A comparison of hotel ratings between verified and non-verified online review platforms.

Paolo Figini,[a, b] Laura Vici,[c] Giampaolo Viglia [d,1]

a. Department of Economics and CAST – Centre for Advanced Studies in Tourism, University of Bologna (Rimini Campus). Via Angherà 22, 47921, Rimini (Italy)

b. TREES, Faculty of Economic & Business Sciences, North-West University (Potchefstroom Campus). Building E3, X6001, Potchefstroom (South Africa)

c. Department of Statistical Sciences and CAST – Centre for Advanced Studies in Tourism, University of Bologna (Rimini Campus). Via Angherà 22, 47921, Rimini (Italy)

d. Department of Marketing, Portsmouth Business School, University of Portsmouth. Richmond Building, Portland st., Portsmouth PO1 3DE (United Kingdom).

[1.] **Corresponding author:** Giampaolo Viglia. E-mail: giampaolo.viglia@port.ac.uk

Declaration of interest: none

Acknowledgements: The authors thank Marco Montanari and Phil Xiang for the precious assistance in data scraping and manuscript positioning, respectively.

A comparison of hotel ratings between verified and non-verified online review platforms

Abstract

Purpose

This study aims to compare the rating dynamics of the same hotels in two online review platforms (Booking.com and Trip Advisor), which mainly differ in requiring or not requiring proof of prior reservation before posting a review (respectively, a verified vs a non-verified platform).

Design/methodology/approach

A verified system, by definition, cannot host fake reviews. Should also the non-verified system be free from "ambiguous" reviews, the structure of ratings (valence, variability, dynamics) for the same items should also be similar. Any detected structural difference, on the contrary, might be linked to a possible review bias.

Findings

Travelers' scores in the non-verified platform are higher and much more volatile than ratings in the verified platform. Additionally, the verified review system presents a faster convergence of ratings towards the long-term scores of individual hotels, whereas the non-verified system shows much more discordance in the early phases of the review window.

Research limitations/implications

The paper offers insights into how to detect suspicious reviews. Non-verified platforms should add indices of scores' dispersion to existing information available in websites and mobile apps. Moreover, they can use time windows to delete older (and more likely biased) reviews. Findings also ring a warning bell to tourists about the reliability of ratings, particularly when only a few reviews are posted online.

Originality/value

The across-platform comparison of single items (in terms of ratings' dynamics and speed of convergence) is a novel contribution that calls for extending the analysis to different destinations and types of platform.

Keywords: online review; rating convergence; verified review platforms; e-word-of-mouth.

1. Introduction

The massive amount of available online information makes it easier than in the recent past to assessing the quality of products to be purchased and consumed. The type and the volume of information searched by consumers depend on the product characteristics, the life-cycle stage, market factors, and the specific context and industry analysed.

The intrinsic nature of the tourism product as an experience good makes it hard for travellers to assess its quality before purchasing it (Woodside & King, 2001). The need to reduce uncertainty and the probability of regretting the decision at a later stage (Park and Nicolau, 2015; Duverger, 2013) leads tourists to search for unbiased and trustworthy information aimed at conveying a true image of what the product looks like (Yoo and Gretzel, 2008). In this context, electronic word-of-mouth (eWOM) plays a growing role in addressing and supporting tourists' decision processes. Its role has been reinforced over time by a rising number of scholars who show how online travel platforms and user-generated contents reduce the uncertainty related to the quality of the tourism product (Goldsmith and Horowitz, 2006; Manes & Tchetchik, 2018; You at al., 2015). Online review platforms – initially based on a community-based model – are now widely offering the possibility to conduct booking transactions in their own website, incorporating reviews as a form of electronic word-of-mouth (Bigné et al., 2019; Gligorijevic, 2016; Yang, 2018). Nonetheless, the quality of online information is highly heterogeneous and often questionable, as it is difficult to discern reliable from redundant or junk information.

Both public discourse and academic investigation have recently tried to address the issue of the so-called fake reviews or deceptive online communication (Hu *et al*., 2011; Luca and Zervas, 2016, Plotkina *et al*., 2019), which hit popular platforms such as Yelp, Amazon and Trip Advisor. These platforms are all taking the issue of fake reviews very seriously and have been developing internal algorithms and review-check systems to identify and delete

suspicious reviews. However, recent cases like the one of the spoof restaurant “The Shed” that became the No. 1 restaurant in London in December 2017 (see https://en.wikipedia.org/wiki/The_Shed_at_Dulwich for an introduction to the case) show how unreliable the review and rating systems can be. The detection and the treatment of fake reviews is beyond the scope of this paper, which, on the contrary, investigates systematically the presence of structural difference in the e-WOM evaluations presented in different platforms.

Consumers have bounded rationality and they are unable to acquire and elaborate massive and heterogeneous amount of data, thus driving them to prefer and rely more on ratings than on textual reviews (Yang, Park and Hu, 2018). We focus on comparing the rating structure and dynamics for the same hotels on different platforms (Booking.com and Trip Advisor). These platforms differ on their review verification system. While in Booking.com users need to undergo a transaction before being allowed to write a review, Trip Advisor does not require any proof of reservation before posting. We thus define the two systems as verified vs. non-verified, respectively, and investigate whether there are systematic differences across different levels of verification.

Our prior is that a verified system cannot host fake reviews by definition. Should also the non-verified system be free from “ambiguous” reviews, the structure of ratings (valence, variability, dynamics) should also be similar. Any detected structural difference, on the contrary, would be linked to a possible review bias. We focus on ratings of hotels, as this is the only product among the ones rated by Trip Advisor (restaurants, destinations, attraction sites), which is also rated in Booking.com.

We expect ratings in non-verified systems to be more inflated and more volatile than ratings in verified systems. Also, the convergence of ratings towards their long-run value should be slower in non-verified systems. These predictions are derived from some recent literature

analysing meta-data features, for which biased reviews tend to have more extreme ratings than genuine reviews, with higher rating deviations in the presence of dubious reviews (Mukherjee *et al*., 2013).

