Who sells knowledge online? An exploratory study of knowledge celebrities in China

Chen, Xiaoyu; Chua, Alton Yeow Kuan; Pee, Loo Geok

2021

Chen, X., Chua, A. Y. K. & Pee, L. G. (2021). Who sells knowledge online? An exploratory study of knowledge celebrities in China. Internet Research.
<https://dx.doi.org/10.1108/INTR-07-2020-0378>

<https://hdl.handle.net/10356/152621>

<https://doi.org/10.1108/INTR-07-2020-0378>

© 2021 Emerald Publishing Limited. All rights reserved. This paper was published in Internet Research and is made available with permission of Emerald Publishing Limited.

Downloaded on 26 Apr 2023 17:29:42 SGT

Who sells knowledge online? An exploratory study of knowledge celebrities in China

Xiaoyu Chen¹, Alton Y. K. Chua², and L. G. Pee²

1. Department of Library, Information and Archives, Shanghai University, China
2. Wee Kim Wee School of Communication and Information, Nanyang Technological University, Singapore

Acknowledgments:

The authors appreciated the comments and suggestions from the editors and review team. An earlier version of the manuscript entitled “*Who sells knowledge to you? Understanding online identities of knowledge influencers on a pay-for-knowledge platform*” was published at Proceedings of 13th China Summer Workshop on Information Management (CSWIM 2019). This research was also benefited from oral presentations at 5th CoHASS Graduate Student Conference at Nanyang Technological University, 17th Chinese Internet Research Conference at National University of Singapore (CIRC 2019), and iConference 2021 Doctoral Colloquium at Renmin University of China.

Corresponding author: Xiaoyu Chen (ORCID: 0000-0003-1741-0674)

Email: xiaoyu001@e.ntu.edu.sg

Who sells knowledge online? An exploratory study of knowledge celebrities in China

Abstract:

Purpose—This study explores identity signaling used by an emerging class of knowledge celebrities in China—“Knowledge Wanghong”—who sell knowledge products on online platforms. Because identity signaling may involve constructing unique online identities and controlling over product-related and seller-related characteristics, the purpose of this study is two-fold: (1) to uncover different online identities of knowledge celebrities; and (2) to examine the extent to which the online identity type is associated with their product-related characteristics, seller-related characteristics and sales performance.

Design/methodology/approach—A unique data set was collected from a Chinese leading pay-for-knowledge platform—Zhihu—which featured the online profiles of tens of thousands of knowledge celebrities. Online identity types were derived from their self-edited content using Latent Dirichlet Allocation (LDA) topic modeling. Thereafter, their product-related characteristics, seller-related characteristics and respective sales performance were analyzed across different identity types using analysis of variance (ANOVA) and multiple-group linear regression.

Findings—Knowledge celebrities are clustered into four distinctive online identities: *Mentor*, *Broker*, *Storyteller*, and *Geek*. Product-related characteristics, seller-related characteristics and sales performance varied across four different identities. Additionally, the online identity type moderated the relationships among their product-related characteristics, seller-related characteristics and sales performance.

Originality/value—As emerging-phenomenon-based research, this study extends related literature by using the notion of identity signaling to analyze a peculiar group of online celebrities who are setting an important trend in the pay-for-knowledge model in China.

Keywords: Digital influencer, Identity signaling, Knowledge celebrity, Knowledge products, Pay-for-knowledge, Sales performance

Paper type: Research paper

Who sells knowledge online? An exploratory study of knowledge celebrities in China

1. Introduction

Online celebrities are “digital influencers” who can attract and maintain a sizable following on the Internet by regularly posting highly engaging and personalized content (Abidin, 2018). Capitalizing on their popularity and expertise, online celebrities are able to shape consumers’ purchasing decision-making via electronic word-of-mouth (Chen and Chua, 2020; Zafar *et al.*, 2020). It is no wonder that many of them have been hired by businesses to review or even subtly promote products and services (Cheah *et al.*, 2019; Djafarova and Rushworth, 2017).

In China, there is an emerging class of online celebrities known as “Knowledge Wanghong” who focuses specifically on knowledge-intensive content (Chen and Fu, 2019). Being knowledgeable online celebrities, they can evoke a strong emotional attachment from their legions of fans (Tao, 2017). However, unlike conventional online celebrities involved in endorsements (Hua *et al.*, 2017), such knowledge celebrities create and sell their own knowledge products directly to fans. Fans often view the outputs from their idols as symbols of affection and thus derive psychological pleasure from the purchase (Duffy and Pooley, 2019).

The knowledge products sold by knowledge celebrities are usually packaged in textual, audio, or video formats and delivered on pay-for-knowledge platforms (iResearch, 2018). Themes of the knowledge products for sale span the gamut from academic and current affairs topics to online game commentaries and practical fixes, reflecting the match between what consumers need and what knowledge celebrities can offer (Tao, 2017). To thrive in the intensely competitive online environment, knowledge celebrities have to carve their own niche markets and differentiate themselves using what is known as *identity signaling*. This is a deliberate communication process by which they may provide cues that convey information

about who they are, what they can do and the products they sell (Schau and Gilly, 2003; Xu *et al.*, 2018). On pay-for-knowledge platforms, the opportunity for identity signaling presents itself most naturally via online profiles (Zhao *et al.*, 2018).

In order to come across as authentic, online celebrities have to set up their online profiles clearly and consistently (Duffy and Pooley, 2019). This means that their profiles must include distinctive details to conjure the identity they wish to portray (Lou *et al.*, 2019; Ren *et al.*, 2020). For example, some online celebrities might choose to create a personable and likable character, while others prefer to be seen as efficient and insightful (Djafarova and Rushworth, 2017). Different identities appeal to different segments of consumers (Kim, 2018). In any case, the goal is to project a reputable image, inspire confidence and ultimately nudge a purchase (Zafar *et al.*, 2020). However, there are currently two gaps in the literature.

First, works on the identity signaling of online celebrities are mostly dominated by qualitative research (e.g., Abidin, 2018; Djafarova and Rushworth, 2017; Duffy and Pooley, 2019). In particular, manual coding has been used to analyze self-authored texts on profile pages (Trammell and Keshelashvili, 2005). Yet, as the number of online celebrities continues to grow quickly, the size of textual datasets has now far exceeded human information-processing capacities (Geva *et al.*, 2019). Moreover, despite methodological guidelines on improving the validity and reliability when analyzing qualitative data with multiple coders, the biases arising from subjective interpretations of the data cannot be completely mitigated (Bao and Datta, 2014). Given this limitation, a topic modeling technique has been recommended (Blei *et al.*, 2003). Not only is this approach well suited to handle a massive amount of data, but it uses statistical associations of words in a text to generate a list of latent topics—clusters of co-occurring keywords that jointly represent a higher-order concept—thereby deepening the understanding of the textual meanings (Geva *et al.*, 2019).

Second, to what extent identity signaling of online celebrities is related to their sales performance still remains unclear. Past research on online marketplaces suggests that sales performance could be affected by product-related characteristics and seller-related characteristics (e.g., Cheung *et al.*, 2014; Guo *et al.*, 2017; Li *et al.*, 2015; Liu *et al.*, 2017; Mavlanova *et al.*, 2012). These characteristics may both reflect and are partially reinforced by the sellers' identity signaling (Verbeke *et al.*, 2011). Additionally, prior literature tends to treat reputable sellers as *homogenous* due to their similarly stellar sales performance (Lee and Li, 2018). However, in reality, they may deliberately adopt different public personae to capitalize on their unique value proposition (Li *et al.*, 2016). Research on a crowdfunding marketplace shows that even among successful project creators, their performance varies across personal identity types (Lin and Boh, 2020; Ryu and Kim, 2018). Nonetheless, there is no empirical investigation conducted in the context of knowledge celebrities.

For the reasons above, the purpose of this study is two-fold. First, it aims to uncover the different online identities of knowledge celebrities. The intent is to enhance the understanding of how knowledge celebrities use identity signaling to differentiate among themselves. To this end, we use a computational approach—Latent Dirichlet Allocation (LDA) topic modeling technique—to identify frequently-occurring topics from their self-edited content in the profiles, which could cluster knowledge celebrities into different identities. Second, this study attempts to examine the extent to which the online identity type [1] is associated with product-related characteristics, seller-related characteristics and sales performance. To do so, we use analysis of variance (ANOVA) to compare the three constructs across different identities. Furthermore, multiple-group linear regression is employed to validate the relationship among the constructs.

Zhihu (www.zhihu.com) was selected as the data source as it is one of China's earliest and leading pay-for-knowledge platforms (Kuang *et al.*, 2019; Wang *et al.*, 2021). Featuring

the online profiles of tens of thousands of knowledge celebrities, it contained the metadata of product-related and seller-related characteristics as well as respective sales performance, which were publicly available (Zhang *et al.*, 2019c; Zhihu, 2021).

This study contributes to research and practice. As emerging-phenomenon-based research, it extends the existing literature by using the notion of identity signaling to analyze a peculiar group of knowledge celebrities—Knowledge Wanghong—who are setting an important trend in the pay-for-knowledge model in China (State Information Center, 2018). It also helps inform practices on pay-for-knowledge platforms. For instance, an aspiring knowledge celebrity aware of the different types of online identities can better position himself or herself more consciously in the marketplace.

The rest of the article proceeds as follows. Section 2 presents related literature on knowledge celebrities, identity signaling used by knowledge celebrities, and relationships among online identity types, product-related characteristics, seller-related characteristics, and sales performance. Section 3 details the data collection and methods of analysis. Next, Section 4 provides the results of the data analyses. Section 5 discussed the findings, the implications for research and practice and offers suggestions for future work. Finally, Section 6 concludes with a summary.

