Strategic Content Generation and Monetization in Financial Social Media

Online Appendix

1 Literature Review

We review the studies on content generation and monetization in Table A1 below to position our study in the relevant literature.

Table A1 Literature Review and Our Paper's Positioning

Stream	Paper	Topic	Data	Analysis Level	DV	IV	Monetization Method
Financial Social Media	Antweiler and Frank (2004), Bagnoli et al. (1999), Chen et al. (2014), Clarkson et al. (2006), Das and Chen (2007), Luo et al. (2013), Sabherwal et al. (2008)	Informational value of social media content	Financial social media (Twitter, Seeking Alpha, online discussion forum, etc.)	Platform or stock	Stock market reactions (trade volume, volatility, returns, etc.)	Social media content (i.e., sentiment)	-
	Kadous et al. (2019)	Why investors rely on social media content	Experiment	Investor	Willingness to invest	Tone, content type (sentiment or advice), source credibility	-
	Chen et al. (2019)	Impact of monetary incentive	Financial social media (Seeking Alpha)	Content contributor	Content quantity, stock recommendation accuracy	Monetary incentive	Platform's ad revenue sharing
Content Generation	Becker et al. (2010), Hsieh et al. (2010), Liu et al. (2014), Liu and Feng (2015), Mason and Watts (2010), Stephen et al.	Impact of monetary incentive	Question & answer platforms; product review; online	Content contributor	Content decisions (e.g., quantity and quality); participation	Introduction or presence of monetary incentive	Platform paying contributors or ad revenue sharing

Stream	Paper	Topic	Data	Analysis Level	DV	IV	Monetization Method
	(2012), Tang et al. (2012), Wang et al. (2012)		community; social media				
	Baek and Shore (2020), Qiu and Kumar (2017), Wei et al. (2020), Bhattacharya et al. (2019), Lu et al. (2017), Guo et al. (2020), Hwang et al. (2015), Cavusoglu et al. (2016), Heimbach and Hinz (2018), Zhao et al. (2016)	Factors affecting content generation	Online community; social media	Content contributor	Content decisions (e.g., quantity and quality); participation	Social network factors; distance factors; content platform policies; cognitive factors	-
	Goes et al. (2014), Nguyen et al. (2021), Shen et al. (2015)	Strategic content generation	Product review	Content contributor (reviewer)	Review characteristics	Ranking system; experience; popularity;	-
Content Monetization	Aral and Dhillon (2020), Lambrecht and Misra (2017)	Design of paid content policy	Sports media website (ESPN)	Platform	Revenue	Price in high/ low demand; free article quantity and sections	Platform's paid content subscription
	Oh et al. (2016), Pattabhiramaiah et al. (2019), Pauwels and Weiss (2008)	Impact of paid content policy	Content websites (e.g., the New York Times)	Platform /newspaper /News articles	Word-of-mouth of articles on social media; readership; reader engagement; print newspaper sales	Implementation of paywall; adoption of paid content policy	Platform's paid content subscription (e.g., paywall)
	Our paper	Strategic content generation and monetization in financial social media	Financial social media (iMaibo)	Content contributor (SMA)	Content generation decision, monetization decision	Time-varying investor preferences	Contributors' paid content subscription

2 Identification Strategy

In our study, various sources of potential endogeneity bias can arise, such as SMAs' content generation and monetization decisions are subject to their knowledge of investor preferences (which we evaluate in our research questions) and to unobserved influences (e.g., omitted variables, self-selection), beyond the focal and control factors accounted for in our proposed model. If this is indeed the case, the investor preference parameters (i.e., $\alpha_{it}^{\bullet}$, $\beta_{it}^{\bullet}$) and SMA response parameters (i.e., $\kappa_{i\bullet}$, $\gamma_{i\bullet}$, $\lambda_{i\bullet}$) are likely biased. To control for potential endogeneity bias, we propose to follow the framework of non-random marketing-mix variables (Bronnenberg and Mahajan 2001, Manchanda et al. 2004), which leverages the full likelihood of the data and the power of Bayesian inference. To elaborate the framework, we look at the example below.

$$y_{it} = \beta * X_{it} + \epsilon_{it} \tag{1}$$

$$\begin{aligned} \ell(\beta_{it}, \theta) &= \prod_{i,t} P(y_{it} \mid X_{it}, \beta) p(X_{it} \mid \theta) \\ &\xrightarrow{\text{if } X_{it}|\theta \perp y_{it}|X_{it},\beta} \prod_{i,t} P(y_{it} \mid X_{it}, \beta) \prod_{it} p(X_{it} \mid \theta) \end{aligned} \tag{2}$$

Equation (1) is a linear model, and equation (2) is the corresponding full likelihood of the data. y_{it} can represent performance metrics (e.g., drug sales, readership, and subscriptions). X_{it} can denote the strategic variable of interest (e.g., pharmaceutical firms' detailing on physicians and the sentiment decision of SMAs' financial social media content). The identification task is to obtain an accurate, unbiased estimate of β. This is often done by maximizing the conditional likelihood (or distribution) of y_{it} given X_{it},

$$y_{it} \mid X_{it}, \beta \tag{3}$$

which is, $\prod_{i,t} P(y_{it} \mid X_{it}, \beta)$ in equation (2). This conditional likelihood approach is valid only under the assumption that the marginal distribution of X_{it} (i.e., $X_{it} \mid \theta$) is independent of the conditional distribution of y_{it} (i.e., $X_{it} \mid \theta \perp y_{it} \mid X_{it}, \beta$). In this case, X_{it} does not provide information about β. Therefore, the full likelihood of the data can be factorized, as in equation (2). Consequently, the conditional likelihood of y_{it} is enough for the estimation of β without resorting to the full likelihood of the data. However, it is possible that X_{it} is endogenous, for example, such that it is chosen strategically by managers (or SMAs in our study). Statistically speaking, there exists a dependence[1] between the conditional distribution $y_{it} \mid X_{it}, \beta$ and the marginal distribution of X_{it}. This dependence is known as endogeneity. Thus, the likelihood function of the model can no longer be factorized in the manner as shown in equation (2). Estimating the model using the conditional likelihood method (i.e., equation (3)) may lead to a biased β.

