The Korean Economic Review
Volume 33, Number 1, Summer 2017, 5-34.

Information Quality of Online Reviews in the Presence of Potentially Fake Reviews*

Wonho Song** · Sangkon Park*** · Doojin Ryu****

Online reviews are important in the evaluation of product quality. This paper seeks to assess information quality of online reviews using the TripAdvisor data for Korean hotels. We first estimate the review model developed by Dai, Jin, Lee, and Luca (2012) and show that high-quality reviews contain most of the information for the quality of hotels. Second, we assess the degree of distortions caused by fake reviews through numerical experiments and show that the distortions of fake reviews are serious. Third, we compare the simple average and weighted average aggregation methods. Weighted average method is better than simple average in finding the quality of hotels but it is more vulnerable to fake reviews. Fourth, we suggest excluding low-quality reviews to deal with fake reviews and show that the benefit of avoiding serious distortions from potentially fake reviews is greater than the cost of losing information from low-quality reviews.

JEL Classification: D70, L83, L86, M37
Keywords: Online Review, Fake Review, Rating, Aggregation, Numerical Experimentation, Tourism Management

I. Introduction

Advances in information technology have rapidly changed the tourism industry in recent decades (Buhalis and Law, 2008; Litvin, Goldsmith, and Pan, 2008). Many online consumer review websites, such as yelp.com and tripadvisor.com, have

Received: Aug. 18, 2016. Revised: Nov. 29, 2016. Accepted: Feb. 17, 2017.

* We thank the anonymous referees for their helpful comments. All remaining errors are of course our own.

** First Author, Associate Professor, School of Economics, Chung-Ang University, 84 Heukseok-ro, Dongjak-gu, Seoul 06974, Republic of Korea. E-mail: whsong@cau.ac.kr, Phone: +82-2-820-5493.

*** Associate Research Fellow, Tourism Policy Research Office, Korea Culture & Tourism Institute, 154 Geumnanghwa-ro, Gangseo-gu, Seoul 07511, Republic of Korea. E-mail: sgpark@kcti.re.kr, Phone: +82-2-2669-8483.

**** Corresponding Author, Professor, College of Economics, Sungkyunkwan University, 25-2, Sungkyunkwan-ro, Jongno-gu, Seoul 03063, Republic of Korea. E-mail: sharpjin@skku.edu, Phone: +82-2-760-0429.

emerged and accumulated masses of information about the quality of products. Given the amount of information, it is increasingly difficult and time-consuming to read and understand all the data. Thus, providing an average of review ratings based on an arithmetic or simple average is a popular shortcut. One important question is whether average ratings correctly reflect the true quality of a product. A lot of social learning literature examines conditions under which social learning takes place, including consistent estimation of the underlying quality of a product. See Ifrach, Maglaras, and Scarsini (2015), Besbes and Scarsini (2015), and the references therein.

Other research suggests ways to aggregate information more efficiently than using the simple average method. Dai, Jin, Lee, and Luca (2012) present one such example, a Bayesian method to optimally aggregate review information and assign different weights to elite and non-elite reviewers. Their intuition is that to optimally aggregate review content, more weight needs to be put on reviews with more information. Thus, it would be more efficient to assign higher weights to experienced reviewers who posted many reviews and to new reviews that might reflect recent changes in product quality. The latter is appropriate for products such as hotel services whose quality can vary over time. Dai et al. (2012) present an econometric model for online reviews and show that their new aggregation method performs much better than the conventional simple average method.

As mentioned above, cumulated reviews together with more efficient aggregation methods are helpful in finding the true quality of a product. However, fake reviews pose a serious obstacle to that. Fake reviews are falsely generated by the sellers themselves to encourage sales or by their competitors to discourage sales (Hu, Bose, Koh, and Liu, 2012). Fake reviews thus severely distort the true quality of a product and lower social welfare by encouraging consumers to buy what they actually do not want to buy (Besbes and Scarsini, 2015). Many studies have worked to detect fake reviews, but the results remain unsatisfactory (see for example Akoglu, Chandy, and Faloutsos, 2013; Li and Hitt, 2008; Mukherjee, Liu, and Glance, 2012; Ott, Choi, Cardie, and Hancock, 2011); it is still very difficult to detect fake reviews.

Therefore, in this paper we add to the literature by examining how harmful fake reviews actually are and suggesting ways to mitigate or prevent those harmful effects without having to detect the fake reviews. Although these questions are practically important, to our knowledge, they remain unanswered. More specifically, we will compare the performances of the simple average method and Dai et al. (2012)'s weighted average method–the optimal aggregation (OA) method hereafter–when fake reviews are present. We examine the OA method as an alternative to the simple average method because it is quite general: it allows group heterogeneities by dividing reviewers into elite/non-elite groups; different motives such as altruism or a simple desire to share a personal experience with a product (as in Besbes and Scarsini, 2015); changes to product quality over time; and individual biases that

emerge from individual experiences. Moreover, it uses review content more efficiently than the simple average method by updating information in a Bayesian way.

We apply the model of Dai et al. (2012) in two ways to estimate the quality of hotels in Korea and analyze the effects of fake reviews. First, we categorize reviewers into three groups in terms of the precision of their reviews: elite, non-elite, and low-quality (defined below) reviewers. The last group is a benchmark group and includes low-quality reviews, single-time reviewers, and potentially manipulated reviews (as pointed out by Feng, Xing, Gogar, and Choi, 2012). One way to partially mitigate the distortions of fake reviews might be assigning lower weights to this third group. Second, we pay special attention to fake or manipulated reviews. If we can identify fake reviews, it is obviously better to eliminate them. The problem is the difficulty in recognizing the existence of fake reviews and identifying them. In those circumstances, a strategy to exclude only fake reviews is not feasible. One possible strategy is to exclude low-quality reviews that might contain fake reviews. We investigate whether that strategy is effective when fake reviews exist and how much efficiency loss results when fake reviews do not exist.

More specifically, we use the OA method to estimate the parameters that best describe the characteristics of reviewers of Korean hotels listed in the TripAdvisor in order to make our numerical experiments more realistic and more practically relevant. With those estimated parameters, we run numerical experiments to investigate how different strategies perform in the estimation of hotel's quality. When only normal reviews are included, the results show that the OA method performs much better than the simple average (SA) method, as already shown in Dai et al. (2012). Strategies that exclude low-quality reviewer groups perform similarly to the ones that include all reviewers, which means that primary information on the quality of hotels is concentrated in the high-quality reviews. That is, information loss from the exclusion of low-quality reviewer groups should be minor. When fake reviews are added to the review ratings, however, the strategies show substantially different performance. Fake reviews can heavily distort quality estimates of hotels, especially when OA methods are used. On the other hand, strategies that exclude low-quality reviews containing fake reviews are not affected by fake reviews and can successfully estimate the quality of hotels with relatively minor information loss.

The rest of this paper is organized as follows. Section 2 surveys the previous literature on issues regarding the aggregation of review contents and fake reviews. Data and estimation models are discussed in Section 3, and the models are estimated in Section 4. We examine various strategies for estimating the true quality of hotels in Section 5. Section 6 concludes.

II. Related Works

Online reviews are gaining popularity and their impact on consumer behavior has been a common focus of recent studies on tourism and hospitality (Filieri and McLeay, 2014; Kim, Chung, and Lee, 2011; Mauri and Minazzi, 2013; Schuckert, Liu, and Law, 2015; Wu, Wall, and Pearce, 2014). Studies of online reviews regarding the estimation of product quality are closely related to literature on social learning.

The literature on social learning started from the seminal papers of Banerjee (1992) and Bikhchandani, Hirshleifer, and Welch (1992). Since then, numerous studies have been done on various issues such as limited information (Celen and Kariv, 2004), heterogeneous consumer types and convergence issues (Goeree, Palfrey, and Rogers, 2006), partial observations (Herrera and Hörner, 2013), and so on. Ifrach et al. (2015) and Crapis et al. (2017) deal with online review issues and discuss the conditions under which social learning occurs.

Besbes and Scarsini (2015) share a similar goal with Dai et al. (2012) in modeling online reviews, but they focus on social learning and present a different structure for reviews. They divide reviewers into naive and sophisticated customers. They analyze an informational setting in which customers observe only the sample mean of past reviews and ask whether customers can recover the true quality of the product based on the feedback they observe. They mainly discuss cases in which customers report their subjective experiences with the product. If consumers are sufficiently heterogeneous, biases can occur, but sophisticated consumers can correct for the selection bias, which leads to social learning. Such self-selection bias is also observed by Li and Hitt (2008) and Racherla, Connolly, and Christodoulidou (2013). Besbes and Scarsini (2015) also discuss the effect of manipulation on the average of reviews in a population of naive customers.

Dai et al. (2012) also divide reviewers into two groups, elite and non-elite reviewers. Unlike Besbes and Scarsini (2015), they consider both altruistic motives and subjective feelings. Here again, individual heterogeneity can generate bias. Thus, they devise an optimal aggregation (OA) method to correct for such biases. The OA method assigns different weights to elite/non-elite reviewers and old/recent reviews. However, Dai et al. (2012) do not consider fake reviews, and the effects of fake reviews on the results from the OA method are not known. Biases from individual heterogeneity are an inherent issue in the aggregation of information, but fake reviews are another type of distortion that demands serious investigation.

Fake review problems are a serious issue, and many studies have been undertaken to find the patterns and characteristics of fake reviews (see Akoglu et al., 2013; Mukherjee et al., 2012, for example). In response to fake reviews, yelp.com devised its own filtering algorithm that has been running for several years (Streitfeld,

2012). Mukherjee, Venkataraman, Liu, and Glance (2013) examined various ways of detecting fake reviews to find out how yelp.com filters fake reviews. Hu et al. (2012) examined textual information available in online reviews by combining sentiment mining techniques with readability assessments and showed that manipulation through a component of writing style that reflects the background of an individual, such as sentiments, is able to significantly affect a consumer's purchase decision. Hu, Liu, and Sambamurthy (2011) studied the temporal patterns of online reviews, and concluded that they cannot rule out manipulation as one of the potential drivers. Hu, Bose, Gao, and Liu (2011) developed discretionary manipulation proxy to study the management of online reviews based upon the discretionary accrual-based earnings management framework and revealed the existence of online review manipulation for books. Luca and Zervas (2016) examined reviews filtered by yelp.com and discussed economic incentives for review fraud.

