ISSN 1330-3651 (Print), ISSN 1848-6339 (Online)
https://doi.org/10.17559/TV-20210121100916
Original scientific paper

An OLS and GMM Combined Algorithm for Text Analysis for Heterogeneous Impact in Online Health Communities

Yunqiu ZHANG, Jack STRAUSS, Hongchang LI*, Lihong LIU

Abstract: The increase of doctors' activity in online health communities (OHCs) plays a decisive role in their development. Although the literature on the determinants of doctors' online activities has received considerable attention, the impact of illness severity on these factors remains rare. A network externality analytical framework is constructed to explain the factors (that is, responsiveness, involvement, word-of-mouth, incentives, price, titles and gender) affecting online doctors' behavior, and assess whether factors differ by. By developing text analysis of 4916 doctors' data from a Chinese OHC, this paper applies ordinary least squares (OLS) and General Method of Moments (GMM) to analyze whether the determinants are equal across serious, moderate, and mild illnesses. Our experiment results find that the determinants affecting doctors' online service activity substantially differ across illness severity. Experiments prove the effectiveness of the proposed OLS and GMM methods and demonstrate that they are applicable in online medical field.

Keywords: doctor activity; GMM; illness severity; network externalities; OLS; online health communities

1 INTRODUCTION

Over the last decade, driven by the implementation of a series of complementary national policies such as "Internet + medical" and "Healthy China 2030", online health communities (OHCs) have developed rapidly and have become an indispensable part of medical services in China. OHCs enable communication and interaction on the platform without restricting geographical location and reduce the transaction cost and information asymmetry between doctors and patients [1]. Further, OHCs may mitigate the uneven distribution of high-quality medical resources and improve resource efficiency by matching patients and doctors. Additionally, through social networks and support systems, OHCs improve a patient's motivation for better health [2].

Fig. 1 and Fig. 2 show that the number of online hospitals and market transactions increased dramatically in recent years. The number of online hospitals increased from 1 in 2014 to 577 in 2020, and the market value of online medical consultation increased from 0.2 billion Yuan in 2009 to nearly 100 billion Yuan in 2020. Due to the impact of Covid-19 in 2020, OHCs' market share is projected to continue over the next few years due to social-distancing and health considerations. Particularly, under the Covid-19 pandemic background, the government and industries around the world are striving to develop online technics and services to offset the negative impacts brought about by Covid-19 and stimulate socio-economic developments. From the perspective of techniques, it is useful to generate practical policy implications through scientific methods application. The conclusions and implications of this paper on detailed methods to develop online health communities are helpful for the development of other online platform service activities including but not limited to food takeout, online education etc., by identifying determinants both from their users generated contents, particularly under the continuous impact of Covid-19.

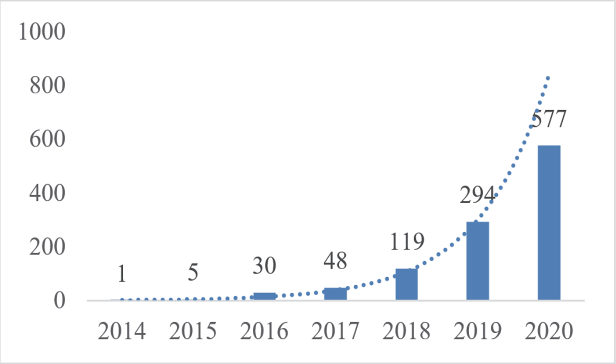

Figure 1 The number of online hospitals from 2014 to 2020 in China

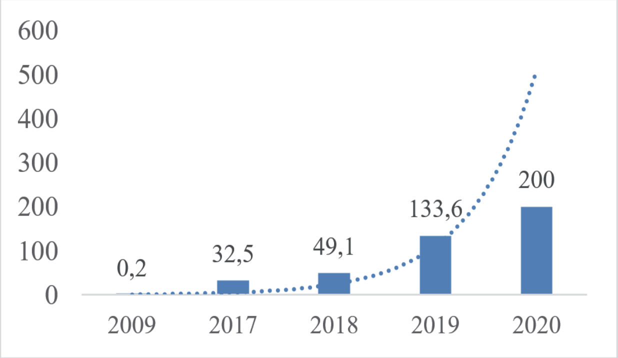

Figure 2 Market transactions for online medical services from 2009 to 2019 in China (unit: billion Yuan)

A successful OHC maintains a relatively large number of active doctors online to offer online services to patients in their working or spare time, and the more active the doctors are, the more attractive the OHC. In turn, the success of an OHC reinforces a doctor's motivation to participate in online medical consultations supply to boost their reputation and yield substantial economic and social rewards. At the same time, benefits such as online word-of-mouth [3] and additional payments from patients who visit the doctor's hospitals offline contribute to the doctor's income. Doctors, patients, and OHCs need to understand what user-generated content (UGC) affects a doctor's online medical consultation activity, because UGC factors can significantly impact the online behavior of doctors [4, 5]. The existence of network externalities [6] implies that doctor user-generated content (DUGC) and patient user-generated content (PUGC) not only impacts a patients' decision to adopt OHCs but also influences a doctors' online medical services [7]. Despite the recent rapid development of OHCs in China, not many doctors are actively involved in online medical consultations. The recent success and positive externalities imply we expect OHCs growth to continue.

The remainder of this paper is organized as follows. Section 2 provides the related research work of this paper. In Section 3, we elaborate the research methods, give an analytical framework and research experiments, and introduce the text analysis technology and process of Python data. Experiment results and discussion are presented in Section 4. Section 5 concludes this paper and highlights directions for future research.

2 RELATED WORK

2.1 Online Health Communities

In recent years, the online health community has increasingly attracted public attention for its convenience, travel expenditure savings, information sharing and support from patients with diseases. Prior literature has discussed the benefits of OHCs to patients from their perspective, such as the online search for health information [2], social support [8], and sharing of treatment experiences. Doctors' motivations to participate in OHCs were mainly to boost their reputations and attract more patients to gain more social and economic returns. In this study, the factors influencing a doctor's online service activity have been widely investigated.

2.2 User-Generated Content

User-generated content created by online users is widely used in industry and academic research. Shim et al. [4] argued that UGC referred to any content submitted by the general public in the digital environment [5]. UGC of online platforms plays a critical role in user decision-making with its open-ended and cost-effective features. Research shows the increasing reliance of consumers on UGC to obtain information about a product or service [9,10] and help or persuade themselves to make the purchase decisions [11]. UGC has significant impacts on user evaluations of products or services [12]. Both qualitative and quantitative dimensions of UGC substantially affect product sales [3, 11], and enterprise online market performance [13, 14].

In OHCs, the content generated by doctors related to their service delivery behavior and that shared online by patients after their diagnosis qualitatively and quantitatively affects doctors' online medical service activity. A rich existing literature has examined the relationship between OHC's information and patient demand or doctor supply. When patients use an OHC, they viewed other patients' feedback to doctors. Additionally, several studies have comprehensively analyzed determinants of OHCs, including both price and non-price factors and their impacts on patient satisfaction, switching between online and offline, and economic returns for physicians [16]. However, these studies have not examined whether a doctor's online activity determinants differ across patient diseases.

