ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/344995248

A Machine Learning Method for Measuring Information Disclosure in Sharing Economy Platforms

Conference Paper · October 2020

CITATIONS
0

READS
316

5 authors, including:

Xin Wei
Tianjin University
8 PUBLICATIONS 8 CITATIONS
SEE PROFILE

Wei (Joy) He
Texas Tech University
28 PUBLICATIONS 906 CITATIONS
SEE PROFILE

xi Zhang
Tianjin University
71 PUBLICATIONS 1,347 CITATIONS
SEE PROFILE

Chuang Zhao
Tianjin University
3 PUBLICATIONS 1 CITATION
SEE PROFILE

Some of the authors of this publication are also working on these related projects:

Project
Knowledge Management View project

Project
Optimal Business Model for IT-enabled Ride-hailing Platform: Pooling, Premier or Hybrid? View project

All content following this page was uploaded by Wei (Joy) He on 30 December 2020.

The user has requested enhancement of the downloaded file.

A Machine Learning Method for Measuring Information Disclosure in Sharing Economy Platforms

Completed Research Paper

Xin Wei
College of Management and Economics, Tianjin University
Tianjin, China
wxin9618@tju.edu.cn

Wei He
Area of ISQS, Rawls College of Business, Texas Tech University
Lubbock, Texas, U.S.A
wei.he@ttu.edu

Xi Zhang
College of Management and Economics, Tianjin University
Tianjin, China
jackyzhang@tju.edu.cn

Chuang Zhao
College of Management and Economics, Tianjin University
Tianjin, China
zhaochuang_tju@163.com

Hongke Zhao
College of Management and Economics, Tianjin University
Tianjin, China
hongke@tju.edu.cn

Abstract

Research on e-commerce, social technologies and privacy has overwhelmingly treated information disclosure as a survey-based, subjective, and unidimensional construct. A few studies employing semantic analysis on objective textual data, on the other hand, are constrained by the manual coding method with limited number of categories and thus prone to bias. Building upon the social penetration theory, we introduce an innovative method of measuring information disclosure using machine learning algorithms in the context of sharing economy platforms. We propose that information disclosure should be examined from two dimensions, i.e., breadth and depth, and machine learning techniques could effectively compute the high-volume factual data of information disclosure. Using 1,200 hosts' self-description data in Airbnb as an example, we report the computational and evaluation processes of operationalizing information disclosure. The research thus provides new theoretical lens and empirical support through which information disclosure in digital age could be better understood and efficiently assessed.

Keywords: Information disclosure, Social Penetration Theory (SPT), breadth, depth, sharing economy platforms, machine learning

Introduction

With the development of Internet and information technologies, more and more businesses are collecting personal information of consumers on a large scale in order to understand their interests and market personalized products and services effectively. Such a ubiquitous practice has inevitably led to concerns on consumers privacy. Despite the progress in information privacy literature, researchers have predominantly examined the various factors that might influence consumers' information disclosure intention or behavior, such as privacy concerns, perceived benefits and risks, and their relationships (Adjerid, Peer, and Acquisti 2018; Dinev and Hart 2006; Smith, Dinev, and Xu 2011; Xu et al. 2011a). There still is a lack of sufficient scholarly attention on information disclosure construct itself (Smith et al. 2011).

Information disclosure, traditionally occurred in interpersonal communication process, refers to individuals' intentional behavior of disclosing their personal information voluntarily (Fisher 1984). It is also called self-information disclosure, self-disclosure, or personal information disclosure (Altman and Taylor 1973). Information disclosure and its associated privacy concerns have been extensively studied in contexts such as online transactions (Awad and Krishnan 2006; Dinev and Hart 2006), social networking sites (Krasnova et al. 2010) and location-aware services (Sun et al. 2015; Xu et al. 2011b). However, less research focuses on information disclosure behavior on emerging sharing economy platforms, which represents unique privacy concerns (Hu, He, and Davis 2020). Specifically, personal information the hosts revealed in the Airbnb platform, for example, may be more private, including personal interests, tastes, marital status and work-related information. Moreover, the motivation and effect of information disclosure in sharing economy context could be distinct (Hu et al. 2020; Ma et al. 2017). Therefore, we need new theoretical lens to understand information disclosure behaviors in a context of sharing economy platforms from the service providers' perspective, not the traditionally consumer-directed perspective that has been widely investigated.

With regard to the operationalization of information disclosure behavior, prior studies can be divided into three categories. The first—probably also the largest—group of research adopts the survey method asking the respondents to self-report their intention or actual behavior of information disclosure as a unidimensional scale (Dinev and Hart 2006; Krasnova et al. 2010; Li 2014; Xu et al. 2011b; Zlatolas et al. 2015). However, the existence of privacy paradox, referring to the phenomenon that consumers readily submit their personal information in most circumstances despite they claim a high level of privacy concerns, poses a severe challenge to the validity of survey-based perceptual or claim-to-be behavioral instruments (Smith et al. 2011). The second group deems disclosure behavior as a multidimensional scale, but still using survey method (Benson, Saridakis, and Tennakoon 2015; Hollenbaugh and Ferris 2014; Knijnenburg, Kobsa, and Jin 2013; Liu et al. 2016). The third group of studies employ purely hard data to compute disclosure by artificial coding and topic classification (Kisilevich, Ang, and Last 2012; Ma et al. 2017). In summary, the survey-based approach, whether unidimensional or multidimensional, has two major limitations: incapability of accurately measuring an object of concern, and the failure of leveraging the vast amount of factual data in e-commerce contexts (Wu, Huang, and Zhao 2019), which has largely constrained current research on self-disclosure behavior and consumer privacy. Early attempts on coding information disclosure based on the text contents shed a light on improving the operationalization of disclosure, but the issue of lacking a theory-directed framework guiding the investigation on information disclosure remains. Innovative research on the conceptualization and operationalization of information disclosure is thus urgently needed.

In order to fill this research gap, building upon Social Penetration Theory (SPT) and self-disclosure literature, the present study focuses on the conceptualization and operationalization of information disclosure in sharing economy context. We propose using machine learning method to calculate the extent of information disclosure from two dimensions: breadth and depth. The rapid development of the social technologies and social commerce arena provides rich sources of objective data logging and technical supports for our research. On one hand, platform logs can intuitively display individuals' disclosure behavior. On the other hand, mature machine learning technology makes it a feasible and promising prospect that the proposed new approach of operationalizing information disclosure could be quickly applied and evaluated in various practical settings.

This study makes several theoretical and practical contributions. First, we provide a new calculation method for the breadth and depth dimensions of self-disclosure using sharing economy platform as an example. Thus, the present study will lay a foundation for further IS research that aims to leverage the incredibly large amounts of data to understand information disclosure behaviors, their explanatory mechanisms and downstream consequences. Second, this research extends the SPT to the sharing economy context by explicitly defining the breadth and depth of information disclosure from the service provider perspective, which enriches both areas of research. Third, we also contribute to the machine learning field by demonstrating not only the development process of computing information disclosure but also the use of manual annotation to verify the effectiveness of the algorithm. The results of this research also have significant implications for practitioners. E-commerce businesses, including the sharing economy platforms, are able to analyze the huge volume of consumer data they have accumulated so as to reveal the economic impacts of users' information disclosure. With the information disclosure behavior analyzed on a factual basis in nature and combining both breadth and depth dimensions, platform operators are likely to generate more insights for decision-making (Smith et al. 2011).

The rest of this paper is organized as follows. First, we introduce the theoretical background about SPT. Next, we report the steps of using machine learning to measure the breadth and depth of self-description and verify it by comparing it with human coding results. Finally, we discuss our main findings.

Theoretical Background and Literature Review

Information Disclosure: The Perspective of Social Penetration Theory

Altman and Taylor (1973) propose the Social Penetration Theory (SPT) to explain the development process of interpersonal relationships. They use the onion model (see Figure 1) as a metaphor to describe that social penetration is realized through self-disclosure. When individuals reveal their personal information from the outer layer that is visible to others to reach the "core self", represented by the wedge penetrating the onion layers, the interpersonal relationships gradually develop from superficial to more intimate. SPT has been applied in various decision-making scenarios, such as target selection, risk identification, relationship development and benefit evaluation (Sultan and Chaudry 2008; West and Turner 2018). In any of these given situations, information disclosure is achieved through conscious or unconscious wills that are designed to achieve a specific set of social goals, and the most significant goal will determine what is disclosed by an individual (Omarzu 2000).