2. Literature Review

2.1 Overview of knowledge celebrities

Online celebrities in China, also called “Wanghong”, describe anyone who manages to cultivate fame on social media platforms (Abidin, 2018). Due to their tremendous popularity, an increasing number of digital businesses have tapped into them for marketing and advertising purposes (Hua *et al.*, 2017). The “Wanghong economy” estimated market value is over US\$ 8.4 billion (State Information Center, 2018). A representative example is *Papi*

1
2
3 *Jiang*, who received around US\$ 1.8 million from venture capitalists to endorse products and
4 services through posting original video content (Tao, 2017). Notably, these online celebrities
5 typically do not receive payments directly from consumers. Sponsorships from marketing and
6 advertising investors as well as online platform operators motivate them to constantly share
7 original content packed with sensational elements (Abidin, 2018).
8
9
10
11
12
13
14

15 The case of Knowledge Wanghong opens the door for generating a new revenue
16 stream from the Internet. Instead of being the front for businesses, these knowledge
17 celebrities create knowledge products and sell them to anyone willing to pay on online
18 platforms like Zhihu (Wang *et al.*, 2021). For each successful transaction, these platforms
19 take a cut from the payment made (Kuang *et al.*, 2019). As the primary creators of knowledge
20 products, knowledge celebrities fuel this business model partly because of their aura in
21 playing a dual role—online sellers and online celebrities. As online sellers, they effectively
22 skip the intermediaries and sell directly to consumers by displaying compelling information
23 on their profiles (Chen and Fu, 2019; Wang *et al.*, 2021). In addition, sharing details about
24 the knowledge products for sale will also tilt consumers' purchase intention in their favor
25 (Mavlanova *et al.*, 2012).
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40

41 Knowledge celebrities are also adored as a peculiar group of online celebrities. As
42 with other micro-celebrities on typical platforms such as YouTube and Instagram (Lou *et al.*,
43 2019; Zafar *et al.*, 2020), they tend to establish appealing and stable images so as to cultivate
44 familiarity and build loyalty among followers (Tao, 2017). They could also further leverage
45 their influence to extract economic benefits from the followers (iResearch, 2018). For
46 example, many followers are willing to donate money to encourage their favorite micro-
47 celebrities to create online content continuously (Wan *et al.*, 2017). However, knowledge
48 celebrities are typically recognized as “semi-professionals” in related domains as their
49
50
51
52
53
54
55
56
57
58
59
60

contributed knowledge are deemed as high-quality products (Tao, 2017). Consequently, consumers may show a strong willingness to pay for knowledge products (Zhao *et al.*, 2020).

There is an emerging body of literature on users' payment for knowledge products (e.g., Cai *et al.*, 2020; Liu *et al.*, 2021; Shi *et al.*, 2020; Zhang *et al.*, 2019a; Zhang *et al.*, 2019c; Zhao *et al.*, 2020). Nevertheless, investigating consumers' willingness to pay for knowledge products is only a part of the pay-for-knowledge model in China. We find a lack of in-depth analysis of the sellers of the knowledge products—knowledge celebrities. After all, if no one like them sold knowledge products, it would be impossible to draw many consumers to pay. This study is partly motivated by the limited understanding of their online identities.

2.2 Identity signaling used by knowledge celebrities

The theoretical foundation of this study is based on identity signaling (Spence, 1973; Xu *et al.*, 2018). It is found to effectively transfer signalers' self-information in marketplaces because it is visible and is designed to communicate (Spence, 2002). In online marketplaces, identity signaling is important for both consumers and sellers (Verbeke *et al.*, 2011; Xu and Liu, 2019). In particular, sellers may use it to highlight who they are and what products they can offer (Pee *et al.*, 2018). Previous research has suggested that sellers' identity helps them develop online relationships with consumers and further improve their sales performance in marketplaces (e.g., Kozlenkova *et al.*, 2017).

For online celebrities, the use of identity signaling is related to creating one's self-branding and highlighting unique selling points (Duffy and Pooley, 2019; Khamis *et al.*, 2017). Self-branding involves someone consciously developing a stable and authentic public persona for commercial gains (Geva *et al.*, 2019). Online celebrities may brand themselves by creating identities that are responsive to target consumers' needs and interests. A

convenient way to do so is to edit one's online profile (Cheah *et al.*, 2019; Ren *et al.*, 2020). Such self-branding is articulated through a subtle blend of self-disclosure and self-promotion (Duffy and Pooley, 2019). Online celebrities could, for example, describe who they are and what expertise they have (Khamis *et al.*, 2017).

For knowledge celebrities, their online identities are created via the personal and professional fields on the profiles provided by most pay-for-knowledge platforms (Jan *et al.*, 2018; Jin *et al.*, 2019). The personal field usually contains details such as personal traits, life experience, and hobbies, while the professional field includes descriptions about education background, working experience, and expertise domains (Zhao *et al.*, 2018; Trammell and Keshelashvili, 2005). To reduce fraudulent claims, almost all pay-for-knowledge platforms have a proprietary verification system to check the authenticity of the self-authored content (Chen and Fu, 2019; Jin *et al.*, 2019).

It is noteworthy that knowledge celebrities' self-constructed identities must necessarily be congruent with the knowledge products they sell. For instance, knowledge celebrities who have established a persona of an expert in a given subject matter are expected to sell products with a high level of professional content related only to that area. Otherwise, if they try to claim mastery over an assortment of disparate disciplines, their trustworthiness will take a hit (Zhao *et al.*, 2018), and consumers will likely be put off (Jan *et al.*, 2018).

2.3 Online identity types, product-related characteristics, seller-related characteristics, and sales performance

In online marketplaces, creators of information goods such as books, ideas, and inventions have been able to sell their works directly to consumers (Lee and Li, 2018). According to identity signaling theory (Spence, 2002; Xu *et al.*, 2018), it is all the more critical for such

sellers to establish a unique online identity to draw consumers and help reduce purchase information asymmetry.

As the demand for transparency and accountability increases, many platforms gradually encourage reputable sellers to qualitatively present themselves using textual content in profiles (Xu *et al.*, 2018). The purpose is further to boost sellers' credibility and reliability and make them more responsible (Luo, 2014). The heterogeneity of sellers' self-disclosure determines their different types (Li *et al.*, 2016). The distinction among different identity types may be rooted in a strategic trade-off that sellers have to face: how to better position themselves to survive in the marketplace (Teodoridis *et al.*, 2019). This decision matters given the competitive nature of knowledge markets. Successful knowledge sellers are unambiguous in advertising their unique expertise and are able to build a large clientele (Luo, 2014).

Sellers' performance may vary across different online identities. For instance, in the context of online crowdfunding, researchers found that crowdfunding projects' performance differed by the creator identity types and suggested platform operators attracting and promoting different project creators (Lin and Boh, 2020; Ryu and Kim, 2018). Although prior research suggests that sellers' identity signaling may predict their sales performance (Verbeke *et al.*, 2011), how different online identities are related to sales performance has not been empirically studied. Moreover, literature on online marketplaces (e.g., Cheung *et al.*, 2014; Guo *et al.*, 2017; Li *et al.*, 2015; Mavlanova *et al.*, 2012; Ryu and Kim, 2018) suggests that prior to transactions, sellers exercise deliberate control over how they present quantitative information on their products (i.e., product-related characteristics) and about themselves (i.e., seller-related characteristics).

Drawing on related literature on users' payment for knowledge products, this study identified four variables of product-related characteristics: (1) diversity refers to the extent to which knowledge products span across different domains (Teodoridis *et al.*, 2019); (2) update frequency refers to the number of times new knowledge products are created within a defined temporal continuum (Jan *et al.*, 2018); (3) prices of knowledge products refer to what a customer is expected to pay (Zhang *et al.*, 2019c), while (4) trustworthiness refers to the degree of reliability of knowledge products (Zhao *et al.*, 2018). Additionally, we identified three variables of seller-related characteristics: (1) information self-disclosure refers to the extent to which knowledge celebrities disclose self-information (Jin *et al.*, 2019); (2) tenure refers to the length of experience as sellers (Cai *et al.*, 2020); (3) identity verification denotes the number of verifications on a knowledge celebrity's identity (Zhihu, 2021).

For sales performance, we focused on two common variables on pay-for-knowledge platforms: sales volume and customer satisfaction (Jin *et al.*, 2019; Wang *et al.*, 2021; Zhao *et al.*, 2018). Sales volume refers to the total number of knowledge products sold by a particular knowledge celebrity. A high transaction volume reflects his or her popularity in knowledge provision (Jan *et al.*, 2018). Customer satisfaction can be judged by consumers' after-sales feedback. Overall positive feedback gives assurance that the seller's product is of good quality (Jin *et al.*, 2019). Past research suggests that product-related characteristics and seller-related characteristics are associated with sales performance (Li *et al.*, 2015; Pee *et al.*, 2018). Thus, the following hypothesis is proposed.

H1: Knowledge celebrities' product-related characteristics and seller-related characteristics are associated with their sales performance

Past research suggests that product-related and seller-related characteristics are also alterable attributes unique to a seller's online identity (Spence, 2002). In an online

crowdfunding context, for example, some identities performed well in terms of the speed of offering products and serving consumers, while others did well in establishing the credibility of their products and developing positive social interactions with buyers (Ryu and Kim, 2018). This implies that the online identity type of knowledge celebrities could moderate the relationships among product-related characteristics, seller-related characteristics, and sales performance. Therefore, the following two hypotheses are proposed. Figure I shows the research model with the three hypotheses.

H2a: Online identity type of knowledge celebrities moderates the relationship between product-related characteristics and their sales performance

H2b: Online identity type of knowledge celebrities moderates the relationship between seller-related characteristics and their sales performance

[Insert Figure 1 here]

3. Methods

3.1 Data source and collection

We chose Zhihu as our research data source for two reasons. First, this platform is currently one of the leading pay-for-knowledge platforms in China. Launched in 2011, it quickly became one of the most frequently visited websites by Chinese users (Chen *et al.*, 2018).

Users can ask and answer questions, write articles, and comment and vote on the answers and articles (Jin *et al.*, 2019; Zhang *et al.*, 2019b). Since 2016, the platform has incorporated an online payment system that enables knowledge creators to monetize the knowledge they shared (Zhao *et al.*, 2018; Zhihu, 2021).

Second, Zhihu had a feature (www.zhihu.com/zhi) [2] that allowed consumers to discover well-recognized knowledge celebrities efficiently through its unique proprietary algorithm (Zhao *et al.*, 2018; Zhihu, 2021). Algorithms to recommend influential content

creators can similarly be found on other websites such as Amazon.com (Yazdani *et al.*, 2018). Researchers are thus able to use the feature to obtain the profiles of all knowledge celebrities from Zhihu.

Moreover, Zhihu provides knowledge celebrities with homepages, enabling each of them to edit the content on a profile, share particulars (e.g., gender, place of residence, and affiliation), and display the metadata of knowledge products being sold (e.g., total volume of transactions and customer feedback) and other related characteristics (e.g., product-related and seller-related characteristics). A screenshot of the profile of a knowledge celebrity is shown in Figure 2.

The collection of the data consisted of two steps. First, using a web crawler developed by a third-party platform (www.bazhuayu.com), we crawled all information of these knowledge celebrities on Zhihu, including URLs of their homepages, content on their profiles, shared particulars, and metadata of knowledge products sold. Such information was stored in the MySQL database. The crawling process took place between January 7 and January 14, 2018. An initial dataset of 2,625 knowledge celebrities was collected.