There are two major ways of finding fake reviews (Mukherjee et al., 2013): analyzing the review text and analyzing the behavior or characteristics of the reviewers. Textual analysis has made some progress with recent developments in computer technology (Hu et al., 2012; Ott et al., 2011; Peng and Zhong, 2014), but it also has limitations because manipulators could learn how to avoid detection by studying observed regularities. Behavioral analysis has been partially successful in identifying fake reviews using reviewer information such as the number of reviews, proportion of positive reviews among total reviews, number of words in reviews, and so on (see for example, Feng et al., 2012; Mukherjee, Liu, Wag, Glance, and Jindal, 2011). However, when fake reviews are mixed among normal reviews, identifying them is still a difficult task (O'Connor, 2008).

Mayzlin, Dover, and Chevalier (2014) try to detect fake reviews indirectly by comparing the review ratings of hotels listed in both expedia.com and tripadvisor.com. At expedia.com, only consumers who actually purchased products can leave reviews, whereas the TripAdvisor has no such restrictions. The researchers thus identify the differences in review ratings between the two websites as differences caused by review manipulation. Mayzlin et al. (2014)'s study successfully uses one site's review ratings as a credible benchmark, but that method is unable to identify exactly which reviews are fake. They also find characteristics common among hotels that try to leave fake reviews, but the finding does not guarantee that such hotels actually left fake reviews.

In this paper, we do not define fake reviews in a precise manner. Our focus is not to isolate fake reviews from normal reviews but to assess the effects of fake reviews on review qualities. Thus, instead of using tightly defined concept of fake reviews, we use a few salient features of fake reviews reported in the previous studies. They are as follows. First, fake reviews pose a serious problem because they are left for a specific hotel over a short time (Mukherjee et al., 2011). Second, the fundamental

reason that hotels leave fake reviews is to change the average review ratings. To do that, fake reviewers are likely to use extreme review ratings such as 5-star or 1-star rather than 3- or 4-stars (Luca and Zervas, 2016). Third, the extreme reviews are left mainly by reviewers with few previous reviews (Luca and Zervas, 2016). Feng et al. (2012) also point out that reviewers with fewer than 10 previous reviews are not credible. In our sample, multi-time reviewers left less-extreme reviews than single-time reviewers. In summary, fake reviews are likely to be extreme and posted within a short time period, mainly by reviewers with few previous reviews. This definition of fake reviews is not tight enough to filter out all the fake reviews on a review platform. However, our main purpose is to assess the degree of distortions caused by review fraud, not to isolate them, and we employ a few known features of fake reviews reported in previous studies to mimic fake reviews in our simulation exercises.

III. Estimation Methods and Data

3.1. Estimation Method

3.1.1. Model

We use the review model of Dai et al. (2012) to describe the estimation process of the quality of hotels and review rating decisions of reviewers, as briefly introduced below. The purpose of this paper is to analyze the effects of the information quality of reviews on average review ratings. That is, this paper investigates how much information is contained in a review when a reviewer efficiently uses information on past reviews and reviewer's attributes and its effects on average review ratings. This is important because review ratings and their information contents are gaining more importance for customers when they make decisions. In this respect, Dai et al. (2012)'s review model provides one good example in which customers use information contents of online reviews efficiently. In their model, a reviewer combine personal experience (signal) and best guess for the true quality of hotels. For details of the model, see Dai et al. (2012).

A reviewer i writes a review for a hotel h at calendar time t_n as the nth reviewer of h. She observes her own signal s_{ht_n} as well as the $n-1$ reviews of h before her, $\{x_{ht_1}, x_{ht_2}, \cdots, x_{ht_{n-1}}\}$. s_{ht_n} is assumed to be an unbiased but noisy signal of the true quality μ_{ht_n} such that $s_{ht_n} = \mu_{ht_n} + \varepsilon_{ht_n}$ where $\varepsilon_{ht_n} \sim N(0, \sigma_i^2)$. Dai et al. (2012) consider two incentives for a reviewer i to report. First, a reviewer gains personal satisfaction from reporting her own signal with certain deviation, which Dai et al. (2012) call stringency $\theta_{hn} \neq 0$. Second, a reviewer tries to make a best guess of hotel quality so that she can earn popularity or reputation in the future.

Combining the two incentives, reviewer i as the nth reviewer of hotel h, chooses her review x_{ht_n} to minimize the following objective function:

$$F_{hn} = (1-\rho_i)(x_{ht_n} - s_{ht_n} - \theta_{hn})^2 + \rho_i (x_{ht_n} - E(\mu_{ht_n} \mid x_{ht_1}, x_{ht_2}, \cdots, x_{ht_{n-1}}, s_{ht_n}))^2$$

where $E(\mu_{ht_n} \mid x_{ht_1}, x_{ht_2}, \cdots, x_{ht_{n-1}}, s_{ht_n})$ is the posterior belief of true quality μ_{ht_n} and ρ_i is the weight that i puts on her popularity on the TripAdvisor. The optimal review that minimize F_{hn} is

$$x_{ht_n} = \lambda_{hn} + (1-\rho_i) s_{ht_n} + \rho_i E(\mu_{ht_n} \mid x_{ht_1}, x_{ht_2}, \cdots, x_{ht_{n-1}}, s_{ht_n}) \quad (1)$$

where $\lambda_{hn} = (1-\rho_i)\theta_{hn}$ represent the stringency or bias of reviewer i for hotel h. Thus, the best expectation of μ_{ht_n} and the nth reviewer's own observed signal s_{ht_n} have the weights ρ_i and $1-\rho_i$, respectively. To those terms, the nth reviewer adds an individual stringency term λ_{hn} to finally determine her review ratings x_{ht_n}. Dai et al. (2012) show that an elite reviewer's ρ_i is larger than that of a non-elite reviewer. That is, an elite reviewer assigns more weight than a non-elite reviewer to the expectation of the true quality of the hotel, which is why elite reviewers are more important in determining the true quality of hotels.

3.1.2. Quality Change

The true quality of a hotel h at time t, μ_{ht}, is assumed to follow a martingale process:

$$\mu_{ht} = \mu_{h,t-1} + \xi_{ht} \quad (2)$$

where $\xi_{ht} \sim \text{iid } N(0, \sigma_\xi^2)$. For the choice of a martingale process over other statistical processes (such as AR(1)), see Dai et al. (2012, p.6).

If $\xi_{ht} = 0$ $(t = 1, 2, \ldots)$, then the quality will not change. However, this assumption is unrealistic except for factory-made products, such as books or TV sets. Service products, such as restaurants or hotels, require assumptions of quality change because a switch of managers can substantially change the service quality.

Signal is the quality observed by a reviewer with the noise ϵ_{ht_n}. Thus,

$$s_{ht_n} = \mu_{ht_n} + \epsilon_{ht_n} \quad (3)$$

where ϵ_{ht_n} is iid normal, and the conditional variance of ϵ_{ht_n} is $\text{Var}(s_{ht_n} \mid \mu_{ht_n}) = \sigma_i^2$. This variance σ_i^2 is the precision of the observation by the nth reviewer. A small σ_i^2 means that the nth reviewer precisely observes the true quality (μ_{ht_n})

of hotels and vice versa. Dai et al. (2012) divided reviewers into two (elite/non-elite) groups. Elite reviewers are good observers of quality and have smaller variance than non-elite reviewers. In this study, we further divide non-elite reviewers into two groups and use the low-quality reviewers as a benchmark group.

3.1.3. Reviewer Heterogeneity

Reviewers are heterogeneous in stringency (λ_{hn}). Reviewers determine ratings based not only on the observable quality of hotels but also on their individual characteristics, such as past experiences, type of travel, etc. Hence, different reviewers can leave different review ratings even if they observe the same signals from hotels. The following stringency terms reflect such individual heterogeneity.

First, the number of previous reviews ($NumRev_{it}$). If a reviewer has more previous experience with other hotels, s/he could be harsher or more generous than new hotel visitors. Second, the frequency of reviews ($FreqRev_{it}$). We treat a person who leaves two reviews within 10 months differently from a person who leaves two reviews within 1 month. Third, the matching difference ($MatchD_{hit}$) between the reviewer and the hotel. This variable indicates how well a reviewer and a hotel are matched. Fourth, the taste variety ($TasteVar_{it}$) of reviewers. This variable measures how many types of hotels a reviewer has experienced so far. If this variable is large, a person visits hotels for various purposes and vice versa. Specific definitions of these variables are detailed in the Appendix A1.

In addition to the above reviewer heterogeneities, we add to the stringency term the number of calendar days (Age_{ht}) since a hotel received its first review. This variable attempts to capture any linear trend in reviews that is missed by the above reviewer heterogeneity variables.

3.1.4. Estimation of Model

The model introduced above is quite general enough to encompass many special models, such as a linear quality model. For example, as mentioned in Dai et al. (2012, p.13), Age_{ht} can be interpreted as a linear trend of changes in true quality. In addition, setting $\lambda_{hn} = 0$ assumes that no individual heterogeneity derives from reviewer stringency. In the model as described, reviews are allowed to be serially dependent, i.e., the current review rating depends on past review ratings. If we set $\rho_i = 0$, the current review rating stops depending on past ratings and becomes independent.

The review model requires estimating the parameters of the model (the weights, ρ_i, for the expected quality of hotels; the precision of signal, σ_i^2; and individual stringency, λ_{hi}) for each type of reviewer (elite, non-elite and benchmark). The stringency term consists of the aforementioned five factors:

$$\lambda_{hi} = \lambda_{0i} + Age_{ht}\,\lambda_{Age,i} + NumRev_{it}\,\lambda_{NumRev,i} + FreqRev_{it}\,\lambda_{FreqRev,i}$$

$$+MatchD_{hit}\,\lambda_{MatchD,i}+TasteVar_{it}\,\lambda_{TasteVar,i} \tag{4}$$

for $i=e,ne,bm$, indicating elite, non-elite and benchmark reviewer groups, respectively. λ_{0i} is a constant. The parameters are estimated using the Maximum Likelihood Estimation method, and the reader is referred to Dai et al. (2012) for the details of the estimation.