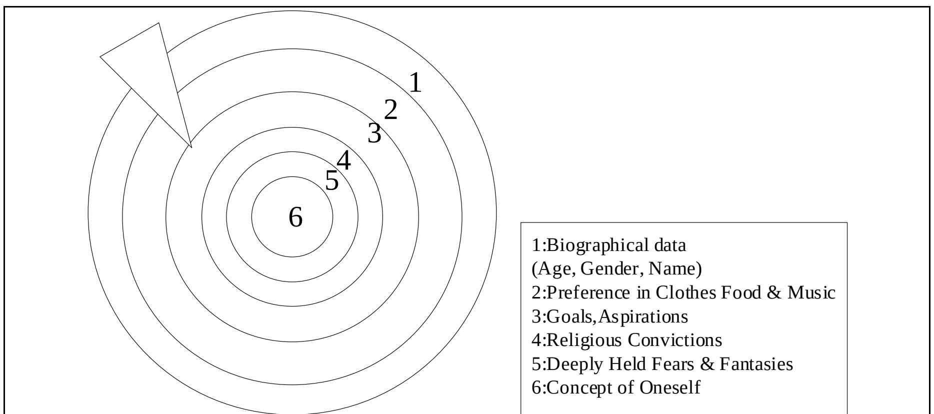

Figure 1. Onion Model

According to SPT, people tend to predict the costs and rewards of disclosure and consider the results of comparisons (Berg 1984), which is consistent with the privacy calculus theory that has widely been adopted in recent privacy literature. Therefore, SPT provides a concrete framework for us to further understand information disclosure (Liu et al. 2016).

SPT believes that there are two main dimensions of information disclosure: *breadth* and *depth*. The breadth of disclosure is the scope of the information areas disclosed, wider or narrower. For example, a person's statement could have different topics, such as age, gender, likes and dislikes of dressing styles and sports, values and thoughts, family conditions and romantic relationship, and so on (Baack, Fogliasso, and Harris 2000). A person may be completely open to disclose his or her information of age or dressing styles while hiding romantic relationships from being accessible to others for a variety of reasons. Having information disclosure ranging over a wide variety of topics is regarded as a high level of disclosure breadth, whereas disclosure centering around one subject of interest (or, topic) is a low level of breadth (Hollenbaugh and Ferris 2014).

On the other hand, the depth of disclosure is more about the quality of information or the level of intimacy (sensitivity) (Collins and Miller 1994; Osatuyi et al. 2018). For example, one's medical record is highly sensitive topic, thus the depth of disclosing emotional depression and medical history will be greater than that of disclosing age and gender information (Cao, Hui and Xu 2018). Because disclosure is prone to loss, information with higher sensitivity is considered to have a higher risk (Lwin, Wirtz, and Williams 2007; Moon 2000).

In the context of sharing economy platforms, service providers such as the Airbnb hosts disclose personal information that might signal their goodwill and trustworthiness (Ma et al. 2017). Therefore, the breadth dimension of information disclosure in a sharing economy context is defined as *the diversity of information content, or, the number of information topics that service providers' self-disclosure has covered and that are relevant to developing relationships with potential consumers*, for example, age, ethnicity, origin or residence, occupation, experience, preference and personality, and so on.

Traditionally, word count (or, the length) has been adopted as a quantitative surrogate of information depth in online review research, which in turn determines persuasiveness of the review messages (Chevalier and Mayzlin 2006; Liang et al. 2019; Liu and Park 2015; Racherla and Friske 2012; Xu, Zeng, and He 2021). Long product reviews on Amazon.com are believed to include more product details and describe more specific usage situations, representing a higher level of review depth than short reviews (Mudambi and Schuff 2010). The context-specific descriptions are likely to be more personal and represent a higher level of intimate disclosure than the surface-level comments on the product features. The length characteristic is also regarded as a signal of information quality in extant literature on various research subjects. For example, the number of words contained in a product's description on the crowdfunding website has a significantly positive impact on successful fundraising (Kim, Por, and Yang 2017). Length of descriptive text has also been examined in the study on financing success of P2P lending sites (Larrimore et al. 2011). Besides the length indicator, we introduce information entropy as another important representation of disclosure depth. Information entropy is originally coined to quantify the expected value of self-information or how "informative" it is (Shannon 1948). Shannon argues that the fundamental problem of communication is for the receiver to be able to identify what data is generated by the source, based on the signal received through the communication channel. Entropy represents the uniqueness of information and thus has been used to measure the quality of various types of information, such as image quality (Soundararajan and Bovik 2012; Gu, Tao, and Qiao 2017) and password quality (Yan 2001). Notably, entropy has recently been applied in an IS study on the helpfulness of online review information to measure the "newness" of information that a review contains (Fresneda and Gefen 2019). To sum up, following the literature, we define the depth dimension of information disclosure in sharing economy platforms as *(1) the length of service providers' self-description (i.e., number of words) and (2) the information entropy of service providers' self-description* which accounts for the quality of disclosed information.

Mandatory versus Voluntary Information Disclosure

There are two different forms of information disclosure: mandatory disclosure and voluntary disclosure (Li, Cheng, and Teng 2020). Mandatory disclosure means that users must provide certain personal information in order to create an account or get certain services, such as filling in name, address, and contact number in order to receive delivery of e-commerce transactions. Voluntary disclosure, on the other hand, refers to users' voluntary disclosure of personal, but not mandatory, information such as posting contents to social networking sites that contain personal photos, moods and self-introduction (Lowry, Cao, and Everard 2011). Voluntary disclosure is more likely to be driven by factors such as perceived economic and convenient benefits, social network size, and personalization (Teubner and Flath 2019; Xu et al. 2009). Mandatory disclosure is usually affected by factors such as the quality of privacy policies of online platforms and perceived risk of disclosing information (Einhorn 2005). Various aspects of the two modes of disclosure, such as the context of occurrence, the types of information content, frequency and purpose of disclosure and who will access to and keep the disclosed information, are fundamentally different (Hu et al. 2020; Li et al. 2020). Previous research on information disclosure does not consider the difference between mandatory and voluntary disclosure, but survey participants may mix these two in responding the survey questions. This leads us to believe that a rigorous, non-perceptual and other than self-reported survey method is necessary for researchers to better examine disclosure behaviors.

Challenges and Opportunities of Measuring Information Disclosure

Scholars have taken different approaches to address the limitations of traditional operationalization of information disclosure. For example, Liu et al. (2016) regard self-disclosure as a multidimensional structure. Although still using the survey method, they measure disclosure from the dimensions of quantity, depth, honesty, intention and potency. Knijnenburg et al. (2013) evaluate information disclosure datasets and demonstrate that information disclosure behaviors are multidimensional, distinct dimensions depending on the specific contexts. Yasui et al. (2010) divide the publicly disclosed information of users into different information units and classify them according to their sensitivity and visibility. Similarly, according to the observations and interviews with 16 nurses, Unhjem et al. (2018) divide the content of self-disclosure into four themes, namely immediate family members, interests and activities, life experience and identity, and explore the reason of self-disclosure of nurses to patients in mental healthcare.

However, the recent advances on the development of computing technologies and business innovations have brought new opportunities for capturing the actual degree of information disclosure not only in a nuanced manner but also being able to analyze mass textual data from consumers (Fresneda and Gefen 2019). Information disclosure research could be enabled by the modern machine learning method to achieve a higher level of accuracy in exploring the privacy related topics. Meanwhile, e-business firms would also be able to unlock huge potentials by identifying the patterns of users' disclosure behavior and further finding its relationship with business growth and economic gains.

Information Disclosure in Sharing Economy

Sharing economy is a new economic phenomenon that the public share their idle resources with others through a third-party platform without obtaining a transfer of ownership (Teubner and Flath 2019). While a number of studies have been conducted to understand why individuals participation in sharing economy (e.g., Hamari, Sjöklint, and Ukkonen 2016; Möhlmann, 2015), research on information disclosure by service providers in the sharing economy platforms is still scarce (Teubner and Flath 2019). Remarkable exceptions include some recent studies on why service providers would be willing to disclose personal information. Some predominant factors include balancing the benefits and privacy concerns (Teubner and Flath 2019) and improving perceived trustworthiness by consumers (Ma et al. 2017). Liang et al. (2019) collect a unique longitudinal dataset from Airbnb to examine the determinants of hosts' decision to disclose information. Their results show that factors of getting more reviews, higher ratings and receiving more informational or readable reviews prompt hosts to disclose more information, especially for less privileged information.

