0.310***
	(0.003)		(0.003)
$\log(Answer\ collections_{t-1})$	0.036***		0.001
	(0.008)		(0.007)
Individual Fixed Effects	√	√	√
Time Fixed Effects	√	√	√
N	34500	34500	34500
F stat	5070.588	795.811	10057.640
Hansen J statistic	17.889	2.124	13.059
(p-value)	(0.001)	(0.145)	(0.000)

Note: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$, standard errors in parentheses. The explanatory variables are the total number of votes/collections for other users by people who voted for the influencer i in period t.

Table 3: First-stage results and IV diagnostics

	Dependent variable: $\log(Anwers_t + 1)$					
	(1)	(2)	(3)	(4)	(5)	(6)
	OLS	OLS	OLS	IV	IV	IV
$\log(Votes_{t-1} + 1)$	0.109***	0.113***	0.088***	0.164***	0.161***	0.128***
	(0.004)	(0.004)	(0.004)	(0.013)	(0.013)	(0.015)
Intercept	0.624***	0.800***	0.817***	0.582***	0.766***	0.789***
	(0.014)	(0.019)	(0.014)	(0.017)	(0.022)	(0.017)
Individual Fixed Effects			√			√
Time Fixed Effects		√	√		√	√
N	34500	34500	34500	34500	34500	34500
R^2	0.012	0.036	0.037	0.012	0.034	0.034

Note: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$, standard errors in parentheses.
Instrumental variables are $\log(Article\ collections_{t-1} + 1)$ and $\log(Article\ votes_{t-1} + 1)$.

Table 4: Instrumental Variable regression results

instrumental variables and fixed components. The four potential instrumental variables are the total number of votes/collections for other influencers by influencers who voted for the influencer i in period t (i.e., the sum of the actions of B to C in Figure 6). The results show that none of the instrumental variables are weak instruments (they all satisfy the reveal condition), and confirm our previous hypothesis that voting for someone else had a positive spillover effect. However, in verifying the exclusion restriction, Hansen J statistic shows that only the voting and collecting for other influencer's articles are the valid variables. This finding suggests that voting up or collecting other influencers' answer may be related to some other unobserved factors. For example, in Figure 6, B votes for both C and A at the same time, probably because C and A are under the same question, which is A hot topic at that time. During the same period, A may have answered many questions related to hot topics. Therefore, we end up choosing only the collection and voting of articles as the instrumental variables.

Table 4 shows the results based on the IV estimates. We find that OLS estimates seriously underestimated the actual effect of "peer recognition" on encouraging con-

tent creation. The true estimated effect based on the instrumental variables is 30-40% higher than the OLS estimates. This means that there are some underlying factors that prevent the "peer recognition" effect from playing out, which ultimately leads OLS to underestimate the effect. This means that there are some underlying factors that prevent the "peer recognition" effect from playing out, which ultimately leads OLS to underestimate the effect. For example, influencers may reduce their contributions for various reasons, such as reputation concern, privacy concern, or drying up of knowledge etc. Understanding which channel, in reality, explains the heterogeneity of the incentive effect of "peer recognition" on content creation on free platforms is crucial to policymaking.

5.3 Spillover effect

Using our IV method, we are natural to study the impact of knowledge spillover in our data. The number of answers is motivated not only by the "peer recognition" of the other influencers' votes but also by the potential impact of the knowledge spillover: the potential number of users who can read the answer because of the other influencers' votes. In Appendix D, we revise all the regression results by changing the variable $\log(Votes_{i,t-1}+1)$ by $\log(Readers_{i,t-1}+1)$. All the other variables remain unchanged, we consider the following regressions:

$$\log(Anwers_{i,t}+1) = \beta_0 + \log(Readers_{i,t-1}+1)\,\beta_r + x'_{i,t}\beta_x + \delta_i + \delta_t + \eta_{i,t}, \quad (4)$$

$$with$$

$$Readers_{i,t-1} = \sum_{influencer\ j\ votes\ i\ during\ t-1} Votes_{j,t-1} \times Follower_{j,0}, \quad (5)$$

where the variable $Readers$ can be also interpreted by the potential traffics brought by the influencers who vote for the answers. The regression results in Table 11

shows that, although by construction, potential readers show a smaller influence on content creation than "received votes from influencers' votes." A 1% increase in the number of the potential number of readers raises the number of answers by 0.1%. The instrumental variables also pass the tests perfectly and prove that OLS greatly underestimated the actual effect of knowledge spillover in the new regression model by almost ten times the OLS coefficient.