2.3 Network Externalities and Online Feedback Mechanism

Online user's activities can generate network externalities and promote a self-reinforcing mechanism on demand and supply behavior. Rohlfs [17] first introduced the concept of externalities into the field of information technology. Direct network externalities refer to the value of a product or service that increases with the rise of its consumption scale [6]. Cross-network externalities mean that the revenue of one side of online platform depends on the total number of users on the other side of the platform [7]. Network externalities enable platform participants to become critical resources. [18-20] argued that social network sites and media have strong direct network externalities. Network externalities further can explain the online demand and supply behaviors of patients and doctors [7, 21].

2.4 Online Users' Activity

Online user activity is often defined as the time spent by users online for some products or services, the frequency of usage of the online platform, and the UGCs through their online feedback. Some defined online user activity as the level of user continued participation in the (social) networks during a specific time; some applied the number of status updates and comments on the other users' posts to measure users' online activity; others used the number of an author's writings during a specific period to evaluate the reviewer's activity.

In this study, we use the UGC by text analysis of 4916 doctors' data from a Chinese online health community to measure the impact of factors on a doctor's online service activity. In contrast to prior work, our paper examines the determinants of online doctors' service activity for different illness severity. Further, we explore the price elasticity response of doctors' online activity using a simultaneous equation framework and evaluate whether the response differs across severity of illness.

3 RESEARCH METHODS

3.1 Research Framework

The greater the prevalence of UGC, the more attractive the OHC will become, due to network externalities effects among patients and doctors [21]. The doctors' online service activity is determined by network externalities of the behaviors of doctors on the supply side as well as cross-network externalities of the patients on the demand side. When the amount of UGC produced by patients and doctors is large, doctors are more likely to obtain useful online medical information to help them form their decisions. In addition, these factors can alter their offline medical service. The influence of direct network externalities, positive online and offline feedback can further impact potential physicians' willingness to adopt and use OHCs [6]. Similarly, when there is more UGC content from patients, doctors react to patients' satisfaction and online state [7], because this is strongly related to a doctors' reputation, self-satisfaction, and social and economic rewards.

Based on the different UGC sources collected by OHCs, this study analyzes the impact of factors on the doctors' online medical supply behavior by categorizing patients into groups with different illness severity and evaluates whether the effects of the variables differ across diseases. Fig. 3 illustrates the factors determining doctors' online activity measured by the doctor and patient-generated content that consists of price factor and non-price factors for patients with different illness severity. The direct network externalities and cross-network externalities among users in OHCs affect the amount of doctor's online activity.

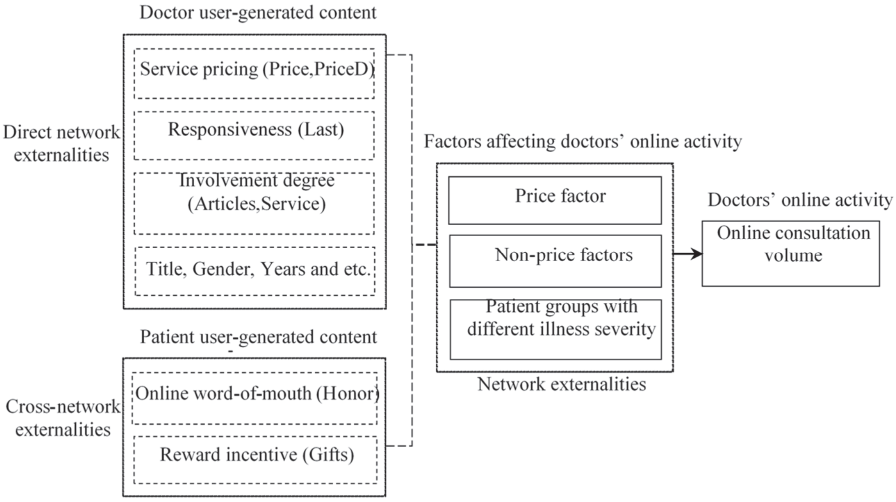
Doctor user-generated content
Service pricing (Price,PriceD)
Responsiveness (Last)
Involvement degree (Articles,Service)
Title, Gender, Years and etc.
Direct network externalities
Patient user-generated content
Online word-of-mouth (Honor)
Reward incentive (Gifts)
Cross-network externalities
Factors affecting doctors' online activity
Price factor
Non-price factors
Patient groups with different illness severity
Network externalities
Doctors' online activity
Online consultation volume

Figure 3 Analytical framework

Nowadays, in the face of the influence of situational factors such as information overload and time pressure, doctors and patients in OHC are more inclined to "selective attention", and the quantified user-generated content in OHC sends a direct signal for them to make a favorable decision.This study obtains digital doctors and patients user-generated content from an open OHC in China to explore their impact on doctors' online activity.

First, we use the timeliness of online responses of doctors proxied by the last time (denoted by Last) they were online on their website to measure responsiveness. Mild to moderate illnesses require more responsive care or patients may switch doctors or not get treatment. Second, the degree of involvement can affect activity as it proxies for platform and persistent participation inclination as well as the investment of time and effort devoted by the users. We use total articles and service type (abbreviated as Articles; Service1, Service2 and Service3) to measure a doctor's involvement. Patients with moderate or mild illness are much more responsive to them. Third, the more online word-of-mouth and reward incentives that doctors receive, the bigger the sales volume of doctors' online healthcare services. In the context of network externalities, it will also attract potential doctors to use the OHC. Patients with serious illness may rely heavily on signal information to select doctors. At the same time, in China's traditional culture, patients with serious illnesses and their family members often give gifts when the doctor provides professional and effective therapies. All of these will further incentivize doctors' online service behavior and have a positive impact on their online activity. Fourth, we examine the effect of reputation measured by doctor's titles. The higher the title, the stronger the signal it sends to the patients on the doctors' service quality. Therefore, this study believes that patients with a serious illness may pay more attention to doctors' titles, because they require more professional examination and treatment. Finally, we examine the effect of service pricing. Doctors can independently set the online service price (such as online written treatment service and telephone consultation prices), implying price and price differences (Price D) considerably vary across doctors and services provided. Medical care is a necessity and hence is relatively price inelastic. Patients will be more willing to pay a higher price and price difference for more severe ailments. Overall, a doctor's semi-flexible pricing strategy may boost a doctor's revenue, positively impacting online service activity. Tab. 1 presents our research experiments.

3.2 Data Collection And Text Analysis

In this study, we directly crawled data from an open and free OHC platform, the Good Doctor Online (https://www.haodf.com) in China. This OHC includes more than 810,000 doctors from 11,033 hospitals in China and has a comprehensive and formal performance reviewing mechanism.

In this paper, Python is used as the tool for crawling data. Python is a powerful and methodical object-oriented programming language. It has a powerful and rich library which can easily combine various modules made in other programming languages (especially C/C++) for efficient development.

