

Alma Mater Studiorum Università di Bologna
Archivio istituzionale della ricerca

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

This is the final peer-reviewed author's accepted manuscript (postprint) of the following publication:

Published Version:

Figini P., Vici L., Viglia G. (2020). A comparison of hotel ratings between verified and non-verified online review platforms. INTERNATIONAL JOURNAL OF CULTURE, TOURISM AND HOSPITALITY RESEARCH, 14(2), 157-171 [10.1108/IJCTHR-10-2019-0193].

Availability:

This version is available at: <https://hdl.handle.net/11585/774561> since: 2021-03-01

Published:

DOI: <http://doi.org/10.1108/IJCTHR-10-2019-0193>

Terms of use:

Some rights reserved. The terms and conditions for the reuse of this version of the manuscript are specified in the publishing policy. For all terms of use and more information see the publisher's website.

This item was downloaded from IRIS Università di Bologna (<https://cris.unibo.it/>).
When citing, please refer to the published version.

(Article begins on next page)

This is the final peer-reviewed accepted manuscript of:

Figini, P., Vici, L. and Viglia, G. (2020), "A comparison of hotel ratings between verified and non-verified online review platforms", *International Journal of Culture, Tourism and Hospitality Research*, Vol. 14 No. 2, pp. 157-171.

The final published version is available online at:

<https://doi.org/10.1108/IJCTHR-10-2019-0193>

Terms of use:

Some rights reserved. The terms and conditions for the reuse of this version of the manuscript are specified in the publishing policy. For all terms of use and more information see the publisher's website.

This item was downloaded from IRIS Università di Bologna (<https://cris.unibo.it/>)

When citing, please refer to the published version.

This is the final peer-reviewed accepted manuscript of:

Figini, P., Vici, L., Viglia, G., 2020. A comparison of hotel ratings between verified and non-verified online review platforms, *International Journal of Culture, Tourism and Hospitality Research*, 14(2):157-171

The final published version is available online at:

<https://doi.org/10.1108/IJCTHR-10-2019-0193>

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

Structured Abstract

Purpose

This study compares the rating dynamics of the same hotels in two online review platforms (Booking.com and Trip Advisor), which mainly differ in requiring or not 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, while the non-verified system shows much more discordance in the early phases of the review window.

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.

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

1. Introduction

The massive amount of information available online helps 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 and King, 2001). The need to reduce uncertainty and the possibility 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 and Tchetchik, 2018; You et al., 2015). Online review platforms – initially built on a community-based model – are now widely offering the possibility to conduct booking transactions in their own websites, incorporating reviews as a form of electronic word-of-mouth (Bigné et al., 2019; 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 (TripAdvisor, 2019). 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, systematically investigates 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 et al., 2018). This study focuses 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 we 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 be similar. Any detected structural difference, on the contrary, would be linked to a possible review bias. The focus is 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.

Ratings in non-verified systems are expected 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).

Findings of this study support these premises. Specifically, travelers' scores in the non-verified platform result to be higher and much more volatile than ratings in the verified platform. More interestingly, 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, this article utilizes data at disaggregated level to scrutinize the rating dynamics of the same individual hotels and to compare reviews posted in the same period across different websites. The issue of comparing products across platforms is indeed a promising avenue for studies on eWOM. Both the novelty of the adopted methodology and the richness of the empirical findings can offer important contributions on the reliability and the trustworthiness of online ratings. Based on these findings, actionable managerial and policy solutions to filter suspicious reviews and increase the transparency of online systems are provided.

The paper is organized as follows: Section 2 briefly reviews the strands of literature on eWOM that are closely related to this investigation. Section 3 introduces the data, the research design and the hypotheses to be tested. 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 to increase the transparency of the information presented.

2. Literature Review

Given the enormous amount of existing research on eWOM, we tightly focused 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 for a comprehensive review we redirect to Babić-Rosario et al., 2016; You et al., 2015; Floyd et al., 2014 (and more specifically for tourism to Yang et al., 2018), a recall on why online ratings and reviews are crucial in consumer decision-making processes is here provided. 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 makes it strongly affected by uncertainty, leading consumers to heavily rely on eWOM (Babić-Rosario et al., 2016). Hence, while the impact of online reviews varies across sectors and contexts, tourism is clearly highly affected (You et al., 2015; Yang et al., 2018).