6 What makes influencers contribute less?

6.1 Differences between commercial and non commercial users

In this section, we mainly explore what affects the incentive of "peer recognition" to the production level of content providers. Different from e-commerce platforms, the badges (e.g., excellent merchants) provided by the platforms eliminate not only information asymmetry, but also stimulate the sales of merchants (e.g. Hui et al. (2016)). On the content based platform, there are two types of influencers: commercial and non-commercial (personal) users. Excellent personal content providers are also responsible for producing products (answers to questions), which may make them less motivated some potential concerns (e.g., reputation or privacy concerns). For commercial users, their purpose on the platform is mainly to promote their products by answering questions, essentially to achieve the purpose of advertising by providing free content. So the incentives for them are not affected by reputation or privacy concerns. We first have check whether commercial and non-commercial influencers behave differently. To do this, we first divided users into two categories based on badge information: commercial influencers and non-commercial influencers.

Table 5 reports the regression results by using *Commertial*, a binary variable indicating whether an influencer is a commercial user. The regression results are in

line with our expectations, and the reduction in the incentive effect OLS estimated was due entirely to "non-commercial users" (i.e., individual users). It shows that after taking into account the cross term between the commercial user and the "peer recognition," the incentive effect received by the business user under OLS estimation is almost the same as the result of using instrumental variables. In particular, we find the same results when we look at the spillover effect on content providing. If we exclude individual influencers and focus on commercial influencers separately, the OLS results are the same as the IV results. We report the corresponding results in Appendix D. Therefore, we will focus on non-commercial influencers in our subsequent studies.[9]

6.2 Reputation and privacy concerns

The above results show that only some of the non-commercial influencers would be "negatively" affected by the votes received from other influencers. We explore here which factors cause them to be less motivated. Based on the information provided in the data, we mainly check two potential factors: reputation or privacy concerns. In our case, reputation concern mainly refers to whether influencers will be less motivated because they are holding the title of high-quality users (i.e., best answerers) on the platform. In the sense of social responsibility, they may seek to provide more rigorous and high-quality answers rather than more answers. Privacy concerns is related to that some influencers, who already provide particularly detailed personal data to the platform, are less motivated because some of them are worried about their speech

[9]In Appendix F the impact of getting a badge without specify the type of badge: it can be either a badge indicating if an influencer has self-authenticated, or a badge showing if an influencer is a best answer of her/his field, or a badge indicates if an influencer is simply a commercial user. Results in Table 15 show that the fact of holding a badge negatively affect on frequency of providing new answers. We know that commercial users are more likely to be motivated than other influencers, which means that the other two types of influencers are far more conservative in their responses to other people's votes.

	Dependent variable: $\log(Anwers_t + 1)$		
	(1)	(2)	(3)
	OLS	OLS	IV
$\log(Votes_{t-1} + 1)$	0.115***	0.114***	0.164***
	(0.004)	(0.004)	(0.013)
$Week$	-0.112***	-0.112***	-0.200***
	(0.024)	(0.024)	(0.023)
$Week^2$	0.006***	0.006***	0.008***
	(0.001)	(0.001)	(0.001)
$\log(Follower_0)$	-0.019	-0.021	-0.035**
	(0.016)	(0.016)	(0.017)
$Commercial$		0.105	
		(0.070)	
$Commercial \times \log(Votes_{t-1} + 1)$		0.052**	
		(0.022)	
Time Fixed Effects	√	√	√
N	34500	34500	34500
R^2	0.036	0.036	0.033

Note: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Standard errors in parentheses.

Commercial is a binary dummy variable.

Table 5: Impact on creating new answers - commercial and non commercial influencers

Dependent variable: $\log(Anwers_t + 1)$				
	(1)	(2)	(3)	(4)
	OLS	OLS	OLS	IV
$\log(Votes_{t-1} + 1)$	0.113***	0.114***	0.124***	0.148***
	(0.004)	(0.004)	(0.004)	(0.013)
$Week$	-0.112***	-0.111***	-0.117***	-0.207***
	(0.024)	(0.024)	(0.024)	(0.023)
$Week^2$	0.006***	0.006***	0.007***	0.009***
	(0.001)	(0.001)	(0.001)	(0.001)
$\log(Follower_0)$	-0.025	-0.026	-0.017	-0.036**
	(0.016)	(0.016)	(0.016)	(0.017)
$Self\ Authenticated$		0.042		
		(0.051)		
$Self\ Authenticated \times \log(Votes_{t-1} + 1)$		-0.002		
		(0.014)		
$Best\ answerer$			-0.086***	
			(0.031)	
$Best\ answerer \times \log(Votes_{t-1} + 1)$			-0.039***	
			(0.009)	
Time Fixed Effects	√	√	√	√
N	33235	33235	33235	33235
R^2	0.035	0.035	0.035	0.033

Note: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Standard errors in parentheses.

Table 6: Effects of reputation and/or privacy concerns (non-commercial users)

being monitored by the platform. In reality, both potential concerns may or may not occur.

We divided all non-commercial influencers into two categories: users who get "best answerer" and users who get "self authenticated." Table 6 reports the regression results. The regression results show that, on average, users who are motivated by other influencers are not affected by submitting personal information to the platform. The main factor that affects motivation depends on whether the user is a "best answerer" awarded by the platform. These content providers may trade-off between quantity and quality of the answers due to the potential "reputation concern" and choose their

content more carefully. [10]

Does the disclosure of information really not affect content incentives? Based on the results of Table 6, we have the following two considerations: first of all, some users have both badges (i.e., they are self authenticated best answerers). Secondly, by disclosing their information on the website and obtaining the website authentication, there are certain restrictions: for example, the user is at least a doctoral student, and the disclosure is limited to work and education. Both of these considerations may affect our regression results.