Moreover, Python has a concise syntax and mature frameworks like scrapy, making it easier to use. Like other scripting languages java, R, Perl, Python can also run the script directly on the command line. In addition, Python has a wide range of application areas. For example, Web development, big data processing, artificial intelligence, web crawlers, cloud computing and game development, etc., Python can be implemented.

Table 1 Research experiment

Experiments	Description
Experiment 1	Responsiveness (Last) has a positive impact on a doctor's online activity, and patients with mild illness are more sensitive to the doctors' last visit.
Experiment 2a	The involvement degree (Articles) has a positive impact on a doctor's online activity. Patients with moderate or mild illness are much more responsive to Articles. [22]
Experiment 2b	The involvement degree (Service) has a positive impact on a doctor's online activity. Patients with moderate or mild illness are much more responsive to service types.
Experiment 3	The online word-of-mouth (Honor) positively impacts a doctor's online activity. Patients with serious illness pay more attention to doctor's comments.
Experiment 4	The reward incentive (Gifts) has a positive impact on a doctor's online activity and patients with serious illness are more responsive.
Experiment 5	The doctor's title has a positive impact, and significantly impacts patients with a serious illness.
Experiment 6	The online service price (Price) or price difference (PriceD) has a negative effect on online doctors' activity and patients with serious illness are more inelastic. [23]

This paper compiles a Python program named "user_data_fetching.py" (Python version 3.6; Python IDE: PyCharm 4.5.4) to automatically download relevant web pages containing our research data from the Good Doctor Online OHC platform (https://www.haodf.com). Through PyCharm programming software, select "Run 'user_data_fetching'" command to run the program. Simulate browser login for data crawling, from which you can see that the Google Chrome browser has been controlled. Data were collected on August 9 and 10, 2019, and the data crawling process lasted for more than two days. After screening and sorting the data, we obtained a total sample of 4,916 doctors. According to illness severity, we further categorize patients into three groups. Serious illnesses include lung cancer, gastric cancer, breast cancer, and uterine fibroid. Moderate illnesses include hypertension, coronary heart disease, diabetes, cerebral infarction. Mild illnesses include headache, cough, depression, anxiety disorder. The distribution of sample data is shown in Fig. 4. Tab. 2 lists part of the source code of the crawler program in this research. Next, modify the URL of the start page in the code to similarly crawl the related data of the second disease. Tab. 3 lists all variables and their descriptions.

Text classification is mainly divided into two processes: training process and testing process. The main goal of the training process is to obtain a text classifier based on the existing training data. The main steps include: (1) Preprocessing the text, removing punctuation marks and spaces in the text, automatically segmenting the text, and removing stop words; (2) Extracting the text through word frequency analysis and part-of-speech analysis of related text (3) Use Bunch type to describe the text; (4) Use TF-IDF value to assign keywords to construct a keyword weight matrix; (5) Use naive Bayes algorithm to classify text. The data in the testing process is processed in the same way and filtered according to the characteristics of the training process. Finally, use the trained classifier to classify the text.

A Bayesian network is a classifier learned from a series of sample instances with class labels. For a given instance X, represented by a feature vector (a_1, a_2, …, a_n), the Bayesian network uses the following equation for classification:

$$c(x) = \arg\max \mathrm{P}(c) P(a_1, a_2, ..., a_m | C) \tag{1}$$

Among them: C is the set of possible categories c, and $c(x)$ is the category of x predicted by the Bayesian network classifier. For known categories, assuming that all attribute conditions are independent of each other, the Bayesian network classifier based on the assumption of attribute condition independence is called a naive Bayes classifier. The naive Bayes algorithm is the simplest form of Bayesian network classifier and one of the most widely used algorithms in the field of classifiers. The expression of the naive Bayes classifier is:

$$c(x) = \arg\max \mathrm{P}(c) \prod_i P(a_i | c_k) \tag{2}$$

The aforementioned probability $P(c)$ and conditional probability $P(a_i | c)$can be calculated by the following equations:

$$P(c) = \frac{\sum_{j=1}^{n} \delta(c_j, c) + 1}{n + l} \tag{3}$$

$$P(a_i | c) = \frac{\sum_{j=1}^{n} \delta(a_{ji}, a_i) \delta(c_j, c) + 1}{\sum_{j=1}^{n} \delta(c_j, c) + n} \tag{4}$$

Among them: n is the total number of instances trained; l is the total number of categories; n_i is the possible value of the i-th attribute; c_j is the category of the j-th instance; $a_{ji}$is the i-th attribute of the first instance;$\delta(\bullet)$is a binary function, If the two parameter values are the same, the function value is 1, otherwise it is 0.

This section adopts the method of short text similarity calculation. First, the extracted events and the events under one category are input into the Chinese pre-training model, and the short text is converted into a vector. After calculating the cosine similarity of the two vectors, the two Short text similarity. Vector cosine similarity calculation formula:

$$similarity = \cos(\theta) = \frac{A \cdot B}{\|A\| \|B\|} = \tag{5}$$
$$= \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2 \times \sum_{i=1}^{n} (B_i)^2}}$$

Take the angle of the vector as the consideration angle, and take the inner product of the vector (multiplying and summing the corresponding elements) to the product of the modulus of the two vectors as the calculation result. So we get the classification results (Fig. 4).

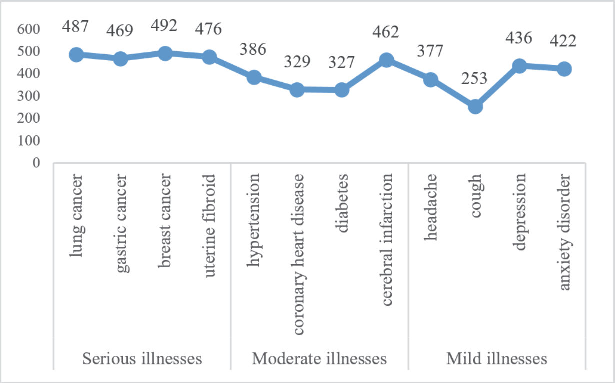

Figure 4 The distribution of sample data based on text analysis

3.3 Models

Ordinary Least Squares (OLS), also known as Least Squares Method, is a mathematical optimization technique. It finds the best function match of the data by minimizing the sum of squares of errors.

Given a set of sample observations(X_i, Y_i), (i = 1, 2, …, n), the judgment standard given by OLS: the sum of the squares of the difference between the two:

$$\text{Min: } Q=\sum_{1}^{n}\left(Y_i-\hat{Y}_i\right)^2=\sum_{1}^{n}\left(Y_i-\left(\hat{B}_0+\hat{B}_1X_i\right)\right)^2 \quad (6)$$

That is, under a given sample observation value, choosing $\left(\hat{B}_0\right)$ and $\left(\hat{B}_1\right)$ can minimize the sum of squares of the difference between Y_i and $\hat{Y}_i$.