Park and Nicolau (2015) show that extreme ratings (positive and negative) are more useful to customers than moderate ratings, and that the effect of reviews in the tourism sector is asymmetric: negative reviews are perceived as more useful as the aim of reducing losses is more salient than increasing gains, in line with Kahneman and Tversky (1979). This applies when rational consumers operate in contexts of uncertainty and risk (Hu, Pavlou and 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 impacts differently 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 and 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 an extremely positive or negative connotation (Agnihotri et al., 2016).

Beyond volume and valence, the variability of online reviews is then a relevant feature that has been generally overlooked by prior literature (Jiménez and 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 large, 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 in ratings increases quality uncertainty and reduces sales (Sun, 2012; Babić Rosario et al., 2016).

With this study, the focus is on the valence, the dynamics, and the variability of rating scores, investigating whether there are structural differences across verified and non-verified platforms.

3. Research Design and Hypotheses

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

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.

Data were collected for the same items (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 for 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 and noise in the rating distributions, only hotels with at least 25 reviews in each of the two platforms at the time of scraping were considered. This final stratified sample included 182 hotels and 68,186 reviews.

As a robustness check, we also considered 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. Results were consistent and are available upon request.

In line with the discussion recalled in the introduction and in the literature review, we formulated the following hypotheses to be tested through the statistical analysis:

HP 1: Average and volatility of ratings are larger in non-verified review systems than in verified review systems.

HP 2: The rating convergence is 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, we tested whether non-verified systems produce higher scores than verified systems for the same entities and whether scores are more volatile (HP1). In order to control for hotels and platforms differing in the number of reviews and in their average rating, we first transformed the data in the same scale (out of 5 points). Then we also used the Coefficient of Variation (CV – the ratio between the standard deviation and the mean) and the Mean Absolute Percentage Deviation (MAPD) to measure dispersion. To measure the so called “review burst” (Gunnemann et al., 2014) effect (HP2), 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 HP1. 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 (Table 1). The breakdown by hotel stars and by location (city centre vs. seaside, often a relevant distinction in sea and sun destinations) shows that 3-star hotels and seaside hotels (hence the segments where competition is higher) 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, we make use of two non-parametric tests for comparing the samples (Kolmogorov-Smirnov and Wilcoxon-Mann-Whitney tests), instead of the traditional t-test (which assumes normal distributions). Results for both tests confirm that the average scores of verified and non-verified systems differ systematically. 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. The differences are significant - according to Kolmogorov-Smirnov and Wilcoxon-Mann-Whitney tests - and the results are robust to the use of alternative indices of dispersion, i.e., the Coefficient of Variation and the Mean Absolute Percentage Deviation (MAPD). The last corrects for the different average of the samples (results for the MAPD are

presented in Table 3). Cohen's effect size values support a significant difference in the dispersion of ratings for the general sample and for all the subsamples.

[Insert Table 1 about here]

[Insert Table 2 about here]

[Insert Table 3 about here]

Results from Tables 1 to 3 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 HP1.

To investigate HP2, the reviews were ranked according to their time stamp (i.e., the review date). 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 were built. 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. Again, results are robust to the use of the cumulative standard deviation.

[Insert Figure 1 about here]

The overall picture is reported in Figure 1. Each line shows the dynamics of the *cvratio* index (the ratio between the cumulative CV and its final value at the time of scraping) for any individual hotel in Booking.com (Figure 1a for 3-star hotels; Figure 1c for 4- and 5-star hotels) and in Trip Advisor (Figure 1b for 3-star hotels; Figure 1d 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 more frequently in

the lower part. Variability is particularly higher in TripAdvisor in the early percentiles of the reviews' distribution. 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. This is 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 at 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. Figure 2 is similar to Figure 1, but with a different breakdown: in Figure 2a (2c) we present the 122 (60) hotels that in Booking.com have a lower (higher) value of 1 in the *cvratio* at the first decile of the rating distribution; in Figure 2b (2d) we present the specular situation for TripAdvisor. As the visual readability of this bunch of lines might be difficult, Figure 3 has been instead built averaging out, for each percentile of the distribution, the *cvratio* of the previous sub-samples of hotels for each provider. Not surprisingly, the averages hide much of the heterogeneity in hotels' dynamics.