We first check if the privacy and reputation concern has some heterogenous effects. We launch regressions for all self-authenticated influencers and all the "best answerers" separately, and evaluate the impact of gaining the "best answer" moniker on those who were supported by self-authenticated influencers. And for those "best answerers," the impact of self-authenticated on them. Results are reported in Table 7. The results confirmed the existence of "reputation concern" again. Whether or not an influencer actively discloses information to the platform and obtains the authentication, being voted "best answerer" will affect the incentive she/he receives from others. However, when we restrict our sample to all the "best answerers," we find that getting a "blue star" (self-authentication), while significantly and negatively affects the total number of questions answered compared with the previous regression results, does not significantly reduce the effect of "peer recognition".

[10] We provide in Appendix E an additional check of the potential multi-homing concern. The best candidates may also receive less incentive from others' votes because other competing platforms may poach them after they are awarded as "best answerers" in their field. Under the assumption that the multi-homing probability is positively correlated with the number of followers, our estimation result supports the existence of multi-homing concern in our sample. However, after controlling the variable "best answerer," we find that for the "best answerers", the increase of followers will have a positive and significant impact on peer incentives. Such finding confirms the existence of the reputation concern and even shows that the reputation concern in our sample dominates the potential multihoming effect.

Dependent variable: $\log(Anwers_t + 1)$	Conditional on self-authenticated influencers			Conditional on best answerers		
	OLS	OLS	IV	OLS	OLS	IV
$\log(Votes_{t-1} + 1)$	0.118***	0.165***	0.105**	0.088***	0.089***	0.128***
	(0.014)	(0.022)	(0.048)	(0.007)	(0.008)	(0.025)
$Week$	-0.201**	-0.200**	-0.261***	-0.186***	-0.186***	-0.224***
	(0.0874)	(0.0873)	(0.0870)	(0.045)	(0.045)	(0.045)
$Week^2$	0.009***	0.009***	0.011***	0.008***	0.008***	0.009***
	(0.003)	(0.003)	(0.003)	(0.001)	(0.001)	(0.001)
$\log(Follower_0)$	0.032	0.028	0.037	0.036	0.034	0.025
	(0.053)	(0.053)	(0.059)	(0.026)	(0.026)	(0.027)
$Self\ Authenticated$					0.120*	
					(0.061)	
$Self\ Authenticated \times \log(Votes_{t-1} + 1)$					-0.006	
					(0.018)	
$Best\ answerer$		0.002			(0.018)	
		(0.091)				
$Best\ answerer \times \log(Votes_{t-1} + 1)$		-0.077***				
		(0.028)				
Time Fixed Effects	√	√	√	√	√	√
N	2530	2530	2530	8326	8326	8326
R^2	0.061	0.063	0.062	0.035	0.035	0.031

Note: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Standard errors in parentheses.

Table 7: Heterogenous effects of reputation and/or privacy concerns (non-commercial users)

6.3 Impact of secondary information disclosure

We also consider influencers who wish to disclose a portion of their information voluntarily but cannot be verified by the platform. We hope to know the impact of "information disclosure" on these influencers. On the platform, influencers are also allowed to disclose their work units, living places and other personal information. These information can be very informal and are voluntarily disclosed by users. Since they cannot be truly verified by the platform, we regard them as "secondary information disclosure."

We categorize these influencers without any badge information into three types based on their level of information disclosure.

1. no badge influencers who have reported both their place of residence, their work address;

2. no badge influencers who have either reported their place of residence or their work address;

3. no badge influencers who do not report any information.

Table 8 report the regression results. For the first type of influencers, who disclosed their residential address and job industry information, the incentive effect is very similar to the IV estimation. Such empirical results explain this phenomenon by a potential "selection effect": Those influencers who are worried about their information will not choose to disclose their information and get the platform authentication. As a result, they will be less motivated when receiving other influencers' votes.

Overall, our results show that reputation concern plays a significant role in non-commercial influencers. The platform badge policies not only encourage commercial influencers to speak more frequently but also limit the incentive for "best answerers"

Dependent variable: $\log(Anwers_t + 1)$					
	(1)	(2)	(3)	(4)	(5)
	All	Type 1	Type 2	Type 3	All-IV
$\log(Votes_{t-1} + 1)$	0.122***	0.144***	0.099***	0.118***	0.152***
	(0.005)	(0.008)	(0.008)	(0.011)	(0.017)
$Week$	-0.084***	-0.068	-0.168***	0.012	-0.199***
	(0.030)	(0.048)	(0.049)	(0.062)	(0.028)
$Week^2$	0.006***	0.005***	0.008***	0.003	0.008***
	(0.001)	(0.001)	(0.001)	(0.002)	(0.001)
$\log(Follower_0)$	-0.049**	-0.060*	0.031	-0.146***	-0.059***
	(0.021)	(0.032)	(0.036)	(0.040)	(0.023)
Time Fixed Effects	√	√	√	√	√
N	23782	9499	8694	5589	23782
R^2	0.036	0.039	0.041	0.037	0.034

Note: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Standard errors in parentheses.