We use ordinary least squares (OLS) method which is widely applied to measure the impact of factors on socio-economic activities to estimate (7):

$$\begin{aligned} Da_i = {} & C_0+\alpha_1\ Last_i+\alpha_2\ln\left(Articles_i\ \right)+\alpha_3\ Honor_i+ \\ & +\alpha_4\ln\left(Gifts_i\ \right)+\alpha_5\ Service_{2i}+\alpha_6\ Service_{3i}+\alpha_7\ DR_{1i}+ \quad (7) \\ & +\alpha_8\ DR_{2i}+\alpha_9\ MALE_i+\alpha_{10}\ Years_i+\varepsilon_i \end{aligned}$$

The subscripts i represent doctor i and ε_i is the error term. Since we are interested in the responsiveness or elasticity, we log the independent variables, gifts and articles; the remaining variables are dummy or years in service. An increase in gifts, for instance, increases online activity by α_4 percent. To avoid multicollinearity, we estimate services 2 and service 3. We follow [24] and add one to Articles and Gifts so that a logarithm can be taken. Similar to these authors, we also tried the Inverse Hyperbolic Sign transformation of [25], and obtain similar results.

To evaluate whether illness severity affects these factors, we estimate (7) for the whole sample, and then examine whether our hypothesized determinants change for different subsamples by illness. For example, we estimate, Last*Serious, Last* Moderate and Last* Mild, for the whole sample and use an F value to evaluate coefficient equality across the sample. A large F implies a rejection of the null equality, indicating that the estimates vary for the three different illnesses.

Table 2 Part of the source code of the crawler program

Part of the source code of the crawler program	
	import requests
	from bs4 import BeautifulSoup
1:	import time
2:	from selenium import webdriver
3:	from selenium.webdriver.chrome.webdriver import Options
4:	import re
5:	from openpyxl import Workbook
6:	from tqdm import tqdm
7:	wb = Workbook()
8:	ws = wb.active
9:	chrome_options = Options()
10:	driver = webdriver.Chrome(options=chrome_options)
11:	for page in tqdm(range(1, 68)):
12:	(part of the code is omitted here)
13:	driver.get(hle_url)
14:	soup4 = BeautifulSoup(driver.page_source, 'lxml')
15:	try:
16:	(part of the code is omitted here)
17:	print(DON)
18:	msg = ['diabetes', DON, RNTW, PTCP,TCP, MA, STN, YGD, SSAC, DT, DG, RH, TP, SDE, HLO]
19:	ws.append(msg)
20:	print(msg)
21:	time.sleep(1)
22:	# time.sleep(3)
23:	wb.save('diabetes.xlsx')

Generalized method of moments (GMM) is a parameter estimation method based on the actual parameters of the model satisfying certain moment conditions. The method is described as follows:

Suppose we have n observations from a statistical model {z_1, z_2, …, z_n} and we know that the following q moment conditions are true:

$$E\left(m_1\left(z_i,\theta\right)\right)=0\ \ldots\ E\left(M_q\left(z_i,\theta\right)\right)=0 \quad (8)$$

Among them, θ is apdimensional unknown parameter of about the statistical model. In addition, define $m(z_i, \theta)$ = (m_1 (z_i, θ), …, m_q (z_i, θ))' as the q dimensional moment function about θ. So, we have the condition

$$E\left(m\left(z_i,\theta\right)\right)=0 \quad (9)$$

Given a weight matrix W of $*q$, we naturally have:

$$E\left(m\left(z_i,\theta\right)'Wm\left(z_i,\theta\right)\right)=0 \quad (10)$$

Thus, the GMM estimator $\hat{\theta}$ for the unknown parameter θ is

$$\hat{\theta}=\arg\min_{\theta\in\Theta}\sum_{i=1}^{n}m\left(z_i,\theta\right)'Wm\left(z_i,\theta\right) \quad (11)$$

Among them, Θ is the value space of parameter θ.

In Eq. (12), we use the General Method of Moments (GMM), since price is endogenously determined. This procedure is similar to two-stage least squares. We estimate price equation with the same instruments as (7), and add location and whether the doctor is private or public.

Table 3 Variables description

Variables	Variables name	Explanation	Abbreviation	Type
Dependent Variables				
Online doctors' activity **Independent Variables**	Doctors' activity	The online total consultation volume of doctors before we collected data	Da	Interval
Doctor user-generated content	Last time online	If the last time online is shown today, then 1; else 0.	Last	Dummy
	Total articles	The total number of articles doctors share on their personal websites.	Articles	Interval
	Service types1	1 if the doctor opens two service types (online written treatment service and telephone consultation), but zero otherwise.	Service1	Dummy
	Service types2	1 if the doctor opens three service types (online written treatment service, telephone consultation, and offline clinical appointment or private doctor service), but zero otherwise.	Service2	Dummy
	Service types3	1 if the doctor opens four service types (online written treatment service, telephone consultation, offline clinical appointment, and private doctor service but zero otherwise.	Service3	Dummy
	Price	Average price of online written treatment service and telephone consultation. The absolute value of price difference between online written treatment service and telephone consultation.	Price	Interval
	Price difference	The number of gifts paid by patients to doctors. Whether to be honored as Good Doctor of the Year.	PriceD	Interval
Patient user-generated content	Virtual gifts	Male 1, female 0.	Gifts	Interval
Control variables	Online word-of-mouth	1 if the doctor's medical title is the attending physician, but zero otherwise.	Honor	Dummy
	Doctor gender	1 if the doctor's medical title is the associate chief physician, but zero otherwise.	Male	Dummy
	Doctor title1	1 if the doctor's medical title is the chief physician, but zero otherwise.	DR1	Dummy
	Doctor title2	How long has the doctor's personal websites existed.	DR2	Dummy
	Doctor title3		DR3	Dummy
	Opening years		Years	Interval

$$\begin{aligned} Da_i = {} & C_0 + \alpha_1\ Last_i + \alpha_2 \ln\left(Articles_i\ \right) + \\ & \alpha_3\ Honor_i + \alpha_4 \ln\left(Gifts_i\ \right) + \alpha_5\ Service_{2i} + \\ & + \alpha_6\ Service_{3i} + \alpha_7\ DR_{1i} + \alpha_8\ DR_{2i} + \\ & + \alpha_9\ MALE_i + \alpha_{10}\ Years_i + \alpha_{10}\ \ln\left(E\,Price_i\right) + \varepsilon_i \end{aligned} \tag{12}$$

where all the variables are the same as in Eq. (7) except for ln(*EPricei*) that represents the logged estimated price for a doctors' online service. We evaluate the impact of estimated price differences similarly.

4 RESULTSAND DISCUSSION

4.1 Experiment Results with OLS Estimation

Columns I-IV in Tab. 4 present experiment results using OLS. Column I presents results for the entire sample, and II-IV for patients with serious, moderate, and mild illnesses, respectively. The main commands for OLS are shown in Tab. 5.

Based on Tab. 4, we further analyzed the online activity relationship between different samples of doctors who diagnose and treat diseases of different severity, as shown in Fig. 5 to Fig. 8. It is not difficult to find that the determinants affecting doctors' online activities differ across illness severity.