3.1.5. Optimal Aggregation Method

Dai et al. (2012) suggest the following optimal aggregation method. The estimated quality of a hotel h is the conditional expectation of quality given all the review ratings information, i.e.,

$$E(\mu_{ht_n}\mid x_{ht_1},x_{ht_2},\cdots,x_{ht_{n-1}},s_{ht_n})=E(\mu_{ht_n}\mid s_{ht_1},s_{ht_2},\cdots,s_{ht_{n-1}},s_{ht_n})=\mu_{n|n}\,.$$

Errors of μ_{ht_n} and s_{ht_n} follow the normal distribution, and the conditional expectation and variance are obtained as follows (see Appendix of Dai et al., 2012):[1]

$$\mu_{n|n}=\frac{\sigma^2_{n|n-1}}{\sigma^2_{n|n-1}+\sigma^2_n}s_n+\frac{\sigma^2_n}{\sigma^2_{n|n-1}+\sigma^2_n}\mu_{n|n-1} \tag{5}$$

$$\sigma^2_{n-1|n-1}=\frac{\sigma^2_{n-1|n-2}\sigma^2_{n-1}}{\sigma^2_{n-1|n-2}+\sigma^2_{n-1}} \tag{6}$$

where $\mu_{n|n-1}=\mu_{n-1|n-1}$ and $\sigma^2_{n|n-1}=\sigma^2_{n-1|n-1}+\Delta t\sigma^2_\xi$. Here, σ^2_n is the variance that indicates the precision of signal of nth reviewer, as mentioned before. This aggregation method assigns higher weights to recent reviews and high-quality reviewers (reviewers with low σ^2_i). Next, finding $\mu_{n|n}$, requires obtaining s_n. Given review ratings x_n and λ_n, computed using reviewers' profiles, s_n is obtained as:

$$s_n=\frac{1}{\left[1-\left(1-\frac{\sigma^2_{n|n-1}}{\sigma^2_{n|n-1}+\sigma^2_n}\right)\rho_n\right]}\left[x_n-\lambda_n-\rho_n\frac{\sigma^2_n}{\sigma^2_{n|n-1}+\sigma^2_n}\mu_{n|n-1}\right]. \tag{7}$$

3.2. Data

Online media are an increasingly important information source for travelers (Liu and Park, 2015; Sparks and Browning, 2011; Xiang and Gretzel, 2010). Among the search engines for travel information, the TripAdvisor is growing in popularity and

[1] We omit the hotel index h hereafter to simplify the notations following Dai et al. (2012).

its data have been widely used for the analysis of online reviews (Ayeh, Au, and Law, 2013; O'Connor, 2010; Wu et al., 2014). Thus, we gather the data used in our research from tripadvisor.com for the period between June 2004 and September 2013 for Korean hotels. In this study, as in Dai et al. (2012) and Besbes and Scarsini (2015), the classification of reviewers plays an important role. The TripAdvisor classifies reviewers into six levels by the total number of reviews they have posted to the website. The number of reviews (Review Count) listed in Table 1 indicates the number of total reviews posted in the whole TripAdvisor website. The other values regarding the number of reviewers and reviews are for Korean hotels only.

[Table 1] Summary Statistics

Panel A: Statistics per Reviewer Levels

Level	Symbol	Review Count	Reviewer No	Review No	Mean	Stan. Dev.
Top Contributor	L1	(50,∞)	2,885	5,567	3.96	0.86
Senior Contributor	L2	(20,50]	4,036	6,337	4.01	0.87
Contributor	L3	(10,20]	3,121	4,423	4.00	0.89
Senior Reviewer	L4	(5,10]	2,539	3,284	4.01	0.94
Reviewer	L5	(2,5]	2,409	2,923	3.99	1.00
None	L6	(0,2]	3,615	3,755	3.90	1.23
Total			18,605	26,289	3.98	0.95

Note: Review count means the number of total previous reviews in the whole TripAdvisor website. The other values are for Korean hotels only.

Panel B: Statistics per Hotels

Variable	Mean	Median	Min	Max	Stan. Dev.
# of reviews	62.15	12	1	1129	133.53
Reviewer level	3.27	3.15	1	6	0.97
Ratings	3.50	3.50	1	5	0.72
Age	1286.46	1237.00	1	3388.00	994.29
Matching	0.27	0.24	0.03	0.99	0.17
Taste variety	0.07	0.05	0.00	0.39	0.06
Frequency	0.25	0.20	0.00	1.00	0.22
# of previous reviews	0.80	0.60	0.00	9.00	0.96

As shown in Panel A of Table 1, a total 18,605 reviewers made 26,289 reviews about 423 Korean hotels between June 2004 and September 2013. Group L2 has the largest number of reviews and reviewers among the groups. The average rating is 3.98, and each level has similar averages. Group L6 has the lowest average at 3.90. Standard deviations become larger as the group levels go down and is largest in L6. Thus, L6 is somewhat different from the other groups in terms of the average and standard deviation of its review ratings. Panel B shows the summary statistics per

hotel. It shows the distributions of the number of reviews, reviewer levels, review ratings and reviewer's heterogeneity variables of hotels.

IV. Estimation Results

Before estimating the review model, we categorize reviewers into three groups. The TripAdvisor divides reviewers into 6 levels by the number of reviews. Thus, we first test how different each two adjacent levels are. Specifically, we test the null hypothesis that two adjacent groups are equal in the sense that they have same parameters, i.e., $H_0 : \sigma_i^2 = \sigma_j^2, \rho_i = \rho_j, \lambda_i = \lambda_j$ ($i, j =$ adjacent levels) against the alternative, $H_1 : \sigma_i^2 \neq \sigma_j^2$ *or* $\rho_i \neq \rho_j$ *or* $\lambda_i \neq \lambda_j$. We ran a likelihood ratio test and got the results shown in Table 2.

[Table 2] Results of Likelihood Ratio Test

	Restricted model	Unrestricted model	LR test
(L1,L2)	-13,507.2	-13,496.7	0.0072
(L2,L3)	-12,349.2	-12,343.0	0.1359
(L3,L4)	-9,012.4	-9,003.3	0.0199
(L4,L5)	-7,685.8	-7,675.4	0.0079
(L5,L6)	-9,268.2	-9,211.0	0.0000

In Table 2, the null hypotheses are rejected except for the pair (L2,L3). That is, each level has different parameters, but levels 2 and 3 share the same parameters. Thus, we put L2 and L3 in the same group. This naturally leads to using L1 as the elite group, L2 and L3 as the non-elite group, and L6 as the benchmark group. To determine where to put L4 and L5 reviewers, we examined the full dataset more closely.

[Table 3] Model Specifications

Model	Group	Stringency term	log likelihood
M1	(1,2,3,4,5,6)	λ_{bm}	-32,733.6
M2	(1)(2,3,4,5,6)	λ_{bm}	-32,636.2
M3	(1)(2,3,4,5,6)	λ_e , λ_{bm}	-32,630.9
M4	(1)(2,3)(4,5,6)	λ_{bm}	-32,412.7
M5	(1)(2,3)(4,5,6)	λ_e , λ_{ne} , λ_{bm}	-32,403.2
M6	(1)(2,3,4)(5,6)	λ_{bm}	-32,337.0
M7	(1)(2,3,4)(5,6)	λ_e , λ_{ne} , λ_{bm}	-32,328.8
M8	(1)(2,3,4,5)(6)	λ_{bm}	-32,305.2
M9	(1)(2,3,4,5)(6)	λ_e , λ_{ne} , λ_{bm}	-32,297.2

Table 3 lists various specifications using the full dataset. M1 is the model in which reviewers are not grouped. M2 divides reviewers into elite and non-elite groups. M3 is a variation of M2 with the stringency terms estimated separately for the two groups. M3 is the one estimated in Dai et al. (2012). M4 and the other models divide reviewers into three groups. M4, M6, and M8 estimate stringency terms only for the benchmark group whereas M5, M7, and M9 estimate stringency terms for each group. The results of a few tests show which specification best fits the data.

[Table 4] Test Results for All Models

Hypothesis	LR Test	p-value
M1 = M2	194.8287	0.0000
M2 = M3	10.5170	0.0618
M2 = M4	447.0519	0.0000
M2 = M6	598.3428	0.0000
M2 = M8	661.9525	0.0000
M4 = M5	18.9926	0.0404
M6 = M7	16.3353	0.0904
M8 = M9	15.9424	0.1013

Table 4 shows the test results for the comparison of models. When we compare M1 and M2, the null hypothesis that the two groups share parameters is strongly rejected. That is, we can divide reviewers into elite and non-elite groups. When we compare M2 and M3, the results show that stringency terms can be estimated separately for each group at the 10% significance level. Similarly, when we compare M2 with M4, M6, and M8, the results show that it is better to divide reviewers into three groups than into two. Separate estimation of stringency terms is overall preferred, although the result between M8 and M9 is slightly less than significant at the 10% level. Finally, among the three-group models, M5, M7, and M9, model M9 has the highest likelihood function value. Therefore, we choose model M9 as the benchmark model for the subsequent analyses.

Estimation results of the review model parameters are shown in Table A1 in the Appendix. The results show that elite reviewers have the smallest variance and the difference in variances between elite and non-elite reviewers is not large. On the other hand, the difference in variance between non-elite and benchmark reviewer groups is large, indicating that benchmark reviewers are different from the other two groups. Thus, the benchmark reviewers are low-quality reviewers in the sense that their variance estimate, σ_{bm}^2, for signal observation is large.

Next, we use the parameters in model M9 and run numerical experiments to investigate which strategy performs best in finding the true quality of hotels under different situations.