[Insert Figure 2 about here]

[Insert Figure 3 about here]

To provide a cleaner picture of the phenomenon at stake, four randomly selected hotels that exemplify different paths of ratings' convergence in the two platforms are presented in Figure

4. Both lines in each graph represent *cvratio*, respectively for TripAdvisor (dashed line) and Booking.com (solid 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 Figures 1 to 3, 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 4c, 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.

[Insert Figure 4 about here]

The same results are confirmed if *cvratio* is plotted against the absolute sequence of ratings, and not their percentiles. Figure 5 reports the first 50 ratings posted in Booking.com (solid line) and TripAdvisor (dashed 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 5c). 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.

[Insert Figure 5 about here]

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 (Bigné et al., 2019). Using the stock of online ratings for hotels located in a popular seaside destination, this study found 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 both 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 the first reviews are posted.

This 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 this approach that cannot be compared with any previous study and that calls for further analysis on different destinations.

The results of this study 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 and Pfeffer, 2014), particularly in non-verified platforms. Although the reasons behind the differences between Trip Advisor and Booking.com cannot be clarified through this 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 present findings. 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 and 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, all platforms (particularly the non-verified ones) should add an index of dispersion of scores to existing information. Most of the popular review systems (e.g. TripAdvisor, Amazon, Google Play) presents the average rating score, the ranking within certain categories, and the absolute number of reviews for each score. However, none of them further elaborates on the rating scores’ dispersion. Although the publication of precise indices (as the standard deviation, the coefficient of variation or the mean absolute percentage deviation) might create difficulties in interpretation for the average user, some graphical derivations would be of great help. For instance, an index of “*disagreement of reviews*“, which takes values from 1 to 5 for each item according to the underlying value of the standard deviation, can be an effective way of presenting such informative trends. Accordingly, users

could check which items have dissimilar ratings, digging into the underlying reasons (for example, going through a careful reading of some of the reviews).

Moreover, platforms 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. On a similar level, platforms might post graphs showing the evolution of the average rating and of the dispersion overtime to let users easily spot upwards or downwards trends. These strategies should positively impact OTAs' reputation and would provide users with more information which will facilitate better informed decisions.

Finally, our contribution rings a warning bell to tourists about the reliability of signals, particularly when there are only a few reviews posted online in non-verified platforms. Together with the average score and the distribution of ratings, the number of reviews is hence a relevant piece of information to be taken into consideration by users when evaluating eWOM. Interfirm connections might be a solution 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 a rich longitudinal sample with 182 hotels was used, we only investigated a single, although popular, destination. Further research will have to test the robustness of these findings to different destinations, characterized by a different tourism mix. Similarly, while the present work focused on hotels, the same approach could in principle be extended to any item and to any platform characterized by different verification rules.

Moreover, we cannot rule out that results are somehow driven by the different profile of reviewers in the two platforms. However, while the presence of sample self-selection could affect average ratings, it would still not explain the suspiciously different convergence path of

ratings over time. Our study opens up for a rich research agenda. An interesting line of investigation is to compare the users' profiles in different rating systems. We believe that this is almost unfeasible with real data, because of privacy reasons and the lack of relevant and reliable information about profiles. This is also challenging in field experiments, because of the difficulty to randomly allocate real customers to different platforms. However, laboratory experiments might rule out a possible consumer's self-selection into platforms. Specifically, the difference in users' profiles is expected to be even higher when comparing traditional online travel agencies with sharing economy operators (Pera et al., 2019).

Finally, users' awareness of limitations in the verification policy of online platforms, and their behavioural consequences should be carefully investigated. On the one hand, the identification of different behavioural patterns of users according to the type of platform would shed light on the true importance of verification systems. On the other hand, as the review system is at the core of TripAdvisor business model, its trustworthiness is of paramount importance, as a recent report demonstrates (TripAdvisor, 2019). How users respond to the effort undertaken by platforms to build trust is an important research direction.