Table 8: Effect of information disclosure on no badge influencers

to contribute content. In the long run, the platform may be flooded with "low-quality" content and lose its charm. In addition, if the "selection effect" makes some users unwilling to disclose their information on the platform and more cautious about providing content, the platform may only be left with users who do not care about information disclosure and lose diversified content. Such findings may partly explain why in reality it is so hard to find objective opinions on a free content platform.[11]

7 Conclusion

Online content provision is undoubtedly a vital topic in the digital era, and there is an increasing number of related empirical studies in recent years. Our paper is the first to identify the impact of "peer recognition" on a platform influencer's content creation.

[11] Just before we finished the paper, Zhihu has launched a more severe "real name" authentication policy: the platform requires every registered user to provide mobile phone and pass the real-name authentication. Users who refuse to contribute will be subject to certain restrictions on what they can say on the platform. Such a policy worries many users, some of whom even refuse to continue exporting content online. We provide an anecdotal evidence in our Appendix G.

Our method of constructing instrumental variables has been proved to be simple and feasible in practice, which solves the endogenous problems in many platform-based empirical studies of digital economics and widens the research boundary.

From an empirical perspective, our results directly provide a quantification of the impact of "peer recognition." It is well known that if influencers are adequately incentivized by platform policies and become more productive, it will bring higher traffic to the platform and higher corresponding advertising revenue. Being able to effectively encourage content provision not only allows content readers to read abundant, higher-quality, and more diverse content but also points the way for policymakers to regulate content platforms: an effective policy should be to eliminate information asymmetry in two-sided markets without compromising the motivation of influencers. In particular, our results also reveal the considerations that platforms should have when formulating policies related to "real-name system": while users are more accountable for each answer provided after obtaining real-name authentication, they may also become less active in providing content for fear of overexposing their privacies. Although we do not find influencers who disclosed to the platform personal information significantly reduce the number of content creation, our empirical results explain this phenomenon by a potential "selection effect": Those influencers who are worried about their information will not choose to disclose their information and get the platform authentication and will be less motivated when receiving other influencers' votes.

The advantage of our data is that we can see very detailed online activity data for each user. However, as China has very strict personal information regulation policies, citizens may care less about personal information disclosure itself than other countries (whether or not users choose to be anonymous on the Internet, the government can easily find out who they are). This may bring some limitations to our research: For example, users may be less sensitive to online disclosure of personal information.

Based on our results, we think there are two possible interesting directions for future research. One is that if we can get more detailed data, such as the daily changes of flowers' number, users' online time, etc., we would be able to establish a dynamic structural model to describe users' dynamic content provision behaviour (e.g., Tang, Gu, and Whinston (2012)). The structural model estimation will enable us to do better quantitative research on the possible platform incentive policies through counterfactual analysis.

Another interesting direction is to delve into the motivations of users who choose to self-authenticate. In our paper, we only studied what kind of users would be motivated by "peer recognition," but we do not know what would happen if they did not choose real-name authentication. Changes in platform/social policies may cause some users to change their choice to submit personal information to the platform, which may affect the quantity/quality of content they provide. These questions require us to communicate with enterprises to obtain more detailed data, and we will try to answer them in future papers.

References

Athey, Susan, Christian Catalini, and Catherine Tucker. 2017. "The digital privacy paradox: Small money, small costs, small talk." Tech. rep., National Bureau of Economic Research.

Bailey, Michael, Drew M Johnston, Theresa Kuchler, Johannes Stroebel, and Arlene Wong. 2019. "Peer Effects in Product Adoption." Working Paper 25843, National Bureau of Economic Research. URL http://www.nber.org/papers/w25843.

Chiou, Lesley and Catherine Tucker. 2017. "Content aggregation by platforms: The

case of the news media." *Journal of Economics & Management Strategy* 26 (4):782–805.

Dasgupta, Amil and Andrea Prat. 2008. "Information aggregation in financial markets with career concerns." *Journal of Economic Theory* 143 (1):83–113.

Fershtman, Chaim and Neil Gandal. 2011. "Direct and indirect knowledge spillovers: the "social network" of open-source projects." *RAND Journal of Economics* 42 (1):70–91.

Goh, Khim-Yong, Kai-Lung Hui, and Ivan PL Png. 2015. "Privacy and marketing externalities: Evidence from do not call." *Management Science* 61 (12):2982–3000.

Goldfarb, Avi and Catherine Tucker. 2012. "Shifts in privacy concerns." *American Economic Review* 102 (3):349–53.