V. Numerical Experiments

In this section, we use the parameters estimated from reviewer data for Korean hotels and generate review ratings under various assumptions on true quality changes. This is to approximate the review structure of Korean hotels as closely as possible in order to make our numerical experiments more realistic and more practically relevant. For given ratings, we examine which strategy correctly estimates the true quality of hotels.

5.1. Data Generating Process

Consider the following data generating processes (DGPs). The true quality of a hotel h is assumed to change according to the following rules.

$$\text{DGP } 1: \mu_n = \mu_{n-1}$$
$$\text{DGP } 2: \mu_n = \mu_{n-1} + \xi_n$$
$$\text{DGP } 3-1: \mu_n = \mu_0 + 0.001 \times n$$
$$\text{DGP } 3-2: \mu_n = \mu_0 - 0.001 \times n$$
$$\text{DGP } 4-1: \mu_n = \begin{cases} \mu_0 & \text{for } n = 1,\ldots,149 \\ \mu_0 - 0.5 & \text{for } n = 150,\ldots,299 \\ \mu_0 + 0.5 & \text{for } n = 300,\ldots,600 \end{cases}$$
$$\text{DGP } 4-2: \mu_n = \begin{cases} \mu_0 & \text{for } n = 1,\ldots,149 \\ \mu_0 + 0.5 & \text{for } n = 150,\ldots,299 \\ \mu_0 - 0.5 & \text{for } n = 300,\ldots,600 \end{cases}$$

for nth review where $\xi_n \sim N(0, 0.0016)$ and $\mu_0 = 3$. DGP1 assumes that the quality does not change. As explained before, this assumption is unrealistic. On the other hand, DGP2 assumes that quality changes randomly and irregularly. In that case, a fixed true quality does not exist, and we present the mean squared error (MSE) instead of the average estimated quality. DGP3 assumes a linear quality change with slope coefficients ± 0.001 estimated from the data. DGP4 assumes that the quality changes discontinuously, which could happen in hotel reviews when, for example, a new manager changes a hotel's management strategy or a hotel's interior is upgraded through renovations. We run numerical experiments 1,000 times and report the average values of the estimated quality of hotels.

To generate the reviewer's stringency term, we randomly choose 600 reviews from among a total of 26,289 reviews. The average number of reviews for hotels with a review number greater than 100 reviews is 295. Here we assume that there are fake reviews among the next 300, resulting in a total of 600 reviews. Using

equation (4) and the chosen reviewer profiles, we generate λ_n. With the quality change introduced above and equation (3), we generate s_n. Then, using equation (5), we generate $\mu_{n|n}$. Finally, we combine s_n, $\mu_{n|n}$, and λ_n to generate x_n using equation (1) and the parameters estimated from model M9. When reviewer n's rating is x_n, we compute simple average (SA) as $(1/N)\Sigma_{n=1}^{N}x_n$ and optimal aggregation (OA) using equation (5) and (7). The SA uses information for $\{x_n\}_{n=1}^{N}$ only, whereas the OA method uses $\{x_n, z_n\}_{n=1}^{N}$, i.e., review ratings x_n and reviewer profiles z_n. Thus, OA has the potential to improve upon SA.

Next, we consider cases in which fake reviews exist among true reviews. Most review platforms such as yelp.com or tripadvisor.com developed filters to detect fake reviews and provide average ratings after filtering out fake reviews. If the filtering process is successful, we will not observe the distortions of fake reviews in the filtered reviews. Thus, we assess the distortions of fake reviews by inserting simulated fake reviews to the above generated reviews. In this paper, we do not attempt to list all the features of fake reviews because our focus is not to isolate fake reviews from normal reviews but to assess the effects of fake reviews on review qualities. Thus, we use a few salient features of fake reviews reported in the previous studies to mimic fake reviews in our simulation exercises.

Hence, first, among the 600 sample reviews, we insert fake reviews after the 300th review with a probability of 30%–70%. The probability of 70% might look large and unlikely, but it could happen in real life. For example, in the analysis of the Yelp dataset, Luca and Zervas (2016) found that the fraction of reviews that are filtered is 79% for the businesses that were caught in the sting. Thus, the ratio of filtered reviews for businesses attempting review fraud could be very high because fake reviews are left for a specific hotel for a short period of time (Mukherjee et al., 2011). Second, we assume that fake reviewers leave extreme review rating such as 5-star or 1-star rather than 3- or 4-stars. The fundamental reason that hotels leave fake reviews is to change average review ratings. To do that, fake reviewers are likely to use extreme review ratings. This is well evidenced in the analysis of Feng et al. (2012) and Luca and Zervas (2016). Third, fake reviewers are mainly level 5 or 6 reviewers. Feng et al. (2012) point out that reviewers with fewer than 10 previous reviews (L4 and higher in our case) are not credible.

Fake reviewers may also manage profiles, thus they may post mediocre reviews (3- or 4-stars) and multiple reviews more than 10 reviews for a long time. However, the marginal cost of generating one fake review will increase as fake reviewers post normal reviews to hide their identity and they would rather post fake reviews using new user id than manage profiles. Thus, we expect that the percentage of fake reviews among a reviewer's total reviews is a decreasing function of the number of previous reviews. This is well evidenced in the analysis of the Yelp dataset by Luca and Zervas (2016, Figure 2(b)). Considering this, we consider two cases for the percentage of fake reviews. In the first, only L6 reviewers leave fake reviews with a

probability of 70%. In the second, L4 reviewers also leave fake reviews with a probability of 30%, L5 reviewers with a probability of 50%, and L6 reviewers with a probability of 70%. In the following simulation exercises, the magnitude of distortions caused by fake reviews depends on the assumptions we take here. However, our qualitative results in the conclusion do not change.

5.2. Strategies

We consider the following strategies to aggregate the information contained in review ratings.

Strategies
[SA1] Simple Average
[SX5] Simple Average excluding L6
[SX3] Simple Average excluding L4–L6
[OA1] Optimal Aggregation with one group
[OA2] Optimal Aggregation with two groups
[OA3] Optimal Aggregation with three groups
[OX5] Optimal Aggregation with three groups excluding L6
[OX3] Optimal Aggregation with three groups excluding L4–L6

SA1 is the usual SA. SX5 is the SA after excluding only L6 reviewers, and SX3 is the SA after excluding L4–L6 reviewers. OA1 is the OA method without division of reviewers. OA2 divides reviewers into two groups, (1)(2,3,4,5,6), as in Dai et al. (2012). OA3 is the OA with three groups, (1)(2,3,4,5)(6). OX5 is like OA3 but excluding L6 reviewers, and OX3 is like OA3 but excluding L4–L6 reviewers.

We check the performances of the different strategies in different ways. First, as mentioned above, DGP is based on the parameters from model M9, which divides the reviewers into three groups. Therefore, we may compare the performances of strategies when they divide the reviewers into one or two groups only. When we use model M3 for DGP, the performances of the OA strategies are similar; thus we omit the results.

Second, when reviews are normal (i.e., no fake reviews), we check the performance of SX5, SX3, OX5, and OX3 that all exclude some review information. Because the reviews are normal, the reviews from L4–L6 reviewers have valid information on the quality of hotels. Excluding those reviews causes information loss, so we check the degree of loss.

Third, we consider a case in which fake reviews exist only among L6 reviewers. In that case, SX5, SX3, OX5, and OX3 are unaffected by fake reviews. On the other hand, SA1, OA1, OA2, and OA3 are affected by fake reviews. We look to see how much distortions the fake reviews cause to the performances of the different

strategies. Here again, SX3 and OX3 exclude some normal reviews and could suffer from information loss. We thus compare the performance of SX3 and OX3 with that of SX5 and OX5. We also compare the performance of SA with those of OA. OA assigns higher weights to more recent information; thus it could be more affected by fake reviews than SA.

Fourth, we consider a case in which fake reviews exist among L4–L6 reviewers. In that case, only SX3 and OX3 are unaffected by fake reviews. SX5 and OX5 fail to eliminate the fake reviews from the L4 and L5 reviewers. Thus, for this case, we compare the performance of SX3 and OX3 with that of SX5 and OX5.

Fifth, fake reviews can be positive (5-star) or negative (1-star). The former is generally done by the hotel itself and the latter by its competitors. We determine the effects of each case on the performances of the various strategies.

5.3. Results

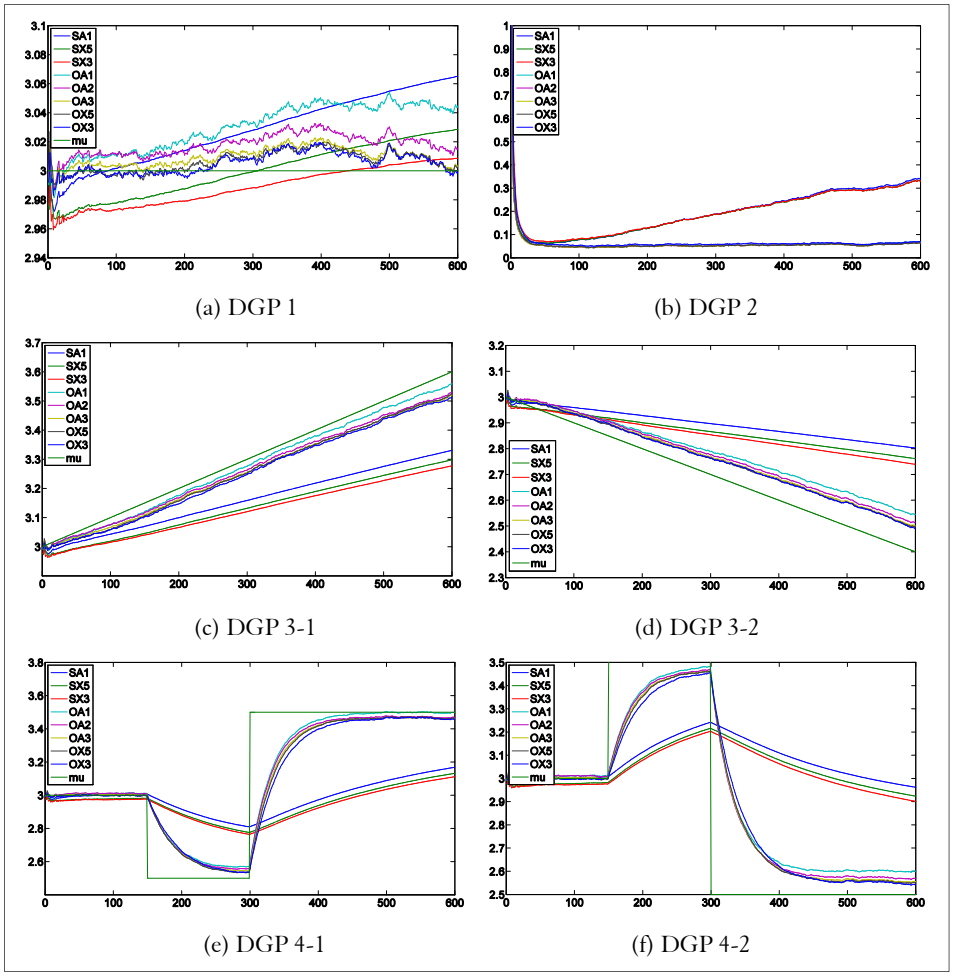
In this subsection, we provide the simulation results for the DGPs considered in Section 5.1. We expect that the strategies that exclude low-quality reviews (strategies SX5, SX3, OX5, and OX3) show worse performance than those that do not because they suffer from information loss. On the other hand, these strategies will not be affected by the existence of fake reviews because the fake reviews contained in the low-quality reviews are all excluded in the aggregation process. Thus, there exist cost and benefit in excluding low-quality reviews and we compare the magnitudes of them in the following simulation experiments.