Some scholars also analyze the impact of service provider disclosure. Tussyadiah and Park (2018) use text mining techniques to analyze a large dataset containing descriptions of Airbnb homeowners in 14 major US cities. They identify two patterns of self-disclosure by service providers: one is to disclose their travel experience and express that they are eager to meet new friends, and the other is to show their profession. The results suggest that consumers would have a higher level of trust and willingness to transact with service providers disclosing good travel experiences.

Ma et al. (2017) manually code the topics Airbnb hosts self-disclose in their profile. Five most popular categories of information broadcasted by hosts are origin or residence, work or education, interests and tastes, travel and demonstrating hospitality. The least revealed information of hosts includes relationships, personality, and life motto and values. The researchers also compare the disclosure topics between on-site hosts and remote hosts and find that their topics released significantly differ. A particular advancement of this study in operationalizing information disclosure is the researchers not only capture the category information but also measure the depth of disclosure, in terms of the number of sentences used in disclosure topics. The results show that longer self-descriptions are considered more credible by consumers.

Ert and Fleischer (2019) find that including photos of the host may affect the listing price of the house and the choice of the guest positively. However, self-disclosure may also come with unexpected consequences. For example, a study on the photos of all New York City landlords on Airbnb and their listing information shows that non-black hosts have about 12% more rent than the black ones (Edelman and Luca 2014).

The above reviewed literature has put forward new approaches and made substantial achievements in capturing the actual disclosure behavior. However, most of the existing research studies disclosure from a categorical perspective and focuses on only one category (e.g., photo) or a very limited number of categories. This may be the result of inherent capacity limitation of traditional methodologies such as manual coding. Building upon the prior work, we propose in this study that information disclosure can be examined from both breadth and depth dimensions, covering a wide range of topics, and utilizing the machine learning techniques to efficiently process high-volume secondary data in various practical settings. We detail the procedures of implementing this method in next section.

Research Method

In the digital age, it is impossible to manually label the innumerable data. Latent Dirichlet Allocation (LDA), a technique for extracting hidden topic structures from documents, provides a solution (Blei 2012). LDA is the main algorithm used in our topic modeling. In the following, we use the LDA algorithm to create a topic model for calculating the depth and breadth of disclosure, followed by human coding experiments to verify the measurement accuracy. Figure 2 draws a flowchart of the processes.

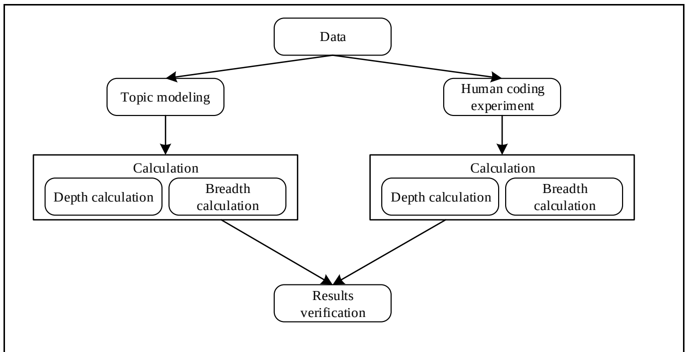

Figure 2. Flow Chart of Breadth and Depth Calculation

Semantic features are one of the measurement dimensions of information quality, such as the semantic relationships between words, topics, and language entities contained in text. According to elaboration likelihood model (ELM), the semantic theme affects the trust of consumers through a central path. This kind of persuasion and influence is stronger and more durable (Petty, Cacioppo, and Schumann 1983). The semantic theme contained in the self-descriptions of the service providers on the sharing economy platforms is the essence of the descriptive text. The larger the number of topics covered by the text, the wider the scope of personal information exposed or the greater the amount of information exposed. These provide us an opportunity to calculate the breadth dimension of service providers' self-disclosure.

As we reviewed earlier, the depth dimension of disclosure will be measured by the length of hosts' self-description as well as the information entropy. We discuss how the two indicators are calculated in the computational processes later.

Data Collection

We illustrate our computation of information disclosure using the hosts' self-description information on Airbnb as an example. Airbnb was established in San Francisco in 2008 as a profitable peer-to-peer accommodation platform (Gunter 2018). The platform has more than 7 million listings and more than 500 million consumers (Airbnb 2019), serving as a good setting for our research.

We retrieved the data from Inside Airbnb (http://insideAirbnb.com/). Inside Airbnb is an Airbnb-independent data collection project created by Murray Cox, which compiles Airbnb listing information for public use. This site includes not only listing details but also 365-day availability calendars and review data for all listings. We firstly identified the London listings data on Inside Airbnb in November 2019. Out of the approximately 80,000 listings, we randomly selected 1,200 listings and their data as the sample of our research. Consistent with the literature (Ma et al. 2017), we focused on the hosts' self-description as the target of investigation. An example of a host's self-description on Airbnb is shown in Figure 3. Hosts who are willing to provide their information to consumers typically introduce themselves in great detail, for example, their family status, hobbies, occupation, and so on, which contains a large amount of information. This rich dataset therefore enables us to study the breadth and depth dimensions of hosts' information disclosure.

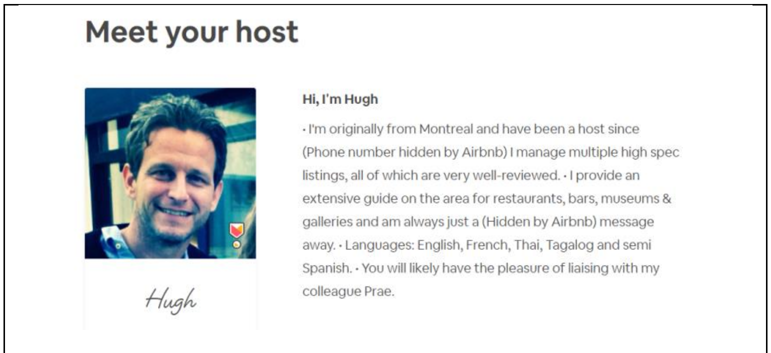

Figure 3. Example of a host's self-description on Airbnb

Data Preprocessing

Before we analyze unstructured text, we removed the mandatory information that the platform required the hosts to fill in. According to the landlord application mechanism in Airbnb, a host must release certain information in order to complete the registration procedure, for example, the phone or email of the host, basic information of the house and basic rental rules. Excluding the information content that must be disclosed from the pool could ensure that we focused on voluntary disclosure only, making the analysis results more accurate.

The most critical step in analyzing unstructured text documents is to convert free-form text into a structured form that can be analyzed. The most popular transformation is to use "word packages" to represent text, that is, a set of documents to be analyzed (usually called a "corpus") using document-term matrix (Kosala and Blockeel 2000). However, in practice, the dimensions of a matrix tend to grow too large, creating computational and memory challenges. Our data preprocessing included six tasks (see Figure 4), namely, (1) converting text to lowercase, (2) removing special characters and tokenizing them into terms, (3) removing stop words, (4) words stemming and pos tagging, (5) removing high-frequency words and low-frequency words, and (6) constructing a document-term matrix.