In the second step, we manually checked for duplicated entries and records with missing fields. Eight problematic cases in the dataset were removed. From the final 2,617 records, we extracted fields related to two aspects of knowledge celebrities, namely, the textual content on their profiles and metadata of product-related and seller-related characteristics and sales performance for further analysis.

[Insert Figure 2 here]

3.2 Document preparation for deriving online identities

To derive the online identities of knowledge celebrities, we analyze the textual content on their profiles comprising personal and professional fields. The personal field allows for

1
2
3 sharing private aspects such as personalities and interests, while the professional field
4
5 displays credentials and other achievements knowledge celebrities wish to highlight. In
6
7 preparation for the analysis, each knowledge celebrity's personal and professional fields were
8
9 combined and treated as one document.
10
11

12
13 The documents were pre-processed for topic modeling analysis as follows. Initially,
14
15 we segmented sentences into words and phrases using a Chinese word segmentation module,
16
17 'JiebaR' (see github.com/qinwf/jiebaR) as all documents were mostly written in Chinese.
18
19 After that, stop words such as pronouns, intonations, prepositions, punctuations, and cohesion
20
21 words were removed. Considering that occurrences of extremely rare and extremely common
22
23 words from the documents would confound the results, we deleted the most frequent words
24
25 (appeared more than 650 times which were found in $> 25\%$ of all documents) and the least
26
27 frequent words (appeared less than 3 times and which were found in $< 0.1\%$ of all
28
29 documents), as suggested by prior work (Lou *et al.*, 2019).
30
31
32
33

34 **3.3 Operationalization of variables**

35
36 We operationalized variables of product-related characteristics, seller-related
37
38 characteristics, and sales performance based on our data set. In our research setting,
39
40 knowledge products contain paid answers and articles offered by knowledge celebrities. They
41
42 can monetize these knowledge products in two ways: (1) by charging fees for responding to
43
44 consumers' queries or giving access to answers to questions previously asked (Jan *et al.*,
45
46 2018); and (2) by receiving donations for articles written (Chen and Fu, 2019).
47
48
49
50

51 First, four variables of product-related characteristics—diversity, update frequency,
52
53 price, and trustworthiness— were operationalized as follows. These knowledge products on
54
55 Zhihu are organized in a taxonomy comprising 20 level-1 topics and 78 level-2 sub-topics
56
57 (Zhihu, 2021). Using this structure, the diversity of knowledge products for each knowledge
58
59
60

celebrity can be measured by considering both the number of knowledge topics covered and the level of each topic in the hierarchical structure. Hence, we calculated the sum of the two levels but assigned different weights to highlight their relative importance (Teodoridis *et al.*, 2019). Considering that level-1 topics are more significant than level-2 topics and that their numbers are about a quarter of those in level-2, the degree of diversity equals the number of level-1 topics covered plus 0.25 times the number of level-2 topics covered. Update frequency was measured by the average number of new knowledge products (i.e., paid answers and articles) per day. The price of the knowledge product was set by knowledge celebrities and could vary over time. Informed by Zhang *et al.* (2019c), we took the average historical price of the data set. Informed by Zhao *et al.* (2018), trustworthiness was measured logarithm numbers of knowledge celebrities' historical knowledge contribution endorsed by Zhihu.

Second, three variables—information self-disclosure, tenure, and identity verification—were operationalized as follows. The extent of information self-disclosure for each knowledge celebrity was measured by the word count of one's disclosed self-information. Tenure was measured by the time period in days between the date the oldest product was first made available and our data collection date. For identity verification, Zhihu provides knowledge celebrities with two verifications (Zhihu, 2021). One is an actual verification mechanism involving documentation checks, and the other is a peer-reviewed process. We calculated the number of identity verification each knowledge celebrity held based on the data set. Those who had both types would be coded as 2; those who had one of the two types would be coded as 1; those who had nothing would be coded as 0.

Third, two variables—sales volume and customer satisfaction—are operationalized as follows. The sales volume of each knowledge celebrity is operationalized as the sum of two measures: the volume of articles that attracted donations and the volume of paid answers.

Customer satisfaction with knowledge products is derived from the consumption feedback (Ghose and Ipeiroitis, 2011; Pavlou and Dimoka, 2006), which comprises two components. The first is the cumulative number of readers who willingly donate after reading articles from a knowledge celebrity. The second component is the overt feedback from consumers who gave positive feedback after receiving answers for which they paid. Hence, customer satisfaction is operationalized as the sum of the two components: the cumulative number of readers who donated and the ratio of satisfied consumers to all consumers of the paid answers. The operationalization of these variables is summarized in Table I.

[Insert Table I here]

3.4 LDA topic modeling for analyzing the prepared documents

Topic modeling is a widely-used machine learning technique for automatically inferring semantic topics from text-based content (Misra *et al.*, 2011). The technique has been regarded as a useful tool for uncovering the thematic structure in vast text archives by comparing word-use and shared patterns (Bao and Datta, 2014). The most common algorithm for topic modeling is Latent Dirichlet Allocation (LDA) (Blei *et al.*, 2003). LDA is a generative probabilistic model with a three-level hierarchical structure based on word, topic, and document. Each document is modeled as a random mixture over a set of latent topics, and each topic is characterized as a probability distribution of words (Geva *et al.*, 2019).

LDA topic modeling analysis primarily consists of three stages. In the first stage, the researchers need to specify the number of topics the model should extract from the documents. In the second stage, LDA randomly assigns every word to a topic according to a Dirichlet distribution, which has been commonly used as a prior distribution in Bayesian statistics. The algorithm updates these topic assignments based on the likelihood that the word would appear in the topic across different documents and how common the topic

appears in the rest of the documents. In the third stage, human judgments are invited to interpret the derived topics and ensure that they are valid and meaningful.

For this study, we conducted a Chinese version of LDA topic modeling using three R packages, JiebaR, lda, and LDAvis. As indicated in Section 3.1.2, the JiebaR package was used to segment the sentences into words. After that, the lda package was used to create a user dictionary to store these segmented Chinese words and convert them into vectors. Finally, the LDAvis package was used to conduct the LDA topic modeling analysis. Following prior research using LDA topic modeling (e.g., Geva *et al.*, 2019; Lin and Boh, 2020), we set the Dirichlet hyperparameters, $\alpha = 0.1$ and $\beta = 0.01$, and the number of iterations to 1,000.

3.5 Statistical analyses of the relationships among online identity types, product-related characteristics, seller-related characteristics, and sales performance

After uncovering different identities of knowledge celebrities using LDA topic modeling analysis, we further use statistical analyses to examine the relationships among online identity types, product-related characteristics, seller-related characteristics, and sales performance.

First, we compare their product-related characteristics, seller-related characteristics, and sales performance across different identities using three ANOVAs. Specifically, the independent variable refers to the online identity type, while dependent variables include product-related characteristics, seller-related characteristics, and sales performance. If results yield three or more online identities, Duncan's post-hoc multiple-range test would be performed to determine where the significant differences lie (viz. which specific identity type significantly differs from another).

Second, we employ multiple-group linear regression to analyze the effects of product-related characteristics and seller-related characteristics on their respective sales performance across different identities. Informed by prior literature (e.g., Guo *et al.*, 2017; Ryu and Kim, 2018), we consider the logarithm values of sales volume and the standardization of customer satisfaction to deal with their large variances. If necessary, a Chow test would be applied to the coefficients for further comparisons (Chow, 1960; Jaccard and Turrisi, 2003). The calculation procedure of the test is presented in Appendix 1.

4. Results

4.1 A typology of online identities of knowledge celebrities

Three steps are involved in the analysis of the outputs of LDA topic modeling. The first is selecting the number of topics, the second is the topic labeling, and the third is the interpretation of each topic.

First, researchers need to determine the number of topics (k) to be extracted from the whole corpus by examining the topic coherence score (Bao and Datta, 2014; Misra *et al.*, 2011). The topic coherence score refers to the extent to which documents under each topic are intrinsically consistent, which is measured by calculating the degree of semantic similarity between high-scoring words in the topic (Geva *et al.*, 2019). As indicated in Section 3.4, we used the LDAvis package to generate a chart to visually display the topic of coherence scores under different k . We found that the topic coherence score reached the highest (0.408) when k was 4, 9, or 15. We thereafter manually evaluated the cohesiveness of the resulting topics when $k=4, 9$, and 15, respectively. We found that when the number of topics increased, many of the top words were repeated across topics. A topic model with many topics probably made the distribution of documents sparse across the topics. Hence,

after keeping the balance between topic coherence score and human assessment, we selected 4 as the optimal number (i.e., $k=4$) to segment the corpus with 2,617 documents.

According to the selected number ($k=4$), the corpus ($n=2,617$ documents) containing profiles of knowledge celebrities was thus segmented into four clusters with different proportions. Specifically, the first cluster contains 611 documents (proportion=23.35%); the second cluster contains 882 documents (proportion=33.70%); the third cluster contains 432 documents (proportion=16.51%); and the fourth cluster contains 692 documents (proportion=26.44%).

In the second step, we continued to use the LDA topic modeling to analyze these documents in each cluster. Following the same approach employed earlier, we examined the topic coherence scores across different clusters of documents. We found that 10 topics could best describe each cluster as their topic coherence scores and human assessments reached a balance. The corresponding number of documents of each cluster is shown in the headlines of Table II (a—d). Since all the words derived from the corpus are Chinese, we translated them into English using Online Bing Dictionary (www.bing.com/dict). To ensure translation accuracy, we enlisted the help of two graduate students majoring in Knowledge Management, both of whom are effectively bilingual in English and Chinese. To ease topic convergences, a label was assigned to each topic based on the list of frequent words (e.g., Bao and Datta, 2014; Geva *et al.*, 2019; Lou *et al.*, 2019). For example, the appearance of frequent words such as “machine,” “programming,” and “processing” suggests that the topic is related to computer science and thus labeled as “computer science.” Table II (a—d) shows representative frequent words and topic labels with their respective frequency covered in the documents of each cluster of knowledge celebrities. Besides, we also conducted a robustness check to assess the validity of the typology of knowledge celebrities with four clusters, as illustrated in Appendix 2.

[Insert Table II (a—d) here]

The third step involves a researcher-led interpretation of the topic labels for each cluster of knowledge celebrities. The purpose is to provide a descriptive summary of the topics by identifying key documents that convey such content. To this end, following Geva et al. (2019), two coders, including one of the authors and another research assistant, were trained to independently examine the top 10% representative documents for each cluster of knowledge celebrities, which were content that had the highest probabilities of containing the topic labels, as identified by the LDA topic modeling analysis. Following good practices based on the LDA topic modeling (e.g., Geva *et al.*, 2019), the credibility of their interpretations was ensured by discussing the potential inconsistent understanding of the documents until there was an explicit agreement. The two coders (i.e., one of the authors and one research assistant) participated in the discussions.