Case 1: Normal Reviews

Figure 1 shows the average quality of a hotel estimated by each strategy when reviews are normal. DGP1 assumes a fixed quality. The SAs suffer from bias caused by reviewers' stringency terms, and their estimates deviate from the true quality, $\mu_t = 3$, and keep going up. OA1 also deviates from the true quality because it is not correctly adjusted for stringency terms. The other strategies show estimates that converge to true quality.

DGP2 clearly shows the superiority of OA over SA strategies. The MSEs of OA strategies remain stable over time, but those of SA increase. More specifically, SA1 is slightly higher than SX5 and SX3, but the difference is only minor. Likewise, the MSE of OX3 is slightly higher than those of the other OAs, but the difference is negligible. Thus, the results show that excluding low-quality reviews in the aggregation does not cause serious information loss. Also, as Dai et al. (2012) noted in their analysis of restaurants, the heterogeneity of each reviewer group did not significantly affect the results of the OA strategy. In the case of hotels, each reviewer group shows heterogeneity, especially between the benchmark group and the other

[Figure 1] Comparison of Strategies: Normal Reviews



Note: For DGP2, MSEs are provided. For the other cases, averages of the estimated hotel quality are presented. SA1: Simple Average; SX5: Simple Average excluding L6; SX3: Simple Average excluding L4–L6; OA1: Optimal Aggregation with one group; OA2: Optimal Aggregation with two groups; OA3: Optimal Aggregation with three groups; OX5: Optimal Aggregation with three groups excluding L6; OX3: Optimal Aggregation with three groups excluding L4–L6.

groups. With normal reviews, however, the heterogeneity does not cause large differences among the OA strategies.

DGP3 assumes linearly increasing or decreasing quality. Comparing SA and OA, OA approaches true quality more quickly than SA. In addition, SA strategies deviate further from true quality as time goes on, unlike OA strategies. Different OA strategies show similar performances, but OA1 displays slightly higher estimates than the other OA strategies.

DGP4 shows discontinuous quality changes. Figure 1 shows that OA quickly approaches the new quality, whereas SA adjusts very slowly. OA1 also approaches the new quality quickly, but its estimates are a little different from the other OA strategies. As before, the estimates of SA1 and OA1 are higher than those of the other strategies. Also, excluding low-quality reviews does not make much differences among the strategies.

We can summarize the results as follows. First, in comparing OX5 and OX3 with OA3, the results are almost same. Thus, excluding the information provided by the L4–L6 reviewers does not cause serious information loss in estimating the true quality of a hotel. The important information for the estimation of true quality is mostly concentrated in the reviews from the L1–L3 reviewers.

Second, as discussed in the previous section, reviewer groups show heterogeneity, and dividing reviewers into three groups is expected to improve the efficiency of the estimators. However, the differences among OA1, OA2 and OA3, are negligible. Thus, as in Dai et al. (2012), reviewer heterogeneity plays a relatively small role, and Bayesian updating is the dominating factor in the superiority of the OA algorithm.

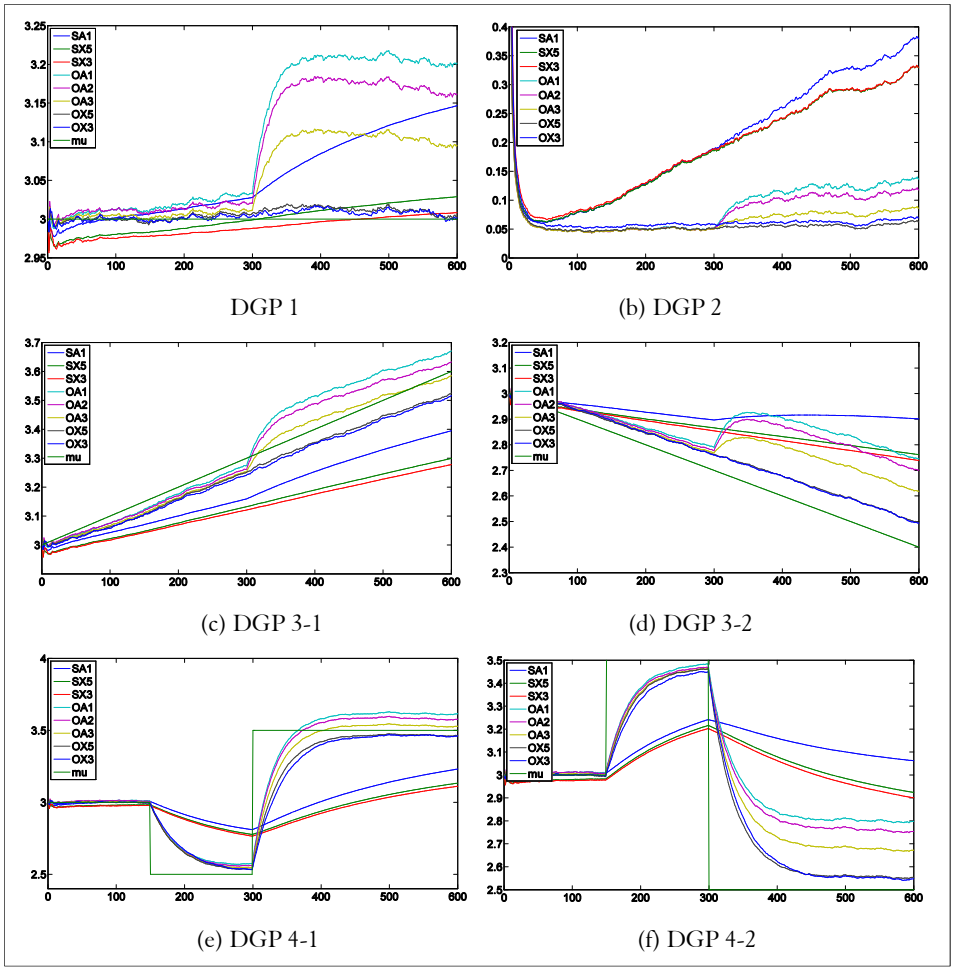
Case 2: Fake Reviews from L6 Reviewers

Figure 2 shows the performances of strategies when 5-star fake reviews are present among the L6 reviewers. For DGP1, all the OA strategies have large biases, even larger than that of SA1. OA1 has the largest bias, and OA2 and OA3 follow in order, which implies that assigning a smaller weight to the benchmark group partially reduces the bias derived from the fake reviews in L6 but fails to fully eliminate them. On the other hand, OX5 and OX3 are unaffected by the fake reviews because they exclude reviews from L6 reviewers. SA1 also suffers from the fake reviews, but SX5 and SX3 exclude L6 reviewers and are thus unaffected.

The results of DGP2 are similar. Thus, assigning a smaller weight to fake reviews is insufficient to prevent damage from them. That is, the large bias in OA3 means that the negative effects of the fake reviews on the estimation of true quality are larger than the positive effects from the normal reviews. On the other hand, OX3 and OX5 converge to true quality, confirming that the important information in estimating the true quality of hotels is mainly contained in the reviews from the L1–L3 reviewers, and the normal reviews from L6 reviewers do not convey additional information. Therefore, excluding the low-quality reviews that contain fake reviews improves the efficiency of estimating true quality.

One of the fundamental differences between the SA and OA strategies is updating information about true quality changes. When quality changes, SA assigns equal weight to old reviews and adjusts to new quality very slowly. On the other hand, OA assigns higher weight to recent reviews and adjusts to new quality

[Figure 2] Comparison of Strategies: 5-Star Fake Reviews in L6



Note: For DGP2, MSEs are provided. For the other cases, averages of the estimated hotel quality are presented. SA1: Simple Average; SX5: Simple Average excluding L6; SX3: Simple Average excluding L4–L6; OA1: Optimal Aggregation with one group; OA2: Optimal Aggregation with two groups; OA3: Optimal Aggregation with three groups; OX5: Optimal Aggregation with three groups excluding L6; OX3: Optimal Aggregation with three groups excluding L4–L6.