4.1.1 Experiment Results for Experiment 1

Results for Last confirm Experiment 1 that a doctor's responsiveness is important. The estimate is significantly positive and indicates that a doctor's online activity is positively related to the last time he or she communicated with a patient. Columns II-IV show the significance of the estimate robust across the seriousness of a patient's illness. The estimates indicate that doctors respond quicker to milder ailments. This occurs because headaches, coughs and anxiety are short-run illnesses, and patients, therefore, demand treatment immediately, or they may switch doctors or not seek treatment. In contrast, cancer treatments are chronic, effective therapies take many weeks, often requiring additional examinations, and there is a long-term doctor-patient relationship. Moderate is the lowest because these illnesses are chronic but require a less urgent response as they are not as life-threatening as cancer.

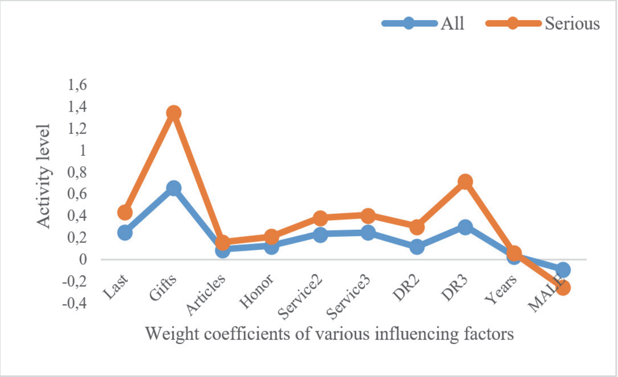

Figure 5 The relationship between the sub-sample of doctors who treat serious diseases and the total sample

Table 4 Regression results for doctor's online activity

Variables	I	II	III	IV
	All	**Serious**	**Moderate**	**Mild**
C	3.305***	3.245***	3.288***	3.358***
	(0.034)	(0.058)	(0.060)	(0.060)
Last	0.249***	**0.189*****	**0.135*****	**0.273*****
	(0.019)	(0.030)	(0.036)	(0.036)
Gifts	0.654***	**0.688*****	**0.663*****	**0.633*****
	(0.008)	(0.012)	(0.014)	(0.014)
Articles	0.093***	**0.064*****	**0.088*****	**0.110*****
	(0.007)	(0.011)	(0.013)	(0.013)
Honor	**0.125*****	**0.083**	**0.096**	**0.165****
	(0.042)	(0.069)	(0.079)	(0.071)
Service2	0.235***	**0.145*****	**0.289*****	**0.259*****
	(0.024)	(0.038)	(0.041)	(0.043)
Service3	0.248***	**0.159*****	**0.306*****	**0.315*****
	(0.029)	(0.044)	(0.053)	(0.054)
DR2	0.118***	**0.189*****	**0.140*****	**-0.004**
	(0.027)	(0.041)	(0.050)	(0.047)
DR3	0.304***	**0.408*****	**0.352*****	**0.140*****
	(0.027)	(0.043)	(0.051)	(0.048)
Years	0.033***	**0.027*****	**0.015****	**0.047*****
	(0.003)	(0.005)	(0.006)	(0.006)
MALE	**−0.088*****	**−0.163*****	**−0.019**	**−0.019**
	(0.019)	(0.031)	(0.035)	(0.034)
Adj. R^2	80.1%	80.2%	79.3%	82.3%

Table 5 Main commands forOLS

Command	
1	**serious
2	**local** z dg dt2 dt3 location years
3	**reg**lnda last lnarticles service2 service3 lnpricelnpricedlngifts `z',r
4	**estatvif**
5	**outreg2** using "yiliao", word tstatdec(3) **append**
6	**moderate
7	**local** z dg dt2 dt3 location years
8	**reg**lnda last lnarticles service2 service3 lnpricelnpricedlngifts `z',r
9	**estatvif**
10	**outreg2** using "yiliao", word tstatdec(3) **append**
11	**mild
12	**local** z dg dt2 dt3 location years
13	**reg**lnda last lnarticles service2 service3 lnpricelnpricedlngifts `z',r
14	**estatvif**
15	**outreg2** using "yiliao", word tstatdec(3) **append**
16	**display _result**(8)
17	**logout,save**(yiliao) word replace dec (3): sum `z0'

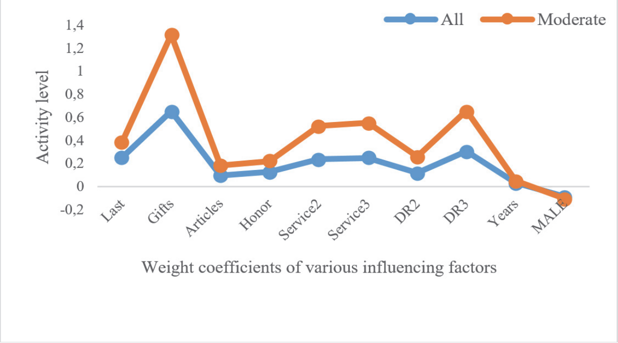

Figure 6 The relationship between the sub-sample of doctors treating moderate diseases and the total sample

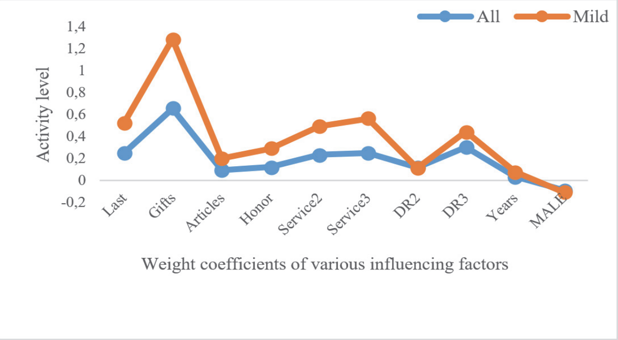

Figure 7 The relationship between the sub-sample of doctors who treat mild diseases and the total sample

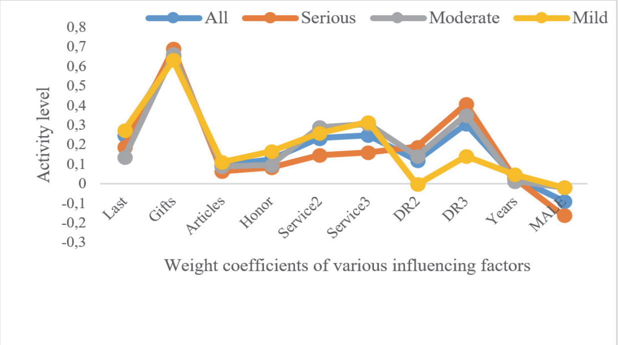

Figure 8 The relationship between the sub-samples and the total sample of doctors treating serious, moderate, and mild diseases

4.1.2 Experiment Results for Experiment 2

Estimates for articles are positive and significant, confirming Experiment 2. The more doctors are involved proxied by articles, the greater their online activity. Patients with mild and moderate illness are more sensitive to articles written by doctors online. Dummy variables are used for Services 2, 3 and they are significant and positive, further confirming Experiment 2. Similar to the response to articles, patients with mild and moderate illness are more sensitive to service types provided by doctors online, because they prefer quick and effective treatments through as many ways or channels as possible. Further, patients with moderate illnesses, such as diabetes and hypertension, coronary heart disease, which are chronic, demand doctors with more online services. Thus, the more service types are, the better the active status and involvement of online doctors.