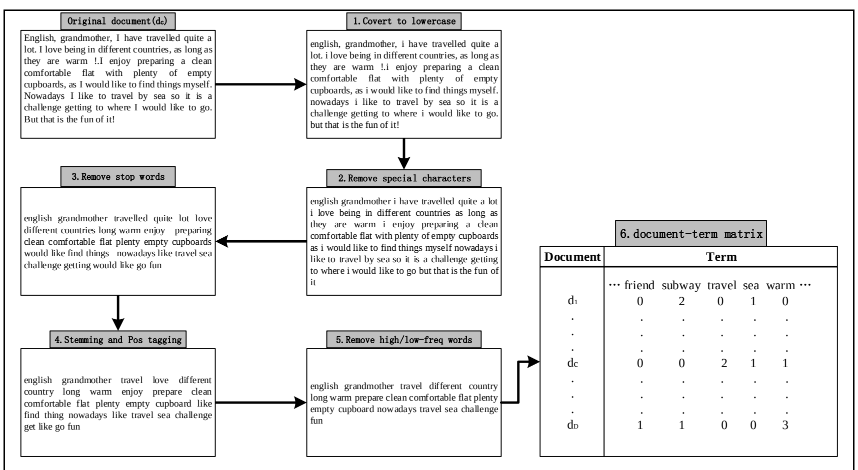

Figure 4. Data preprocessing tasks with a host's self-description with ID

Specifically, in step 1, we converted all the documents into lowercase to provide a consistent format for analysis. Step 2 was removing the special characters, including punctuation marks (e.g., ! % $ # & *?, / .;

"\) and numbers and use regular expressions to exclude phone, date and time and mail, as they never contribute to our text mining result, and then tokenized the document into terms. In step 3, the major step of data preprocessing, stop words—generally a set of commonly used words in any language (here English)—were removed from the document in order to make the focus on important words instead (Mankad et al. 2016). The stop words in the current scenario consisted of some general stop words (for example, “a”, “the”, “I”, etc.) and other stop words customized in the context, depending on special vocabularies related to platform functions or business scenarios (such as place names, scenic spots, etc.) (Wu et al. 2019). The goal of step 4 stemming analysis was to reduce the variation of text data by converting words to their common base form. Since verbs, nouns and adjectives have strong subject meanings in sentences, we filtered the whole speech and only analyzed the words of these three parts of speech. Since high-frequency and low-frequency words in the original documents would have a great obstacle to the correct extraction of topics, in step 5, TF-IDF algorithm was used for word frequency statistics to remove these extreme words (Salton and McGill 1983). The last step was to build a document-term matrix, in which each row corresponded to a document and each column corresponded to a term. The values in the matrix cells represented the frequency of each item in each document. Preprocessing is a very important step in our analysis because it reduced the noise in the data that may seriously affect the performance of LDA and ensured the accuracy of topic modeling.

Latent Dirichlet Allocation (LDA)

Topic modeling refers to a class of algorithms that identify information in a large-scale document set or corpus by discovering hidden “topics” or topics discussed in a set of documents. LDA uses a Bayesian estimation framework for given text data to infer topics (distributions over words) and decompose each document into a mixture of topics. The document to the topic, the topic to the words are all subject to polynomial distribution. The underlying generative process of LDA is illustrated in Figure 5.

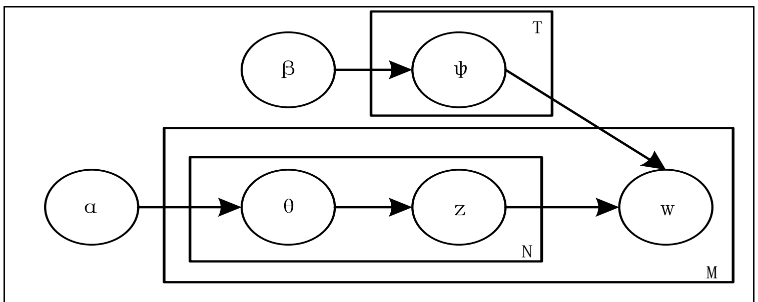

Figure 5. Latent Dirichlet Allocation (LDA) for topic modeling

As shown in Figure 5, the LDA text modeling process consists of two parts: one is $\alpha \rightarrow \theta \rightarrow z$, the other is $\beta \rightarrow \varphi_k \rightarrow w_{m,n}$. In the first part, the topic distribution θ_i of document i is randomly generated according to the Dirichlet distribution α. Then, the topic z_i of the j_{th} word of document i is randomly generated from the topic distribution θ_i. In the second part, the word distribution $\emptyset z_{i,j}$ of topic $z_{i,j}$ is randomly generated according to the Dirichlet distribution β, followed by the random generation of words $w_{i,j}$ from the word distribution $\emptyset z_{i,j}$. In simple terms, LDA can learn the unobservable variables based on the number of topics we input and two prior distributions, and calculate the posterior distribution through Bayesian learning, so that we can clarify the semantic structure hidden in the document (Blei et al., 2003).

According to our settings, LDA outputs two distributions, $P(document|topic)$, which is the probability of the topic output under a given document, and $P(word|topic)$, which is the probability of a word output under the given topic. We calculate the number of topics and topic entropy based on these results.

Computational Processes

Both data preprocessing and LDA analysis were conducted in Python. Regular expressions and the NLTK library were used to preprocess the hosts’ self-description according to the tasks described in the previous section (Joakim 2012). The result of the preprocessing was a document-term matrix, which was imported into the genism (http://radimrehurek.com/gensim) library for LDA analysis. As shown in Figure 5, α, β and the number of topics T were of great significance to the construction of the LDA model. α was a topic smoothing parameter and β was a word smoothing parameter. As suggested by the topic modeling literature, the topic had more semantic meaning when the values of both hyperparameters were 0.1 (Bastani, Namavari, and Shaffer 2019; Hoffman, Bach, and Blei 2010; Kaplan and Vakili 2015). This has also been verified by our numerical analysis, that is, when the values of the hyperparameters α and β equal to 0.1, there turned to have more meaningful outcomes.

The LDA algorithm also needed to input the number of topics T. The literature that used LDA for analysis mostly used the perplexity curve to determine the number of topics (Vu, Li, and Law 2019; Xiang et al. 2017). Perplexity means that in text analysis, the trained model has uncertainty about which topics are contained in certain documents. Therefore, the lower the perplexity, the smaller the uncertainty, and the better the final clustering result. The calculation of the perplexity is shown in the following equation (1):

$$perplexity(D) = exp\left(-\frac{\sum \log P(w)}{\sum_{d=1}^{M} N_d}\right) \tag{1}$$

where N is the total number of words contained in the data set, and $P(w)$ refers to the frequency of different words in the data set. The calculation formula is $P\ (w)\ =\ P\ (z \mid d)\ *P\ (w \mid z)$. Where $P\ (z \mid d)$ represents the probability of certain topics in different documents, and $P\ (w \mid z)$ represents the probability of certain words in different topics. The perplexity curve we drew is shown in Figure 6, and the magnitude of perplexity is reasonable (Mankad et al. 2016). We found that the degree of confusion is the minimum when the number of topics was 50. The optimal number of topics was 50.

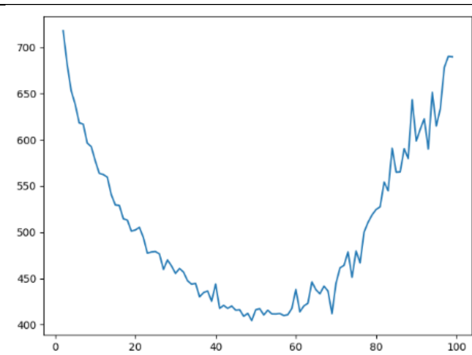

Figure 6. Perplexity under different number of topics

Then, we used the LDAVIS package to visually analyze the feature words under 50 topics. As shown in Figure 7, the circles in the coordinates represented the topics contained in the document, and the area of the circle was proportional to the number of documents containing a certain topic. The red and blue bars represented words that appeared most frequently in one or all documents, and the red part was how often words appeared under a certain topic.

Another output of LDA is $P\ (topic \mid document)$, which is the probability of the topic output under a given document. However, topics with low probability cannot express the meaning of documents well. We used $1/number\ of\ topics$ (i.e. 0.02) as the threshold to filter out topics with probability lower than this threshold. The document-topic matrix was shown in Table 1.

As we mentioned in the research method section, the breadth of self-disclosure was calculated using the number of topics. The depth of self-disclosure was calculated using sentence length (i.e. original sentence length) and topic entropy. According to Table 1, the number of topics contained in each document could be calculated. Also, it is simple to count the length of each piece of self-description. Table 1 showed the corresponding probability of each topic included in a host's self-disclosure. The formula for calculating topic entropy was as follows:

$$Entropy(d) = -\sum_{i=1}^{n} p_i \log p_i \tag{2}$$

where $Entropy\,(d)$ represented the topic entropy of self-description d, n was the number of topics contained in each description based on LDA, and p_i was the probability that self-description d belonged to topic i. The depth and breadth calculation results were shown in Table 2.