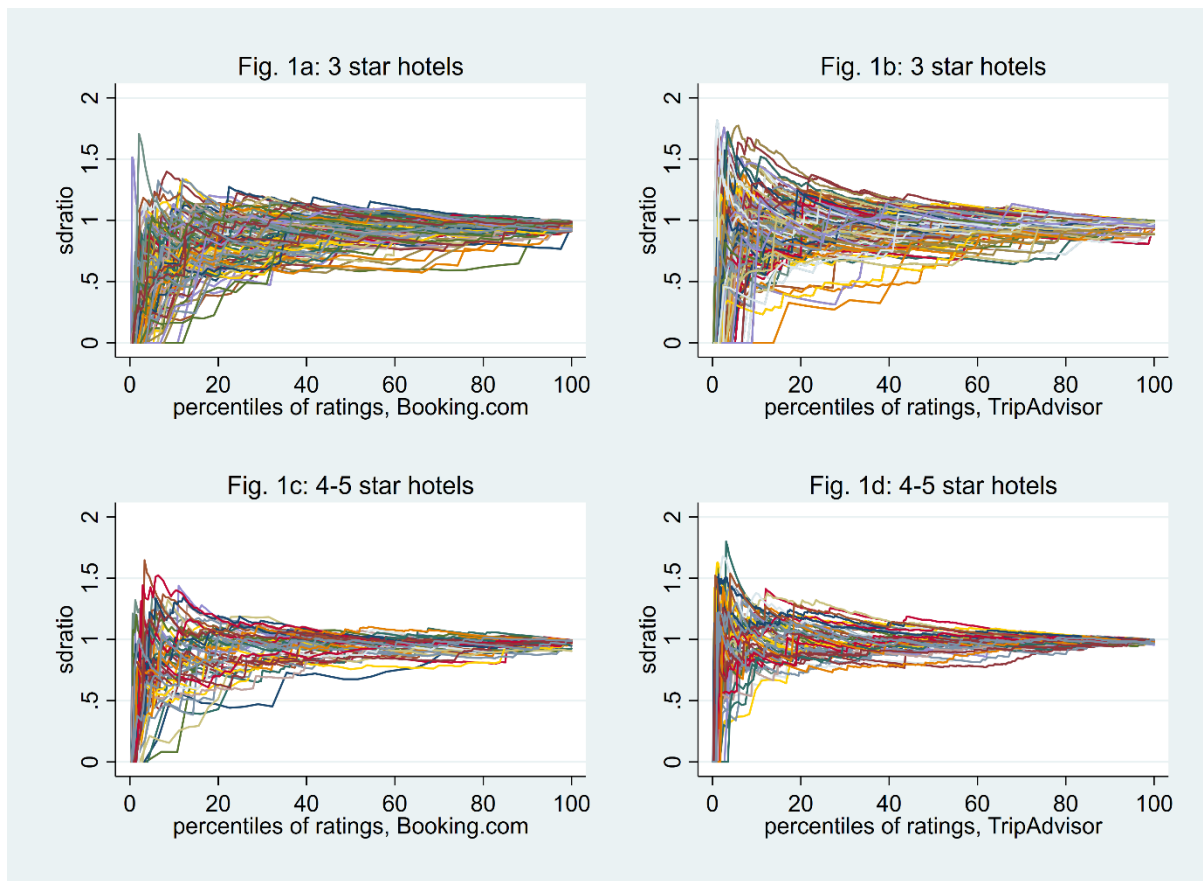
References

- Abrate, G., Bruno, C., Erbetta, F., Fraquelli, G. (2019). Which Future for Traditional Travel Agencies? A Dynamic Capabilities Approach. *Journal of Travel Research*, 0047287519870250.
- Agnihotri, A., Bhattacharya, S. (2016). Online review helpfulness: Role of qualitative factors. *Psychology and 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 and 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
- Goldsmith, R.E., Horowitz, D. (2006). Measuring Motivations for Online Opinion Seeking. *Journal of Interactive Advertising*, 6 (2), 2–14.
- Günemann, S., Günemann, N., Faloutsos, C. (2014). Detecting anomalies in dynamic rating data: a robust probabilistic model for rating evolution. *In Proceedings of the 20th 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 manuscript*.
- 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.