Instead of relying on the conditional likelihood of y_{it}, the framework of non-random marketing-mix variables utilizes the full likelihood. Specifically, the dependence between the conditional distribution $y_{it} \mid X_{it}, \beta$ and the marginal distribution of X_{it} is explicitly modeled to establish the full model. The framework recommends making the marginal distribution of X_{it} depend on the response (or preference) parameter in the conditional model (i.e., β). The full model can be summarized below,

$$y_{it} \mid X_{it}, \beta \text{ and } X_{it} \mid \beta, \tau \tag{4}$$

Modeling how X_{it} depends on β (via τ) entails three benefits (Manchanda et al. 2004): 1) removal of possible endogeneity biases since the dependence is explicitly accounted for; 2) more precise estimate of β by exploiting the information in both the conditional distribution $y_{it} \mid X_{it}, \beta$ and the marginal distribution of $X_{it} \mid \beta, \tau$; 3) detection of the existence of strategic behaviors through assessing the extent of how the information of β affects the generation of X_{it} (via the significance and magnitude of τ). However, traditional frequentist approaches have a hard time estimating the full model in equation (4). Bayesian inference, on the other hand, is ideally geared toward handling the above model structure. This method of inference is fully Bayesian and does not rely on the existence of valid instruments and asymptotic approximations (Manchanda et al. 2004). This framework has been adopted by many papers to handle the strategic behaviors

[1] This dependence has other interpretations such as omitted variables (including self-selection, reverse causality and simultaneity) or violation of the independence between the regression residual error term and the independent variables. See a detailed discussion in Manchanda et al. (2004).

of firms and others, such as Li et al. (2011), Luo and Kumar (2013), Musalem et al. (2008), Nair et al. (2017), Schweidel and Knox (2013), Van Diepen et al. (2009).

Given the structure of our model, it is important to provide some intuition for the identification of the parameters. Our simulation result in the online appendix's Section 4 also shows that our proposed model can recover the parameters with a good degree of accuracy.

The parameters in our model can be grouped into two categories: 1) the regression coefficients in the demand-side observation equations ($\rho_i^{\bullet}$), demand-side state equations ($\delta_i^{\bullet}$, $SMA_i^{\bullet}$), and supply-side equations ($\kappa_{i\bullet}$, $\gamma_{i\bullet}$, $\lambda_{i\bullet}$, $\eta_i^{\bullet}$), and the associated (co)variance for the error term of each equation; 2) the investor preferences ($\alpha_{it}^{\bullet}$, $\beta_{it}^{\bullet}$). The identification of the regression coefficients and (co)variance for the error terms is identical to that in simple regressions, i.e., determined by the relative rate of change between dependent variables and independent variables and by the difference between dependent variables and the predicted values.

The identification of investor preferences is a bit complicated. We note that the investor preferences have both subscripts *i* and *t*, that is, the number of one type of investor preference (e.g., subscribers' preference for free content sentiment) is equal to the number of observations. Thus, without additional structure, the investor preferences cannot be identified. It is useful to understand the investor preferences first without the time dimension. Without the time dimension, the investor preferences are similar to normal regression coefficients. The difference is that two pieces of information are used jointly to pin down the investor preference: 1) the relation between $Reader_{it}$ ($Subscriber_{it}^{*}$) with $FreeSentiment_{it}$ ($FreeSentiment_{it}$, $PaidSentiment_{it}$), and 2) the relation between $FreeSentiment_{it}$ ($PaidSentiment_{it}$) and SMAs' degree of catering to investor preferences, γ_{i1} and γ_{i2} (λ_{i1}). When estimating the investor preferences, we essentially "borrow" additional observations from the equations of $FreeSentiment_{it}$ and $PaidSentiment_{it}$. With the time dimension, the investor preferences cannot be identified without extra structures. Infinite investor preference values can fit our model equally well (i.e., investor preferences can vary freely), since we have the same amount of unknown parameters and data. The state equations provide such a structure that specifies the evolution pattern of investor preferences or how adjacent investor preferences are connected. With this structure, when estimating the investor preference at time *t*, in addition to the two pieces of information mentioned above, the investor preference at time *t-1* offers another equation. Thus, the investor preferences can no longer vary freely, but by the pattern specified in the state equations, such that they can be validly identified.

3 MCMC Procedure for the Proposed Model

The proposed model is estimated using a hierarchical Bayesian framework. We estimate the model parameters by iteratively sampling from their full conditional distributions, which proceeds as follows,

- Step 1, draw potential subscription, $Subscriber_{it}^{*}$

$$Subscriber_{it}^{*} \sim TruncatedNormal_{[Subscriber_{it},\ +\infty]}(M,S)\text{, if } Monetize_{it}=0$$

$$Subscriber_{it}^{*} = Subscriber_{it}\text{, if } Monetize_{it}=1$$

where

$$M = \beta_{i0} + \beta_{it}^{f} FreeSentiment_{it} + \beta_{it}^{p} PaidSentiment_{it} * 1(Monetize_{it}=1) + \rho_{i}^{s} Control$$

$$S = V_{(1,1)}$$

- Step 2, draw latent variable, $Monetize_{it}^{*}$

$$Monetize_{it}^{*} \sim TruncatedNormal_{[-\infty,0]}(M,S) \text{ if } Monetize_{it}=0$$

$$Monetize_{it}^{*} \sim TruncatedNormal_{[0,\infty]}(M,S) \text{ if } Monetize_{it}=1$$

where

$$M = \kappa_{i0} + \kappa_{i1} E(Reader_{it}) + \kappa_{i2} E(Subscriber_{it}) + \eta_{i}^{m} Control$$

$$S = \Omega_{(1,1)}$$

- Step 3, draw investor preferences, $\alpha_{it}^{\bullet}$ and $\beta_{it}^{\bullet}$

Let us first re-arrange the model in the following format:

Observation equations

$$Y_{it} = \mu_{it} F_{it} + \omega_{it}$$

State equations

$$\mu_{it} = G_i \mu_{it-1} + \overline{\mu}_{it} + \omega_{it}$$

where

$$Y_{it} = \begin{bmatrix} y_{1it} \\ y_{2it} \\ y_{3it} \\ y_{4it} \\ y_{5it} \end{bmatrix}$$

$$y_{1it} = Reader_{it} - \rho_i^r Control - \alpha_{i0}$$

$$y_{2it} = Subscriber_{it}^* - \rho_i^s Control - \beta_{i0}$$

$$y_{3it} = Monetize_{it}^* - \kappa_{i0} - \eta_i^m Control$$

$$y_{4it} = FreeSentiment_{it} - \gamma_{i0} - \eta_i^f Control$$

$$y_{5it} = PaidSentiment_{it} - \lambda_{i0} - \eta_i^p Control$$

$$\omega_{it} \sim MVN(0, \Pi),\ \Pi = \begin{bmatrix} V & \\ & \Omega \end{bmatrix}$$

$$\mu_{it} = [\alpha_{it}^{\bullet}, \beta_{it}^{\bullet}],\ F_{it} = \begin{bmatrix} [FreeSentiment_{it}] & & \\ & [FreeSentiment_{it}, PaidSentiment_{it} * 1(Monetize_{it} = 1)] & \\ [\kappa_{i1}, \kappa_{i2}] & & \\ & & [\gamma_{i1}, \gamma_{i2}] \\ & & [\lambda_{i1}] \end{bmatrix}$$

$$G_i = \begin{bmatrix} \delta_i^{\alpha^{\bullet}} & \\ & \delta_i^{\beta^{\bullet}} \end{bmatrix}, \overline{\mu}_{it} = SMA_i^{\bullet} + \tau_t^{\bullet}$$

The draw of investor preferences is done using the Kalman filtering technique which consists of two steps: forward filtering and backward sampling. The derivation of the distributions is in Lindsten and Schön (2013).