4.1.3 Experiment Results for Experiment 3

Experiment 3 shows a positive and significant estimate for Honor. The results show a 1% increase leads to a 0.125% increase in online healthcare activity. However, the results only partially support Experiment 3, since patients with mild illness are much more sensitive to word-of-mouth than those with serious illness. Two reasons explain the partial rejection of Experiment 3. First, patients with a serious ailment value a doctor's qualification through official titles (D_{r2} and D_{r3} have very large *t* statistics) rather than online honors. Second, patients with mild illnesses may seek a doctor with honors as a signal that he/she can provide quick, routine and efficient treatment.

4.1.4 Experiment Results for Experiment 4

The incentive structure measured by gifts is strongly related to online activity and supports experiment 4. Patients in OHC increase a doctor's online activity through the payment of gifts after online treatment to express their gratitude to the doctor and maintain the doctor-patient relationship. Gifts have the largest coefficient estimate in the model as well as the largest t statistic, implying that this variable explains a substantial portion of the variance. A simple OLS regression of doctor's online activity on gifts generates an R2 over 70%, with a coefficient estimate of .77 and a *t* statistic of 126. This incentive structure provides strong incentives for doctors with a good reputation to supply more online health treatment. The gifts coefficient is significantly larger for patients with serious illnesses. These patients often give relatively expensive gifts to express their gratitude towards their doctor, which provides strong motivations for doctors to provide more online services.

4.1.5 Experiment Results for Experiment 5

The titles of doctors (attending physician, associate chief physician, or chief physician) have a strong positive relationship on the dependent variable, and confirm experiment 5. Further, the impact of titles differs significantly across illnesses. Patients with severe diseases rely on the signal that a title provides more than patients with a moderate or mild ailment; e.g., the coefficient and t statistics for patients with serious illnesses are several times higher than those with mild illnesses. In contrast, patients with less severe ailments rely on honors to signal a doctor's quality.

4.1.6 Experiment Results for Other Factors

Our controls of gender difference and office years are significant. There is research on the impact of gender differences in OHCs. We find that the Male dummy is negative and significant for the whole sample; its significance is driven by patients with serious illness, as the doctor's sex is not significant for moderate and mild illnesses. The negative coefficient for serious illness is likely due to women patients who have breast cancer or uterine fibroids who prefer female doctors.

4.2 Experiment Results with GMM Estimation

The main commands for GMM are shown in Tab. 6. Tab. 7 reports results for online activity using a GMM approach. We use this methodology to evaluate the effects of price elasticity and prices differences, which are endogenously determined. We use location and whether a doctor is private as instruments in addition to the variables used in Tab. 4. Columns I-IV present results using price, V-VIII price differences. The estimated price regression (not shown for conciseness) indicates price and private doctors are highly significant instruments. Location has a *t* statistic exceeding 28 and private doctors has a *t* statistic above six, implying these instruments significantly and substantially explain price elasticity. The adjusted R^2 for the log price equation is 33%, and indicates no weak instrument problem (e.g., the *F* statistic exceeds 200). Cragg-Donaldson statistics proposed by [26] strongly reject weak instruments. For price differences, the *t* statistics for location and private doctor's dummies are ten and five, respectively; further indicating these variables substantially and significantly explain price differences. Cragg-Donaldson statistics also for price differences show no weak instrument problem. Since the *J* statistics are insignificant, they support our two-stage specification [27]. After adding the price and price difference, the changes in online activity among different doctor samples are shown in Fig. 9 and Fig.10.

Table 6 Maincommands for GMM

Command	
1	**ivregressgmm**da last lnarticles service2 service3 gifts honor dr1 dr2 male years (price = location private), igmm
2	**estatoverid**
3	**ivregressgmm**da last lnarticles service2 service3 gifts honor dr1 dr2 male years (priced = location private), igmm
4	**estatoverid**
5	**ivregress 2sls** da last lnarticles service2 service3 gifts honor dr1 dr2 male years (price = location private)
6	**ivregress 2sls** da last lnarticles service2 service3 gifts honor dr1 dr2 male years (price = location private), **r first**
7	**estatfirststage, all forcenonrobust**
8	**ivregress 2sls** da last lnarticles service2 service3 gifts honor dr1 dr2 male years (priced = location private)
9	**ivregress 2sls** da last lnarticles service2 service3 gifts honor dr1 dr2 male years (priced = location private), **r first**
10	**estatfirststage, all forcenonrobust**

Table 7 Simultaneous regression results for doctor's online activity with price and price differences

Variables	Price				Price Difference			
	I	II	III	IV	V	VI	VII	VIII
	All	Serious	Moderate	Mild	All	Serious	Moderate	Mild
C	4.125***	3.759***	4.441***	4.291***	3.962***	3.607***	3.978***	4.330***
	(0.125)	(0.157)	(0.234)	(0.293)	(0.133)	(0.136)	(0.242)	(0.452)
Last	0.266***	**0.196*****	**0.133*****	**0.281*****	0.210***	**0.175*****	**0.056**	**0.164****
	(0.020)	(0.030)	(0.038)	(0.037)	(0.024)	(0.032)	(0.051)	(0.069)
Gifts	0.675***	**0.696*****	**0.709*****	**0.656*****	0.652***	**0.685*****	**0.671*****	**0.650*****
	(0.009)	(0.014)	(0.017)	(0.017)	(0.009)	(0.013)	(0.018)	(0.019)
Articles	0.081***	**0.054*****	**0.070*****	**0.102*****	0.098***	**0.067*****	**0.091*****	**0.113*****
	(0.007)	(0.011)	(0.014)	(0.013)	(0.008)	(0.011)	(0.016)	(0.016)
Honor	0.215***	**0.138***	**0.267****	**0.248*****	0.138**	**0.089**	**0.148***	**0.140***
	(0.046)	(0.072)	(0.089)	(0.077)	(0.050)	(0.074)	(0.098)	(0.092)
Service2	0.265***	**0.180*****	**0.353*****	**0.301*****	0.220***	**0.158*****	**0.271*****	**0.252*****
	(0.030)	(0.045)	(0.057)	(0.055)	(0.035)	(0.047)	(0.066)	(0.076)
Service3	0.223***	**0.151*****	**0.270*****	**0.217*****	0.191***	**0.129*****	**0.232*****	**0.184*****
	(0.024)	(0.038)	(0.043)	(0.046)	(0.029)	(0.041)	(0.054)	(0.066)
DR2	0.161***	**0.227*****	**0.224*****	0.022	0.164***	**0.223*****	**0.193*****	**0.050**
	(0.028)	(0.043)	(0.055)	(0.048)	(0.033)	(0.045)	(0.063)	(0.065)
DR3	0.379***	**0.477*****	**0.483*****	**0.183*****	0.386***	**0.467*****	**0.443*****	**0.245*****
	(0.030)	(0.048)	(0.060)	(0.051)	(0.036)	(0.050)	(0.069)	(0.079)
Years	0.040***	**0.031*****	**0.029*****	**0.052*****	0.048***	**0.035*****	**0.036*****	**0.063*****
	(0.004)	(0.006)	(0.007)	(0.006)	(0.005)	(0.006)	(0.010)	(0.010)
MALE	−0.099***	**−0.167*****	**−0.022**	**−0.034**	−0.067***	**−0.161*****	**0.062**	**−0.003**
	(0.020)	(0.031)	(0.037)	(0.035)	(0.023)	(0.033)	(0.050)	(0.045)
Price	**−0.231*****	**−0.150****	**−0.344*****	**−0.237*****	**−0.253**	**−0.149*****	**−0.286*****	**−0.322*****
	(0.034)	(0.043)	(0.067)	(0.073)	(0.049)	(0.050)	(0.095)	(0.147)
J Stat	2.12	1.83	1.21	0.909	2.12	1.83	1.21	0.909