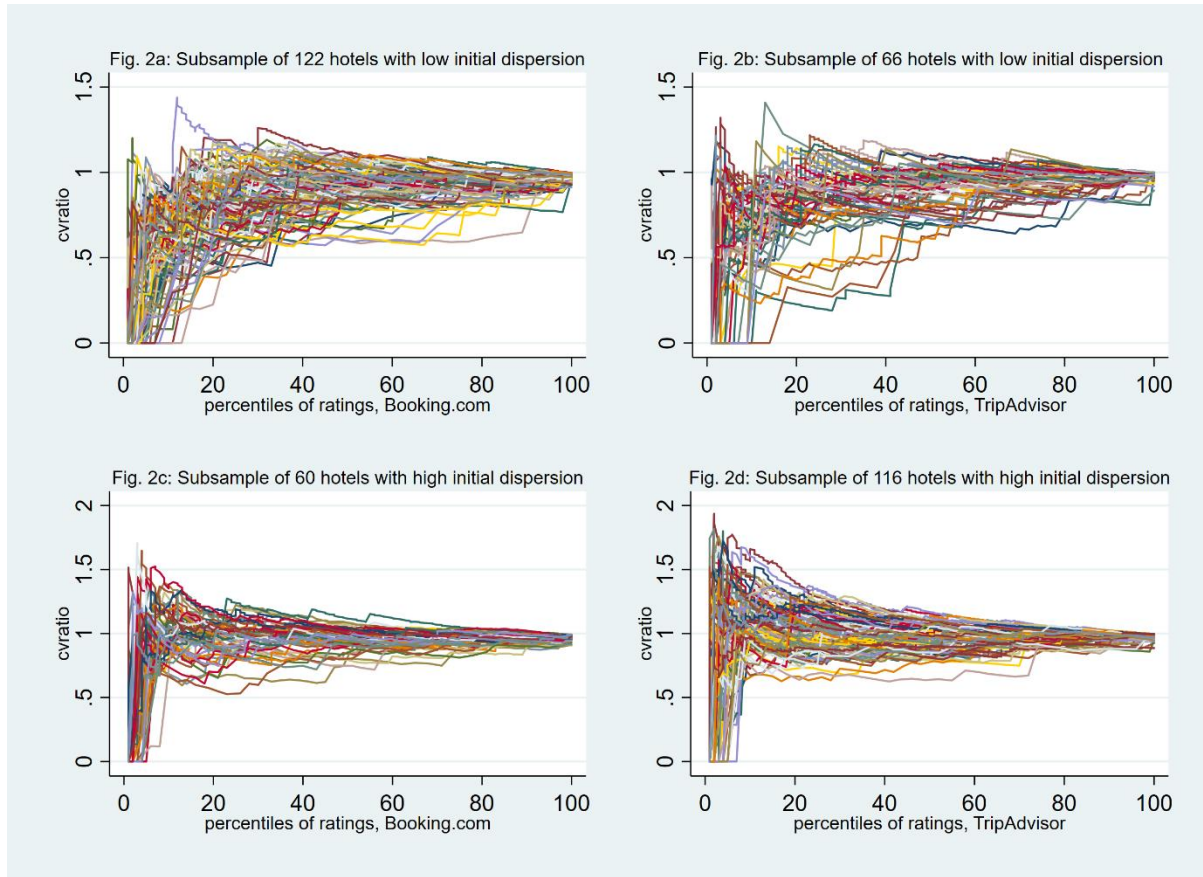
- Park, S., Nicolau, J.L. (2015). Asymmetric effects of online consumer reviews, *Annals of Tourism Research*, 50, 67-83.
- Pera, R., Viglia, G., Grazzini, L., & Dalli, D. (2019). When empathy prevents negative reviewing behavior. *Annals of Tourism Research*, 75, 265-278.
- Plotkina, D., Munzel, A., 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 Product Ratings Matter?, *Management Science*, 58(4), 696-707.
- TripAdvisor (2019). 2019 TripAdvisor Review Transparency Report. https://www.tripadvisor.com/TripAdvisorInsights/wp-content/uploads/2019/09/2147_PR_Content_Transparency_Report_6SEP19_US.pdf
- 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 and 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 and 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.