At the filtering step, the objective is to obtain the filtering density,

$$\mu_{it} \sim MVN(m_{it}, C_{it})$$

where

$$m_{it} = a_{it} + A_{it} e_{it} \ , \ C_{it} = \mathrm{R}_{it} - A_{it} Q_{it} A_{it}^{'}$$

$$e_{it} = Y_{it} - f_{it} \, , \ A_{it} = \mathrm{R}_{it} F_{it} Q_{it}^{-1}$$

$$\mathrm{R}_{it} = G_i C_{it-1} G_i^{'} + W \ , \ Q_{it} = F_{it}^{'} \mathrm{R}_{it} F_{it} + \Pi$$

$$a_{it} = G_i m_{it-1} + \overline{\mu}_{it} \, , \ f_{it} = F_{it}^{'} a_{it}$$

Note that we assume the initial states $\mu_{i0} \sim MVN(m_0, C_0)$ where $m_0 = 0$ and $C_0 = 100I$.

At the sampling step, we draw μ_{it} recursively from the following distribution starting backward from μ_{iT} to μ_{i1}:

$$\mu_{it} \sim MVN(\hat{\mu}_{it}, M_{it})$$

where

$$\hat{\mu}_{it} = m_{it} + C_{it} G_i^{'} (W + G_i C_{it} G_i^{'})^{-1} (\mu_{it+1} - (G_i m_{it} + \overline{\mu}_{it}))$$

$$M_{it} = C_{it} - C_{it} G_i^{'} (W + G_i C_{it} G_i^{'})^{-1} G_i C_{it}$$

Note that the last state μ_{iT} is drawn directly from the filtering density $\mu_{iT} \sim MVN(m_{iT}, C_{iT})$.

- Step 4, draw SMA-specific regression coefficients $\phi = [\rho_i^{\bullet}, \delta_i^{\alpha^{\bullet}}, \delta_i^{\beta^{\bullet}}, SMA_i^{\bullet}, \kappa_{i\bullet}, \gamma_{i\bullet}, \lambda_{i\bullet}, \eta_i^{\bullet}]$

$$Y_{it} = \begin{bmatrix} y_{1it} \\ y_{2it} \\ \vdots \\ y_{7it} \end{bmatrix}$$

$$y_{1it} = Reader_{it} - \alpha_{i0} - \alpha_{it}^{f} FreeSentiment_{it}$$

$$y_{2it} = Subscriber_{it}^{*} - \beta_{i0} - \beta_{it}^{f} FreeSentiment_{it} - \beta_{it}^{p} PaidSentiment_{it} * 1(Monetize_{it} = 1)$$

$$y_{3it} = \alpha_{it}^{\bullet} - \tau_t^{\alpha^{\bullet}}$$

$$y_{4it} = \beta_{it}^{\bullet} - \tau_t^{\beta^{\bullet}}$$

$$y_{5it} = Monetize_{it}^{*} - \kappa_{i1} E(Reader_{it}) - \kappa_{i2} E(Subscriber_{it})$$

$$y_{6it} = FreeSentiment_{it} - \gamma_{i1} \alpha_{it}^{f} - \gamma_{i2} \beta_{it}^{f}$$

$$y_{7it} = PaidSentiment_{it} - \lambda_{i1}\beta_{it}^{p}$$

$$X_{it} = \begin{bmatrix} Control & & & & & & \\ & Control & & & & & \\ & & [\alpha_{it-1}^{\bullet}, 1] & & & & \\ & & & [\beta_{it-1}^{\bullet}, 1] & & & \\ & & & & [1, Control] & & \\ & & & & & [1, Control] & \\ & & & & & & [1, Control] \end{bmatrix}$$

$$\Sigma = \begin{bmatrix} V & & \\ & W & \\ & & \Omega \end{bmatrix}$$

$$\overline{b_i} = \overline{\phi} Z_i , \Sigma_0 = \Gamma$$

Then, $b_i \sim MVN(A_i, B_i)$, where

$$A_i = B_i[X_{it}{}' \Sigma Y_{it} + \Sigma_0^{-1} \overline{b_i}]$$

$$B_i = [X_{it}{}' \Sigma X_{it} + \Sigma_0^{-1}]^{-1}$$

- Step 5, draw $\tau_t^{\bullet}$

  $\tau_t^{\bullet}$ are time-fixed effects. The draw of $\tau_t^{\bullet}$ follows a similar procedure to Step 4. Here, we take $\alpha_{it}^{\bullet}$ and $\beta_{it}^{\bullet}$ as the dependent variables, group the observations at the same time period *t* as separate regressions, and draw $\tau_t^{\bullet}$ as the coefficient of the intercept in each regression.

- Step 6, draw population-level parameters $\overline{\phi}$

$$X = \begin{bmatrix} I & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & I \end{bmatrix}$$

$$a = 0 , \ \Sigma_0 = I$$

Then, $\overline{\beta}_n \sim MVN(A, B)$, where

$$A = B[X' \Gamma_i \phi_i + \Sigma_0^{-1} a]$$

$$B = [X'\Gamma_i X + \Sigma_0^{-1}]^{-1}$$

- Step 7, draw the error term variance for the demand-side observation equations, V

$$Y_{it} = \begin{bmatrix} y_{1it} \\ y_{2it} \end{bmatrix}$$

$$y_{1it} = Reader_{it} - (\alpha_{i0} + \alpha_{it}^{f} FreeSentiment_{it} + \rho_i^{r} Control)$$

$$y_{2it} = Subscriber_{it}^{*} - (\beta_{i0} + \beta_{it}^{f} FreeSentiment_{it} + \beta_{it}^{p} PaidSentiment_{it} * 1(Monetize_{it} = 1) + \rho_i^{s} Control)$$

$$V_0 = 100\mathbf{I}, \nu_0 = 5$$

Then, $V_{it} \sim InvertedWishart(\sum_i \sum_t Y_{it}' Y_{it} + V_0, N + \nu_0)$, where N is the total number of data observations.

The draw of W_{it}, U_t and Γ_i follow the same procedure, in which we just need to replace Y_{it} with the errors of corresponding equations.