Guerrieri, Veronica and Péter Kondor. 2012. "Fund managers, career concerns, and asset price volatility." *American Economic Review* 102 (5):1986–2017.

Han, Xintong and Lei Xu. 2018. "Technology adoption in input-output networks." *NET Institute Working Paper 18-05* .

Han, Xintong and Pu Zhao. 2019. "Pay for content or pay for referral? An Empirical Study on Content Pricing." *NET Institute Working Paper 19-03* .

Hui, Xiang, Maryam Saeedi, Zeqian Shen, and Neel Sundaresan. 2016. "Reputation and regulations: evidence from ebay." *Management Science* 62 (12):3604–3616.

Hui, Xiang, Maryam Saeedi, and Neel Sundaresan. 2018. "Adverse Selection or Moral Hazard, An Empirical Study." *The Journal of Industrial Economics* 66 (3):610–649.

Kuang, Lini, Ni Huang, Yili Hong, and Zhijun Yan. 2019. "Spillover Effects of Financial Incentives on Non-Incentivized User Engagement: Evidence from an Online Knowledge Exchange Platform." *Journal of Management Information Systems* 36 (1):289–320.

Lerner, Josh and Jean Tirole. 2002. "Some simple economics of open source." *Journal of Industrial Economics* 50 (2):197–234.

Penney, Jonathon. 2016. "Chilling effects: Online surveillance and Wikipedia use." *Berkeley Tech. LJ* 31:117.

———. 2019. "Chilling effects and transatlantic privacy." *European Law Journal* 25 (2):122–139.

Roberts, James W. 2011. "Can warranties substitute for reputations?" *American Economic Journal: Microeconomics* 3 (3):69–85.

Saeedi, Maryam. 2014. "Reputation and adverse selection, theory and evidence from eBay." .

Sun, Monic, Xiaoquan Zhang, and Feng Zhu. 2019. "U-shaped conformity in online social networks." *Marketing Science* .

Sun, Monic and Feng Zhu. 2013. "Ad revenue and content commercialization: Evidence from blogs." *Management Science* 59 (10):2314–2331.

Tang, Huan. 2019. "The Value of Privacy: Evidence from Online Borrowers." *Working Paper* .

Tang, Qian, Bin Gu, and Andrew B Whinston. 2012. "Content contribution for revenue sharing and reputation in social media: A dynamic structural model." *Journal of Management Information Systems* 29 (2):41–76.

Tirole, Jean. 2019. "Digital Dystopia." .

Wu, Yanhui and Feng Zhu. 2019. "Competition, contracts, and creativity: Evidence from novel writing in a platform market." .

Xu, Lei, Tingting Nian, and Luis Cabral. 2019. "What makes geeks tick? a study of stack overflow careers." *Management Science* .

Zhang, Xiaoquan Michael and Feng Zhu. 2011. "Group size and incentives to contribute: A natural experiment at Chinese Wikipedia." *American Economic Review* 101 (4):1601–15.

Zuboff, Shoshana. 2018. *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. 1st ed.