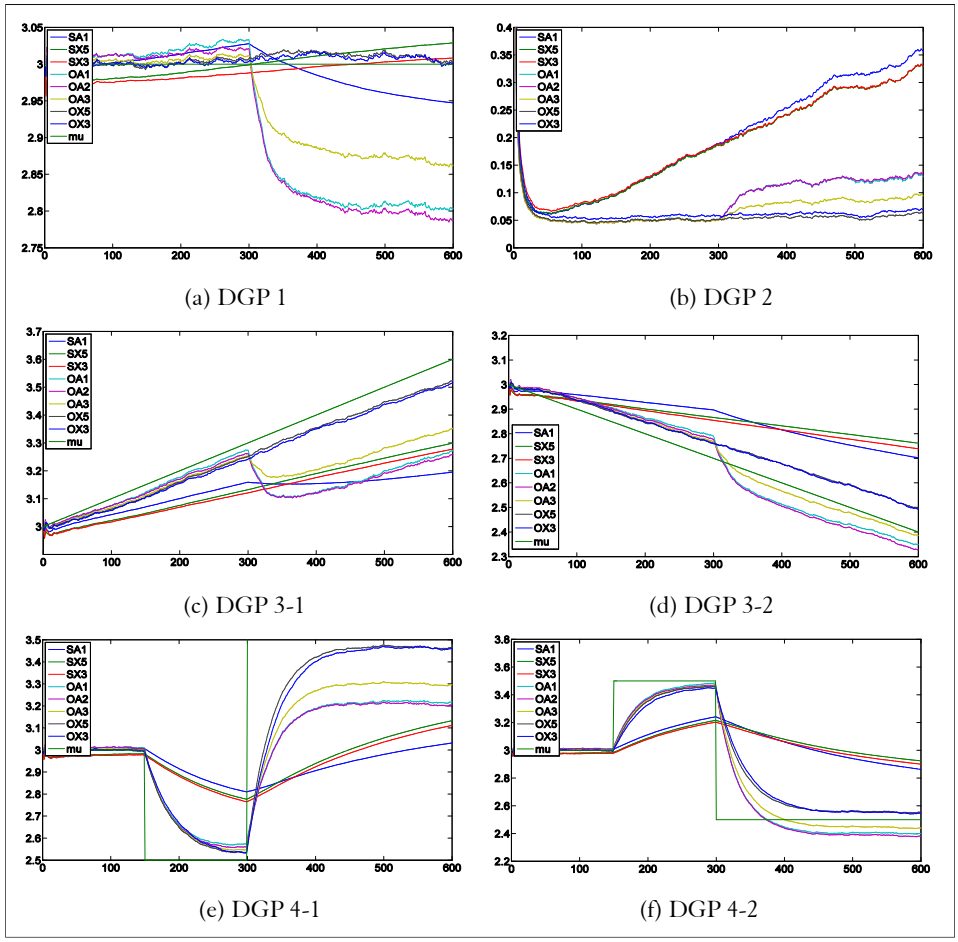
very quickly. However, this advantage of OA turns to critical weakness in the presence of fake reviews. That is, OA updates the inaccurate information from the fake reviews, and thus the distortions get even larger. Therefore, eliminating inaccurate information from fake reviews is crucially important in estimating true quality.

DGP3 and DGP4 present similar results. It is also worth noting that when true quality increases, bias from 5-star fake reviews is not easily noticeable. When true

quality decreases, however, bias from 5-star fake reviews becomes large, and the estimates deviate from the true quality by more than 0.5-star.

Our results are robust when fake reviews are inserted before the 300th reviews rather than after the 300th reviews (the results are not shown here to save space). In this case, the effects on both SA and OA are huge when fake reviews exist. After the 300th reviews, however, OA method returns to true quality very quickly after the 300th reviews, as we have seen in the case 1 where only normal reviews exist, but SA method does not return to true quality for a long time.

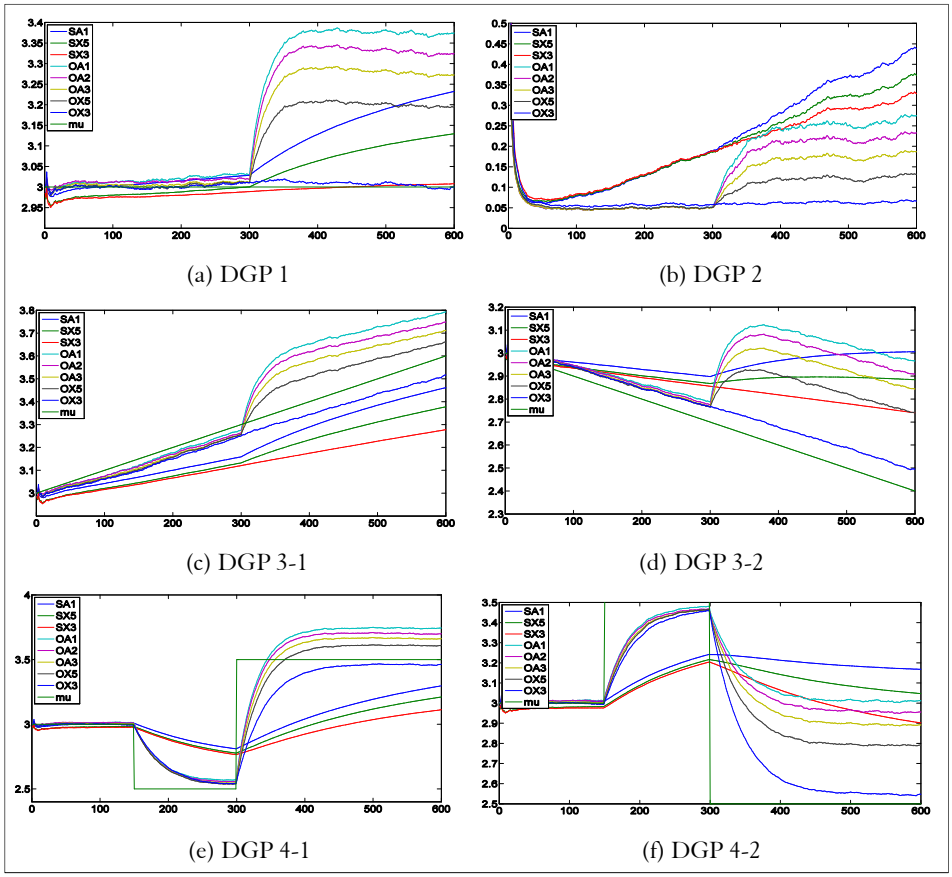
[Figure 3] Comparison of Strategies: 1-Star Fake Reviews in L6



Note: For DGP2, MSEs are provided. For the other cases, averages of the estimated hotel quality are presented. SA1: Simple Average; SX5: Simple Average excluding L6; SX3: Simple Average excluding L4–L6; OA1: Optimal Aggregation with one group; OA2: Optimal Aggregation with two groups; OA3: Optimal Aggregation with three groups; OX5: Optimal Aggregation with three groups excluding L6; OX3: Optimal Aggregation with three groups excluding L4–L6.

Figure 3 shows the performance of each strategy when 1-star fake reviews are present in reviews from L6 reviewers. As before, SX5, SX3, OX5, and OX3 are unaffected by the fake reviews. The other strategies suffer from biases caused by the fake reviews, especially when true quality increases. Individual bias terms also increase slightly as shown in Figure 1 (a). However, they are small in magnitude and they do not significantly affect the results of each strategy when 1-star fake reviews are present. Thus, the overall results are similar to those in Figure 2 except that the direction of the bias is downward.

[Figure 4] Comparison of Strategies: 5-Star Fake Reviews in L4–L6

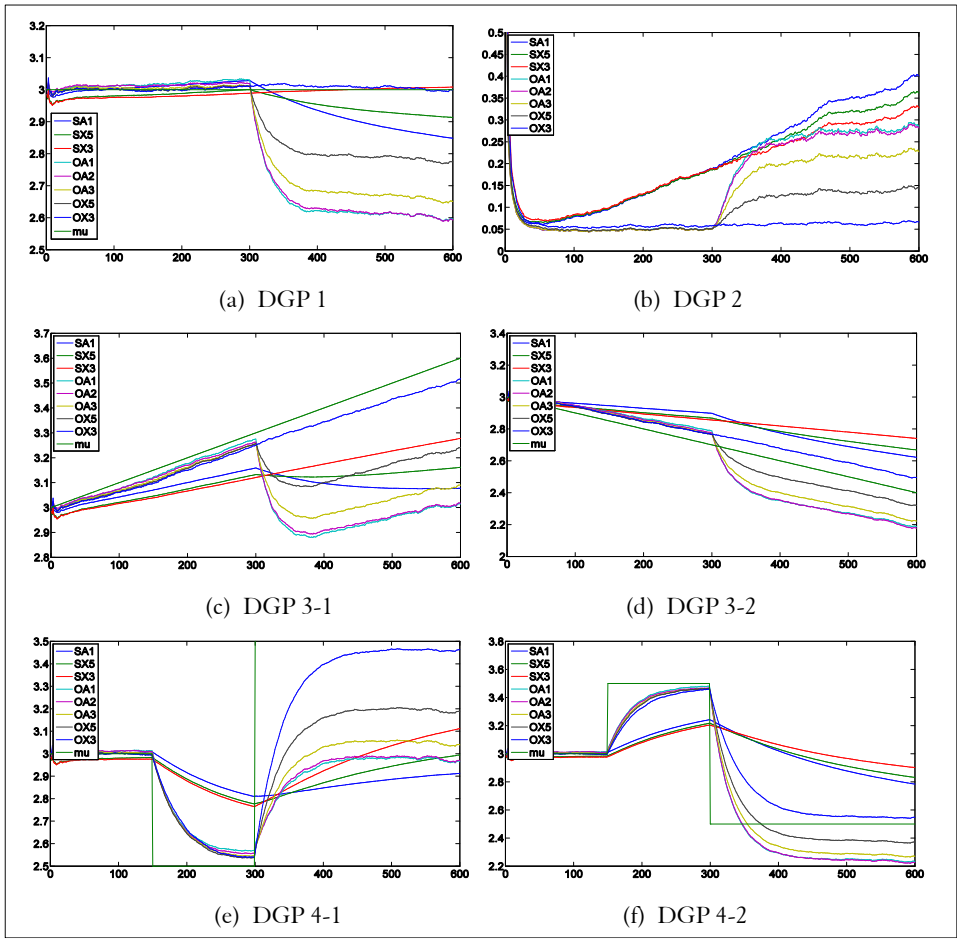


Note: For DGP2, MSEs are provided. For the other cases, averages of the estimated hotel quality are presented. SA1: Simple Average; SX5: Simple Average excluding L6; SX3: Simple Average excluding L4–L6; OA1: Optimal Aggregation with one group; OA2: Optimal Aggregation with two groups; OA3: Optimal Aggregation with three groups; OX5: Optimal Aggregation with three groups excluding L6; OX3: Optimal Aggregation with three groups excluding L4–L6.