Our findings support most of our premises. Specifically, travelers' scores in the non-verified platform result to be higher and much more volatile than ratings in the verified platform. The verified review system also presents a faster convergence of ratings towards their long-term values for individual hotels, while the non-verified system shows much more discordance particularly in the early phases of the review window.

To the best of our knowledge only one paper (Bigné et al., 2019) provides a thorough comparison of hotels' review performance dynamics over time across different platforms. Contrarily to that paper, our article utilizes data at disaggregated level to scrutinize the rating dynamics of the same individual hotels and to compare reviews posted in the same periods across different websites. The issue of comparing products across platforms is indeed a promising avenue for studies on eWOM. We believe that both the novelty of the adopted methodology and the richness of the empirical findings offer important contributions on the reliability and the trustworthiness of online ratings. Based on these findings, we provide actionable managerial and policy solutions to filter our suspicious reviews and increase the transparency of online systems.

The paper is organized as follows: Section 2 briefly reviews the strands of literature on eWOM that are closely related to our investigation, i) the impact of eWOM on consumer decisions and ii) the reliability of eWOM. Section 3 introduces the data, the research design and the research questions to be addressed by our study. Section 4 presents the main results of the investigation. Finally, Section 5 discusses the findings, linking them back with the theoretical development. This part also offers some suggestions to reduce the presence of fake and ambiguous reviews and increase the transparency of the information presented.

2. Literature Review

Given the enormous amount of eWOM research, we tightly focused our literature on prior work investigating i) the impact of online reviews on consumers' purchasing decision and on firms' sales and ii) the issue of reliability and trustworthiness of reviews, including the detection of fake reviews.

For what concerns the first topic, while we redirect to Babic-Rosario et al., 2016; You et al., 2015; Floyd et al., 2014 (and more specifically to Yang et al., 2018 in tourism) for a comprehensive review, we want to highlight why online ratings and reviews are crucial in consumer decision-making processes. The functional risk and the degree of uncertainty related to the quality of a product are generally higher for services than for tangible goods; for hedonic rather than utilitarian products; for new rather than old goods (Murray and Schacter, 1990). The experiential nature of the tourism product, being a hedonic service, makes it highly affected by uncertainty and leads consumers to heavily rely on online reviews (Babić Rosario et al., 2016). Hence, while the impact of eWOM varies across sectors and contexts, tourism is clearly highly affected by eWOM (You et al., 2015; Yang et al., 2018).

Park and Nicolau (2015) show that extreme ratings (positive and negative) are more useful than moderate ratings, and that the effect of reviews in the tourism sector is asymmetric: negative reviews are perceived as more useful (in line with Kahneman and Tversky (1979), people consider negative reviews more useful than positive reviews as they perceive the aim of reducing losses more salient than increasing gains; his applies when rational consumers operate in contexts of uncertainty and risk, see Hu, Pavlou & Zhang, 2007), whereas positive reviews are crucial for enjoyment aspects and purchasing decisions. The impact of eWOM on sales and consumer decisions is also moderated by the life cycle stage and by specific characteristics of the product (You et al., 2016). This means that eWOM impact changes

across destinations (new vs. mature destinations), type of service (hotels vs. attraction sites) and other factors.

eWOM has been measured in different and interchangeably ways. Volume and valence are the main used levers, as they have been shown to be related to corporate sales and to the reduction of consumer's uncertainty. Volume is an index of market popularity and delivers information on the number of people who experienced or used the product. It helps reduce consumers' uncertainty and it is generally associated to an increase in sales (Chen et al., 2011; Chintagunta et al., 2010; Park et al., 2012). Valence is related to the sentiment of online reviews and is indicative of product reputation and expected quality (Kim and Gupta, 2012). Indicators of valence are, among others, the average score, the share of positive posts and the percentage of one-star scores. The exposure to comments' sentiments in the form of rating sign and magnitude thus highly affects consumers' preferences. While in several sectors there is no significant difference in using volume and valence to predict consumer decisions (Babić Rosario et al, 2016), in the tourism sector the effect of valence on sales is much larger than the effect of volume (Yang et al., 2018).

For what concerns the second topic, the quality of data sources is critical to make accurate inferences and inform consumers. In this regard, there are long-standing concerns about the reliability of online reviews (Luca & Zervas, 2016; Mayzlin *et* al., 2014; Park and Nicolau, 2015, Plotkina *et al*., 2019; Zhuang *et al*., 2018). A recent strand of computer science literature has been investigating fake reviews by exploring different dimensional information of data (Wu *et al*., 2017), from textual features to metadata burst features (Fontanarava *et al*., 2017). A review burst is an abrupt concentration of reviews in a limited period of time. It presents precise characteristics that stem from the sudden popularity of a reviewed object or from spam attacks. There is evidence that reviews in the same burst tend to have the same nature (Fei *et al*., 2013), thus suggesting that is possible to identify fake reviews by analysing

the timing and other features of the burst (Gunnemann *et al*., 2014; Lim *et al*., 2010; Ye *et al*., 2016).