Appendix

A A comparison of Zhihu and Quora's user interface

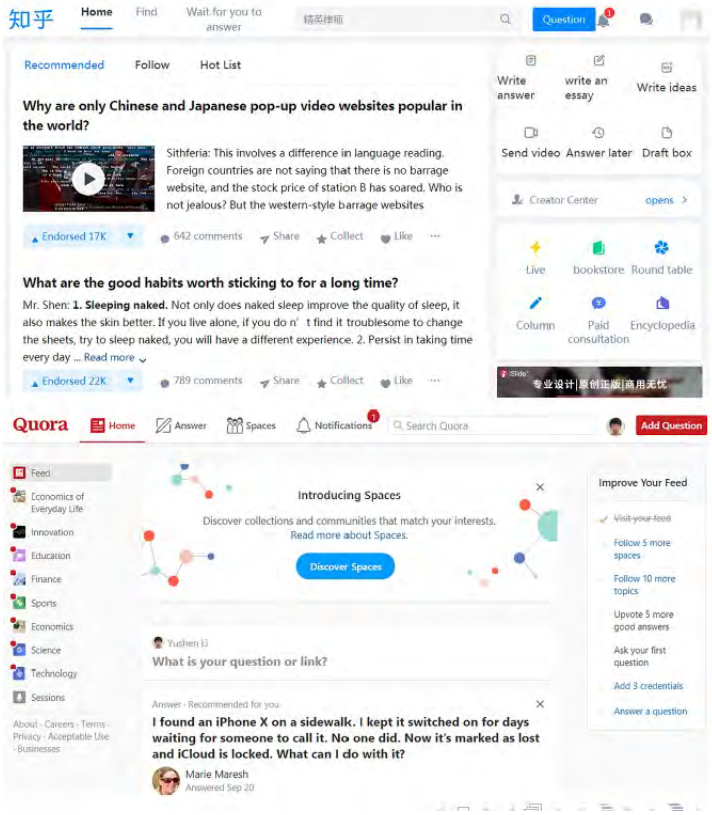

Figure 7: A comparison of Zhihu and Quora's user interface

B Sample comparisons

Table 9 reports a comparison of relevant statistical values between the full sample and 6-month samples. We find that the statistical data of the two samples were highly similar, which means that our samples based on the last six months were very representative. Besides, we also check the coincidence of the influencers selected based on our criteria in the two samples. We find that 66.50% of the 388 influencers excluded in the two samples are matched. We also have concerns about whether users who had been active for two years in the original sample might have been excluded in the last six months because they were less active than in the previous 18 months. So we also double-checked user activity frequencies in the original (two-year based) and selected (six-month based) sample. Figure 8 reports the comparison, we find that there is a high degree of overlap in activity frequencies between the two samples and that users who frequently answered questions in the full sample were also positive respondents in the six-month sample.

	Two-year full sample (81.77% of total followers)	Six-month sample (82.55% of total followers)
Total number of influencers	1888/1500	1888/1500
Increasing rate of followers	0.224	0.251
(min, max)	(-0.014, 2.268)	(-0.001, 2.268)
Avg. number of answers created	1.902	2.331
(min, max)	(0.011, 102.483)	(0.125, 71.375)
Avg. number of votes received	2.985	3.545
(min, max)	(0.013, 64.688)	(0.042, 68.917)

Note: 1. Influencers are the influential users in the platform, each of them has at least 10,000 followers;

2. Increasing rate of followers is measured in last six months data. It is calculated approximately by $\Delta log(follower) = log(follower_{t+1}) - log(follower_t)$;

3. Avg. # of answers created is measured by week, Avg. # of votes received only contains the votes from other influencers;

4. 66.50% matching rate for unselected users (388) between two years and six months datasets.

Table 9: Summary of statistics

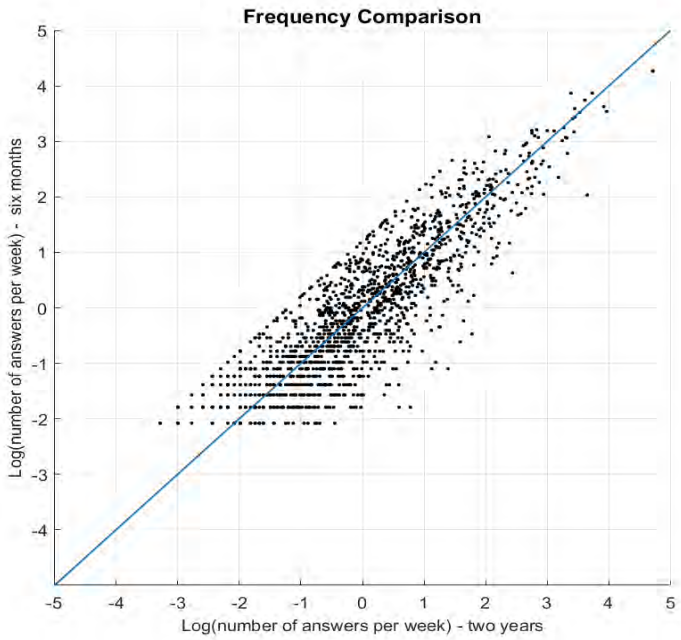

Figure 8: Comparing answer created per week per user between six-month and two-year data

C Cross-sectional data checks

In the data, however, we do not directly observe the weekly increase in the number of votes from all followings, nor do we observe the weekly increase in the number of votes. We can only observe the changes in the number of all votes before and after six months in the data. Therefore we check the following regression model:

$$\log\left(Anwers_i\right) = \beta_0 + \log\left(Votes_i\right)\beta_v + \eta_{i,t},$$
$$\log\left(Anwers_i\right) = \beta_0 + \Delta\log\left(All\ Votes_i\right)\beta_v + \eta_{i,t}.$$

Regression results in the Table 10 indicate that there is almost no additional effect on the incentive of creating answers when replacing the votes received from influencers by all users. The correlation coefficients from our two regressions are almost exactly the same, which means that the vote from other influencers is a good proxy for the

Dependent var.: $\log(Anwers)$		
	(2) OLS	(3) OLS
$\log(Votes)$	0.449***	
	(0.019)	
$\log(All\ Votes)$		0.451***
		(0.019)
Intercept	1.618***	1.585***
	(0.071)	(0.071)
N	1497	1500
R^2	0.288	0.294

Note: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.
Standard errors in parentheses

Table 10: The impact of "all votes" and "votes" on the answer creations

"peer recognition" effect. We also note that the regression coefficient of the cross-sectional data is much higher than the regression result of the panel data, which well proves that the use of lag variable can help us eliminate many potential endogenous problems.