The results showed that each cluster's topic labels were roughly consistent with the content of related documents. For instance, with regard to Cluster 4, as shown in Table II (d), the topic labels indicate that this cluster of knowledge celebrities may concern working in the science-intensive or tech-intensive domains. Also, related documents of this cluster of knowledge celebrities provided content describing their ability to create knowledge about some domains such as Internet and digital technologies, data analysis and prediction, and science popularization. A descriptive summary of related documents and examples for each cluster is illustrated in Table III.

[Insert Table III here]

Using the LDA topic modeling technique, we have developed a typology of knowledge celebrities with four clusters. With respect to the typology, we further assign one unique online identity to each cluster of knowledge celebrities by combining the

1
2
3 interpretation of related topic labels and the specific content of related documents.
4

5 Robustness check of the assignments can be referred to Appendix 3, which suggests that the
6
7 online identity assignments hold satisfactory reliability.
8
9

10 For Cluster 1 (n= 611), the content of related documents tended to present knowledge
11 celebrities' educational backgrounds and professional experience so as to enhance the
12 reputation to handle and teach in some domains. Additionally, most of their online profiles
13 (n=482, ratio= $482/611=78.90\%$) obtained at least one type of identity verification from
14 Zhihu, which might allow them to inform their credentials and expertise. Thus, we assign
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
“mentor” as their identity.

For Cluster 2 (n= 882), the content of related documents tended to present their broad
knowledge contribution in different domains or show off their diverse work experience. A
typical example is that someone highlighted her rich experience in keeping fit and doing
makeup, so she effectively communicated with others who want to master relevant basic
knowledge. Thus, we assign “broker” as their identity.

For Cluster 3 (n= 432), the content of related documents tended to present their
popularity among readers and underline the attractiveness in terms of their content generation
on Zhihu. A typical example is that someone was good at making compelling stories and
meanwhile embedded some knowledge into the stories. This strategy could draw others to
read and digest the informative content. Thus, we assign “storyteller” as their identity.

For Cluster 4 (n= 692), the content of related documents tended to present their high-
end levels and professionalism in some high-tech or tech-intensive domains, such as
computer science and engineering. A typical example is that someone interested in
technologies such as blockchain and artificial intelligence had a passion for writing articles

and answering paid questions about such digital technologies with others. Thus, we assign “geek” as their identity.

4.2 Results of statistical analyses

First, we reported the results of comparing variables of product-related characteristics, seller-related characteristics, and sales performance across four clusters (i.e., Mentor, Broker, Storyteller, and Geek) shown in Table IV. Additional analyses including the correlations of variables and regression with the full sample are listed in Appendix 4 and 5.

With regard to product-related characteristics, there was a significant difference in diversity ($F(3, 2613) = 8.66, p < 0.001$) and trustworthiness ($F(3, 2613) = 7.01, p < 0.001$) across the four identities. Specifically, Broker ($M = 36.13, SD = 2.46$) and Storyteller ($M = 33.74, SD = 2.51$) sold knowledge products covering more varied knowledge topics, while knowledge products sold by Mentor ($M = 13.83, SD = 4.56$) and Geek ($M = 12.36, SD = 4.01$) were of more trustworthiness.

With regard to seller-related characteristics, there was a significant difference in information self-disclosure ($F(3, 2613) = 19.67, p < 0.001$) and identity verification ($F(3, 2613) = 4.74, p < 0.01$) across the four identities. Specifically, Mentor had the highest extent of information self-disclosure ($M = 43.64, SD = 12.32$), while Storyteller had the lowest extent of information self-disclosure ($M = 18.22, SD = 19.77$). Meanwhile, Geek ($M = 1.32, SD = 0.16$) and Mentor ($M = 1.28, SD = 0.23$) had more identity verifications than Broker ($M = 1.03, SD = 0.17$) and Storyteller ($M = 0.85, SD = 0.37$).

With regard to sales performance, there was a significant difference in sales volume [$F(3, 2613) = 18.34, p < 0.001$] and customer satisfaction [$F(3, 2613) = 5.21, p < 0.01$] across the four identities. Specifically, Mentor sold the most quantity of knowledge products ($M =$

78.80, $SD = 103.90$). Likewise, Mentor had the highest customer satisfaction ($M = 13.83$, $SD = 14.56$).

[Insert Table IVhere]

Second, we reported the effects of variables of product-related and seller-related characteristics on sales performance, as shown in Table V. For Cluster 1 (Mentor), update frequency ($\beta = 0.121$, $p < 0.001$), trustworthiness ($\beta = 0.206$, $p < 0.001$), and information self-disclosure ($\beta = 0.063$, $p < 0.01$) were statistically significant to their sales volume, whereas update frequency ($\beta = 0.117$, $p < 0.001$) and trustworthiness ($\beta = 0.128$, $p < 0.001$) were statistically significant to their customer satisfaction.

For Cluster 2 (Broker), update frequency ($\beta = 0.274$, $p < 0.001$) and trustworthiness ($\beta = 0.075$, $p < 0.001$) were statistically significant to their sales volume, whereas update frequency ($\beta = 0.113$, $p < 0.01$), trustworthiness ($\beta = 0.078$, $p < 0.01$), and information self-disclosure ($\beta = 0.069$, $p < 0.05$) were statistically significant to their customer satisfaction.

For Cluster 3 (Storyteller), diversity ($\beta = 0.057$, $p < 0.001$), update frequency ($\beta = 0.338$, $p < 0.001$), price ($\beta = -0.087$, $p < 0.05$), and information self-disclosure ($\beta = 0.079$, $p < 0.01$) were statistically significant to their sales volume, whereas diversity ($\beta = 0.039$, $p < 0.05$), update frequency ($\beta = 0.241$, $p < 0.001$), trustworthiness ($\beta = 0.163$, $p < 0.001$), information self-disclosure ($\beta = 0.125$, $p < 0.001$), tenure ($\beta = 0.090$, $p < 0.001$), and identity verification ($\beta = 0.078$, $p < 0.01$) were statistically significant to their customer satisfaction.

For Cluster 4 (Geek), diversity ($\beta = -0.044$, $p < 0.01$), update frequency ($\beta = 0.109$, $p < 0.001$), and information self-disclosure ($\beta = 0.103$, $p < 0.001$) were statistically significant to their sales volume, whereas diversity ($\beta = -0.035$, $p < 0.05$), update frequency ($\beta = 0.106$, $p < 0.001$), information self-disclosure ($\beta = 0.093$, $p < 0.01$) and tenure ($\beta = 0.048$, $p < 0.05$) were statistically significant to their customer satisfaction.

[Insert Table V here]

5. Discussion

5.1 Key findings

According to the results above, knowledge celebrities have been clustered into four online identities—*Mentor*, *Broker*, *Storyteller*, and *Geek*. Product-related characteristics, seller-related characteristics and sales performance differed across the four different identities. Furthermore, the online identity type played a moderating role in the relationships among the three constructs. Results of hypothesis testing are summarized in Table VI.

[Insert Table VI here]

The comparisons among product-related characteristics, seller-related characteristics, and sales performance across four different identities presented several significant findings (see Table IV). We discussed the findings based on prior literature.

First, in terms of product-related characteristics, the diversity of knowledge products varied across four online identities. Broker ($M_{diversity} = 36.13, SD = 2.46$), Storyteller ($M_{diversity} = 33.74, SD = 2.51$), and Geek ($M_{diversity} = 28.99, SD = 7.61$) seemed to sell more diverse knowledge products than the Mentor ($M_{diversity} = 27.21, SD = 7.10$). We conjectured that this was probably because different online identities of knowledge celebrities could have some discretion about allocating their effort in offering knowledge products. According to related literature (e.g., Teodoridis *et al.*, 2019), knowledge creators who focus on a narrow domain tend to contribute deeper and more content than those who can cross diverse domains. Knowledge celebrities with the identity of Mentor were recognized as “experts” in some domains. Hence, they were more likely to draw consumers with relevant knowledge demand. Other online identities might attract consumers by selling knowledge products that cover more diverse domains, reducing the disadvantage of lacking in-depth

knowledge vis-à-vis Mentor. Moreover, the trustworthiness of knowledge products varied across four online identities. Mentor ($M_{trustworthiness} = 13.83, SD = 4.56$) and Geek ($M_{trustworthiness} = 12.36, SD = 4.01$) both offered knowledge products of higher trustworthiness, while Broker ($M_{trustworthiness} = 10.10, SD = 3.12$) and Storyteller ($M_{trustworthiness} = 8.25, SD = 2.75$) lagged. This result was similar to the comparison in their customer satisfaction. According to the literature (Pavlou and Dimoka, 2006), more credible sellers performed better post-purchase feedback. Our results showed that different online identities were related to their credibility variances, eventually reflecting their difference in customer satisfaction.

Second, in terms of seller-related characteristics, the extent of information self-disclosure varied across four online identities. Broker ($M_{info\ self - disclosure} = 43.64, SD = 12.32$) disclosed the most self-information while Geek ($M_{info\ self - disclosure} = 28.75, SD = 13.19$) disclosed the second most self-information. In comparison, Broker ($M_{info\ self - disclosure} = 25.44, SD = 16.75$) and Storyteller ($M_{info\ self - disclosure} = 18.22, SD = 19.77$) disclosed less self-information. This result was similar to the comparison in the trustworthiness of knowledge products and customer satisfaction. Consistent with prior works (e.g., Lee and Li, 2018; Zhao *et al.*, 2018), the extent to which sellers disclosed self-information appears to correlate with their product trustworthiness. Meanwhile, past research also suggests that the extent to which online sellers presented themselves in a marketplace seemed related to customer satisfaction (Li *et al.*, 2015; Li *et al.*, 2016). Moreover, the degree of identity verification varied across four online identities. Both Geek ($M_{identity\ verification} = 1.32, SD = 0.16$) and Mentor ($M_{identity\ verification} = 1.28, SD = 0.23$) had more identity verification than Broker ($M_{identity\ verification} = 1.03, SD = 0.17$) and Storyteller ($M_{identity\ verification} = 0.85, SD = 0.37$). This result was

similar to the comparison in their customer satisfaction. In line with past research (Jin *et al.*, 2019; Shi *et al.*, 2020; Zhao *et al.*, 2018), sellers with more verifications were more likely to be perceived as high quality and ultimately obtained more positive post-purchase feedback.