- Step 8, draw the error term variance for the supply-side observation equations, Ω.

Note that the $Monetize_{it}$ equation is a binary probit model. For identification purpose, we need to set its variance (i.e., the first element of Ω) to 1. Let us assume

$$\Omega = \begin{bmatrix} 1 & \rho \\ \rho & \pi \end{bmatrix}$$

$$y_{1it} = Monetize_{it}^{*} - (\kappa_{i0} + \kappa_{i1} E(Reader_{it}) + \kappa_{i2} E(Subscriber_{it}) + \eta_i^{m} Control)$$

$$y_{2it} = \begin{bmatrix} FreeSentiment_{it} - (\gamma_{i0} + \gamma_{i1}\alpha_{it}^{f} + \gamma_{i2}\beta_{it}^{f} + \eta_i^{f} Control) \\ PaidSentiment_{it} - (\lambda_{i0} + \lambda_{i1}\beta_{it}^{p} + \eta_i^{p} Control) \end{bmatrix}'$$

Then $\pi \sim InvertedWishart(v_1, s_1^2)$ where

$$v_1 = v_0 + N, s_1^2 = \frac{v_0 s_0^2 + (y_{2it} - \rho y_{1it})'(y_{2it} - \rho y_{1it})}{v_0 + N},$$

$$v_0 = 4, s_0^2 = 1I$$

$\rho \sim N((y'_{1it} y_{1it} + A)^{-1}(y'_{1it} y_{2it} + A\overline{\beta}), \pi(y'_{1it} y_{1it} + A)^{-1})$ where

$$A = 100, \bar{\beta} = 0$$

- Step 9, draw expected readership $E(Reader_{it})$ and expected subscriptions $E(Subscriber_{it})$

We follow the approximation method used in Albuquerque et al. (2012) (see Page 417) to obtain $E(Reader_{it})$ and $E(Subscriber_{it})$. Specifically, we use the estimated parameters and the explanatory variables (in equations (1) and (2)) to compute the deterministic part of $E(Reader_{it})$ and $E(Subscriber_{it})$. We then draw unobserved shocks from V and add the shocks to the deterministic part of $E(Reader_{it})$ and $E(Subscriber_{it})$. Finally, we average the obtained quantities to compute the expectations, that is, $E(Reader_{it})$ and $E(Subscriber_{it})$.

4 Simulation Study and Result

To ensure that our model is fully identified, we first create a simulated dataset that resembles the structure of the real data we have. To be specific, we simulate a dataset of 500 SMAs, each with 100 observations. For illustration purposes, we only include two control variables for each observation equation and only include the *Monetize* and *FreeSentiment* equations on the supply side. The true parameter values are stated in the following tables, along with the estimated parameter values. We run the model on the simulated dataset with 20,000 MCMC iterations and use the last 5,000 iterations to compute the mean and standard deviation of the parameters. If our model is fully identified, the estimated parameter values should be close to the true parameter values that we use to create the simulated dataset. We find that the estimated parameter values are indeed quite close to the true values, showing that our model is validly identified well.

Table A2 Population-level Estimates for Observation Equations

Equation	Variable	True Value	Estimated Value
Reader	*Covariate1*	-3	-3.147*** (0.115)
	Covariate2	-3	-3.068*** (0.108)
Subscriber	*Covariate3*	0	-0.003 (0.122)
	Covariate4	-3	-3.079*** (0.130)

Table A3 Population-level Estimates for SMA-specific Effects (i.e., $SMA_i^{\bullet}$)

Parameter	True Value	Estimated Value
SMA^{α^f}	3	3.155*** (0.168)
SMA^{β^f}	-3	-3.007*** (0.089)

Table A4 Carryover Effects of Investor Preferences

Parameter	True Value	Estimated Value
$\delta_i^{\alpha^f}$	0.5	0.473*** (0.015)
$\delta_i^{\beta^f}$	0.4	0.385*** (0.009)

Table A5 Supply-side Population-level Estimates

Equation	Variable	True Value	Estimated Value
Monetize	$E(Reader_{it})$	1	1.063*** (0.066)
	$E(Subscriber_{it})$	-2	-2.072*** (0.071)
FreeSentiment	α_{it}^f	1	0.959*** (0.042)
	β_{it}^f	-1	-1.008*** (0.044)

Table A6 Supply-side Population-level Estimates (Control)

Equation	Variable	True Value	Estimated Value
Monetize	*Covariate5*	1	0.968*** (0.075)
	Covariate6	-2	-2.011*** (0.091)
FreeSentiment	*Covariate7*	2	2.011*** (0.049)
	Covariate8	3	3.047*** (0.078)

5 Additional Data Description

Here, we present more information about our data context. The data for our empirical analysis comes from a China-based financial social media platform, iMaibo, which hosts around five million registered users. iMaibo employs a Twitter-like following structure, that is, upon a user (i.e., the follower) following another user (i.e., the followee), the followee's content will automatically appear on the follower's timeline in a reverse chronological order. Figure A1 presents a sample screenshot of an iMaibo user's timeline. On iMaibo, we observe two types of users: SMAs and retail investors. SMAs are platform-verified advisors, with special badges signaling their status (the blue circle with white "V" next to "SMA a" in Figure A1). SMAs can publish both free and paid tweets (i.e., VIP tweets in Figure A1, the paid content in our study) to retail investors (i.e., their followers) to consume. Such content mainly involves investment advice (e.g., stock recommendations and asset allocations) for the Chinese stock market.

Paid tweets employ the subscription model. There are three types of subscription contracts with prices set independently by the SMAs: daily, monthly, and quarterly. Based on the contract a retail investor chooses, she can view all the paid (as well as free) tweets authored by a SMA for the coming 1 day, 30 days (monthly), or 90 days (quarterly). The only button for the subscription payment (i.e., "*Subscribe to View VIP Content*" in Figure A1) is embedded in paid tweets, and the button is deactivated the following day at 12 am midnight subsequent to a paid tweet's publication. Therefore, an investor cannot make a subscription to a SMA if the SMA does not produce paid content on that day. Moreover, the renewal option (i.e., new subscription) is only available to the investor after his current contract expires. In Table A7, we present additional summary statistics for variables not mentioned in the main paper.