D The impact of knowledge spillover

In this section, we use the data to evaluate the impact of potential knowledge spillover. Table 11 report results of estimation from both OLS and IV models. The regression results showed that, although by construction, potential readers show a smaller influence on content creation than "received votes from influencers' votes." A 1% increase in the number of the potential number of readers raises the number of answers by 0.1%. The instrumental variables also pass the test perfectly and prove that OLS greatly underestimated the actual effect of knowledge spillover in the new regression model by almost ten times the OLS coefficient. The instrumental variables also pass the tests perfectly and prove that OLS greatly underestimated the actual effect of knowledge spillover in the new regression model by almost ten times the OLS coefficient. We report the tests related to the first stage of IV method as well as the diagnostic tests in Table 12. Specifically, this time, we construct our instruments variables by the corresponding readers brings to other articles:

$$Readers\ article\ votes_i = \sum_{influencer\ j\ votes\ i\ during\ t-1} Article\ votes_{j,t-1} \times Follower_{j,0},$$

$$Readers\ article\ collections_i = \sum_{influencer\ j\ votes\ i\ during\ t-1} Article\ collections_{j,t-1} \times Follower_{j,0},$$

where $Articlecollections_{j,t-1}$ and $Articlecollections_{j,t-1}$ are the total number of other articles that an influencer j, who votes for i's answers, votes for and collects during the period $t-1$.

Finally, we also report results by controlling if an influencer is a commercial user. Table 13 shows that if we exclude individual users and look at business users separately, the OSL results are exactly the same as the IV results.

	Dependent variable: $\log(Anwers_t + 1)$										
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
	OLS	OLS	OLS	OLS	OLS	OLS	IV	IV	IV	IV	IV
$\log(Readers_{t-1} + 1)$	0.013***	0.013***	0.010***	0.010***	0.013***	0.013***	0.081***	0.095***	0.073***	0.089***	0.101***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.011)	(0.013)	(0.013)	(0.016)	(0.014)
$Week$					-0.204***	-0.204***					-0.198***
					(0.023)	(0.023)					(0.029)
$Week^2$					0.008***	0.008***					0.008***
					(0.000)	(0.000)					(0.001)
$\log(Follower_0)$						-0.005					-0.152***
						(0.017)					(0.030)
Intercept	0.635***	0.488***	0.652***	0.504***	0.863***	0.921***	0.251***	0.034	0.291***	0.067	1.924***
	(0.014)	(0.019)	(0.004)	(0.014)	(0.048)	(0.174)	(0.067)	(0.079)	(0.075)	(0.094)	(0.249)
Individual Fixed Effects			√	√					√	√	
Time Fixed Effects		√		√	√	√		√		√	√
N	34500	34500	34500	34500	34500	34500	34500	34500	34500	34500	34500
R^2	0.007	0.031	0.007	0.032	0.031	0.031	0.007	0.013	-0.444	-0.444	0.013

Table 11: The impact of potential knowledge spillover

Dependent variable: $\log(Readers_{t-1}+1)$	
Instrument	(1)
$\log(Readers\ article\ votes_i+1)$	0.041***
	(0.009)
$\log(Readers\ article\ collections_i+1)$	0.037***
	(0.013)
Individual Fixed Effects	√
Time Fixed Effects	√
N	34500
R^2	-0.444
Kleibergen-Paap Wald stat	35.636
(p-value)	(0.000)
Hansen J statistic	0.855
(p-value)	(0.355)
F stat	35.636

Table 12: First stage results and diagnostic Test

	Dependent variable: $\log(Anwers_t+1)$		
	(1)	(2)	(3)
	OLS	OLS	IV
$\log(Readers_{t-1}+1)$	0.013***	0.013***	0.101***
	(0.001)	(0.001)	(0.014)
$Week$	-0.204***	-0.204***	-0.198***
	(0.023)	(0.023)	(0.029)
$Week^2$	0.008***	0.008***	0.008***
	(0.000)	(0.000)	(0.001)
$\log(Follower_0)$	-0.005	-0.006	-0.152***
	(0.016)	(0.016)	(0.029)
$Commercial$		0.076	
		(0.072)	
$Commercial \times \log(Readers_{t-1}+1)$		0.012***	
		(0.003)	
Time Fixed Effects	0.921***	0.928***	1.924***
N	(0.174)	(0.174)	(0.249)
R^2	√	√	√

Note: * $p<0.1$, ** $p<0.05$, *** $p<0.01$

Standard errors in parentheses.

Commercial is a binary dummy variable.

Table 13: Impact on creating new answers - commercial and non commercial influencers

E The concern of potential multi-homing

The best candidates may also receive less incentive from others' votes because other competing platforms may poach them after they are awarded as "best answerers" in their field. Under the assumption that the multi-homing probability is positively correlated with the number of followers, our estimation results in Table 14 show that the magnitude of the effect of multi-homing on is relatively small. In most of the columns, the intersection of followers and votes is not significant. Except in the last column, the estimates indicate that the total number of followers negatively ." However, after controlling the variable "best answerer," we find that for the "best answerers", the increase of followers will have a positive and significant impact on peer incentives. Such finding confirms the existence of the reputation concern and even shows that the reputation concern in our sample dominates the potential multihoming effect.