Figure 1 – The dynamics of ratings’ variability between TripAdvisor and Booking.com



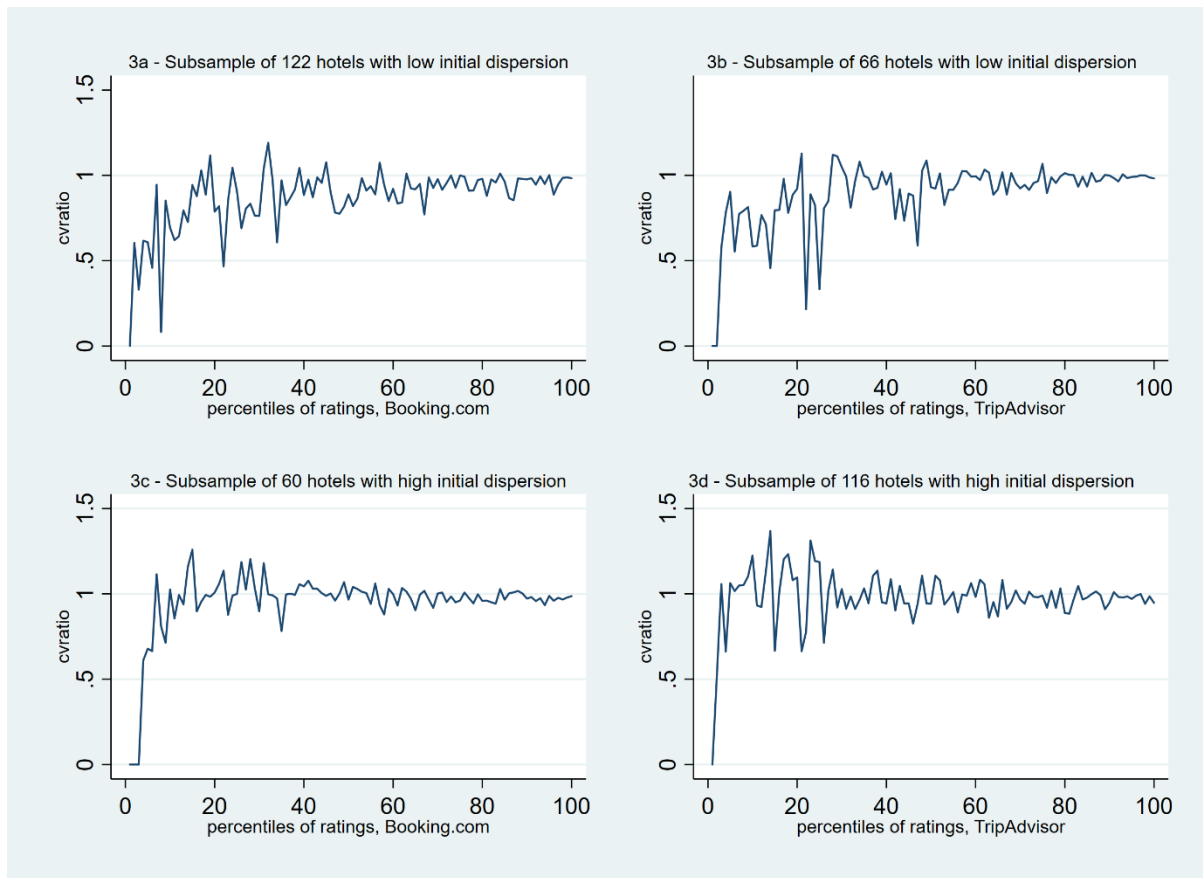
Notes: cvratio is the ratio between the cumulative standard deviation and its final value, computed for each hotel and each platform; ratings are sorted according to their review date and normalized in percentiles. 3-star hotels are reported in (1a) and (1b); 4- and 5-star hotels are reported in (1c) and (1d); Booking.com ratings are reported in (1a) and (1c); TripAdvisor ratings are reported in (1b) and (1d).