Figure A1 An iMaibo User's Timeline

Table A7 Additional Summary Statistics

Variable	Definition	Mean	S.D.	Min	Max
TotalTweetCount	Total number of tweets generated by SMA *i* on day *t*	10.754	16.065	1	169
Price	Per-day price of SMA *i*'s subscription package	183.384	312.943	0	1000
FreeAccuracy1D	Next 1-day prediction accuracy of SMA *i*'s free content on day *t*	0.125	0.284	0	1
FreeAccuracy1M	Next 1-month prediction accuracy of SMA *i*'s free content on day *t*	0.096	0.257	0	1
FreeDiversity	Diversity of the stock coverage of SMA *i*'s free content on day *t*	6.745	17.975	0	410
FreeNovelty	Novelty (number of stocks not mentioned by other SMAs) of the stock coverage of SMA *i*'s free content on day *t*	0.433	3.798	0	185
PaidTweetCount	Number of paid tweets generated by SMA *i* on day *t*	2.055	9.331	0	148
PaidAccuracy1D	Next 1-day prediction accuracy of SMA *i*'s paid content on day *t*	1.001	5.056	0	227
PaidAccuracy1M	Next 1-month prediction accuracy of SMA *i*'s paid content on day *t*	0.190	2.573	0	342
PaidDiversity	Diversity of the stock coverage of SMA *i*'s paid content on day *t*	13.232	76.474	0	1231.435
PaidNovelty	Novelty (number of stocks not mentioned by other	0.039	0.172	0	1

	SMAs) of the stock coverage of SMA *i*'s paid content on day *t*				

6 Demand-side Results

6.1 Demand-side Model

Table A8 Population-level Estimates for Observation Equations

Variable	*Reader*	*Subscriber*
CommentLag	2.756*** (0.115)	1.437*** (0.077)
LikeLag	0.017 (0.020)	0.019 (0.021)
FollowerLag	0.013 (0.017)	0.020 (0.015)
Tenure	-0.111 (0.426)	0.068 (0.223)
isTradeDay	0.364*** (0.070)	-0.159*** (0.057)
TradeVolume	-0.001 (0.012)	-0.001 (0.012)
VIX	-0.010 (0.013)	-0.002 (0.012)
TotalTweetCount	0.081*** (0.015)	0.042** (0.018)
Price	-0.005 (0.031)	-0.001 (0.030)
FreeAccuracy1D	0.021 (0.034)	0.013 (0.038)
FreeAccuracy1M	0.019 (0.037)	0.041 (0.048)
FreeDiversity	0.011 (0.017)	0.020 (0.018)
FreeNovelty	-0.031 (0.032)	-5e-04 (0.033)
PaidTweetCount	-0.001 (0.059)	-0.223*** (0.067)
PaidAccuracy1D		-0.011 (0.088)

PaidAccuracy1M		0.045 (0.111)
PaidDiversity		-0.023 (0.062)
PaidNovelty		-0.021 (0.075)

Table A9 Population-level Estimates for State Equations (i.e., Equations (4) and (5))

Investor Preference	Variable	Estimate
α_{it}^{f}	SMA^{α^f}	-0.009 (0.053)
	StocksReturn^f	0.508 (0.535)
β_{it}^{f}	SMA^{β^f}	-1.143*** (0.053)
	StocksReturn^f	1.128* (0.747)
β_{it}^{p}	SMA^{β^p}	-0.069 (0.088)
	StocksReturn^p	0.891 (0.567)

6.2 Time-varying Nature of Investor Preferences for Content

Our paper's focus is to identify the strategic behaviors of SMAs in the sense that SMAs may cater their content generation and monetization decisions to investors' changing preferences. An important premise here is that investors have time-varying preferences for financial content. To ascertain that this is the case, we present some demand-side results that can demonstrate the time-varying nature of investor preferences.

Recall that we assume the investor preferences for content ($\alpha_{it}^{\bullet}$ and $\beta_{it}^{\bullet}$) are SMA- and time-specific. However, given that there are more than 500 SMAs, it is impossible to plot the evolution pattern of investor preferences for each SMA. We hereby plot the time-specific effects of each investor preference (i.e., $\tau_t^{\bullet}$s) in Figure A2. From these plots, we can discern that investor preferences all change considerably over time.

In addition, we show the carryover effects in investor preferences ($\delta_i^{\alpha^*}$ and $\delta_i^{\beta^*}$ in equations (4) and (5)) in Table A10. If these carryover effects are not equal to one, the investor preferences at time t can be systematically different from their previous values. As we can see, at most 51.3% of investor preferences from the last period would carry over to the next period, suggesting that investor preferences across consecutive periods are rather different. Overall, the findings below confirm that investor preferences for SMAs' content indeed vary greatly across time.

Further, we find that the carryover effect of free readers' preference for free sentiment (51.3%) is larger than that of paid subscribers' preference for free sentiment (12.7%) in magnitude. These results suggest that free readers' preferences are more persistent than paid subscribers' in SMAs' free content sentiment decision. In addition, the correlation between readers' and subscribers' preferences in free content sentiment is 0.289. Therefore, free readers' and paid subscribers' preferences are substantively different.

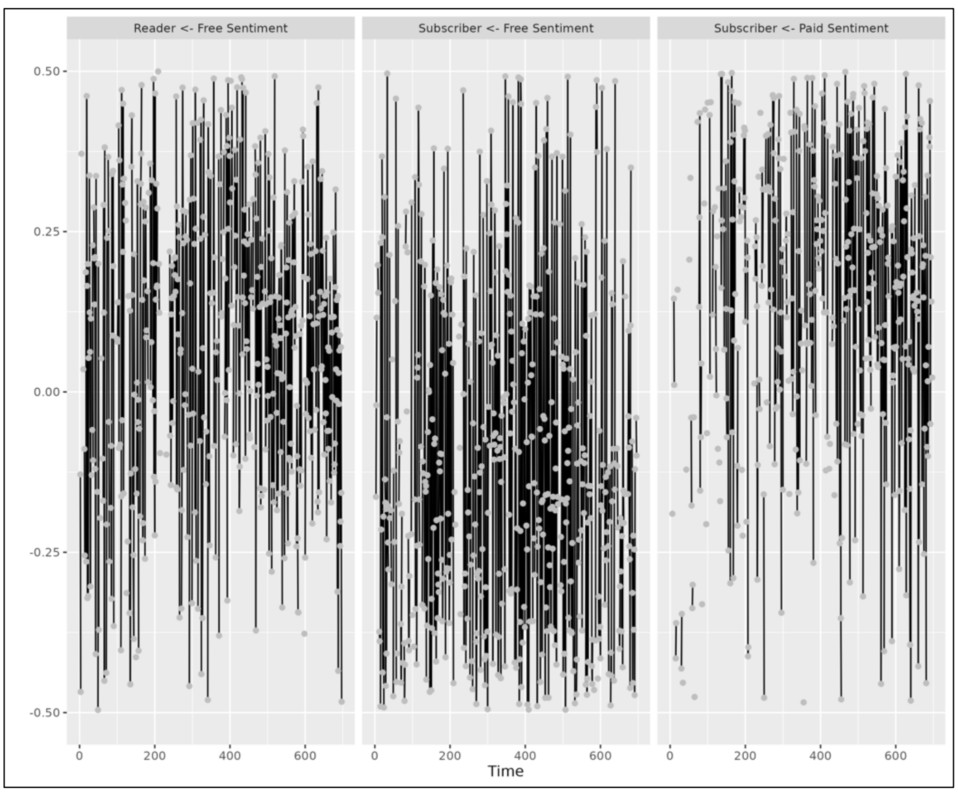


Figure A2 Population-level Time-specific Trends of Investor Preferences (i.e., $\tau_t^{\bullet}$s)

Table A10 Carryover Effects of Investor Preferences

Parameter	Carryover Effects
$\delta_i^{\alpha^f}$	0.513*** (0.020)
$\delta_i^{\beta^f}$	0.127*** (0.016)
$\delta_i^{\beta^p}$	0.320*** (0.043)
1. Standard deviation in parenthesis. 2. Estimates significant at 90%, 95%, 99% are marked by *, **, *** respectively.	