	Dependent variable: $\log(Anwers_t + 1)$							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	OLS	OLS	OLS	OLS	OLS	OLS	OLS	OLS
$\log(Votes_{t-1} + 1)$	0.107***	0.111***	0.081***	0.085***	0.085***	0.180***	0.165***	0.283***
	(0.004)	(0.004)	(0.004)	(0.004)	(0.004)	(0.054)	(0.054)	(0.064)
$Week$					-0.208***	-0.116***	-0.100***	-0.110***
					(0.023)	(0.024)	(0.026)	(0.026)
$Week^2$					0.009***	0.007***	0.006***	0.006***
					(0.001)	(0.001)	(0.001)	(0.001)
$\log(Follower_0)$						-0.020	-0.037*	-0.027
						(0.017)	(0.020)	(0.020)
$\log(Follower_0) \times \log(Votes_{t-1} + 1)$						-0.006	-0.004	-0.015**
						(0.005)	(0.005)	(0.006)
$Best$							-0.892**	-0.550
							(0.375)	(0.388)
$\log(Follower_0) \times Best$							0.078**	0.044
							(0.036)	(0.037)
$\log(Votes_{t-1} + 1) \times Best$							-0.040***	-0.447***
							(0.009)	(0.120)
$\log(Follower_0) \times \log(Votes_{t-1} + 1) \times Best$								0.039***
								(0.011)
Individual Fixed Effects			√	√	√			
Time Fixed Effects		√		√	√	√	√	√
N	33235	33235	33235	33235	33235	33235	33235	33235
R^2	0.011	0.035	0.011	0.036	0.036	0.035	0.035	0.036

Note: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Standard errors in parentheses.

Table 14: Impact of number followers on answers created (non-commercial influencers)

F The impact of getting a badge

We check in this section the effect of badge issued by the platform on influencers' motivation for providing answers. In particular, we do not specify the types of badge. Holding a badge indicates that an influencer belongs to at least one of the following categories: 1. a self-authenticated user; 2. a platform awarded "best answerer"; 3. a commercial user. Results in Table 15 show that the fact of holding a badge negatively affect on frequency of providing new answers.

	Dependent variable: $\log(Anwers_t + 1)$		
	(1)	(2)	(3)
	OLS	OLS	IV
$\log(Votes_{t-1} + 1)$	0.115***	0.123***	0.164***
	(0.004)	(0.005)	(0.013)
$Week$	-0.112***	-0.116****	-0.200***
	(0.024)	(0.024)	(0.023)
$Week^2$	0.006***	0.007***	0.008***
	(0.001)	(0.001)	(0.001)
$\log(Follower_0)$	-0.019	-0.014	-0.035**
	(0.016)	(0.016)	(0.017)
$Badge$		-0.055*	
		(0.028)	
$Badge \times \log(Votes_{t-1} + 1)$		-0.021**	
		(0.008)	
Time Fixed Effects	√	√	√
N	34500	34500	34500
R^2	0.036	0.036	0.033

Note: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$
Standard errors in parentheses.
$Commercial$ is a binary dummy variable.

Table 15: Impact on creating new answers - badge and no badge influencers

G An anecdotal evidence of privacy concern

In the late December 2019, Zhihu has launched a more severe "real name" authentication policy: the platform requires every registered user to provide mobile phone and pass the real-name authentication. Users who refuse to contribute will be subject to certain restrictions on what they can say on the platform. Such a policy worries many users, some of whom even refuse to continue exporting content online. Figure 10 shows an article by Mather King, one of the "best answerers" under math questions. The article received more than three thousand votes within two days of publication. In the article, the author made clear his potential concerns about personal privacy disclosure after being asked to provide real-name information. He said that he would henceforth stop contributing any academic content to the platform.

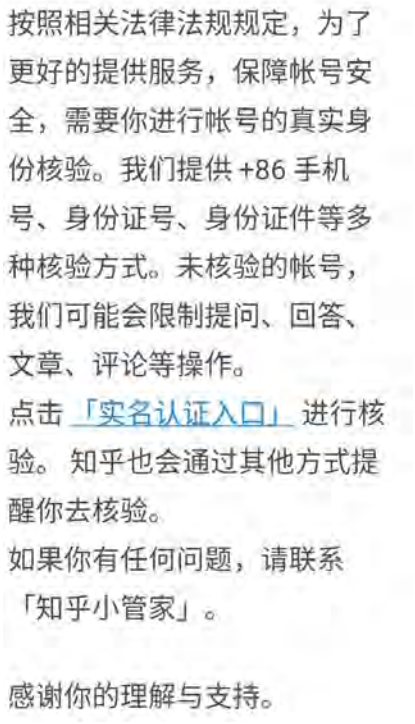

Figure 9: An anecdotal evidence of the raising privacy concerns in the platform

Figure 10: An anecdotal evidence of the raising privacy concerns in the platform