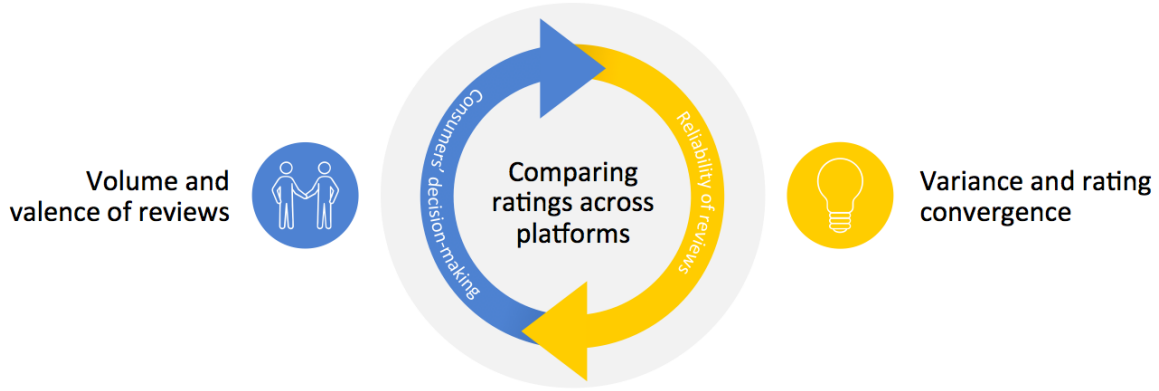
The business literature has started to assess information quality comparing different online platforms (Xiang *et al*., 2017). In this domain, big data analytics can help examining two or more distinct datasets. Specifically, statistical tools facilitate predictions and generalized understandings about the phenomenon at hand (Wu *et al.*, 2014). In the literature, there are recent papers and meta-analysis controlling for the potential bias produced by different platforms (Babić Rosario et al., 2016; You et al., 2015). In particular, Bigné et al. (2019) and Yang et al. (2018) find no statistical difference in the dynamics and valence of reviews across platforms.

The issue of reliability is intertwined with the impact on consumers' decisions, as the presence of fake reviews reduces the trustworthiness of eWOM. Given that consumers perceive as more valuable (useful) extreme (positive or negative) information (Park and Nicolau, 2015), reliability is strongly threatened by fake reviews, which generally have exactly an extremely positive or negative connotation (Agnihotri *et al*., 2016).

Beyond volume and valence, the variability of online reviews is a relevant feature that has been generally overlooked by prior literature (Jiménez & Mendoza, 2013). Measures of eWOM variability capture the heterogeneity in consumer opinions. A low variability in ratings characterizes consistent evaluations of products. A broad consensus among consumers lowers functional risk and uncertainty and, especially if review volume is huge, may influence new consumers' ratings, triggering a bandwagon effect which tends to keep variability low (Cicognani *et al*., 2016). On the other hand, high variability ratings increase quality uncertainty and reduce sales (Sun, 2012; Babić Rosario et al., 2016).

With our study, we focus on the valence, the dynamics, and the variability of rating scores, investigating whether there are structural differences across verified and non-verified platforms. Figure 1 presents our conceptual framework graphically, showing the reader a simple depiction of our contribution to the literature. We focus on the intersection of the two proposed stands, i.e., consumers' decision-making processes and reliability of reviews over time.

Figure 1 – Conceptual framework



Note: The proposed comparison is between online review platforms that differ in requiring (or not) proof of prior reservation before posting a review, i.e., verified vs. non-verified platforms.

3. Research Design and Research Questions

As mentioned in the introductory part of this work, we assess the dynamics of ratings of the same hotels across platforms characterized by different levels of verification of posted reviews. Bigné et al. (2019) have recently proposed a comprehensive and well-executed comparison across platforms, but analysing aggregated data at destination level or at hotel class level. On the contrary, in our study we compare and analyse the same individual hotel (at the micro level) across two platforms: Booking.com and TripAdvisor, which differ in requiring or not proof of prior reservation before allowing for rating the service (verified vs. non-verified platforms).

The rationale for selecting TripAdvisor and Booking.com is that the former is the largest community-based site in the world while the latter is the largest OTA where it is possible to write a review only after undergoing a transaction on the website. Given their importance, academic research has recently been using their posted data. This offered clear insights on how online reviews and rating scores affect the accommodation industry (Banerjee and Chua, 2016; Cezar and Ögüt,2016; Mariani and Borghi, 2018). For instance, Yang et al. (2018) control for the role played by TripAdvisor in estimating the effects of eWOM on sales.

We collected data for the same entities (hotels) on both platforms to closely examine and compare the distributions of ratings. This approach is widely accepted in empirical literature (Cavallo, 2017). Ten years review data were collected through a scraper in May 2016. The database consists of 103,423 reviews posted up to April 2016 on 872 hotels in Rimini, a renowned Italian seaside destination hosting more than 10 million overnight stays every year. Reviews were equally divided between TripAdvisor (51,036 reviews, 49.35%) and Booking.com (52,387 reviews, 50.65%). To avoid too much dispersion in the rating distributions, we only considered those hotels with at least 25 reviews in both platforms at the time of scraping. This final stratified sample included 182 hotels and 68,187 reviews.

Consistently with the discussion recalled in the introduction, we formulated the following research questions to guide the statistical analysis:

RQ 1: Are average and volatility of ratings larger in non-verified review systems than in verified review systems?

RQ 2: Is the rating convergence slower in non-verified review systems compared to verified review systems?

By focusing on the first two moments of the rating distribution, the average and the standard deviation (RQ1), we explored whether non-verified systems produce higher scores than verified systems for the same entities, and whether scores are more volatile. In order to control for hotels and platforms differing in the number of reviews and in their average rating, we used the Coefficient of Variation (CV), the ratio between the standard deviation and the mean. To measure the so called "review burst" (Gunnemann et al., 2014) effect (RQ2), we analysed the timing of the reviews in relative and in absolute terms.

4. Results

Tables 1 and 2 present descriptive statistics that are fit to answer RQ1. Table 1 compares the average scores of verified and non-verified platforms to analyse whether they differ significantly. Table 2 compares dispersion measures to investigate rating volatility. For the majority of hotels (116 out of 182, the 63.74%), the average rating score was higher in TripAdvisor than in Booking.com. The breakdown by hotel stars and by location (city centre vs. seaside, often a relevant distinction in sea & sun destinations) shows that 3-star hotels and seaside hotels mainly drive the difference in averages between TripAdvisor and Booking.com.