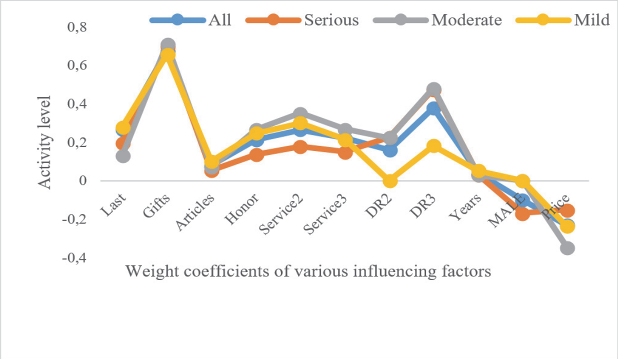

Figure 9 After adding the price, the changes of online activity among different doctor samples

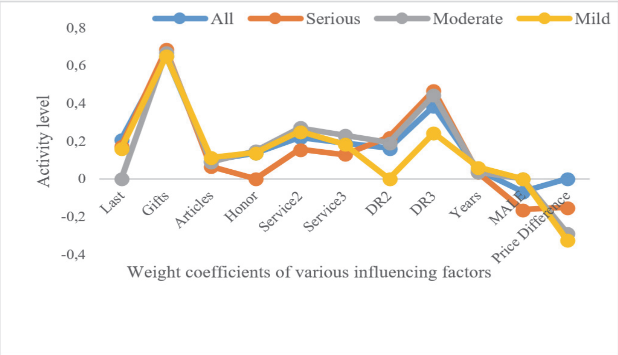

Figure 10 After adding the price difference, the changes of online activity among different doctor samples

4.2.1 Experiment Results for Experiment 6: Price Effect

Results for log price in column I indicate a significant coefficient of −0.231. The price elasticity for serious illness, moderate and mild illnesses is −0.150, and −0.344 and −0.237, respectively. The more inelastic coefficient for serious illness is consistent with economic theory that medical care is a necessity. Whereas, if the disease is less severe, patients may shop around more for doctors with a lower price. The other coefficients, articles, gifts, honor and services are relatively similar to the OLS estimates in Tab. 3. Thus, results for price supports our experiment 6 as well as price elasticity differences across severity of illnesses.

4.2.2 Experiment Results for Experiment 6: Price Difference Effect

Results for price differences presented in column V for the entire sample are −0.253, and −0.149, −0.286 and −0.322 for serious, moderate and mild illnesses, respectively. The price difference elasticity is also more inelastic for more serious illness. Variables such as articles, honor, gifts, and services are all significant for the whole sample and subsamples with magnitudes relatively similar to Tab. 4. Based on our estimation in Tab. 4 and Tab. 7, we summarize results for our 6 experiments in Tab. 8.

Table 8 Experiment results for the 6 experiments

Experiments	**Experiment Results**		**Note**
	General Conclusions	**Coefficient Sign**	
E1: responsiveness (Last)	√	+	Patients with mild illness are more sensitive
E2: involvement degree (both Articles & Service)	√	+	Patients with mild or moderate illness are more sensitive
E3: online word-of-mouth (Honor)	**Partially proved**	+	Patients with mild and moderate illness have higher sensitivity
E4: reward incentive (Gifts)	√	+	Gifts have the strongest impact on doctors' behavior; patients with serious illness are more sensitive
E5: doctors' titles	√	+	Patients with serious illness have higher sensitivity
E6: price or price difference	√	-	Patients with serious illness are less elastic to price or price difference changes.

5 IMPLICATIONS, FUTURE RESEARCH AND CONCLUSION

5.1 Implications

Our research has important policy implications for developing online service platforms through improvement of concerning determinants. Taking the positive externality into consideration, the government can also consider giving subsidy to encourage online service activity development.

First, doctors' online service behavior in OHCs is affected by price, price differences, response, involvement, online word-of-mouth and reward incentives. In response to these determinants, OHCs can adopt comprehensive measures to increase the attractiveness of online medical service and expand their online market segment. Second, the illness severity can affect the price elasticity and price differences of online medical consultation services, which means that regulators of OHCs should pay attention to the potential market power of theirs online doctors. Third, monetary and reputation rewards in OHCs have the strongest impact on the doctors' online activity. Therefore, government policy makers, OHC operators and other stakeholders should promote online gift-giving and voting behaviors to attract more doctors and patients to OHCs. Lastly, network platforms explain OHCs relationships, implying considerable network externalities among users. Large OHCs effectively reduce information asymmetry, better match patients and doctors, and save online transaction costs due to network economies. The healthy mergers and acquisitions of OHCs may benefit doctors, patients, and OHCs and generate positive spillover effects to the broader health care system.

5.2 Future Research

There are several avenues for future research. First, we use cross-sectional data crawled from one OHC to verify our experiments. More data can be collected from a wide range of OHCs across China to analyze the robustness of our conclusions in future research. Second, in addition to quantifiable UGC, text mining technology and questionnaire survey methods can be used in further research to study more OHC content. Third, this paper explains the network externality among users through an analytical framework. The factors used in our model reflect the network externality; however, interactions and possible nonlinear relationship among patients and doctors can be further explored.

5.3 Conclusion

Based on user-generated content (UGC) by text analysis of 4916 doctors from one of the largest Chinese online health communities, Good Doctor Online, this paper explains how factors derived from UGC affect doctors' online service in online health communities (OHCs). We present an analytical framework in which network externalities among doctors and patients are prevalent. The direct and cross-network externalities mitigate information asymmetry and boost utility through interaction and communication among users, support among patients. Better matches between patients and doctors through online reputation and signal selection mechanism further contribute to these externalities. The positive feedback effects in OHCs lead to the rapid expansion of online doctors' service volume. We show a patient's illness severity significantly affects the determinants of a doctor's online activity, and patients with severe illnesses are substantially more price inelastic.

Acknowledgement

This work was supported by the National Social Science Fund of China (No. 19FJYB042).