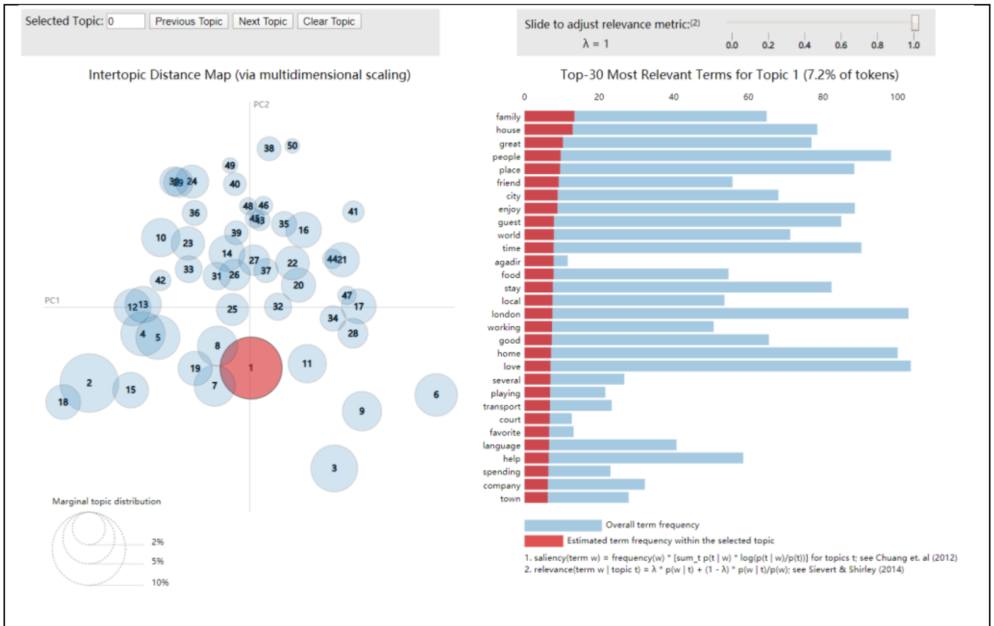

Figure 7. Topic distribution and word distribution

Table 1. Examples of Document-topic Matrix	
Document (D)	Topic (T)
D1	0.727*T7+0.225*T11
D2	0.186*T2+0.081*T5+0.057*T6+0.145*T34+0.301*T44+0.191*T46
D3	0.303*T1+0.537*T45
D4	0.260*T1+0.550*T9+0.089*T18+0.068*T37
D5	0.674*T17+0.299*T46
D6	0.020*T2+0.220*T13+0.040*T16+0.036*T27+0.051*T34+0.042*T35+0.516*T44*0.046*T45+0.020*T49
D7	0.262*T7+0.084*T13+0.075*T15+0.089*T24+0.085*T25+0.342*T46
D8	0.804*T5
D9	0.520*T20+0.384*T31
D10	0.082*T1+0.045*T2+0.274*T5+0.028*T13+0.022*T17+0.082*T18+0.116*T27+0.026*T33+0.021*T36+0.020*T41+0.021*T43+0.031*T45+0.059*T46

Table 1. Example of Document-Topic Matrix

In Table 2, *Breadth* represents the number of topics, *Depth1* means the number of words of the hosts' self-description, and *Depth2* indicates the topic entropy calculated according to the formula 2. As shown in Figure 8, they all show a positive correlation. In the next experiment, we use human coding to determine whether the hypothesis is true.

Table 2. Examples of Breadth and Depth Calculation Results

Document (D)	Breadth	Depth1 (Length)	Depth2 (Entropy)
D1	2	41	0.567
D2	6	37	1.638
D3	2	12	0.696
D4	4	67	1.079
D5	2	67	0.627
D6	9	227	1.510
D7	6	34	1.545
D8	1	12	0.175
D9	2	33	0.707
D10	13	424	2.162

Table 2. Examples of Breadth and Depth Calculation Results

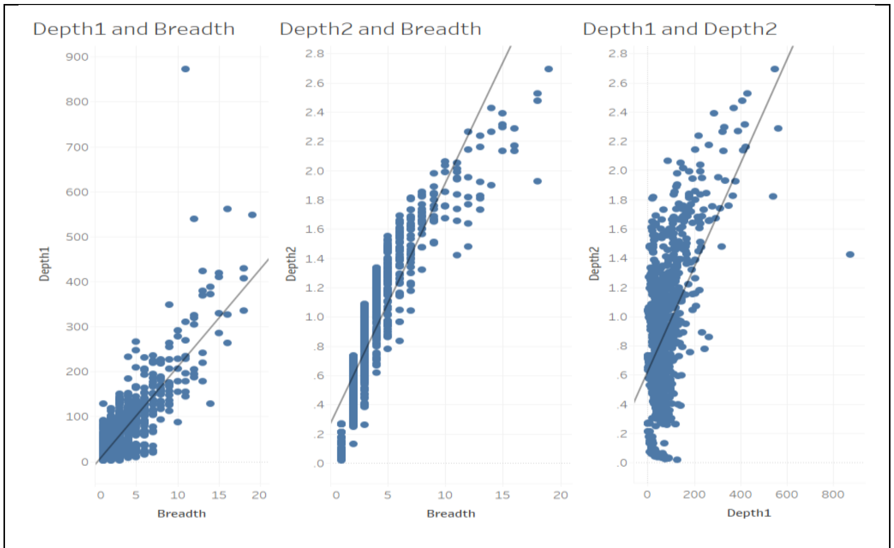

Figure 8. The pairwise relationship between the three indicators

Evaluation of Computational Processes Using Human Coding

In order to test the effectiveness of our algorithm, we conducted three human coding experiments to check the validity and accuracy of topic model. We classified the data based on the results of machine learning calculations. All experiments used the median as the standard (i.e., the data with the median index value as the example data), the median above defined as high and the median below defined as low.

Three coders were invited to manually code the breadth and depth of hosts' self-description. Two coders were male, and all three were between 20-22 years old, college degree and proficient in English. Before the actual coding, we used a small amount of data to train the coders. After all coders had a consistent result on the extent of breadth and depth dimensions of self-description information in the training set, they independently coded 200 self-descriptions (other than the ones in the training set). Krippendorff's Alpha value was used as the reliability statistics between encoders (Hayes and Krippendorff 2007). We used the majority principle to resolve disagreements between coders. We used the precision, recall, and F1-score values to measure the matching results of human coding experiment and algorithm calculations (Wu et al. 2019; Zhao et al. 2014, 2019).

The first experiment used breadth (median was 3) as an indicator in the results of our machine learning classification. We first selected a piece of data with a breadth value of 3 as the example data. Secondly, we asked the coders to rate 20 host self-descriptions for training. The reliability calculation result was 0.861 (see Table 3), which was of high consistency.

Table 3. Krippendorff's Alpha Reliability Estimate

	Alpha	LL95%CI	UL95%CI	Units	Observers	Pairs	Bootstrap samples
Ordinal	0.8613	0.7759	0.9377	20.0000	3.0000	60.0000	10000

Table 3. Krippendorff's Alpha Reliability Estimate

Then, we randomly selected 100 pieces of high-breadth self-description (breadth> 3) and 100 pieces of low-breadth ones (breadth <3) for manual labeling. Examples of labeling were shown in Figure 9.