Since the normality of the ratings' distribution is rejected by both the Shapiro-Wilk and the Shapiro-Francia tests at the 1% significance level, two non-parametric tests for comparing the samples (Kolmogorov-Smirnov and Wilcoxon-Mann-Whitney tests) complement the traditional t-test, which assumes normal distributions. Results confirm that average scores of verified and non-verified systems systematically differ. The magnitude of differences is assessed through effect size measures (Cohen's effect size), which values suggest small (for 4-5 stars and city-centre subsamples) or small to moderate (for the total sample and for 3 stars and seaside subsamples) practical significance for the difference in the average ratings.

More striking is the difference in the dispersion of ratings between the two platforms (Table 2), which highlights how scores of verified systems are less volatile. Standard deviation is larger for the great majority of hotels in TripAdvisor (145 out of 182, the 79.67%) than in Booking.com. Results are robust to non-parametric tests (Kolmogorov-Smirnov and Wilcoxon-Mann-Whitney) and to different samples of the original population (for instance, similar results are found on the population of 73,590 reviews for the 221 hotels with at least 10 reviews in each of the two platforms at the time of scraping). Cohen's effect size values

support a significant difference in the standard deviation of ratings for the general sample and for all the subsamples.

Table 1 – The difference in average rating scores between TripAdvisor (TA) and Booking.com (BC)

Sample	Obs	Mean TA	Mean BC	t-test	KS test	WMW test	Prob mean TA > mean BC	Effect size (Cohen's d)	
Total	182	4.20	4.11	|4.119|***	0.232***	|52.55|***	63.7%	0.252	SM
1 and 2 stars	7	4.30	4.27	|0.347|	-	-	-	0.118	S
3 stars	123	4.19	4.07	|4.128|***	0.240***	|47.84|***	65.8%	0.311	SM
4 and 5 stars	52	4.21	4.17	|1.032|	0.240***	|23.76|***	61.5%	0.113	S
City Centre	85	4.17	4.15	|0.485|	0.192***	|29.16|***	52.9%	0.046	S
Seaside	97	4.22	4.07	|5.136|***	0.298***	|44.53|***	73.2%	0.398	SM

*Notes: * = significant at the 5% level, ** = significant at the 1% level, *** = significant at the 1‰ level. KS = Kolmogorov-Smirnov test; WMW = Wilcoxon-Mann-Whitney test. In the last column the effect size, according to Cohen (1988) is reported: S = small; SM = small-medium.*

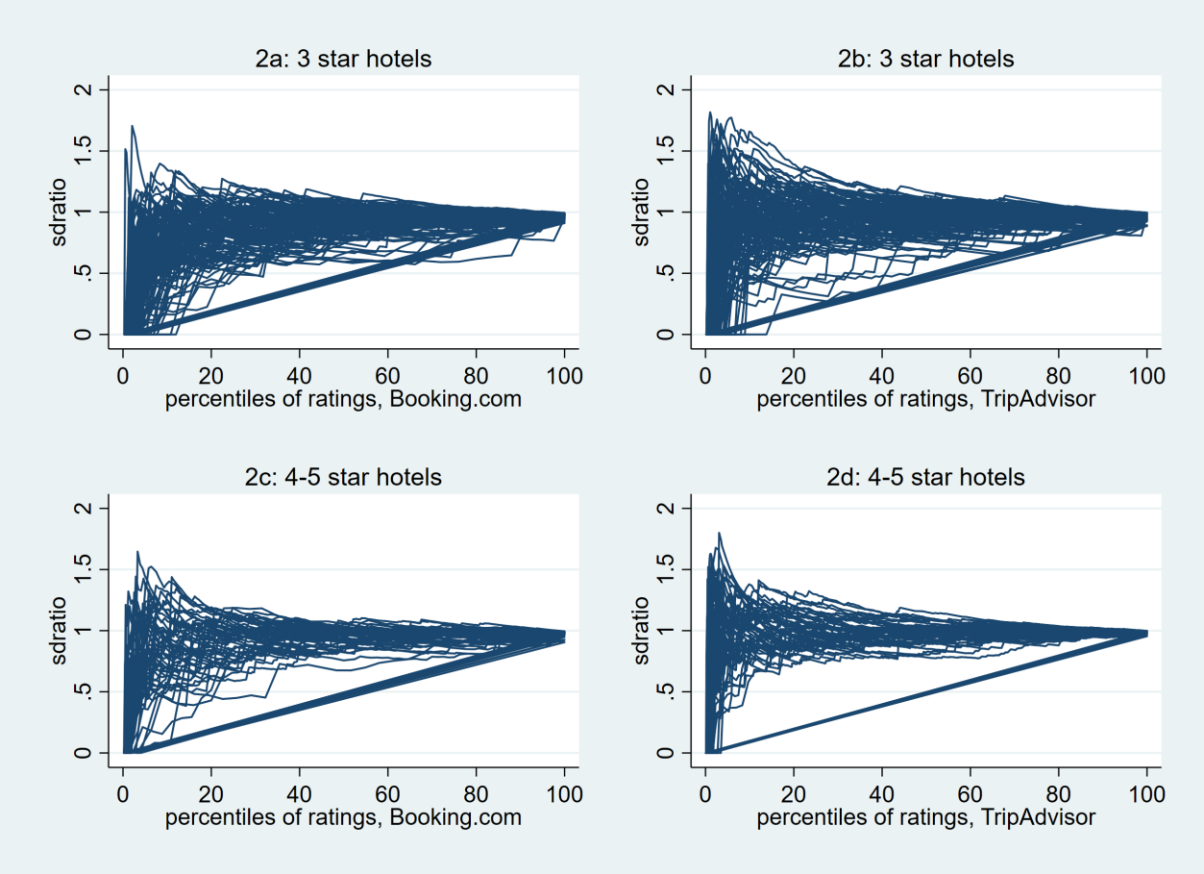
Table 2 – The difference in ratings dispersion between TripAdvisor (TA) and Booking.com (BC)