Third, sales volume and customer satisfaction varied across four online identities in terms of sales performance. Notably, Mentor ($M_{sales\ volume} = 78.80, SD = 103.90$; $M_{customer\ satisfaction} = 13.83, SD = 14.56$) performed best in both aspects, while Storyteller ($M_{sales\ volume} = 63.52, SD = 97.46$; $M_{customer\ satisfaction} = 8.25, SD = 10.23$) lagged behind other identities in both aspects. Broker performed better in sales volume ($M_{sales\ volume} = 74.93, SD = 105.22$), while Geek performed better in customer satisfaction ($M_{customer\ satisfaction} = 12.36, SD = 13.01$). The results supported the arguments in prior literature that different online identities may present heterogeneous sales performance (Ryu and Kim, 2018; Verbeke *et al.*, 2011).

The effects of product-related characteristics and seller-related characteristics on sales performance across four different identities also presented interesting findings (see Table V). We discussed the important findings based on prior literature.

First, for product-related characteristics, the moderating role of online identity type seemed significant to the effects of diversity, prices, and trustworthiness on sales performance. Specifically, the diversity of knowledge products strengthened Storyteller's sales volume ($\beta = 0.057, p < 0.001$) and customer satisfaction ($\beta = 0.039, p < 0.05$), while weakened Geek's sales volume ($\beta = -0.044, p < 0.01$) and customer satisfaction ($\beta = -0.035, p < 0.05$). This indicated that Storyteller and Geek were able to effectively use the diversity to inform their sales performance but for different purposes. The price of knowledge products only weakened Storyteller's sales volume ($\beta = -0.087, p < 0.05$), suggesting that Storyteller was more sensitive to the price than other identities. Furthermore,

the trustworthiness of knowledge products was less likely to be significant to Storyteller's sales volume ($\beta = 0.002, n.s$), Geek's sales volume ($\beta = -0.008, n.s$) and customer satisfaction ($\beta = 0.014, n.s$). Following prior literature on online marketplaces (e.g., Cheung *et al.*, 2014; Li *et al.*, 2015), our results further suggested that sellers with different identities could selectively use certain product-related characteristics to highlight their respective advantages in sales performance.

Second, for seller-related characteristics, the moderating role of online identity type seemed significant to the effects of the information self-disclosure, tenure, and identity verification on sales performance. The extent of information self-disclosure roughly helped promote sales performance of knowledge celebrities, while it did not work for Broker's sales volume ($\beta = -0.007, n.s$) and Mentor's customer satisfaction ($\beta = 0.027, n.s$), indicating that the two identities' information self-disclosure might contribute to different aspects of sales performance. Tenure strengthened Storyteller's customer satisfaction ($\beta = 0.090, p < 0.001$) and Geek's customer satisfaction ($\beta = 0.048, p < 0.05$). This suggested that the two identities' lengths of experience in selling knowledge products could affect consumers' post-purchase feedback. The impact of the Storyteller's tenure was more pronounced than Geek's tenure ($F_{Chow-test} = 66.23, p < 0.001$). Identity verification strengthened Storyteller's customer satisfaction ($\beta = 0.078, p < 0.01$), indicating that this identity could depend more on external verifications to improve post-purchase feedback. Informed by literature on identity signaling (Spence, 2002; Xu *et al.*, 2018), we argued that different identities of knowledge celebrities might emphasize distinctive aspects of seller-related characteristics and highlight respective sale performance strategically. The key findings of this study are summarized in Table VII.

[Insert Table VII here]

5.2 Implications for literature

By addressing two gaps proposed in the part of Introduction, we move forward the literature in four ways. First, we employed a popular computational approach—the LDA topic modeling technique—to develop a typology of online identities of knowledge celebrities by analyzing their online profiles. Past research mostly employs qualitative content analysis to derive online celebrities' identities from their profiles. This study complements the field by providing quantitative evidence. In particular, the LDA topic modeling enables us to compare the distributions of topic clusters covered in self-edited profiles of knowledge celebrities across different identity types and thus offer insights into their identity differentiation by comparing topic clusters included in the profiles.

Second, we compared product-related characteristics, seller-related characteristics and sales performance across different identities of knowledge celebrities. Past research lacks such comparisons because research on identity signaling between online celebrities and the literature regarding relationships among product-related characteristics, product-related characteristics and sales performance in online marketplaces are mostly separated. Given the dual role of knowledge celebrities, our work contributes to both areas. Specifically, we found that the online identity type of knowledge celebrities was statistically significant in accounting for the differences in terms of (i) diversity and trustworthiness of knowledge products, (ii) the information self-disclosure and identity verification of sellers, and (iii) sales volume and customer satisfaction. The findings extend extant literature by showing that identity signaling could contribute in part to the variances of product-related and seller-related characteristics and sale performance of knowledge celebrities.

Third, we studied the moderating role of the online identity type in the relationships among product-related characteristics, product-related characteristics and sales performance. The multiple-group linear regression analyses suggested that some identity could strengthen

the relationship between product-related characteristics and sales performance, while some identity could weaken the relationship. Moreover, some identity could strengthen the relationship between seller-related characteristics and sales performance, while some identity did not affect the relationship. The findings extend the existing literature by showing that the identity signaling of a group of sellers could partially explain why some variables of product-related characteristics or seller-related characteristics affect their sales performances *disproportionately*.

Last, we unveiled a unique monetization strategy used by knowledge celebrities. They monetize their efforts in both ways: (1) obtaining monetary donations from their fans who are fond of their knowledge products; (2) earning money through transactions with buyers of their knowledge products. Prior literature suggests three distinct ways in which online celebrities can gain economic incentives. The first income source comes from being hired by brands to do product and/or service endorsements (e.g., Djafarova and Rushworth, 2017; Lou *et al.*, 2019). The second is from donations of their fans who want to encourage them to generate online content (Duffy and Pooley, 2019; Wan *et al.*, 2017). The third is selling physical goods such as clothes and cosmetics (Cheah *et al.*, 2019; Hua *et al.*, 2017).

5.3 Implications for practices

This study also provides several implications for knowledge celebrities, consumers, pay-for-knowledge platform operators, and policymakers of the digital economy. Our results suggest that it is important for knowledge celebrities to link their online identity type to sales performance. Hence, when setting up their profiles, they have to consider how to better position themselves and highlight their unique advantages in offering knowledge products and sales performance.

Consumers of knowledge products are encouraged to leverage the feedback mechanisms offered by the pay-for-knowledge platforms (e.g., ratings and comments) after consumption to render recognitions for those knowledge celebrities with a particular identity. This also helps potential consumers leverage both online identities and sales performance to look for appropriate sellers to purchase.

The pay-for-knowledge platform operators should provide suitable features that enable knowledge celebrities to manage and strengthen their constructed online identities. For instance, they can develop an identity reputation mechanism for knowledge celebrities to endorse their images. Additionally, by uncovering the relationship between online identity type and sales performance, operators may understand the market positioning strategies behind each type of them. Accordingly, they may assign appropriate resources and support to knowledge celebrities based on their distinctive sales performance in the marketplace.

In the digital economy, online identity has become everyone's virtual wealth (Xu *et al.*, 2018). We suggest that the digital economy policymakers need to scrutinize identity signaling among knowledge celebrities. In particular, the policymakers are encouraged to regulate the self-branding activities among knowledge celebrities to ensure their identity authenticity on the pay-for-knowledge platforms. This helps the healthy development of pay-for-knowledge platforms.

5.4 Research limitations and future work

There are some limitations to this study. First, the variances of economic gains among knowledge celebrities are not considered. Future research may investigate how identity signaling and sales performance are associated with their actual earnings. Second, the content characteristics of knowledge products sold by knowledge celebrities are not explored. Future research may investigate some specific textual attributes (e.g., readability, writing style, and

professionalism) (Banerjee and Chua, 2017; Ghose and Ipeiritis, 2011; Zhang *et al.*, 2019b) and compare them across different identities to obtain more fine-grained findings. Third, this study did not consider the market structure of pay-for-knowledge practices, given that the site of data collection was confined to Zhihu. Future research can explore knowledge celebrities across different marketplaces. It is interesting to incorporate other factors in different market structures such as market maturity, competitive forces, and additional features, including recommendation systems and customer retention incentive schemes.

6. Conclusion

This study explored knowledge celebrities in China— Knowledge Wanghong – from two aspects: (1) to study identity signaling among knowledge celebrities by developing a typology of online identities; (2) to associate their online identity type with product-related characteristics, seller-related characteristics and sales performance. We developed a typology with four distinctive online identities: Mentor, Broker, Storyteller, and Geek. After that, we showed the variances of product-related characteristics, seller-related characteristics, and sales performance across different online identities for the second aspect. Moreover, their online identity type moderated the relationships among product-related characteristics, seller-related characteristics and sales performance. By focusing on an emerging phenomenon of online celebrities who sell knowledge products in China, this study offers a fresh reference point for discussing identity signaling and sales performance of online celebrities in a global context.

Notes:

[1] When referring to the independent variable, we use “online identity type” instead of “online identity types.”

[2] The URL was available during our data collection period. However, it has been changed to the current URL: <https://www.zhihu.com/consult> when we finalized the manuscript.