7 SMA Heterogeneity

The random coefficients framework we have employed permits us to further recover the individual SMA-level parameters. We plot the individual-level parameter estimates in Figure A3. We find that strategic SMAs are considerable in numbers. If we differentiate various aspects of strategic behaviors in content generation and monetization for free readers and paid subscribers, we find that around 100% of SMAs rely on the strategic use of expected readership for content monetization and, similarly, 100% of SMAs do so with expected subscriptions for content monetization. About 25% (100%) of SMAs make strategic use of free readers' (paid subscribers') preferences for free content sentiment in the content generation decision and likewise, 77% of SMAs do so with paid subscribers' preferences for paid content sentiment in the content generation decision.

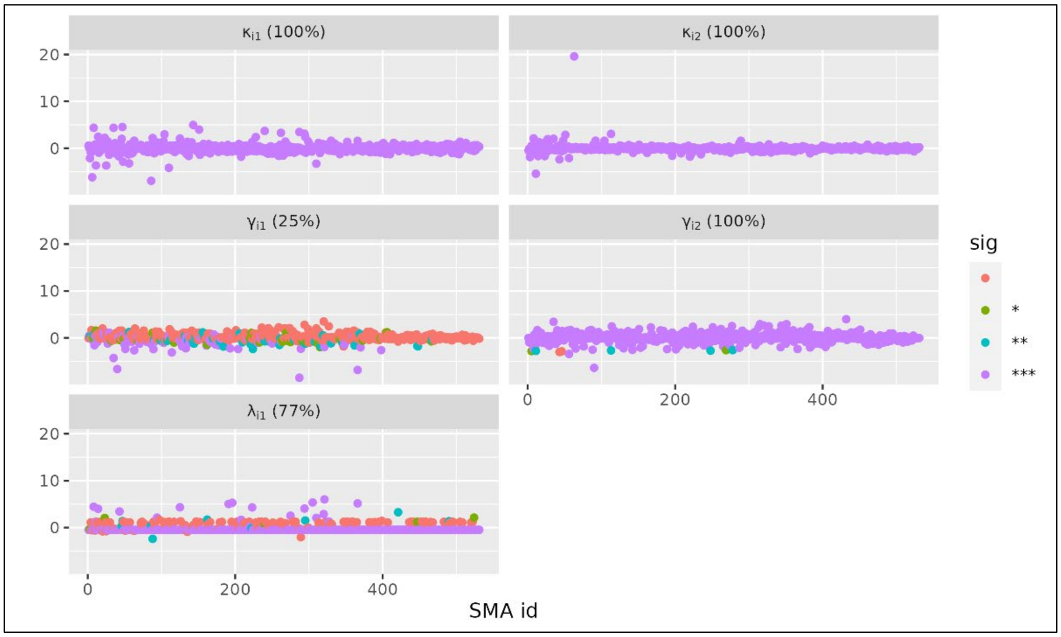


Figure A3 Individual SMA's Heterogenous Strategic Behaviors[2]

8 SMAs' Readership and Subscription Optimization Strategies

In this section, we provide some analyses on SMAs' readership and subscription optimization strategies.

In each period, we simulate the readership and subscriptions under different content monetization and generation decisions based on the estimated model coefficients. We then identify the content monetization and generation decisions that will achieve the maximal readership in each period. We term these decisions as the readership-maximizing strategies. Similarly, we can obtain the subscription-maximizing strategies. After that, we compare the observed content monetization and generation decisions of SMAs to those based on the optimization strategies identified above.

We find that for 97.57% of observations, SMAs' content generation and monetization decisions are neither consistent with the readership-optimizing strategies nor the subscription-maximizing strategies. This result, combined with our main findings, suggests that for the majority of the time, SMAs' content generation and

[2] SMAs whose individual parameters are significant are regarded as "strategic SMAs". The percentage of strategic SMAs are shown in parenthesis in Figure A3. We note that the values in the plots are scaled (significance levels are not affected) for the purpose of presentation.