Sample	Obs	St. dev. TA	St. dev. BC	t-test	KS test	WMW test	Prob sd TA > sd BC	Effect size (Cohen's d)	
Total	182	0.92	0.76	|11.728|***	0.289***	|76.80|***	79.7%	0.819	L
1 and 2 stars	7	0.96	0.67	|3.929|**	-	-	-	1.481	L
3 stars	123	0.92	0.77	|9.048|***	0.276***	|49.97|***	76.4%	0.787	L
4 and 5 stars	52	0.92	0.76	|6.826|***	0.324***	|51.79|***	86.5%	0.797	L
City Centre	85	0.93	0.75	|11.032|***	0.357***	|69.08|***	90.6%	1.074	L
Seaside	97	0.91	0.78	|6.657|***	0.280***	|42.68|***	70.1%	0.648	ML

*Notes: * = significant at the 5% level, ** = significant at the 1% level, *** = significant at the 1‰ level. KS = Kolmogorov-Smirnov test; WMW = Wilcoxon-Mann-Whitney test. In the last column the effect size, according to Cohen (1988) is reported: ML = medium-large; L = large.*

Results from Tables 1 and 2 hence suggest that reviews in the non-verified review system (TripAdvisor) generally produce higher averages and larger dispersion of ratings for the same hotels under observation, in line with RQ1.

Figure 2 – The dynamics of ratings' variability between TripAdvisor and Booking.com



Notes: cvratio is the ratio between the cumulative coefficient of variation and its final value, computed for each hotel and each platform; ratings are sorted according to their review date and normalized in percentiles. 3-star hotels are reported in (2a) and (2b); 4- and 5-star hotels are reported in (2c) and (2d); Booking.com ratings are reported in (2a) and (2c); TripAdvisor ratings are reported in (2b) and (2d).

To investigate RQ2, we ranked the reviews according to their time stamp (i.e., the review date) and we built the cumulative average (i.e., the mean value of the scores received by hotels over time) and the cumulative CV of scores over time for each hotel in both platforms. These measures allow to evaluate how the mean score and its volatility change when the stock of new reviews adds up. The cumulative CV was preferred to the cumulative standard

deviation to control for the different average scores that hotels might have in the two platforms. However, results are robust to the use of the cumulative standard deviation.

The overall picture is reported in Figure 2. Each line shows the dynamics of the *cvratio* index, i.e., the ratio between the cumulative CV and its final value, for any individual hotel in Booking.com (Figure 2a for 3-star hotels; Figure 2c for 4- and 5-star hotels) and in Trip Advisor (Figure 2b for 3-star hotels; Figure 2d for 4- and 5-star hotels). The rating dynamics is normalized to the final value of the index (that is, the value at the time of scraping the data), hence it always converges to 1. It is visible that lines for TripAdvisor are mainly clustered in the upper part of their graphs while lines for Booking.com are mainly in the lower part. Variability is particularly higher in TripAdvisor in the early reviews. In other words, the process of convergence appears to be slower in TripAdvisor than in Booking.com, showing much more discordance in the hotels' rating when the early reviews are posted and when hotels are more vulnerable, since it is more likely that negative or positive informational cascades may occur (Banerjee, 1992). The mere availability of other consumers' previous reviews might in fact have an influence on other consumers (regardless of whether they are positive or negative), who may also disregard the prior information they have on the products (Banerjee, 1992; Van den Bulte and Lilien, 2001; Xiong and Bharadway, 2014).

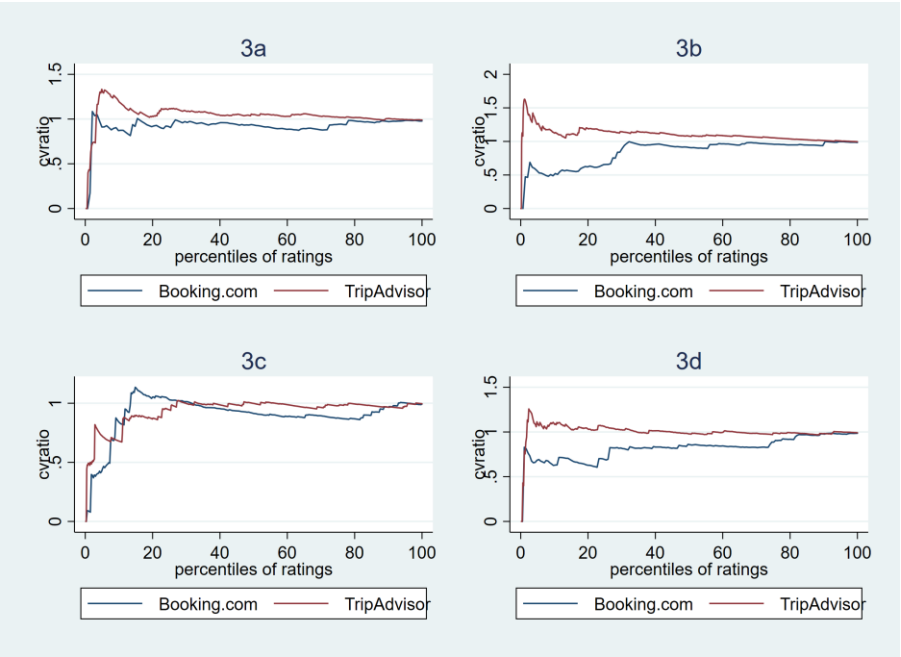
Inferential tests on these trends are supportive. In TripAdvisor 116 hotels out of 182 (63.74%) have a larger dispersion in the first decile than in the remaining part of the distribution of reviews. The respective share for Booking.com is only 32.97% (60 hotels). This difference blurs proceeding along the distributions. The divergent behavioural pattern is particularly strong for 3-star hotels.