Case 3: Fake Reviews from L4–L6 Reviewers

Figure 4 shows the performance of each strategy when 5-star fake reviews are present in reviews from L4–L6 reviewers. Because fake reviews are also in L4 and L5, SX5 and OX5 are not unbiased anymore and suffers from bias. The strategies immune to fake reviews in this case are SX3 and OX3, and they display unbiased results. When there are no fake reviews, SX3 and OX3 perform similarly to the other strategies, and even with fake reviews from three levels of reviewers, SX3 and

[Figure 5] Comparison of Strategies: With 1-Star Fake Reviews in L4-L6



Note: For DGP2, MSEs are provided. For the other cases, averages of the estimated quality of hotel are presented. SA1: Simple Average; SX5: Simple Average excluding L6; SX3: Simple Average excluding L4–L6; OA1: Optimal Aggregation with one group; OA2: Optimal Aggregation with two groups; OA3: Optimal Aggregation with three groups; OX5: Optimal Aggregation with three groups excluding L6; OX3: Optimal Aggregation with three groups excluding L4–L6.

OX3 are not affected from fake reviews. Among the strategies considered, however, OX3 performs best in various circumstances.

With fake reviews in L4–L6, the bias in the other strategies becomes much larger. In DGP1, the OA methods perform even worse than the SA strategies. In DGP4-2, for example, the bias is so large that the OA strategies stay at 3-stars when the true quality is 2.5 stars. The OX5 strategy that excludes L6 reviewers performs slightly better than the other OA strategies, but it still suffers from bias.

Figure 5 shows the performance of each strategy when 1-star fake reviews are present in reviews from L4–L6 reviewers. Here again, SX3 and OX3 are unbiased estimator of true quality. The other strategies suffer severely from the fake reviews. Among the strategies considered, OX3 performs best. For example, in DGP1 and DGP4-1, the estimated qualities of the other strategies are off true quality by 0.5-star.

We summarize the lessons from our numerical experiments as follows. First, with normal reviews only, the performances of all the OA strategies are similar. OA1, with only one reviewer group, performs slightly worse than the other OA strategies, but OA2 and OA3 with two and three groups, respectively, perform similarly. Excluding low-quality reviewers, such as L6 reviewers, does not make a large difference in estimating the true quality of hotels. Second, with fake reviews present, the strategies that do not exclude the groups containing fake reviews suffer severely from bias caused by fake reviews. Only the strategies that exclude the groups containing fake reviews remain unbiased. Third, OA methods are critically affected by the presence of fake reviews. Thus, with fake reviews, OA strategies can perform even worse than SA strategies.

These results imply that excluding fake reviews is important in aggregating review ratings. To avoid fake reviews, website managers can allow only actual visitors to leave reviews, as expedia.com does, or develop filters to detect fake reviews, as yelp.com does. However, neither of those options is easy. Moreover, because yelp.com does not make its filtering algorithm public, third-party websites cannot use it. Many studies are being undertaken to develop ways to detect fake reviews, but none of the resulting methods is satisfactory. The results of this study imply that the benefit of avoiding distortions from fake reviews by excluding low-quality reviews is greater than the cost of losing information from low-quality reviews.

VI. Conclusion

6.1. Implications and Suggestions for Future Research

Aggregating online review ratings and determining the true quality of hotels are important for both hotel managers and consumers. However, few studies examined

the quality of reviews in aggregating review ratings. Moreover, the existence of fake or manipulated reviews pose a serious obstacle to effective use of online reviews. In this paper, we seek to address this question by investigating the TripAdvisor data for Korean hotels and estimating the information quality of online reviews. To model the structure of online reviews we employ the Bayesian method used by Dai et al. (2012) and also consider various aggregation strategies for online review and compare their performance.

Our results are as follows. First, we divide the reviewers by the number of previous reviews and find that the information quality of reviews positively depends on them. Second, important information on hotel quality is concentrated in high-quality reviews and excluding low-quality reviews does not cause serious information loss. Third, distortions from fake reviews are serious and excluding low-quality reviews (in which most fake reviews are contained) successfully eliminate such distortions. Fourth, weighted average aggregation methods, giving more weights to recent reviews, are more efficient than simple average methods in that they update recent information more quickly. However, they suffer more from the existence of fake reviews because weighted average methods give more weights to inaccurate information from recently posted fake reviews.

We have the following implications and suggestions for future research. First, excluding low-quality reviews in aggregating online review ratings provides a practical guideline because the benefit of avoiding serious distortions from potential fake reviews is greater than the cost of losing information from low-quality reviews. This issue is further complicated because developing a good spam review filter costs a lot and such filters are imperfect because fake reviewers try to find a way to avoid being filtered out. Second, one challenge to implementing above guideline is that businesses may feel that removing legitimate reviews is unfair. Since every review contains valuable information on the quality of hotels, it will be more efficient to use all the reviews in evaluating the quality of hotels. Thus, instead of excluding the low-quality reviews from aggregation, we may assign low weights to low-quality reviews. Finding optimal weights for reviewer groups may be a future research topic. Third, aggregation method assigning higher weights to recent reviews is better than simple average method in that the former updates recent change of hotel's quality more quickly. However, weighted average methods suffer from the existence of fake reviews. Thus, we have a tradeoff between information efficiency and vulnerability from fake reviews. This issue requires attention in developing filtering algorithms. Fourth, we showed that the number of previous reviews is an important factor that determines the information quality of reviews. We examined it because it is one of the most easily observable features of a reviewer. We may carry out similar exercises, for example, to the proportion of positive reviews out of total reviews and determine which group has higher quality information. This is left for future research.

6.2. Limitations

This paper has a few limitations. First, Dai et al. (2012)'s review model used in this paper has a few limitations (Dai et al., 2012, p.24). For example, 1) they do not account for the selection of consumers who decide to leave a review. Depending on whether their experiences are good or bad, consumers may decide to leave a review or not. 2) Also, the incentives of elite reviewers to place higher values on popularity are not explicitly modeled. 3) Individual heterogeneity is modeled separately from estimating signals of the hotel's quality. However, different reviewers may capture different signals from their experience, hence interaction terms between individual heterogeneity and signals may better fit the data. 4) Fundamentally, one may argue whether the model can uniquely identify the quality of hotels. If the quality of hotels depends on reviewers' ratings, then the quality of hotels may not be uniquely identified because different groups of reviewers may evaluate the quality of hotels differently.

Second, this paper assesses the effects of fake reviews using the review model of Dai et al. (2012). In this model, fake reviewers enter the model exogenously unlike elite or non-elite reviewers and their behavioral equations are not accounted for in the model. This is the theoretical limitation and practical limitation is that the TripAdvisor does not report filtered reviews in the website unlike the Yelp. Hence, the determinants of fake reviews and their parameters cannot be estimated inside the model. Under this circumstance, instead of fully modeling and precisely defining fake reviews, we assume other determinants of fake reviews fixed and use a few popularly known features of fake reviews reported in the previous studies to mimic the behaviors of fake reviews in the numerical experiments. Also, since fake reviewers enter the model exogenously, we do not attempt to model economic incentives of fake reviewers unlike Luca and Zervas (2016) and Mayzlin et al. (2014).

Appendix

A1. Calculation of Reviewer Heterogeneity

We explain how to calculate individual heterogeneity terms here.

First, the number of previous reviews ($NumRev_{it}$). We calculate the number of previous reviews posted for Korean hotels only for the sample period.

Second, the frequency of reviews ($FreqRev_{it}$). We treat a person who leaves two reviews within 10 months differently from a person who leaves two reviews within 1 month. We calculate this variable as a reciprocal of the number of months between two adjacent reviews. For example, if 20 days have passed since the last review, then 1/1=1. If 40 days have passed since the last review, 1/2=0.5. If a very long time has passed since the last review, the variable will be close to zero, as if the reviewer is leaving a first review. This definition and that for $MatchD_{hit}$ below are different from those of Dai et al. (2012).

Third, the matching difference ($MatchD_{hit}$) between the reviewer and the product. This variable indicates how well a reviewer and a product are matched. We first define a hotel's characteristics. The purpose of a visit to a hotel is categorized as family /couple/ friend/ solo/ business. If two visitors indicate family as the purpose and one visitor indicates friend as the purpose, then the hotel's characteristic is (0.66/0/0.33/0/0). When a person visits a hotel, that hotel's characteristic is recorded in the person's profile. If the person visits another hotel, the second hotel's characteristic values are added to the person's profile and averaged. That average is the reviewer's characteristic. Now, when that person visits a new hotel, the root of the squared sum of the differences between each item of the hotel's characteristic and the person's characteristic is $MatchD_{hit}$. A small value for this variable means that the newly visited hotel is not much different from the hotels the person previously visited and vice versa.

Fourth, the taste variety ($TasteVar_{it}$) of reviewers. This variable measures how many types of hotels a reviewer has experienced before, calculated as the root of the sum of the variances of each item in a person's characteristic. If this variable is large, a person visits hotels for various purposes and vice versa.

A2. Estimation Results of Review Model

The following Table A1 shows the estimation results of the review model introduced in Section 3.