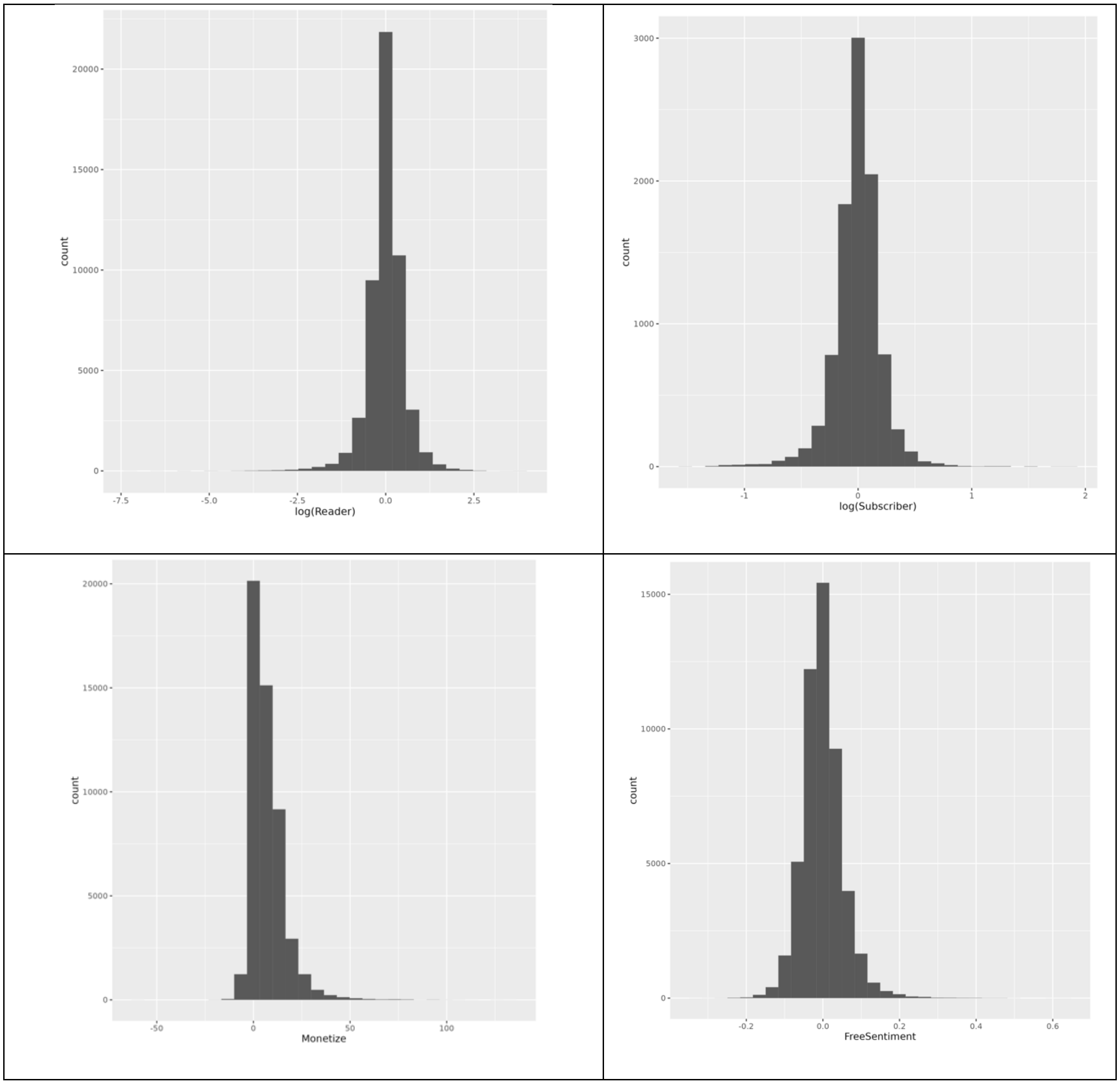
monetization decisions are not performance-optimizing in nature, but are mainly strategic responses to the changes in the demand side (e.g., investor preferences).

Among the 531 SMAs, we find that about 251 SMAs exercise the readership-maximizing strategy at least once, and 27 SMAs exercise the subscription-maximizing strategy at least once (out of which 22 exercise both the readership-maximizing and subscription-maximizing strategy at least once). We find that the median timing for exercising the readership-maximizing strategy is at the 40th percentile (the mean is at the 34th percentile) of a SMA's tenure at the platform. Correspondingly, the median timing for exercising the subscription-maximizing strategy is at the 57th percentile (the mean is at the 55th percentile). The timing difference between the readership-maximizing and subscription-maximizing strategies shows that it is common that SMAs first maximize or build up the free readership, and then adopt monetization actions later on the acquired investor or readership base.

9 Normality Assumption of Error Terms

In our econometric model, we assume that the error terms ($\epsilon_{it}^{\bullet} \sim MVN(0,V)$, $\pi_{it}^{\bullet} \sim MVN(0,\Omega)$) follow the normal distribution. We note that the normality assumption of error terms is widely used in Bayesian models for its tractability (e.g., enabling conjugacy and Gibbs sampling) and ease of statistical sampling and simulation in Markov Chain Monte Carlo-based methods (Rossi and Allenby 2003). Nevertheless, we here present some evidence that the error terms are indeed normally distributed, and thus verify that our assumption is reasonable.

We draw the histogram plots of the residuals of each equation and show the histogram plots in Figure A4. We find that the model residuals are all approximately normally distributed and bell-shaped. Therefore, the assumption of the normal distribution is reasonable.


count
log(Reader)
count
log(Subscriber)
count
Monetize
count
FreeSentiment

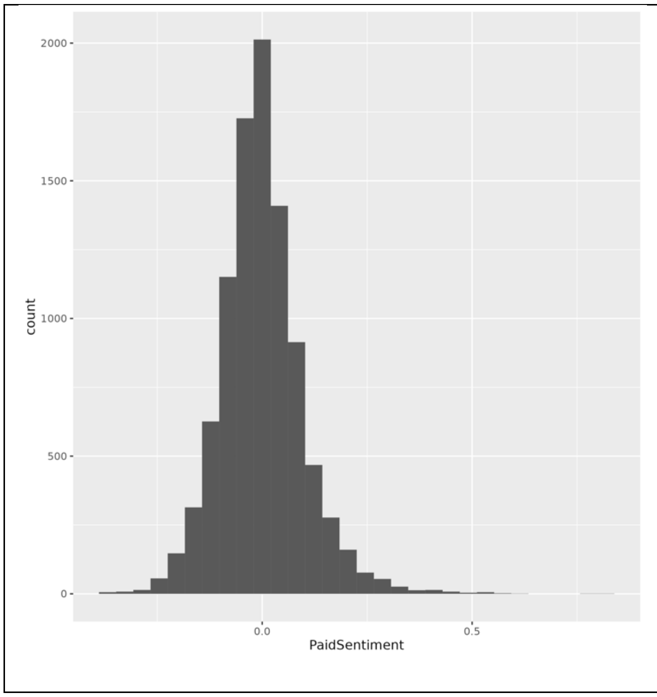


Figure A4 Histograms of Residuals

Reference

Antweiler W, Frank MZ. (2004) Is All That Talk Just Noise? The Information Content of Internet Stock Message Boards. *The Journal of Finance* 59(3):1259-1294.

Aral S, Dhillon P. (2020) Digital Paywall Design: Implications for Content Demand & Subscriptions. *Management Science*.

Baek J, Shore J. (2020) Forum Size and Content Contribution per Person: A Field Experiment. *Management Science* 66(12).

Bagnoli M, Beneish MD, Watts SG. (1999) Whisper forecasts of quarterly earnings per share. *J. Account. Econ.* 28(1):27-50.

Becker JU, Clement M, Schaedel U. (2010) The impact of network size and financial incentives on adoption and participation in new online communities. *Journal of Media Economics* 23(3):165-179.

Bhattacharya P, Phan TQ, Bai X, Airoldi EM. (2019) A Coevolution Model of Network Structure and User Behavior: The Case of Content Generation in Online Social Networks. *Inform Syst Res* 30(1):117-132.

Bronnenberg BJ, Mahajan V. (2001) Unobserved retailer behavior in multimarket data: Joint spatial dependence in market shares and promotion variables. *Marketing Science* 20(3):284-299.

Cavusoglu H, Phan TQ, Cavusoglu H, Airoldi EM. (2016) Assessing the Impact of Granular Privacy Controls on Content Sharing and Disclosure on Facebook. *Inform Syst Res* 27(4):848-879.