To provide a cleaner picture of the phenomenon at stake, four hotels that exemplify the main paths of ratings' convergence in the two platforms are presented in Figure 3. Both lines in

each graph represent *cvratio*, the ratio between the cumulative CV and its final value at the time of data scraping, respectively for TripAdvisor (red line) and Booking.com (blue line). Lines start at 0 (when hotels get the first rating) and finish at 1 (when the cumulative CV equals the final one). Consistently with Figure 2, TripAdvisor ratings exhibit a higher variability in the early stages of the pattern and a slower speed of convergence with respect to Booking.com: the only exception to this trend is the hotel presented in Figure 3c, which well represents the minority of hotels not conforming to the main pattern. It is important to recall that this picture does not show higher scores in TripAdvisor than in Booking.com, but that in TripAdvisor there is much more discordance among scores in the early phases of review. The first ratings usually show a sequence of high and low scores, which is more pronounced than in Booking.com.

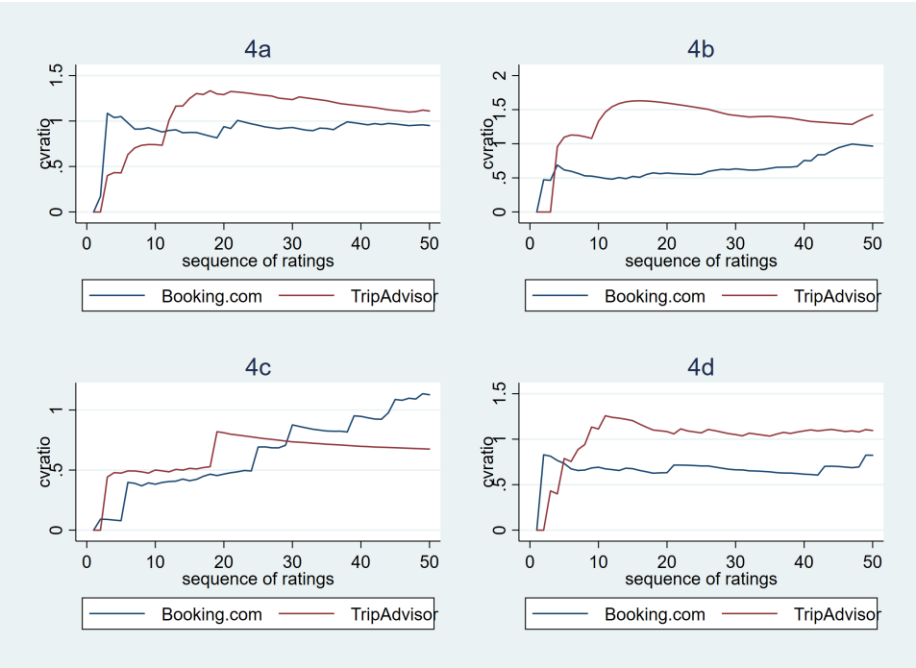
The same results are confirmed if *cvratio* is plotted against the absolute sequence of ratings, and not their percentiles. Figure 4 reports the first 50 ratings posted in Booking.com (blue line) and TripAdvisor (red line). Noticeably, the variability in rating scores in TripAdvisor increases around the tenth posted review (again, with the exception of the hotel presented in Figure 4c). This is arguably the moment when the hotel starts being visible in the TripAdvisor ranking and when marketing strategies aimed at influencing the ranking are particularly effective. If consumers feel that the average score is strongly not in line with perceived quality, they are more prone to leave a review. This is consistent with the so-called "review burst" phenomenon.

Figure 3 – The dynamics of ratings' variability between TripAdvisor and Booking.com for 4 randomly selected hotels



Notes: cvratio is the ratio between the cumulative coefficient of variation and its final value, computed for each platform; the percentiles of ratings are ranked according to their review date.

Figure 4 – The dynamics of ratings' variability between TripAdvisor and Booking.com for 4 randomly selected hotels, absolute sequence of ratings



Notes: cvratio is the ratio between the cumulative coefficient of variation and its final value, computed for each platform; ratings are ranked according to their review date.

5. Discussion and conclusions

Information technology has created new operators and challenges for traditional markets. This research contributes to the ongoing discussion on data quality in review platforms (Xiang *et al.*, 2017) and adds to the existing literature on review-centric features (Fontanarava *et al*., 2017) by comparing the rating features and dynamics of the same entities across different platforms. Expanding knowledge on online platforms (Casalo *et al*., 2015; Bore *et al.*, 2017), the novelty of our contribution lies in unpacking the differences between platforms that facilitate eWOM.

Results show the existence of structural differences in the ratings of the same entities, depending on the characteristics of the review platform. To the best of our knowledge, such analysis at micro level was not assessed before, as previous studies have analyzed differences across platforms by using aggregated data at destination level or service level. Using the stock of online ratings for hotels located in a popular seaside destination, we find that scores in a verified review system such as Booking.com (a platform where reviewers need to undergo a transaction before posting a review) are generally lower and much less volatile than ratings in a non-verified review system such as Trip Advisor. Booking.com also presents a faster convergence of ratings towards their long-term values, since the standard deviation and the coefficient of variation converge towards their final value in a shorter time span, while Trip Advisor shows much more discordance in the early phases of the review window, when only few reviews are posted.

Our approach differs from previous research (Bigné *et al*., 2019), which made use of aggregated data at destination level. Hence, the analysis of the ratings' dynamics and of the speed of convergence of ratings overtime is a novel contribution of our approach that cannot be compared with any previous study and that calls for further analysis on different destinations.

Our results indeed suggest that not all the review websites show similar ratings for the same hotels, indicating the presence of potential biases in social media data (Ruths & Pfeffer, 2014), particularly in non-verified platforms. Although the reasons behind the differences between Trip Advisor and Booking.com cannot be clarified through our study, a possible explanation revolves around the presence of fake reviews or of marketing initiatives to boost ratings in non-verified platforms. This phenomenon is particularly relevant when the marginal impact of each score (and hence the effectiveness of the rating) is larger. This happens in the early phases of the "review window", something that is supported by our findings about RQ2. Such high variability of ratings in non-verified systems is suspicious, thus calling for further research in this area.