6 REFERENCES

[1] Vennik, F. D., Adams, S. A., Faber, M. J., & Putters, K. (2014). Expert and experiential knowledge in the same place: Patients'experiences with online communities connecting patients and health professionals. *Patient Education & Counseling*, *95*(2), 265-270. https://doi.org/10. 1016/j.pec.2014.02.003

[2] Ba, S. & Wang, L. (2013). Digital health communities: The effect of their motivation mechanisms. *Decision Support Systems*, *55*(4), 941-947. https://doi.org/10.1016/j.dss.2013.01.003

[3] Shankar, A., Jebarajakirthy,C., & Ashaduzzaman, M. (2020). How do electronic word of mouth practices contribute to mobile banking adoption? *Journal of Retailing and Consumer Services*, *52*. https://doi.org/10.1016/j.jretconser.2019.101920

[4] Shim, S. & Lee, B. (2009). Internet portals' strategic utilization of UCC and Web 2.0 Ecology. *Decision Support Systems*, *47*(4), 415-423. https://doi.org/10.1016/j.dss.2009.04.008

[5] Müller, J. & Christandl, F. (2019). Content is king-But who is the king of kings? The effect of content marketing, sponsored content & user-generated content on brand responses. *Computers in Human Behavior*, *96*, 46-55. https://doi.org/10. 1016/j.chb.2019.02.006

[6] Katz, M. L. & Shapiro, C. (1985). Network externalities, Competition and Compatibility. *American Economic Review*, *75*(3), 424-440. https://www.jstor.org/stable/1814809

[7] Armstrong, M. (2006). Competition in two-sided markets. *The RAND Journal of Economics*, *37*(3), 668-691. https://doi.org/10.1111/j.1756-2171.2006.tb00037.x

[8] Nambisan, P. (2011). Information seeking and social support in online health communities: impact on patients' perceived empathy. *Journal of the American Medical Informatics Association: JAMIA*, *18*(3), 298-304. https://doi. org/10.1136/amiajnl-2010-000058

[9] Dilogini, K., Shivany, S., & Kumara, A. (2019). Analysis of relation between customer behavior and information technology market. *Journal of System and Management Sciences*, *9*(1), 87-104.

[10] Christina, T. (2020). Comparison of the impact of information and communication technology between bilateral trade in goods and services. *Journal of System and Management Sciences*, *10*(1), 1-31.

[11] Chen, Y. & Xie, J. (2008). Online consumer review: Word-of-mouth as a new element of marketing communication mix. *Management science*, *54*(3), 477-491. https://doi.org/10.1287/mnsc.1070.0810

[12] Dellarocas, C. (2003). The digitization of word of mouth: Promise and challenges of online feedback mechanisms. *Management science*, *49*(10), 1407-1424.6 https://doi.org/10. 1287/mnsc.49.10.1407.17308

[13] Goh, K. Y., Heng, C. S., & Lin, Z. (2013). Social media brand community and consumer behavior: Quantifying the relative impact of user-and marketer-generated content. *Information systems research*, *24*(1), 88-107. https://doi.org/ 10.1287/isre.1120.0469

[14] Dragota, V., Oprea, D. S., & Brasoveanu, L. O. (2019). Market efficiency, predictions and returns. *Economic Computation and Economic Cybernetics Studies and Research*, *53*(3), 59-76. https://doi.org/10.24818/18423264/53.3.19.04

[15] Batta, H. E. & Iwokwagh, N. S. (2015). Optimising the digital age health-wise: Utilisation of new/social media by Nigerian teaching hospitals. *Procedia - Social and Behavioral Sciences*, *176*(20), 175-185. https://doi.org/10. 1016/j.sbspro.2015.01.459

[16] Shah, A. M., Yan, X., Shah, S. A. A., & Mamirkulova, G. (2019). Exploring the impact of online information signals in leveraging the economic returns of physicians. *Journal of Biomedical Informatics*, *98*. https://doi.org/10.1016/j.jbi. 2019.103272

[17] Rohlfs, J. (1974). A Theory of Interdependent Demand for a Communications Service. *The Bell Journal of Economics and Management*, *5*(1), 16-37. https://doi.org/10.2307/3003090

[18] Sun, M. & Tse, E. (2009). The Resource-Based View of Competitive Advantage in Two-Sided Markets. *Journal of Management Studies*, *46*(1), 45-64. https://doi.org/10.1111/j. 1467-6486.2008.00796.x

[19] Joon-Hee, K., Myeong-Suk, K., Ryung-Kee, H., & Jong-Wook, K. (2019). Continuous use intention of corporate mobile SNS users and its determinants: Application of extended technology acceptance model. *Journal of System and Management Sciences*, *9*(4), 12-28.

[20] Catalin, B., Alin, Z., Madalina, Z., & Bogdan, I. (2019). User behavior profiling in social media applications. *Economic Computation and Economic Cybernetics Studies and Research*, *53*(1), 21-38. https://doi.org/10.24818/18423264

[21] Pontiggia, A. & Virili, F. (2010). Network effects in technology acceptance: Laboratory experimental evidence. *International Journal of Information Management*, *30*(1), 68-77. https://doi.org/10.1016/j. ijinfomgt.2009.07.001

[22] Zhang, D. & Yoon, S. (2018). Social media, information presentation, consumer involvement, and cross-border adoption of pop culture products. *Electronic Commerce Research & Applications*, *27*, 129-138. https://doi.org/10. 1016/j.elerap.2017.12.005

[23] Venkatesan, R., Mehta K., & Bapna, R. (2006). Understanding the confluence of retailer characteristics, market characteristics and online pricing strategies. *Decision Support Systems*, *42*(3), 1759-1775. https://doi.org/ 10.1016/j.dss.2006.03.012

[24] Criscuolo, C., Martin, R., Overman, H. G., & John, V. R. (2019). Some causal effects of an industrial policy. *American Economic Review*, *109*(1), 48-85. https://doi.org/10. 1257/aer.20160034

[25] Card, D. & Dellavigna, S. (2020). What do editorsmaximize? Evidence from four economics journals. *The Review of Economics and Statistics*, *102*(1), 195-217. https://doi.org/10.1162/rest_a_00839

[26] Stock, J. H. & Yogo, M. (2005). Testing for Weak Instruments in Linear IV Regression. *Nber Technical Working Papers*, *14*(1), 80-108. https://doi.org/10.1017/CBO9 780511614491.006

[27] Hansen, L. P. (1982). Large sample properties of Generalized Method of Moments estimators. *Econometrica*, *50*(4), 1029-1054. https://doi.org/10.2307/ 1912775

Contact information:

Yunqiu ZHANG
School of Economics and Management,
Beijing Jiaotong University,
No. 3,Shangyuancun, Haidian District, Beijing, China
E-mail: 14113150@bjtu.edu.cn

Jack STRAUSS, professor
Reiman School of Finance,
University of Denver,
2101 S. University Blvd., Denver, USA
E-mail: jack.strauss@du.edu

Hongchang LI, professor
(Corresponding author)
School of Economics and Management,
Beijing Jiaotong University,
No. 3, Shangyuancun, Haidian District, Beijing, China
E-mail: hchli@bjtu.edu.cn

Lihong LIU, lecturer
Department of Economics,
Party School of the Beijing Municipal Committee,
No. 6, Chegongzhuang Street, Xicheng District, Beijing, China
E-mail: liulihong@bac.gov.cn