S/N	host_about	Ecoding values options 1	2	3	4	5	values
153	I am a professional artist and film maker.... well travelled food and music fanatic. Please let me know if you have any questions or queries.	√					1
129	Hello, I'm an Australian living in London. I live in Wanstead which is a lovely area with a village feel, surrounded by lots of green spaces but a few train stops away from the hustle and bustle.	√					1
77	We are a college librarian and a writer who live in a two-story Victorian terrace house with our three cats. We like living in a traditional English house in a historic neighbourhood that's smack in the middle of a vibrant urban environment, it's nice to be able to enjoy all London has to offer, then come home to a calm, relaxing environment. We're happy to show you where to find the hidden gems of London, off the beaten tourist path, and hope that you'll see our house as your home away from home.		√				2
10	I am a principal of an international architecture firm specialising in the design of healthcare facilities. I have always had a keen interest in quality living environments and this led to specialising in the design of healing environments, which major on views of nature, daylight, fresh air and natural materials. That's why I love travelling in Italy which is where you see me in the photo!Building is a passion too. I renovated the first house I lived in myself with two architectural colleagues. We bought all the materials and did all the work, with the exception of some very specialised trades. Through this I have developed skills which allow me to manage the properties I rent (there are only two actually!) and provide an excellent speedy service when things go wrong. I am very particular about keeping things in good condition and repair, so they rarely do go wrong in fact.I am a family man - married with two sons. My wife, Gilla, is a psychotherapist and works for the NHS. Asher, my younger son, is a qualified carpenter-joiner and works for a property maintenance company looking after buildings and clients in the Grosvenor Estate in London. Eliot is at University College London completing Russian Studies. He has lived in Moscow and has a passion for 20th century Russian film.My office is nearby Kestrel House, so I am always available to provide advice and assistance as well as sort out any problems should they arise. I am an accredited landlord with the National Landlords Association (NLA).The reason I bought the apartment is because it epitomises urban living and is fantastically well located to exploit all of the best that London has to offer. It is usually rented to students during the academic year, which is why it is only available through Airbnb during summer - the best time to visit London! I think it offers very good value for money as well as a quintessential 'London' experience. I hope you do too.				√		4
118	Busy working mum who loves living in Hackney. You won't regret booking with us - we're friendly, helpful and give our guests the privacy they require.Happy warning: we have a dog - but he's a big soft border collie, who isn't allowed upstairs.	√					1
137	I am a happy, friendly person who enjoys meeting new people. I work in the health profession as a nurse. I enjoy travel, spa breaks, eating out, socialising with friends, cinema and theatre musicals.	√					1
28	English, Fran?ais, Deutsche, Espa?ol, Italiano, Portugues, 中文 **Hello and welcome. We are experienced landlords and hosted some 10,000+ international guests over many years. We are also keen travelers.As resident landlords we are not subject to London 90 day annual letting limit and your BOOKING IS SAFE and WILL NEVER BE CANCELLED by the authorities because we are letting legally and we pay taxes. Our location is one of the best in London and transport links are the best in London to and from our accommodations. We have detailed policies and house rules so that you know exactly what to expect. In the accommodation you will find further information, maps and instructions for appliances. We enjoy meeting people from around the world, and, time permitting exchanging experiences with our guests. We met many lovely people and made some friends. **Bonjour et bienvenue. Nous sommes des propriétaires expérimentés et avons accueilli plus de 10 000 invités internationaux au cours de nombreuses années. Nous sommes aussi de grands voyageurs.Nous sommes des locateurs résidents vivant dans une autre unité du même batiment. En tant que locateurs ré sidents, nous ne sommes pas soumis à la limite de location annuelle de Londres de 90 jours et votre réservation est s?re et ne sera jamais annul ée par les autorités car nous louons légalement et nous payons des imp?ts. Notre emplacement est l'un des meilleurs de Londres et les liaisons de transport sont les meilleures de Londres vers et depuis nos hébergements. Nous avons des politiques détaillées et des règles internes afin que vous sachiez exactement à quoi vous attendre. Vous trouverez dans l'hébergement des informations supplémentaires, des cartes et des instructions pour les appareils.Nous aimons rencontrer des gens du monde entier et le temps nous permet d'échanger des expériences avec nos invités. Nous avons rencontré beaucoup de gens charmants et nous nous sommes fait des amis.**Hallo und Willkommen. Wir sind erfahrene Vermieter und haben über viele Jahre mehr als 10.000 internationale G?ste empfangen. Wir sind auch begeisterte Reisende.Wir sind ans?ssige Vermieter, die in einer anderen Wohneinheit im selben Geb?ude wohnen. Da wir in London ans?ssige Vermieter sind, unterliegen wir nicht dem j?hrlichen Vermietungslimit von 90 Tagen. Ihre Buchung ist sicher und wird von den Beh?rden niemals storniert, da wir legal vermieten und Steuern zahlen. Unsere Lage ist eine der besten in London und die Verkehrsanbindung von und zu unseren Unterkünften ist die beste in London. Wir haben detaillierte Richtlinien und Hausregeln, damit Sie genau wissen, was Sie erwartet. In der Unterkunft finden Sie weitere Informationen, Karten und Anleitungen für Ger?te.Wir lieben es, Menschen aus der ganzen Welt zu treffen und Erfahrungen mit unseren G?sten auszutauschen. Wir haben viele nette Leute getroffen und ein paar Freunde gefunden.**Hola y bienvenido. Somos propietarios experimentados y hemos acogido a más de 10,000 invitados internacionales durante muchos a?os. También somos viajeros entusiastas.Somos propietarios residentes que viven en otra unidad en el mismo edificio. Como somos propietarios residentes, no estamos sujetos al límite de arrendamiento anual de 90 días de Londres y su RESERVA ES SEGURA y NUNCA SERá CANCELADA por las autoridades porque estamos alquilando legalmente y pagamos impuestos. Nuestra ubicación es una de las mejores en Londres y los enlaces de					√	5

Figure 9. Examples of Labeling Result

Finally, we matched the results of human coding with the classification results of the algorithm. The calculation results were reported in Table 4.

Table 4. Result Analysis of Breadth

Evaluation index	High-breadth	Low-breadth
Precision	100%	81%
Recall	77%	100%
F1-Score	87%	89%

Table 4. Result Analysis of Breadth

In the second experiment, we used depth1 (median was 53) as an indicator and repeated the same steps as in the first experiment. The reliability calculation result was 0.762, which had high consistency. The results of human coding with the classification results of the algorithm were presented in Table 5.

Table 5. Result Analysis of Depth1		
Evaluation index	High-depth1	Low-depth1
Precision	98%	100%
Recall	100%	98%
F1-Score	99%	99%

Table 5. Result Analysis of Depth1

In the third experiment, we used depth2 (median was 0.8767) as an indicator and took the same training and coding. The reliability calculation result was 0.869, which again had high consistency. The results of human coding with the classification results of the algorithm were in Table 6.

Table 6. Result Analysis of Depth2		
Evaluation index	High-depth2	Low-depth2
Precision	98%	89%
Recall	88%	99%
F1-Score	93%	93%

Table 6. Result Analysis of Depth2

Altogether, the calculation results of three experiments showed that the precision, recall, and F1-score values were all high, which demonstrated the effectiveness of the algorithm.

Discussion

This study refines the conceptualization of information disclosure from the SPT perspective, develops a new method of operationalizing information disclosure using the machine learning technique and further evaluates the proposed computation in the context of Airbnb platform. Specifically, using the data from 1,200 Airbnb listings in London, the United Kingdom, we decomposed the hosts' self-description through topic modeling, using the number of topics to measure the breadth dimension of disclosure and the sentence length and topic entropy to measure its depth dimension. We validated the proposed method by comparing the results of the algorithm with human-coded scores in three experiments. The validation results proved the effectiveness of the algorithm.

Implications for Theory and Practice

Our study contributes to information disclosure and privacy research by introducing a theory-driven approach novel approach of examining information disclosure, i.e., measuring both breadth and depth dimensions of information disclosure using the machine learning method. Prior literature has overwhelmingly treated disclosure as a survey-based, subjective, and unidimensional construct. We argue that online users differ in their privacy sensitivity and may exhibit quite deviating behavioral patterns from the privacy preferences that they claim to have (Cao et al. 2018; Smith et al. 2011), which has challenged the use of survey methodology in measuring users' information disclosure. Based on the factual data of users' voluntary disclosure, the proposed new method of operationalization can substantially improve the validity and reliability of information disclosure measurement.

Second, as an important part of the calculation process for measuring breadth and depth of information disclosure, comprehensive data preprocessing is performed and described in detail. In particular, the step of constructing a stop words lexicon is very important. Our study considers not only common stop words but also the context-specific words that are mandated disclosure by the sharing economy platforms, shedding a light on future research of developing machine learning algorithms. Finally, we also provide clear evaluation information that could be referenced by future studies.

Practically, our approach could be scaled up easily to compute online users' information disclosure in various e-commerce settings. Businesses will be able to accurately understand their consumers' disclosure behavior or privacy concerns by calculating information disclosure breadth and depth automatically and precisely on a very large scale. With this understanding, they could potentially cluster the consumers and explore corresponding initiatives to better engage with consumers and boost the business. For the hosts of short-term rental platforms, since the prior studies show that a higher level of information disclosure may increase the perceived trustworthiness of a host (Ma et al. 2017), they may leverage the findings of this study to improve their self-description. For example, based on our results, hosts can add more nouns, adjectives, verbs to increase the number of topics (i.e., breadth) of their self-disclosure. Last but not least, while the current study focuses on the example of Airbnb, the proposed operationalization method is equally applicable to different kinds of businesses like online sellers (regarding how they describe the products or services), online dating and Internet-based broker services, etc.