[Table A1] Estimation Results

	M1	M3	M9
σ_{bm}^2	1.4548 (0.0748)a	1.4942 (0.0975)a	2.0204 (0.2404)a
σ_{ne}^2			1.1970 (0.0990)a
σ_e^2		1.1930 (0.1666)a	1.0778 (0.1263)a
ρ_{bm}	0.3192 (0.0194)a	0.3072 (0.0234)a	0.2364 (0.0470)a
ρ_{ne}			0.2791 (0.0307)a
ρ_e		0.3413 (0.0431)a	0.3047 (0.0432)a
σ_ξ^2	0.0019 (0.0005)a	0.0019 (0.0004)a	0.0016 (0.0003)a
$(\mu+\lambda_{bm})_{NumRev}$	-0.0116 (0.0050)b	-0.0227 (0.0071)a	0.0636 (0.1000)
$(\mu+\lambda_{bm})_{Freq}$	-0.0454 (0.0190)b	-0.0566 (0.0287)b	-0.0978 (0.1308)
$(\mu+\lambda_{bm})_{Match}$	-0.0569 (0.0773)	-0.0433 (0.0972)	0.2316 (0.1694)
$(\mu+\lambda_{bm})_{Taste}$	0.1721 (0.0765)b	0.3209 (0.1013)a	-0.0568 (0.3413)
$(\mu+\lambda_{bm})_{Age}$	0.0000 (0.1310)	0.0000 (0.0000)	0.0000 (0.0158)
$(\lambda_{ne}-\lambda_{bm})_0$			0.1730 (0.0760)b
$(\lambda_{ne}-\lambda_{bm})_{NumRev}$			0.0885 (0.0999)
$(\lambda_{ne}-\lambda_{bm})_{Freq}$			0.0321 (0.1348)
$(\lambda_{ne}-\lambda_{bm})_{Match}$			-0.3044 (0.1774)c
$(\lambda_{ne}-\lambda_{bm})_{Taste}$			0.3447 (0.3815)
$(\lambda_{ne}-\lambda_{bm})_{Age}$			-0.0000 (0.0774)
$(\lambda_e-\lambda_{bm})_0$		0.0791 (0.0457)c	0.2184 (0.0754)a
$(\lambda_e-\lambda_{bm})_{NumRev}$		0.0168 (0.0084)b	-0.0693 (0.1003)
$(\lambda_e-\lambda_{bm})_{Freq}$		0.0390 (0.0410)	0.0789 (0.1293)
$(\lambda_e-\lambda_{bm})_{Match}$		-0.0280 (0.1306)	-0.2832 (0.1706)c
$(\lambda_e-\lambda_{bm})_{Taste}$		-0.3605 (0.1037)a	0.0126 (0.3375)
$(\lambda_e-\lambda_{bm})_{Age}$		-0.0000 (0.0000)b	-0.0000 (0.4114)

Note: a, b, and c indicate statistical significance at the 1%, 5%, and 10% levels, respectively. Standard errors are in the parenthesis. e, ne, and bm indicates elite, non-elite, and benchmark reviewer groups, respectively.

Table A1 shows the estimation results for models M1, M3, and M9 defined in Table 3. Elite reviewers have the smallest variance and largest ρ, as already pointed out in Dai et al. (2012). In M9, the difference in variance between elite and non-elite reviewers is not large, whereas that between non-elite and benchmark reviewers is large, greater than 0.3-star. Hence, benchmark reviewers are different from the other two groups. The benchmark reviewers are low-quality reviewers in the sense that their variance estimates, σ_{bm}^2, for signal observation are large. σ_ξ^2 is significant, and each stringency term is significant.

References

Akoglu, L., R. Chandy, and C. Faloutsos (2013), "Opinion Fraud Detection in Online Reviews by Network Effects," Proceedings of the Seventh International AAAI Conference on Web and Social Media.

Ayeh, J. K., N. Au, and R. Law (2013), ""Do We Believe in TripAdvisor?" Examining Credibility Perceptions and Online Travelers' Attitude Toward Using User-generated Content," *Journal of Travel Research*, 52(4), 437-452.

Banerjee, A. V. (1992), "A Simple Model of Herd Behavior," *Quarterly Journal of Economics*, 107(3), 797-817.

Besbes, O., and M. Scarsini (2015), "On Information Distortions in Online Ratings," Working Paper.

Bikhchandani, S., D. Hirshleifer, and I. Welch (1992), "A Theory of Fads, Fashion, Custom, and Cultural Change in Informational Cascades," *Journal of Political Economy*, 100(5), 992-1026.

Buhalis, D., and R. Law (2008), "Progress in Information Technology and Tourism Management: 20 Years on and 10 Years after the Internet - The State of eTourism Research," *Tourism management*, 29(4), 609-623.

Celen, B., and S. Kariv (2004), "Observational Learning under Imperfect Information," *Games and Economic Behavior*, 47(1), 72-86.

Crapis, D., B. Ifrach, C. Maglaras, and M. Scarsini (2017), "Monopoly Pricing in the Presence of Social Learning," *Management Science*, forthcoming.

Dai, W., G. Z. Jin, J. Lee, and M. Luca (2012), "Optimal Aggregation of Consumer Ratings: An Application to Yelp.com," NBER Working Paper, No. 18567, National Bureau of Economic Research.

Feng, S., L. Xing, A. Gogar, and Y. Choi (2012), "Distributional Footprints of Deceptive Product Reviews," Proceedings of the Sixth International AAAI Conference on Web and Social Media.

Filieri, R., and F. McLeay (2014), "E-WOM and Accommodation: An Analysis of the Factors that Influence Travelers' Adoption of Information from Online Reviews," *Journal of Travel Research*, 53(1), 44-57.

Goeree, J. K., T. R. Palfrey, and B. W. Rogers (2006), "Social Learning with Private and Common Values," *Economic theory*, 28(2), 245-264.

Herrera, H., and J. Hörner (2013), "Biased Social Learning," *Games and Economic Behavior*, 80, 131-146.

Hu, N., I. Bose, Y. Gao, and L. Liu (2011), "Manipulation in Digital Word-of-Mouth: A Reality Check for Book Reviews," *Decision Support Systems*, 50(3), 627-635.

Hu, N., I. Bose, N. S. Koh, and L. Liu (2012), "Manipulation of Online Reviews: An Analysis of Ratings, Readability, and Sentiments," *Decision Support Systems*, 52(3), 674-684.

Hu, N., L. Liu, and V. Sambamurthy (2011), "Fraud Detection in Online Consumer Reviews," *Decision Support Systems*, 50(3), 614-626.

Ifrach, B., C. Maglaras, and M. Scarsini (2015), "Bayesian Social Learning from Consumer

Reviews," Working Paper.

Kim, M.-J., N. Chung, and C.-K. Lee (2011), "The Effect of Perceived Trust on Electronic Commerce: Shopping Online for Tourism Products and Services in South Korea," *Tourism Management*, 32(2), 256-265.

Li, X., and L. Hitt (2008), "Self-Selection and Information Role of Online Product Reviews," *Information Systems Research*, 19, 456-474.

Litvin, S. W., R. E. Goldsmith, and B. Pan (2008), "Electronic Word-of-Mouth in Hospitality and Tourism Management," *Tourism Management*, 29(3), 458-468.

Liu, Z., and S. Park (2015), "What Makes a Useful Online Review? Implication for Travel Product Websites," *Tourism Management*, 47, 140-151.

Luca, M., and G. Zervas (2016), "Fake It till You Make It; Reputation, Competition, and Yelp Review Fraud," *Management Science*, 62(12), 3412-3427.

Mauri, A. G., and R. Minazzi (2013), "Web Reviews Influence on Expectations and Purchasing Intentions of Hotel Potential Customers," *International Journal of Hospitality Management*, 34, 99-107.

Mayzlin, D., Y. Dover, and J. Chevalier (2014), "Promotional Reviews: An Empirical Investigation of Online Review Manipulation," *American Economic Review*, 104(8), 2421-2455.

Mukherjee, A., B. Liu, and N. Glance (2012), "Spotting Fake Reviewer Groups in Consumer Reviews," Proceedings of the 21st International Conference on World Wide Web, 191-200.

Mukherjee, A., B. Liu, J. Wag, N. Glance, and N. Jindal (2011), "Detecting Group Review Spam," Proceedings of the 21st International Conference Companion on World Wide Web, 93-94.

Mukherjee, A., V. Venkataraman, B. Liu, and N. Glance (2013), "What Yelp Fake Review Filter Might Be Doing?," Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media.

O'Connor, P. (2008), "User-Generated Content and Travel: A Case Study on Tripadvisor.com," *Information and Communication Technologies in Tourism*, 47-58.

O'Connor, P. (2010), "Managing a Hotel's Image on TripAdvisor," *Journal of Hospitality Marketing & Management*, 19(7), 754-772.

Ott, M., Y. Choi, C. Cardie, and J. Hancock (2011), "Finding Deceptive Opinion Spam by Any Stretch of the Imagination," Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, 309-319.

Peng, Q., and M. Zhong (2014), "Detecting Spam Review Through Sentiment Analysis," *Journal of Software*, 9(8), 2065-2072.

Racherla, P., D. J. Connolly, and N. Christodoulidou (2013), "What Determines Consumers' Ratings of Service Providers? An Exploratory Study of Online Traveler Reviews," *Journal of Hospitality Marketing & Management*, 22(2), 135-161.

Schuckert, M., X. Liu, and R. Law (2015), "Hospitality and Tourism Online Reviews: Recent Trends and Future Directions," *Journal of Travel & Tourism Marketing*, 32(5), 608-621.

Sparks, B. A., and V. Browning (2011), "The Impact of Online Reviews on Hotel Booking

Intentions and Perception of Trust," *Tourism Management*, 32(6), 1310-1323.

Streitfeld, D. (2012), "Buy Reviews on Yelp, Get Black Mark," *New York Times*, http://www.nytimes.com/2012/10/18/technology/yelp-tries-to-halt-deceptive-reviews.html (last accessed on August 3, 2015).

Wu, M. Y., G. Wall, and P. L. Pearce (2014), "Shopping Experiences: International Tourists in Beijing's Silk Market," *Tourism Management*, 41, 96-106.

Xiang, Z., and U. Gretzel (2010), "Role of Social Media in Online Travel Information Search," *Tourism Management*, 31(2), 179-188.