References

- Abidin, C. (2018), *Internet Celebrity: Understanding Fame Online*, Emerald Publishing Limited, Howard House, Wagon Lane, UK.
- Banerjee, S. and Chua, A. Y. K. (2017), "Theorizing the textual differences between authentic and fictitious reviews: Validation across positive, negative and moderate polarities", *Internet Research*, Vol. 27 No.2, pp. 321-337.
- Bao, Y. and Datta, A. (2014), "Simultaneously discovering and quantifying risk types from textual risk disclosures", *Management Science*, Vol. 60 No.6, pp.1371-1391.
- Blei, D. M., Ng, A. Y. and Jordan, M. I. (2003), "Latent Dirichlet allocation", *Journal of Machine Learning Research*, Vol. 3, pp. 993-1022.
- Cai, S., Luo, Q., Fu, X. and Fang, B. (2020) "What drives the sales of paid knowledge products? A two-phase approach", *Information & Management*, Vol. 57 No. 5, 103264.
- Cheah, J.H., Ting, H., Cham, T.H. and Memon, M.A. (2019), "The effect of selfie promotion and celebrity endorsed advertisement on decision-making processes", *Internet Research*, Vol. 29 No. 3, pp. 552-577.
- Chen, X., Chua, A.Y.K. and Deng, S. (2018), "Comparing the web and mobile platforms of a social Q&A service from the user's perspective", *Aslib Journal of Information Management*, Vol. 70 No.2, pp.176-191.
- Chen, X. and Chua, A.Y.K. (2020), "Reviewing the landscape of research in influencer-generated content", in the *6th International Conference on Information Management (ICIM 2020)*, pp. 244-248, IEEE: March 27-30, London, UK.
- Chen, X. and Fu, S. (2019), "Who sells knowledge to you? Understanding online identities of knowledge influencers on a pay-for-knowledge platform", paper presented at the 13th China Summer Workshop on Information Management (CSWIM 2019), June 29-30, Shenzhen, China, available at: <http://2019.cswimworkshop.org/wp-content/uploads/2017/01/CSWIM2019-Proceedings-3.pdf> (accessed 8 July 2021).
- Cheung, C. M., Xiao, B. S. and Liu, I. L. (2014), "Do actions speak louder than voices? The signaling role of social information cues in influencing consumer purchase decisions", *Decision Support Systems*, Vol. 65, pp. 50-58.
- Chow, G.C. (1960) "Tests of equality between sets of coefficients in two linear regressions", *Econometrica*, Vol. 28 No.3, pp.591-605.
- Chua, A. Y. K. and Banerjee, S. (2013), "So fast so good: An analysis of answer quality and answer speed in community Q&A sites", *Journal of the Association for Information Science and Technology*, Vol. 64 No. 10, pp. 2058-2068.
- Djafarova, E. and Rushworth, C. (2017), "Exploring the credibility of online celebrities' Instagram profiles in influencing the purchase decisions of young female users", *Computers in Human Behavior*, Vol. 68, pp. 1-7.
- Donath, J. (2007), "Signals in social supernets", *Journal of Computer-Mediated Communication*, Vol. 13 No.1, pp. 231-251.
- Duffy, B.E. and Pooley, J. (2019), "Idols of promotion: The triumph of self-branding in an age of precarity", *Journal of Communication*, Vol. 69 No.1, pp.26-48.
- Geva, H., Oestreicher-Singer, G. and Saar-Tsechansky, M. (2019), "Using retweets when shaping our online persona: Topic modeling approach", *MIS Quarterly*, Vol. 43 No. 2, pp. 501-524.
- Ghose, A. and Ipeirotis, P. G. (2011), "Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics", *IEEE Transactions on Knowledge and Data Engineering*, Vol. 23 No. 10, pp. 1498-1512.
- Guo, S., Guo, X., Fang, Y. and Vogel, D. (2017), "How doctors gain social and economic returns in online health-care communities: a professional capital perspective", *Journal of Management Information Systems*, Vol. 34 No.2, pp.487-519.
- Hua, W., Leong, C. and Yu, J. (2017), "Exploring user-created digital content ecosystem: A study of China's digital celebrity industry", paper presented at the *38th International Conference on Information Systems (ICIS 2017)*, December 10-13, Seoul, South Korea, available at:

- <http://aisel.aisnet.org/cgi/viewcontent.cgi?article=1088&context=icis2017> (accessed 8 July 2021).
- iResearch. (2018), "Online paid knowledge platforms seek for more business model", available at http://www.iresearchchina.com/content/details7_43746.html (accessed 8 July 2021).
- Jaccard, J., and Turrissi, R. (2003), *Interaction effects in multiple regression*, Thousand Oaks, CA: Sage.
- Jan, S. T., Wang, C., Zhang, Q. and Wang, G. (2018), "Analyzing payment-driven targeted Q&A systems", *ACM Transactions on Social Computing*, Vol.1 No.3, Article 12, available at: <https://dl.acm.org/citation.cfm?doid=3297860.3281449> (accessed 8 July 2021).
- Jin, Y., Zheng, J., Huang, J. and Tan, Y. (2019), "Knowledge is like money: A demand estimation for paid-knowledge products", paper presented at the *13th China Summer Workshop on Information Management (CSWIM 2019)*, June 29-30, Shenzhen, China, available at: <http://2019.cswimworkshop.org/wp-content/uploads/2017/01/CSWIM2019-Proceedings-3.pdf> (accessed 8 July 2021).
- Kozlenkova, I. V., Palmatier, R. W., Fang, E., Xiao, B. and Huang, M. (2017), "Online relationship formation", *Journal of Marketing*, Vol.81 No.3, pp.21-40.
- Khamis, S., Ang, L. and Welling, R. (2017), "Self-branding, 'micro-celebrity' and the rise of social media influencers", *Celebrity Studies*, Vol.8 No.2, pp.191-208.
- Kim, J. Y. (2018), "A study of social media users' perceptual typologies and relationships to self-identity and personality", *Internet Research*, Vol. 28 No.3, pp. 767-784
- Kuang, L., Huang, N., Hong, Y. and Yan, Z. (2019), "Spillover effects of financial incentives on non-incentivized user engagement: Evidence from an online knowledge exchange platform", *Journal of Management Information Systems*, Vol. 36 No. 1, pp. 289-320.
- Lee, B. H. C. and Li, X. (2018), "Impact of online word of mouth on channel disintermediation for information goods", *Journal of Management Information Systems*, Vol. 35 No. 3, pp. 964-993.
- Li, H., Fang, Y., Wang, Y., Lim, K. H. and Liang, L. (2015), "Are all signals equal? Investigating the differential effects of online signals on the sales performance of e-marketplace sellers", *Information Technology & People*, Vol. 28 No. 3, pp. 699-723.
- Li, W., Chen, H. and Nunamaker Jr, J. F. (2016), "Identifying and profiling key sellers in cyber carding community: AZSecure text mining system", *Journal of Management Information Systems*, Vol. 33 No. 4, pp. 1059-1086.
- Lin, Y., and Boh, W. F. (2020), "How different are crowdfunders? Examining archetypes of crowdfunders", *Journal of the Association for Information Science and Technology*, Vol. 71 No.11, pp. 1357-1370.
- Liu, F., Xiao, B. S., Lim, E. T., and Tan, C. W. (2017), "The art of appeal in electronic commerce: Understanding the impact of product and website quality on online purchases", *Internet Research*, Vol. 27 No.4, pp. 752-771.
- Liu, Z., Zhao, Y. C., Chen, S., Song, S., Hansen, P. and Zhu, Q. (2021) "Exploring askers' switching from free to paid social Q&A services: A perspective on the push-pull-mooring framework", *Information Processing & Management*, Vol. 58 No.1, 102396.
- Lou, C., Tan, S. S. and Chen, X. (2019), "Investigating consumer engagement with influencer-vs. brand-promoted ads: The roles of source and disclosure", *Journal of Interactive Advertising*, Vol. 19 No. 3, pp. 169-186.
- Luo, H. (2014), "When to sell your idea: Theory and evidence from the movie industry", *Management Science*, Vol. 60 No. 12, pp. 3067-3086.
- Mavlanova, T., Benbunan-Fich, R. and Koufaris, M. (2012), "Signaling theory and information asymmetry in online commerce", *Information & Management*, Vol. 49 No. 5, pp. 240-247.
- Misra, H., Yvon, F., Cappé, O. and Jose, J. (2011), "Text segmentation: A topic modeling perspective", *Information Processing & Management*, Vol. 47 No.4, pp.528-544.
- Pavlou, P. A. and Dimoka, A. (2006), "The nature and role of feedback text comments in online marketplaces: Implications for trust building, price premiums, and seller differentiation", *Information Systems Research*, Vol. 17 No. 4, pp. 392-414.

- 1
 - 2
 - 3
 - 4
 - 5
 - 6
 - 7
 - 8
 - 9
 - 10
 - 11
 - 12
 - 13
 - 14
 - 15
 - 16
 - 17
 - 18
 - 19
 - 20
 - 21
 - 22
 - 23
 - 24
 - 25
 - 26
 - 27
 - 28
 - 29
 - 30
 - 31
 - 32
 - 33
 - 34
 - 35
 - 36
 - 37
 - 38
 - 39
 - 40
 - 41
 - 42
 - 43
 - 44
 - 45
 - 46
 - 47
 - 48
 - 49
 - 50
 - 51
 - 52
 - 53
 - 54
 - 55
 - 56
 - 57
 - 58
 - 59
 - 60
- Pee, L.G., Jiang, J. and Klein, G. (2018), "Signaling effect of website usability on repurchase intention", *International Journal of Information Management*, Vol. 39, pp.228-241.
- Ren, R., Yan, B. and Jian, L. (2020), "Show me your expertise before teaming up: sharing online profiles predicts success in open innovation", *Internet Research*, Vol. 30 No. 3, pp. 845-868.
- Ryu, S. and Kim, Y. G. (2018), "Money is not everything: A typology of crowdfunding project creators", *Journal of Strategic Information Systems*, Vol. 27 No. 4, pp. 350-368.
- Schau, H.J. and Gilly, M. C. (2003), "We are what we post? Self-presentation in personal web space", *Journal of Consumer Research*, Vol. 30 No. 3, pp. 385-404.
- Shi, X., Zheng, X. and Yang, F. (2020) "Exploring payment behavior for live courses in social Q&A communities: An information foraging perspective", *Information Processing & Management*, Vol. 57 No. 4, 102241.
- Spence, M. (1973), "Job market signaling", *Quarterly Journal of Economics*, Vol. 87 No. 3, pp.355-374.
- Spence, M. (2002), "Signaling in retrospect and the informational structure of markets", *American Economic Review*, Vol. 92 No. 3, pp.434-459.
- State Information Center. (2018), "The annual report on the development of China's sharing economy in 2018", available at <http://www.sic.gov.cn/archiver/SIC/UpFile/Files/Default/20180320144901006637.pdf> (accessed 8 July 2021).
- Tao, L. (2017), "Why money can't buy you knowledge—Except in China", available at <https://www.sixthtone.com/news/1001161/why-money-cant-buy-you-knowledge-%E2%80%94except-in-china> (accessed 8 July 2021).
- Teodoridis, F., Bikard, M. and Vakili, K. (2019), "Creativity at the knowledge frontier: The impact of specialization in fast-and slow-paced domains", *Administrative Science Quarterly*, Vol. 64 No. 4, pp.894-927.
- Trammell, K. D. and Keshelashvili, A. (2005), "Examining the new influencers: A self-presentation study of A-list blogs", *Journalism & Mass Communication Quarterly*, Vol. 82 No. 4, pp. 968-982.
- Verbeke, W., Dietz, B. and Verwaal, E. (2011), "Drivers of sales performance: a contemporary meta-analysis. Have salespeople become knowledge brokers?" *Journal of the Academy of Marketing Science*, Vol. 39 No.3, pp.407-428.
- Wan, J., Lu, Y., Wang, B. and Zhao, L. (2017), "How attachment influences users' willingness to donate to content creators in social media: A socio-technical systems perspective", *Information & Management*, Vol. 54 No.7, pp.837-850.
- Wang, J., Li, G. and Hui, K. L. (2021), "Monetary incentives and knowledge spillover: Evidence from a natural experiment", *Management Science*, available at://doi.org/10.1287/mnsc.2021.4048 (accessed 8 July 2021).
- Xu, H., Chen, J. and Whinston, A. B. (2018), "Identity management and tradable reputation", *MIS Quarterly*, Vol. 42 No.2, pp. 577-593.
- Xu, P. and Liu, D. (2019) "Product engagement and identity signaling: The role of likes in social commerce for fashion products", *Information & Management*, Vol. 56 No.2, pp.143-154.
- Yazdani, E., Gopinath, S. and Carson, S. (2018), "Preaching to the choir: the chasm between top-ranked reviewers, mainstream customers, and product sales", *Marketing Science*, Vol. 37 No. 5, pp. 838-851.
- Zafar, A.U., Qiu, J. and Shahzad, M. (2020), "Do digital celebrities' relationships and social climate matter? Impulse buying in f-commerce", *Internet Research*, Vol. 30 No. 6, pp. 1731-1762.
- Zhang, F., Hu, Q. and Fang, X. (2019a), "Why pay? An empirical study of paid-for SQA sites in China", *Online Information Review*, Vol. 43 No. 7, pp. 1302-1315.
- Zhang, Y., Lu, T., Phang, C. W. and Zhang, C. (2019b), "Scientific knowledge communication in online Q&A communities: Linguistic devices as a tool to increase the popularity and