Limitations and Future Research

This study does have certain limitations. First, we follow the prior literature to employ length and entropy as proxies of the depth of information disclosure. Although both are pervasive indicators of information quality widely accepted in academia, they do not capture the degree of intimacy of the topic disclosed or information sensitivity. Caution must be exercised when this depth measure is used practically. We encourage future studies to incorporate additional methods (such as semantic analysis) to further improve the depth measure. Secondly, this study only uses the Airbnb housing data in London. Data from other cities, particularly other countries, or culturally diverse regions, may exhibit somewhat different patterns. Future studies could consider including more countries or examining data from other types of business websites. Third, due to the limitation of manpower, we cannot manually code more data in the evaluation procedure, thus the accuracy of the algorithm might need more thorough check. Finally, as the current research aims to propose and validate a new method of measuring information disclosure, we concentrate on the methodological procedures themselves rather than the antecedents of information disclosure or its impacts on consumers. However, when zooming out to a broad picture of behavioral research, various future directions seems promising. For example, by investigating personal characters of hosts (such as gender and cultural differences), actual content of information disclosure by the hosts, and their trustworthiness evaluated by the consumers are likely to generate meaningful findings that may help hosts to write a high-quality self-description. To do so, it is necessary to firstly establish the validity of the proposed measures by following the procedures of behavioral research. It will also be interesting to replicate prior research on information disclosure using the newly developed operationalization and see if different methodologies would matter to the empirical results.

Conclusion

This study applies the SPT to understand information disclosure behavior in the sharing economy platforms. Using the machine learning technique, we provide a new method of measuring the breadth and depth of self-disclosure, which not only has superior reliability and validity compared to the traditional survey method but also makes it possible to rapidly collect the disclosure data at a very large scale. We call for more research on further improving our understanding on information disclosure and its associated mechanisms.

Acknowledgements

The study is supported by funds from National Natural Science Foundation of China (No. 71722005, No. 71790590 and No. 71790594) and from Natural Science Foundation of Tianjin (No. 18JCJQJC45900).

References

Adjerid, I., Peer, E., and Acquisti, A. 2018. "The Paradox of Wanting Privacy but Behaving as If It Didn't Matter," *LSE Business Review*. Retrieved from https://blogs.lse.ac.uk/businessreview/2018/04/19/the-paradox-of-wanting-privacy-but-behaving-as-if-it-didnt-matter/.

Airbnb. 2019. "Building on our commitment to trust," *Airbnb Newsroom*. Retrieved from https://news.airbnb.com/building-on-our-commitment-to-trust/.

Altman, I., and Taylor, D. A. 1973. *Social Penetration: The Development of Interpersonal Relationships*. NY: Holt, Rinehart & Winston.

Awad, N. F., and Krishnan, M. S. 2006. "The Personalization Privacy Paradox: An Empirical Evaluation of Information Transparency and the Willingness to Be Profiled Online for Personalization," *MIS Quarterly* (30:1), pp. 13-28.

Baack, D., Fogliasso, C., and Harris, J. 2000. "The Personal Impact of Ethical Decision: A Social Penetration Theory," *Journal of Business Ethics* (24:1), pp. 39-49.

Bastani, K., Namavari, H., and Shaffer, J. 2019. "Latent Dirichlet Allocation (Lda) for Topic Modeling of the Cfpb Consumer Complaints," *Expert Systems with Applications* (127), pp. 256-271.

Benson, V., Saridakis, G., and Tennakoon, H. 2015. "Information Disclosure of Social Media Users," *Information Technology & People* (28:3), pp. 426-441.

Berg, J. H. 1984. "Development of Friendship between Roommates," *Journal of Personality and Social Psychology* (46:2), p. 346-356.

Blei, D. M. 2012. "Probabilistic Topic Models," *Communications of the ACM* (55:4), pp. 77-84.

Blei, D. M., Ng, A. Y., and Jordan, M. I. 2003. "Latent dirichlet allocation," *Journal of machine Learning research*(*3*:1), pp. 993-1022.

Cao, Z., Hui, K.-L., and Xu, H. 2018. "An economic analysis of peer disclosure in online social communities," *Information Systems Research* (28:3), pp. 546-566.

Chevalier J A, Mayzlin D. 2006. "The effect of word of mouth on sales: Online book reviews," *Journal of Marketing Research*, (43:3), pp. 345-354.

Collins, N. L., and Miller, L. C. 1994. "Self-Disclosure and Liking: A Meta-Analytic Review," *Psychological bulletin* (116:3), p. 457.

Dinev, T., and Hart, P. 2006. "An Extended Privacy Calculus Model for E-Commerce Transactions," *Information systems research* (17:1), pp. 61-80.

Edelman, B. G., and Luca, M. 2014. "Digital Discrimination: The Case of Airbnb.com," *Harvard Business School NOM Unit Working Paper*:14-054. Retrieved from https://www.hbs.edu/faculty/Publication%20Files/Airbnb_92dd6086-6e46-4eaf-9cea-60fe5ba3c596.pdf.

Einhorn, E. 2005. "The Nature of the Interaction between Mandatory and Voluntary Disclosures," *Journal of Accounting Research* (43:4), pp. 593-621.

Ert, E., and Fleischer, A. 2019. "The Evolution of Trust in Airbnb: A Case of Home Rental," *Annals of Tourism Research* (75), pp. 279-287.

Fisher, D. 1984. "A Conceptual Analysis of Self-Disclosure," *Journal for the Theory of Social Behaviour* (14:3), pp. 277-296.

Fresneda, J. E., and Gefen, D. 2019. "A Semantic Measure of Online Review Helpfulness and the Importance of Message Entropy," *Decision Support Systems (125)*, 113-117, DOI: 10.1016/j.dss.2019.113117.

Gu, K., Tao, D., and Qiao, J. F. 2017. "Learning a no-reference quality assessment model of enhanced images with big data," *IEEE transactions on Neural Networks and Learning Systems (29:4)*, pp.1301-1313.

Gunter, U. 2018. "What Makes an Airbnb Host a Superhost? Empirical Evidence from San Francisco and the Bay Area," *Tourism Management* (66), pp. 26-37.

Hamari, J., Sjöklint, M., and Ukkonen, A. 2016. "The Sharing Economy: Why People Participate in Collaborative Consumption," *Journal of the Association for Information Science and Technology* (67:9), pp. 2047-2059.

Hayes, A. F., and Krippendorff, K. 2007. "Answering the Call for a Standard Reliability Measure for Coding Data," *Communication methods and measures* (1:1), pp. 77-89.

Hoffman, M., Bach, F. R., and Blei, D. M. 2010. "Online Learning for Latent Dirichlet Allocation," In *Proceedings of Advances in Neural Information Processing Systems*, J. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta (eds.), 23, pp. 856–864.

Hollenbaugh, E. E., and Ferris, A. L. 2014. "Facebook Self-Disclosure: Examining the Role of Traits, Social Cohesion, and Motives," *Computers in Human Behavior* (30), pp. 50-58.

Hu, J., He, W., and Davis, F. 2020. "When Hosts Disclose Their Private Information on Accommodation Sharing Platforms: An Information Commercialization Perspective," in *Proceedings of the 53rd Hawaii International Conference on System Sciences*. DOI: 10.24251/HICSS.2020.102

Joakim, C. 2012. "Explore Python, Machine Learning, and the Nltk Library," *IBM Developer Works*. Retrieved from https://developer.ibm.com/articles/os-pythonnltk/.

Kaplan, S., and Vakili, K. 2015. "The Double-Edged Sword of Recombination in Breakthrough Innovation," *Strategic Management Journal* (36:10), pp. 1435-1457.

Kim, T., Por, M. H., and Yang, S.-B. 2017. "Winning the Crowd in Online Fundraising Platforms: The Roles of Founder and Project Features," *Electronic Commerce Research and Applications* (25), pp. 86-94.

Kisilevich, S., Ang, C. S., and Last, M. 2012. "Large-Scale Analysis of Self-Disclosure Patterns among Online Social Networks Users: A Russian Context," *Knowledge and Information Systems* (32:3), pp. 609-628.

Knijnenburg, B. P., Kobsa, A., and Jin, H. 2013. "Dimensionality of Information Disclosure Behavior," *International Journal of Human-Computer Studies* (71:12), pp. 1144-1162.

Kosala, R., and Blockeel, H. 2000. "Web Mining Research: A Survey," *ACM Sigkdd Explorations Newsletter* (2:1), pp. 1-15.