These findings also provide managerial and practical implications regarding the reliability and trustfulness of online rating systems. Recently, it has been shown how trust is the main determinant of travellers' adoption of user-generated contents (Ukpabi & Karialuoto, 2018). An excessive dispersion of ratings can create uncertainty, affecting the level of trust in the review system. While the use of internal algorithms to spot and delete fake reviews can certainly be effective, non-verified platforms should add a clear index of dispersion of scores to existing information on average scores, ranking, and number of scores. Moreover, they can use time windows to delete older (and more likely biased) reviews. On this line, Booking.com has recently changed its policy by eliminating reviews that are older than two years. These strategies should positively impact OTAs' reputation.

Finally, our contribution rings a warning bell to tourists about the reliability of signals, particularly when there are only a few reviews posted online. Together with the average score and the distribution of ratings, the number of reviews is a relevant piece of information to be taken into consideration by users when evaluating eWOM. A possible alternative solution is proposing interfirm connections to increase the quality and richness of the information

presented to customers (see Abrate et al., 2019). In this sense, popular online platforms like Trivago have started introducing reviews from multiple platforms.

The main limitation of this study is that, although we have a rich longitudinal sample with 182 hotels, we observe a single popular seaside destination. Further research will have to test the robustness of our findings to different destinations, characterized by a different tourism mix. This research does not focus on what would engender a generalised positive shift in review values. In this sense, recent evidence suggests the need to focus on consumer emotions (Prayag et al., 2017) or CSR elements (Andreu et al., 2015). Finally, field experimental research, thus almost unfeasible in this domain because of the difficulty to randomly allocate real customers to different platforms, would rule out a possible consumer's self-selection into platforms. However, while the presence of sample selection could affect average ratings, it would still not explain the suspiciously different convergence of ratings over time.

References

Abrate, G., Bruno, C., Erbetta, F., & Fraquelli, G. (2019). Which Future for Traditional Travel Agencies? A Dynamic Capabilities Approach. *Journal of Travel Research*, 0047287519870250.

Andreu, L., Casado-Díaz, A. B., & Mattila, A. S. (2015). Effects of message appeal and service type in CSR communication strategies. *Journal of Business Research*, *68*(7), 1488-1495.

Agnihotri, A., & Bhattacharya, S. (2016). Online review helpfulness: Role of qualitative factors. *Psychology & Marketing*, 33(11), 1006-1017.

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

Banerjee, S., Chua, A.Y. (2016). In search of patterns among travellers' hotel ratings in TripAdvisor. *Tourism Management*, *53*, 125-131.

Babić R.A., Sotgiu F., De Valck K., & Bijmolt T.H. (2016). The Effect of Electronic Word of Mouth on Sales: A Meta Analytic Review of Platform, Product, and Metric Factors, *Journal of Marketing Research 53(3)*, 297-318.

Bigné E., William E., & Soria-Olivas E. (2019). Similarity and Consistency in Hotel Online Ratings across Platforms, *Journal of Travel Research*, 1-17.

Bore, I., Rutherford, C., Glasgow, S., Taheri, B., & Antony, J. (2017). A systematic literature review on eWOM in the hotel industry: Current trends and suggestions for future research. *Hospitality & Society*, *7*(1), 63-85.

Casalo, L. V., Flavian, C., Guinaliu, M., & Ekinci, Y. (2015). Do online hotel rating schemes influence booking behaviors?. *International Journal of Hospitality Management*, 49, 28-36.

Cavallo, A. (2017). Are online and offline prices similar? Evidence from large multi-channel retailers. *American Economic Review*, *107*(1), 283-303.

Cezar, A., Ögüt, H. (2016). Analyzing conversion rates in online hotel booking: The role of customer reviews, recommendations and rank order in search listings, *International Journal of Contemporary Hospitality Management*, 28, 286-304.

Chen, Y., Wang, Q., Xie, J. (2011). Online Social Interactions: A Natural Experiment on Word of Mouth Versus Observational Learning, *Journal of Marketing Research*, 48, 238–54.

Chintagunta, P.K., Gopinath, S., Venkataraman, S. (2010). The Effects of Online User Reviews on Movie Box Office Performance: Accounting for Sequential Rollout and Aggregation Across Local Markets, *Marketing Science*, 29 (5), 944–57.

Cicognani, S., Figini, P., Magnani, M. (2016). Social Influence Bias in Online Ratings: a Field Experiment. *Quaderni – Working Paper n. 1060*, Department of Economics, University of Bologna.

Duverger P. (2013). Curvilinear effects of user-generated content on hotels' market share: A dynamic panel-data analysis, *Journal of Travel Research*, 52 (4), 465-478.

Fei, G., Mukherjee, A., Liu, B., Hsu, M., Castellanos, M., Ghosh, R. (2013). Exploiting Burstiness in Reviews for Review Spammer Detection, In *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media*, AAAI, 175-184.

Floyd, K., Freling, R., Alhoqail, S., Cho, H.Y., Freling T. (2014). How Online Product Reviews Affect Retail Sales: A Meta-analysis, *Journal of Retailing*, 90(2), 217-232.

Fontanarava, J., Pasi, G., Viviani, M. (2017). Feature analysis for fake review detection through supervised classification. *IEEE International Conference on Big Data*, 658-666

Gligorijevic, B. (2016). Review platforms in destinations and hospitality. In *Open Tourism* (pp. 215-228). Springer Berlin Heidelberg.

Goldsmith, R.E., Horowitz, D. (2006). Measuring Motivations for Online Opinion Seeking. *Journal of Interactive Advertising*, 6 (2), 2–14.