- perceived professionalism of knowledge contributions”, *Journal of the Association for Information Systems*, Vol. 20 No. 8, pp. 1129-1173.
- Zhang, J., Zhang, J., and Zhang, M. (2019c) “From free to paid: Customer expertise and customer satisfaction on knowledge payment platforms”, *Decision Support Systems*, Vol. 127, 113140.
- Zhao, Y., Peng, X., Liu, Z., Song, S. and Hansen, P. (2020) “Factors that affect asker's pay intention in trilateral payment-based social Q&A platforms: From a benefit and cost perspective”, *Journal of the Association for Information Science and Technology*, Vol. 71 No. 5, pp. 516-528.
- Zhao, Y., Zhao, Y., Yuan, X. and Zhou, R. (2018), “How knowledge contributor characteristics and reputation affect user payment decision in paid Q&A? An empirical analysis from the perspective of trust theory”, *Electronic Commerce Research and Applications*, Vol. 31, pp. 1-11.
- Zhihu. (2021). “Zhihu Help Center” (in Chinese), available at: <https://www.zhihu.com/help-center/category/7/question/150> (accessed 8 July 2021).

Appendix:

1. About Chow test

By definition, the Chow test is used to determine whether the true coefficients in two linear regressions on different data sets are equal (Chow, 1960). The Chow test statistic follows the F-distribution with k and $(N_1 + N_2 - 2k)$ degrees of freedom. The specific statistic is below.

$$\frac{(S_c - (S_1 + S_2))/k}{(S_1 + S_2)/(N_1 + N_2 - 2k)}$$

Where S_c is the sum of squared residuals from the combined data; S_1 is the sum of squared residuals from the first group; S_2 is the sum of squared residuals from the second group. N_1 and N_2 are the number of observations in each group and k is the total number of parameters.

2. Robustness check: the clustering of knowledge celebrities

We did a robustness check to ensure the effectiveness of the clustering of Knowledge Wanghong (four clusters). As suggested by Ryu and Kim (2018), we conducted the ANOVA test to compare the differences across clusters in terms of the number of their shared particulars on Zhihu. The shared particulars include gender, place of residence, contact information, affiliation, professional institutions, and educational institutions (Max score=6, Min score=0). The results showed significant differences in the obtained scores [$F(3, 2613) = 6.69, p < 0.001$]. Moreover, Duncan's post hoc multiple-range test was used and confirmed that any two categories differ in the number of shared personal particulars (significant level is under 0.05). Hence, these results confirmed the validity of the four clusters of Knowledge Wanghong.

3. Robustness check: the online identity assignments

We did another robustness check to ensure the reliability of the online identity assignments.

Informed by Chua and Banerjee (2013) (see pp.2064 – pp. 2065), we did a manual assessment to ensure the reliability of the assignments. Specifically, three annotators were invited to read the topic labels across Knowledge Wanghong clusters without knowing the respective online identities. After that, we asked the annotators to independently assign the topic labels to an online identity that they think most relevant based on their understanding. The annotators can choose one topic label to one respective online identity. The accuracy for the first annotator is 87.5% (35/40), 82.5% for the second annotator (33/40), and 92.5% for the third annotator (37/40). The high accuracy across three annotators showed that the fit between the topic labels and respective online identities is acceptable.

4. Correlation coefficients of all variables and VIF values

Variables	VIF	1	2	3	4	5	6	7	8	9
1. Diversity	1.43	1.00								
2. Update frequency	1.51	0.15	1.00							
3. Price	1.39	-0.05	-0.16	1.00						
4. Trustworthiness	1.17	-0.03	0.37	0.07	1.00					
5. Information self-disclosure	2.62	0.12	0.24	0.12	0.37	1.00				
6. Tenure	1.77	0.07	0.32	0.05	0.08	0.15	1.00			
7. Identity verification	2.02	0.02	0.06	0.05	0.47	0.33	0.19	1.00		
8. Sales volume	n/a	0.22	0.41	-0.09	0.27	0.29	0.45	0.13	1.00	
9. Customer satisfaction	n/a	0.18	0.27	-0.06	0.32	0.27	0.16	0.35	0.62	1.00

Note: All the correlations are significant at $p < 0.01$; VIF (variance inflation factor) is used to assess how much the square of the estimate's standard deviation of a regression coefficient would change due to collinearity; it is satisfactory when a VIF value ranges from 0 to 3.

5. Effects on sales performance in the full sample

Factors	Effect on sales volume	Effect on customer satisfaction
Product-related characteristics		
Diversity	0.013 (0.018)	0.008 (0.016)
Update frequency	0.232*** (0.022)	0.185*** (0.020)
Price	-0.030 (0.018)	-0.010 (0.016)
Trustworthiness	0.061*** (0.016)	0.129*** (0.014)
Seller-related characteristics		
Information self-disclosure	0.049** (0.018)	0.098*** (0.022)
Tenure	0.008 (0.021)	0.047 (0.033)
Identity verification	-0.002 (0.013)	0.030** (0.012)
Constant	0.258***	0.175***
Adj R^2	0.413	0.436

Note: number of observations = 2,617; *** $p < 0.001$; ** $p < 0.01$; the values in parentheses were standard errors while outside parentheses were coefficients.

Figure: **Figure 1.** Research model

Figure 2. A screenshot of online profile of Knowledge Wanghong on Zhihu

Tables:

Table I. Operationalization of variables

Variable	Description	Operational definition	Supporting literature
Product-related characteristics			
Diversity	The extent to which a knowledge celebrity can offer knowledge products that cover varied domains	Number of unique topics covered in knowledge products (i.e., paid answers and articles)	Teodoridis <i>et al.</i> , 2019
Update frequency	Number of times a knowledge celebrity creates new knowledge products within a defined temporal continuum	Average number of new knowledge products (i.e., paid answers and articles) created per day	Jan <i>et al.</i> , 2018
Price	Price of knowledge products set by a knowledge celebrity	Average historical price of knowledge products for sale	Zhang <i>et al.</i> , 2019c
Trustworthiness	Reliability of knowledge products for sale	Logarithm numbers of a knowledge celebrity's historical knowledge contribution endorsed by Zhihu	Zhao <i>et al.</i> , 2018
Seller-related characteristics			
Information self-disclosure	The length of information self-disclosure of a knowledge celebrity	Word count of information self-disclosure	Jin <i>et al.</i> , 2019
Tenure	The length of experience in selling knowledge products	Number of days since a knowledge celebrity first sold knowledge products	Cai <i>et al.</i> , 2020
Identity verification	Verified identity for a knowledge celebrity	0: No verifications from peers or Zhihu 1: Verification from peers or Zhihu 2: Verification from both peers and Zhihu	Zhihu, 2021
Sales performance			
Sales volume	Transacted volume of knowledge products	Sum of the number of paid answers and the number of articles received monetary donations	Zhang <i>et al.</i> , 2019c
Customer satisfaction	Consumption feedback from actual consumers	Cumulative average ratings of paid answers plus the number of monetary donations to articles	Jan <i>et al.</i> , 2018

Note: the unit of analysis is each knowledge celebrity ; number of observations = 2,617

Table II (a). Topic labels of Cluster 1 (n=611)

Representative frequent words ¹	Topic label	Count (%) ²
如何(how-to), 参考(reference), 权衡(trade-off)	Decision-making	303 (49.59%)
主意(ideas), 想法(thoughts), 提供(supply)	Idea provision	265 (43.37%)
规划(plan), 统筹(coordinate), 整理(arrange)	Collection and arrangement	253 (41.41%)
知识(knowledge), 受众(audience), 互动(interaction)	Knowledge sharing	206 (33.72%)
指导(guide), 技能(technique), 经验(experience)	Expertise and specialization	198 (32.41%)
教育(education), 传授(impart), 方式(ways)	Teaching and supervision	189 (30.93%)
阅读(read), 写作(write), 记忆(memorize)	Reading and learning	173 (28.31%)
医生(doctor), 治疗(cure), 疾病(disease)	Safety and healthcare	105 (17.18%)
客观(objective), 理性(rationality), 包容(tolerance)	Critical thinking	101 (16.53%)
意见(suggestion/advice), 看法(views), 方案(project)	Suggestions and advice	93 (15.22%)

Note: 1. Representative frequent words were randomly chosen from a list of frequent words of each topic because of the presentation limitation of this paper. Full lists of frequent words can be available on request.

2. Count refers to the number of documents containing the topic label. Percentage in the parentheses (%) refers to the frequency of the topic labels relative to the total topics embedded in each cluster. Since LDA topic modeling can assign multiple topics to a certain document, the sum of the percentages for all topic labels does not equal 100% (it should be > 100%).