Chen H, De P, Hu YJ, Hwang B-H. (2014) Wisdom of crowds: The value of stock opinions transmitted through social media. *Review of Financial Studies* 27(5):1367-1403.

Chen HL, Hu YJ, Huang S. (2019) Monetary Incentive and Stock Opinions on Social Media. *Journal of Management Information Systems* 36(2):391-417.

Clarkson PM, Joyce D, Tutticci I. (2006) Market reaction to takeover rumour in Internet Discussion Sites. *Accounting and Finance* 46(1):31-52.

Das SR, Chen MY. (2007) Yahoo! for Amazon: Sentiment extraction from small talk on the web. *Management Science* 53(9):1375-1388.

Goes PB, Lin M, Au Yeung C-m. (2014) "Popularity Effect" in User-Generated Content: Evidence from Online Product Reviews. *Inform Syst Res* 25(2):222-238.
Guo C, Kim TH, Susarla A, Sambamurthy V. (2020) Understanding Content Contribution Behavior in a Geosegmented Mobile Virtual Community: The Context of Waze. *Inform Syst Res* 31(4):1398-1420.
Heimbach I, Hinz O. (2018) The Impact of Sharing Mechanism Design on Content Sharing in Online Social Networks. *Inform Syst Res* 29(3):592-611.
Hsieh G, Kraut RE, Hudson SE. (2010) *Why pay?: Exploring how financial incentives are used for question & answer*. ACM.
Hwang EH, Singh PV, Argote L. (2015) Knowledge Sharing in Online Communities: Learning to Cross Geographic and Hierarchical Boundaries. *Organ Sci* 26(6):1593-1611.
Kadous K, Mercer M, Zhou YD. (2019) *Do Individual Investors Understand How Social Media Advice Influnces Their Investment Decision?* , SSRN.
Lambrecht A, Misra K. (2017) Fee or Free: When Should Firms Charge for Online Content? *Management Science* 63(4):1150-1165.
Li S, Sun B, Montgomery AL. (2011) Cross-selling the right product to the right customer at the right time. *Journal of Marketing Research* 48(4):683-700.
Lindsten F, Schön TB. (2013) Backward simulation methods for Monte Carlo statistical inference. *Foundations and Trends® in Machine Learning* 6(1):1-143.
Liu TX, Yang J, Adamic LA, Chen Y. (2014) Crowdsourcing with All-Pay Auctions: A Field Experiment on Taskcn. *Management Science* 60(8):2020-2037.
Liu Y, Feng J. (2015) *Can Monetary Incentives Increase UGC Contribution? The Motivation and Competition Crowding Out*. Fort Worth.
Lu Y, Singh PV, Sun B. (2017) IS A CORE-PERIPHERY NETWORK GOOD FOR KNOWLEDGE SHARING? A STRUCTURAL MODEL OF ENDOGENOUS NETWORK FORMATION ON A CROWDSOURCED CUSTOMER SUPPORT FORUM. *Mis Quart* 41(2):607-+.
Luo A, Kumar V. (2013) Recovering hidden buyer–seller relationship states to measure the return on marketing investment in business-to-business markets. *Journal of Marketing Research* 50(1):143-160.
Luo X, Zhang J, Duan W. (2013) Social media and firm equity value. *Inform Syst Res* 24(1):146-163.
Manchanda P, Rossi PE, Chintagunta PK. (2004) Response modeling with nonrandom marketing-mix variables. *Journal of Marketing Research* 41(4):467-478.
Mason W, Watts DJ. (2010) Financial incentives and the performance of crowds. *ACM SigKDD Explorations Newsletter* 11(2):100-108.
Musalem A, Bradlow ET, Raju JS. (2008) Who's got the coupon? Estimating consumer preferences and coupon usage from aggregate information. *Journal of Marketing Research* 45(6):715-730.
Nair HS, Misra S, Hornbuckle WJ, Mishra R, Acharya A. (2017) Big Data and Marketing Analytics in Gaming: Combining Empirical Models and Field Experimentation. *Marketing Science* 36(5):699-725.
Nguyen P, Wang X, Li X, Cotte J. (2021) Reviewing Experts' Restraint from Extremes and Its Impact on Service Providers. *J. Consum. Res.* 47(5):654-674.
Oh H, Animesh A, Pinsonneault A. (2016) Free Versus For-A-Fee: The Impact of a Paywall on the Pattern and Effectiveness of Word-of-Mouth via Social Media. *Mis Quart* 40(1):31.
Pattabhiramaiah A, Sriram S, Manchanda P. (2019) Paywalls: Monetizing online content. *Journal of Marketing* 83(2):19-36.
Pauwels K, Weiss A. (2008) Moving from Free to Fee: How Online Firms Market to Change Their Business Model Successfully. *Journal of Marketing* 72(3):14-31.
Qiu L, Kumar S. (2017) Understanding Voluntary Knowledge Provision and Content Contribution Through a Social-Media-Based Prediction Market: A Field Experiment. *Inform Syst Res* 28(3):529-546.
Sabherwal S, Sarkar SK, Zhang Y. (2008) Online talk: does it matter? *Managerial Finance* 34(6):423-436.
Schweidel DA, Knox G. (2013) Incorporating direct marketing activity into latent attrition models. *Marketing Science* 32(3):471-487.
Shen W, Hu YJ, Rees J. (2015) Competing for Attention: An Empirical Study of Online Reviewers' Strategic Behavior. *Management Information Systems Quarterly* 39(3):683-696.

Stephen A, Bart Y, Du Plessis C, Goncalves D. (2012) Does Paying For Online Product Reviews Pay Off? the Effects of Monetary Incentives on Content Creators and Consumers. *NA-Advances in Consumer Research* 40.

Tang Q, Gu B, Whinston AB. (2012) Content Contribution for Revenue Sharing and Reputation in Social Media: A Dynamic Structural Model. *Journal of Management Information Systems* 29(2):41-75.

Van Diepen M, Donkers B, Franses PH. (2009) Dynamic and competitive effects of direct mailings: A charitable giving application. *Journal of Marketing Research* 46(1):120-133.

Wang J, Ghose A, Ipeirotis P. (2012) *Bonus, disclosure, and choice: What motivates the creation of high-quality paid reviews?* , Orlando.

Wei Z, Xiao M, Rong R. (2020) Network Size and Content Generation on Social Media Platforms. *Prod. Oper. Manag.*

Zhao L, Detlor B, Connelly CE. (2016) Sharing Knowledge in Social Q&A Sites: The Unintended Consequences of Extrinsic Motivation. *Journal of Management Information Systems* 33(1):70-100.