Günnemann, S., Günnemann, N., Faloutsos, C. (2014). Detecting anomalies in dynamic rating data: a robust probabilistic model for rating evolution. *In Proceedings of the 20h international conference on Knowledge discovery and data mining*, *ACM SIGKDD*, 841-850.

Hu, N., Pavlou, P.A., & Zhang, J. (2007). Why do online product reviews have a J-shaped distribution? Overcoming biases in online Word-of-Mouth communication. *Unpublished manuscrip*t.

Hu, N., Liu, L., & Sambamurthy, V. (2011). Fraud detection in online consumer reviews. *Decision Support Systems*, 50(3), 614–626.

Kahneman, D., Tversky, A. (1979). Prospect theory: An analysis of decision under risk, *Econometrica*, 47 (2), 263-292.

Kim, J., Gupta, P. (2012). Emotional expressions in online user reviews: How they influence consumers' product evaluations, *Journal of Business Research*, 65 (7), 985-992.

Jiménez, F.R., & Mendoza, N.A. (2013). Too popular to ignore: The influence of online reviews on purchase intentions of search and experience products. *Journal of Interactive Marketing*, 27(3), 226-235.

Lim, E.P., Nguyen, V.A., Jindal, N., Liu, B., Lauw, H.W (2010). Detecting product review spammers using rating behaviors. *In Proceedings of the 19th ACM international conference on Information and knowledge management*. ACM: 939-948.

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

Luca, M., & Zervas, G. (2016). Fake it till you make it: Reputation, competition, and Yelp review fraud. *Management Science*, *62*(12), 3412-3427.

Manes, E., & Tchetchick, A. (2018). The role of electronic word of mouth in reducing information asymmetry: An empirical investigation of online hotel booking, *Journal of Business Research*, 85, 185-196.

Mariani, M.M., Borghi, M. (2018). Effects of the Booking.com rating system: bringing hotel class into the picture. *Tourism Management*, 66:47-52.

Mayzlin, D., Dover, Y., & Chevalier, J. (2014). Promotional reviews: An empirical investigation of online review manipulation. *American Economic Review*, 104(8), 2421–2455.

Mukherjee, A., Kumar, A., Liu, B., Wang, J., Hsu, M., Castellanos, M., Ghosh, R. (2013). Spotting opinion spammers using behavioral footprints. *In Proceedings of the 19th ACM*

SIGKDD international conference on Knowledge discovery and data mining. ACM: 632–640.

Murray, K.B. (1991). A Test of Services Marketing Theory: Consumer Information Acquisition Activities, *Journal of Marketing*, 55, 10–25.

Park, J.H. , Gu, B. , Young, L.H. (2012). The Relationship Between Retailer-Hosted and Third-Party Hosted WOM Sources and Their Influence on Retailer Sales, *Electronic Commerce Research and Applications*, 11 (3), 253–61.

Park, S., Nicolau, J.L. (2015). Asymmetric effects of online consumer reviews, *Annals of Tourism Research*, 50, 67-83.

Prayag, G., Hosany, S., Muskat, B., & Del Chiappa, G. (2017). Understanding the relationships between tourists' emotional experiences, perceived overall image, satisfaction, and intention to recommend. *Journal of Travel Research*, *56*(1), 41-54.

Plotkina, D., Munzel, A., & Pallud, J. (2019). Illusions of truth. Experimental insights into human and algorithmic detection of fake online reviews, *Journal of Business Research*, https://doi.org/10.1016/j.jbusres.2018.12.009

Ruths, D., & Pfeffer, J. (2014). Social media for large studies of behavior. *Science*, *346*(6213), 1063-1064.

Sun, M. (2012). How Does the Variance of Proudct Ratings Matter?, *Management Science*, 58(4), 696-707.

Ukpabi, D. C., & Karjaluoto, H. (2018). What drives travelers' adoption of user-generated content? A literature review. *Tourism Management Perspectives*.

Van den Bulte, C., Lilien, G.L. (2001). Medical Innovation Revisited: Social Contagion Versus Marketing Effort, *American Journal of Sociology*, 106 (5), 1409–1435.

Woodside, A.G., King, R.I. (2001). An updated model of travel and tourism purchase-consumption systems. *Journal of Travel & Tourism Marketing*, 10(1), 3–27.

Wu, X., Zhu, X., Wu, G. Q., & Ding, W. (2014). Data mining with big data. *IEEE transactions on knowledge and data engineering*, *26*(1), 97-107.

Wu, X., Dong, Y., Tao, J., Huang, C., Chawla, N.V. (2017). Reliable fake review detection via modeling temporal and behavioral patterns. *IEEE International Conference on Big Data*, 494-499.

Xiang, Z., Du, Q., Ma, Y., & Fan, W. (2017). A comparative analysis of major online review platforms: Implications for social media analytics in hospitality and tourism. *Tourism Management*, *58*, 51-65.

Xiong, G., Bharadwaj, S. (2014). Prerelease Buzz Evolution Patterns and New Product Performance, *Marketing Science*, 33 (3), 401–421.

Yang, Y., Park, S., Hu, X. (2018). Electronic Word of Mouth and Hotel Performance: A Meta-Analysis, *Tourism Management*, 67, 248-260.

Ye, J., Kumar, S., Akoglu, L. (2016). Temporal opinion spam detection by multivariate indicative signals. *In Proceedings of the Tenth International AAAI Conference on Web and Social Media*, *ICWSM16*, 743–746.

Yoo, K.H., Gretzel, U. (2008). What motivates consumers to write online travel reviews? *Information Technology & Tourism*, 10, 283–295.

You, Y., Gautham G.V., Joshi A.M. (2015). A Meta-Analysis of Electronic Word-of-Mouth Elasticity, *Journal of Marketing,* 79 (2), 19-39.

Zhuang, M., Cui, G., & Peng, L. (2018). Manufactured Opinions: the Effect of Manipulating Online Product Reviews, *Journal of Business Research*, 87, 24-35.