Table II (b). Topic labels of Cluster 2 (n=882)

Representative frequent words	Topic label	Count (%)
解决(solve), 实用(practical), 效率(efficacy)	Problem-solving	374 (42.40%)
经验(experience), 实习(intern/internship), 合作(cooperation)	Work-experience sharing	280 (31.75%)
烹饪(cooking), 健身(fitness), 锻炼(exercise)	Food and fitness	231 (26.19%)
协商(negotiate), 分析(analysis), 维权(rights protect)	Negotiation and discussion	225 (25.51%)
工作(job), 公司(company), 行业(industry)	Career and job	194 (21.99%)
交流(communication), 倾听(listen), 讲述(description)	Online communication	189 (21.43%)
商业(business), 创业(start-up), 企业(enterprise)	Management and business	182 (20.63%)
国家(nation), 社会(society), 新闻(news)	News and public opinions	177 (20.06%)
辩论(debate), 技巧(skill), 立论(argument)	Debate skills	170 (19.27%)
咨询(consult), 参考(reference), 客户(customer)	Consultation	140 (15.87%)

Table II (c). Topic labels of Cluster 3 (n=432)

Representative frequent words	Topic label	Count (%)
天下(world), 历史(history), 文明(civilization)	History and culture	167 (38.66%)
故事(story), 创作(compose), 小说(novel)	Stories and narratives	152 (35.19%)
音乐(music), 电影(movie), 剧情(plot)	Music and movie	146 (33.80%)
生活(life), 意义(meaning), 家庭(family)	Lifestyles	140 (32.41%)
游戏(game), 玩(play), 英雄(hero)	Enjoyment and entertainment	125 (28.93%)
搞笑(amusement), 笑话(joke), 开心(happy)	Humorous and irony	122 (28.24%)
民族(ethnics), 人文(humanity), 社会(society)	Humanity and society	122 (28.24%)
旅行(travel), 爱好(hobby), 美食(gourmet)	Hobbies and interests	107 (24.77%)
婚姻(marriage), 感情(emotions), 分享(sharing)	Life-experience sharing	101 (23.38%)
照片(photo), 摄影(photography), 画(painting)	Photographer and painting	71 (16.44%)

Table II (d). Topic labels of Cluster 4 (n=692)

Representative frequent words	Topic label	Count (%)
科学(science), 谣言(rumor), 素养(literacy)	Science popularization	243 (35.12%)
网络(internet), 在线(online), 社交(social)	Internet and social media	241 (34.83%)
机器(machine), 编程(programming), 处理(processing)	Computer science	218 (31.50%)
搜索(search), 科技(sci-tech), 信息(information)	Technology and engineering	200 (28.90%)
上市(IPO), 风投(venture capital), 众筹(crowdfunding)	Entrepreneurship	196 (28.32%)
数据(data), 分析(analysis), 预测(prediction)	Data analysis and prediction	181 (26.16%)
科幻(sci-fi), 光年(light year), 大爆炸(big bang)	Science-fiction	177 (25.58%)
博士(Ph.D.), 研究(research), 阅读(reading)	Ph.D. and scientific research	174 (25.14%)
花卉(flowers), 宠物(pet), 照料(care for)	Animals and plants	162 (23.41%)
创新(innovation), 服务(service), 发展(develop)	Innovation and service	162 (23.41%)

Table III. Descriptive summary and examples across clusters

Cluster	Descriptive summary	Examples
Cluster 1	Related documents provided content disclosing their recognized educational backgrounds and/or rich professional experience; They also highlighted their ability to handle and teach in some domains, in which relevant certificates and professional experience are needed	<p>“(I’m) a doctoral student at the University of Pennsylvania... (I) ever worked as a journalist at a well-known news agency in China... (I’m) good at advising professionals, researchers, and students in the journalism and communication area... ” [K84]</p> <p>“(I) obtained the national certificate for lawyers (of China) and (I’m) good at lawsuits regarding economic cases and civil cases, (especially) for divorce cases...contact me via ... and I’m glad to supervise your case...” [K232]</p>
Cluster 2	Related documents provided content describing their broad knowledge contribution in different domains such as consultations and guidance in terms of learning, working and daily life	<p>“(I) have rich working experience in several consulting companies, like Accenture.cn and Bain.cn... (I’m) now working at a start-up focusing on online education... (I’m) good at assisting you in career planning and related information collection and analysis...” [K904]</p> <p>“I have 7-year bodybuilding experience and (I can) help you keep fit and healthy by sharing my information about food and exercises...I (also) provide other important information sources (with you) in terms of keeping healthy and good lifestyles...” [K1266]</p>
Cluster 3	Related documents provided content describing their ability to generate compelling and well-recognized content; Also, they highlighted that they gained popularity among their readers	<p>“(I) do very well in writing... (because I) obtained many upvotes and likes (on Zhihu) due to my prior writing works... (I’m) interested in many topics such as history and movies, for which I wrote lots of commentaries...” [K1635]</p> <p>“(I) like photography and also (published) a book about photography... (I) can tell you details about my creative works and stories behind these photos...” [K1678]</p>
Cluster 4	Related documents provided content describing their ability to create knowledge about sciences and digital technologies, as well as their expertise in some knowledge-intensive or tech-intensive domains such as blockchain and artificial intelligence	<p>“(I) worked as an engineer at several tech companies, like Ali (baba) and Tencent... Recently (I’m) focusing on deep learning and Bitcoins...” [K2030]</p> <p>“I major in computer science and software engineering... (I) practice 1000-line codes per day... Welcome to discuss... (related technologies) and algorithm optimization issues...” [K2615]</p>

Note: Brackets in the third column represent an ID number of each knowledge celebrity in the dataset. E.g., [K84] means the knowledge celebrity with ID 84 in the data set.

Table IV. Comparisons of variables across clusters

Variable	Online identity types of knowledge celebrities				Total	F-value	Pairwise contrast
	Cluster 1 (n = 611)	Cluster 2 (n = 882)	Cluster 3 (n = 432)	Cluster 4 (n = 692)	All clusters (n = 2,617)		
	Mentor	Broker	Storyteller	Geek			
	Mean ± SD	Mean ± SD	Mean ± SD	Mean ± SD	Mean ± SD		
Product-related characteristics							
Diversity	27.21±7.10	36.13±2.46	33.74±2.51	28.99±7.61	32.75±4.42	8.66***	2>3>4&1
Update frequency	4.67±2.28	4.40±2.36	4.63±2.32	4.29±2.75	4.49±2.43	0.58	n/a
Price	24.13±22.58	23.86± 22.70	23.44±22.01	23.51±21.97	23.74±22.31	1.37	n/a
Trustworthiness	13.83±4.56	10.10±3.12	8.25±2.75	12.36±4.01	12.14±3.61	7.01***	1&4>2>3
Seller-related characteristics							
Information self-disclosure	43.64±12.32	25.44 ±16.75	18.22 ±19.77	28.75±13.19	29.22±15.51	19.67***	1>4>2>3
Tenure	89.37±23.74	87.35±23.58	85.64±24.24	86.19±25.37	88.54±25.36	1.65	n/a
Identity verification	1.28±0.23	1.03±0.17	0.85±0.37	1.32±0.16	1.11±0.22	4.74**	4&1>2>3
Sales performance							
Sales volume	78.80±103.90	74.93±105.22	63.52±97.46	66.43±88.54	72.92±86.28	18.34***	1>2>4>3
Customer satisfaction	13.83±14.56	10.10±9.76	8.25±10.23	12.36±13.01	11.75±12.89	5.21**	1>4>2>3

Note: *** $p < 0.001$; ** $p < 0.01$; results of pairwise contrast are based on Duncan's post-hoc multiple-range test ($\alpha = 0.05$)

Table V. Effects of related variables on sales performance across clusters

Independent variable	Effect on sales volume				Effect on customer satisfaction			
	Cluster 1 (Mentor)	Cluster 2 (Broker)	Cluster 3 (Storyteller)	Cluster 4 (Geek)	Cluster 1 (Mentor)	Cluster 2 (Broker)	Cluster 3 (Storyteller)	Cluster 4 (Geek)
Product-related characteristics								
Diversity	0.018 (0.022)	0.005 (0.028)	0.057*** (0.009)	-0.044** (0.012)	0.002 (0.025)	0.004 (0.013)	0.039* (0.019)	-0.035* (0.016)
Update frequency	0.121*** (0.026)	0.274*** (0.018)	0.338*** (0.013)	0.109*** (0.015)	0.117*** (0.023)	0.113** (0.033)	0.241*** (0.019)	0.106*** (0.018)
Price	-0.009 (0.037)	-0.003 (0.025)	-0.087* (0.016)	-0.007 (0.033)	0.005 (0.014)	-0.012 (0.020)	-0.033 (0.041)	0.006 (0.013)
Trustworthiness	0.206*** (0.026)	0.075*** (0.013)	0.002 (0.028)	-0.008 (0.016)	0.128*** (0.015)	0.078** (0.024)	0.163*** (0.004)	0.014 (0.017)
Seller-related characteristics								
Information self-disclosure	0.063** (0.022)	-0.007 (0.029)	0.079** (0.019)	0.103*** (0.016)	0.027 (0.033)	0.069* (0.027)	0.125*** (0.008)	0.093** (0.020)
Tenure	0.015 (0.021)	0.002 (0.037)	0.006 (0.015)	-0.003 (0.017)	0.004 (0.013)	0.011 (0.050)	0.090*** (0.022)	0.048* (0.019)
Identity verification	0.001 (0.022)	-0.012 (0.018)	-0.072 (0.038)	0.014 (0.008)	-0.002 (0.025)	0.011 (0.027)	0.078** (0.023)	0.003 (0.025)
No. of Obs.	611	882	432	692	611	882	432	692
Adj. R²	0.452	0.422	0.317	0.404	0.439	0.381	0.276	0.551

Notes: ***p<0.001; **p<0.01; *p<0.05; the values in parentheses were standard errors while outside parentheses were coefficients.

Table VI. Results of hypothesis testing

Hypotheses		Results
H1	Knowledge celebrities' product-related characteristics and seller-related characteristics are associated with their sales performance	Partially supported
H2a	Online identity type of knowledge celebrities moderates the relationship between product-related characteristics and their sales performance	Partially supported
H2b	Online identity type of knowledge celebrities moderates the relationship between seller-related characteristics and their sales performance	Partially supported

Table VII. Summary of key findings

Comparisons of related variables	Moderating role of online identity type
Diversity of knowledge products: Broker > Storyteller > Geek > Mentor	Diversity strengthened Storyteller's sales volume and customer satisfaction, while weakened Geek's sales volume and customer satisfaction;
Trustworthiness of knowledge products: Mentor > Geek > Broker > Storyteller	Price weakened Storyteller's sales volume;
Information self-disclosure: Mentor > Geek > Broker > Storyteller	Trustworthiness was less likely significant to Storyteller's and Geek's sales volume, and Geek's customer satisfaction;
Identity verification: Geek > Mentor > Broker > Storyteller	Information self-disclosure was less likely significant to Broker's sales volume and Mentor's customer satisfaction;
Sales volume: Mentor > Broker > Geek > Storyteller	Tenure strengthened Storyteller's customer satisfaction;
Customer satisfaction: Mentor > Geek > Broker > Storyteller	Identity verification strengthened Storyteller's customer satisfaction