Krasnova, H., Spiekermann, S., Koroleva, K., and Hildebrand, T. 2010. "Online Social Networks: Why We Disclose," *Journal of information technology* (25:2), pp. 109-125.

Larrimore, L., Jiang, L., Larrimore, J., Markowitz, D., and Gorski, S. 2011. "Peer to Peer Lending: The Relationship between Language Features, Trustworthiness, and Persuasion Success," *Journal of Applied Communication Research* (39:1), pp. 19-37.

Li, K., Cheng, L., and Teng, C.-I. 2020. "Voluntary Sharing and Mandatory Provision: Private Information Disclosure on Social Networking Sites," *Information Processing & Management* (57:1), p. 102128.

Li, Y. 2014. "A Multi-Level Model of Individual Information Privacy Beliefs," *Electronic Commerce Research and Applications* (13:1), pp. 32-44.

Liang, S., Li, H., Liu, X., and Schuckert, M. 2019. "Motivators Behind Information Disclosure: Evidence from Airbnb Hosts," *Annals of Tourism Research* (76), pp. 305-319.

Liu, Z., Min, Q., Zhai, Q., and Smyth, R. 2016. "Self-Disclosure in Chinese Micro-Blogging: A Social Exchange Theory Perspective," *Information & Management* (53:1), pp. 53-63.

Liu, Z., and Park, S. 2015. "What makes a useful online review? Implication for travel product websites," *Tourism management* (47), pp. 140-151.

Lowry, P. B., Cao, J., and Everard, A. 2011. "Privacy Concerns Versus Desire for Interpersonal Awareness in Driving the Use of Self-Disclosure Technologies: The Case of Instant Messaging in Two Cultures," *Journal of Management Information Systems* (27:4), pp. 163-200.

Lwin, M., Wirtz, J., and Williams, J. D. 2007. "Consumer Online Privacy Concerns and Responses: A Power–Responsibility Equilibrium Perspective," *Journal of the Academy of Marketing Science* (35:4), pp. 572-585.

Ma, X., Hancock, J. T., Lim Mingjie, K., and Naaman, M. 2017. "Self-Disclosure and Perceived Trustworthiness of Airbnb Host Profiles," in *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, pp. 2397-2409.

Mankad, S., Han, H. S., Goh, J., and Gavirneni, S. 2016. "Understanding Online Hotel Reviews through Automated Text Analysis," *Service Science* (8:2), pp. 124-138.

Mudambi, S. M. , and Schuff, D. 2010. "What makes a helpful online review? a study of customer reviews on amazon.com". *MIS Quarterly* (*34*:1), pp. 185-200.

Möhlmann, M. 2015. "Collaborative consumption: Determinants of satisfaction and the likelihood of using a sharing economy option again," *Journal of Consumer Behaviour* (14:3), pp. 193-207.

Omarzu, J. 2000. "A Disclosure Decision Model: Determining How and When Individuals Will Self-Disclose," *Personality and Social Psychology Review* (4:2), pp. 174-185.

Osatuyi, B., Passerini, K., Ravarini, A., and Grandhi, S. A. 2018. "'Fool Me Once, Shame on You... Then, I Learn.' An Examination of Information Disclosure in Social Networking Sites," *Computers in Human Behavior* (83), pp. 73-86.

Petty, R. E., Cacioppo, J. T., and Schumann, D. 1983. "Central and Peripheral Routes to Advertising Effectiveness: The Moderating Role of Involvement," *Journal of Consumer Research* (10:2), pp. 135-146.

Racherla, P., and Friske, W. 2012. "Perceived 'usefulness' of online consumer reviews: An exploratory investigation across three services categories," *Electronic Commerce Research and Applications* (11:6), pp. 548-559.

Salton, G., and McGill, M. J. 1983. *Introduction to Modern Information Retrieval*. NY: McGraw-Hill.

Shannon, C. E. 1948. "A Note on the Concept of Entropy," *Bell System Technical Journal* (27:3), pp. 379-423.

Smith, H. J., Dinev, T., and Xu, H. 2011. "Information Privacy Research: An Interdisciplinary Review," *MIS Quarterly* (35:4), pp. 989-1016.

Soundararajan, R. and Bovik, A. C. 2012. "RRED Indices: Reduced Reference Entropic Differencing for Image Quality Assessment," *IEEE Transactions on Image Processing (21:2),* pp. 517-526.

Sultan, S., and Chaudry, H. 2008. "Gender-Based Differences in the Patterns of Emotional Self-Disclosure," *Pakistan Journal of Psychological Research* (23:2), pp. 107-122.

Sun, Y., Wang, N., Shen, X.-L., and Zhang, J. X. 2015. "Location Information Disclosure in Location-Based Social Network Services: Privacy Calculus, Benefit Structure, and Gender Differences," *Computers in Human Behavior* (52), pp. 278-292.

Teubner, T., and Flath, C. M. 2019. "Privacy in the Sharing Economy," *Journal of the Association for Information Systems* (20:3), pp. 213-242.

Tussyadiah, I. P., and Park, S. 2018. "When Guests Trust Hosts for Their Words: Host Description and Trust in Sharing Economy," *Tourism Management* (67), pp. 261-272.

Unhjem, J. V., Vatne, S., and Hem, M. H. 2018. "Transforming Nurse-Patient Relationships—A Qualitative Study of Nurse Self-Disclosure in Mental Health Care," *Journal of Clinical Nursing* (27:5-6), pp. e798-e807.

Vu, H. Q., Li, G., and Law, R. 2019. "Discovering Implicit Activity Preferences in Travel Itineraries by Topic Modeling," *Tourism Management* (75), pp. 435-446.

West, R., and Turner, L. H. 2018. *Introducing Communication Theory: Analysis and Application*. NY: McGraw-Hill.

Wu, J., Huang, L., and Zhao, J. L. 2019. "Operationalizing Regulatory Focus in the Digital Age: Evidence from an E-Commerce Context," *MIS Quarterly* (43:3), pp. 745-764.

Xiang, Z., Du, Q., Ma, Y., and Fan, W. 2017. "A Comparative Analysis of Major Online Review Platforms: Implications for Social Media Analytics in Hospitality and Tourism," *Tourism Management* (58), pp. 51-65.

Xu, H., Dinev, T., Smith, J., and Hart, P. 2011a. "Information Privacy Concerns: Linking Individual Perceptions with Institutional Privacy Assurances," *Journal of the Association for Information Systems* (12:12), Article 1.

Xu, H., Luo, X. R., Carroll, J. M., and Rosson, M. B. 2011b. "The Personalization Privacy Paradox: An Exploratory Study of Decision Making Process for Location-Aware Marketing," *Decision Support Systems* (51:1), pp. 42-52.

Xu, H., Teo, H.-H., Tan, B. C., and Agarwal, R. 2009. "The Role of Push-Pull Technology in Privacy Calculus: The Case of Location-Based Services," *Journal of Management Information Systems* (26:3), pp. 135-174.

Xu, X., Zeng, S., and He, Y. 2021. "The impact of information disclosure on consumer purchase behavior on sharing economy platform Airbnb," *International Journal of Production Economics* (231), DOI: 10.1016/j.ijpe.2020.107846.

Yan, J. J. 2001. "A Note on Proactive Password Checking," In *Proceedings of 2001 ACM New Security Paradigms Workshop,* ACM Press, pp.127-135.

Yasui, R., Kanai, A., Hatashima, T., and Hirota, K. 2010. "The Metric Model for Personal Information Disclosure," in Proceedings of *Fourth International Conference on Digital Society*, IEEE, pp. 112-117.

Zlatolas, L. N., Welzer, T., Heričko, M., and Hölbl, M. 2015. "Privacy Antecedents for SNS Self-Disclosure: The Case of Facebook," *Computers in Human Behavior* (45), pp. 158-167.

Zhao, H., Wu, L., Liu, Q., Ge, Y., and Chen, E. 2014. "Investment Recommendation in P2P Lending: A Portfolio Perspective with Risk Management," *International Conference on Data Mining*, pp. 1109-1114. DOI: 10.1109/ICDM.2014.104.

Zhao, H., Jin, B., Liu, Q., Ge, Y., Chen, E., Zhang, X., and Xu, T. 2019. "Voice of Charity: Prospecting the Donation Recurrence & Donor Retention in Crowdfunding," *IEEE Transactions on Knowledge and Data Engineering* (1:1), DOI: 10.1109/TKDE.2019.2906199.

View publication stats