Figure 2 – The dynamics of ratings’ variability in different sub-samples of hotels, TripAdvisor and Booking.com



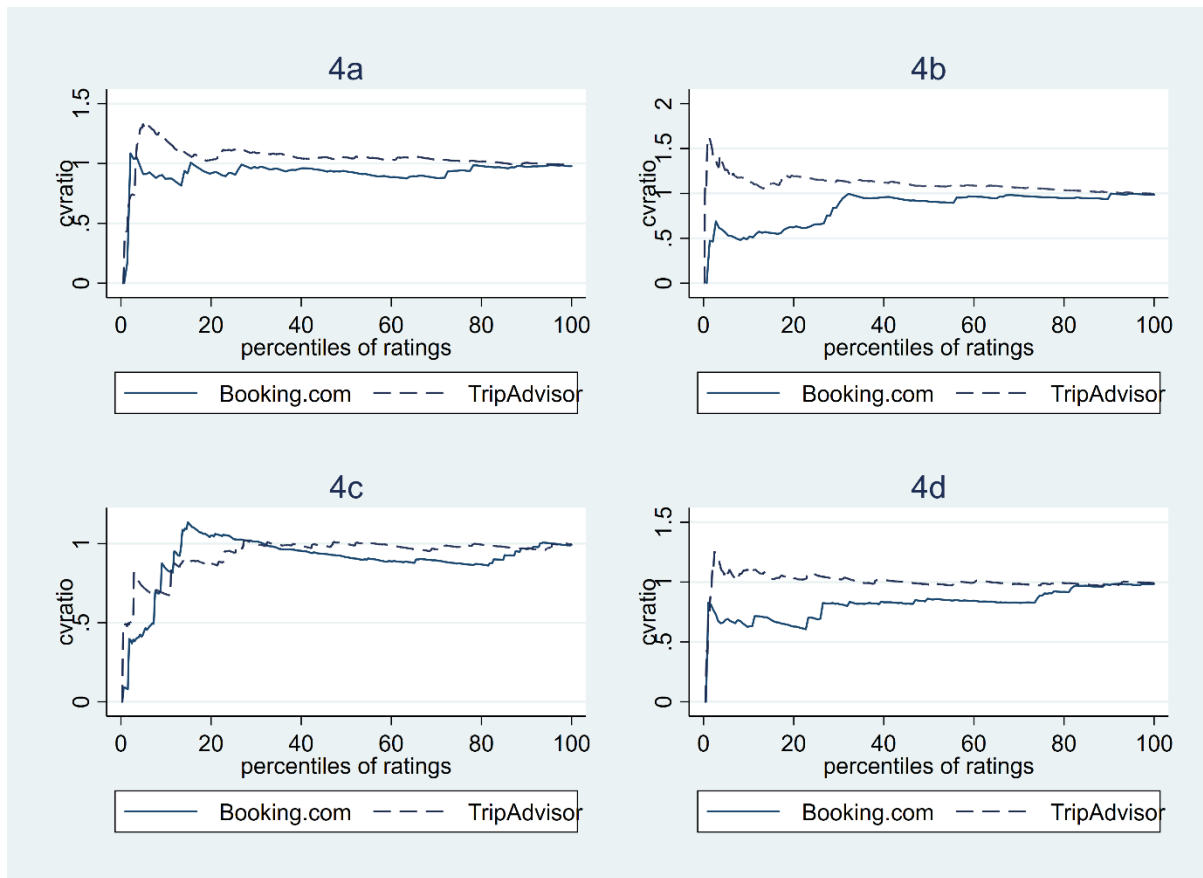
Notes: cvario is the ratio between the cumulative coefficient of variation and its final value, computed for each hotel and each platform; ratings are sorted according to their review date and aggregated for each percentile of the distribution. The sub-samples of hotels for which cvario is lower than 1 in the first decile of the distribution are reported in (2a) for Booking.com and (2b) for TripAdvisor. The sub-samples of hotels for which cvario is higher than 1 in the first decile of the distribution are reported in (2c) for Booking.com and (2d) for TripAdvisor.

Figure 3 – The average dynamics of ratings’ variability in different sub-samples of hotels, TripAdvisor and Booking.com



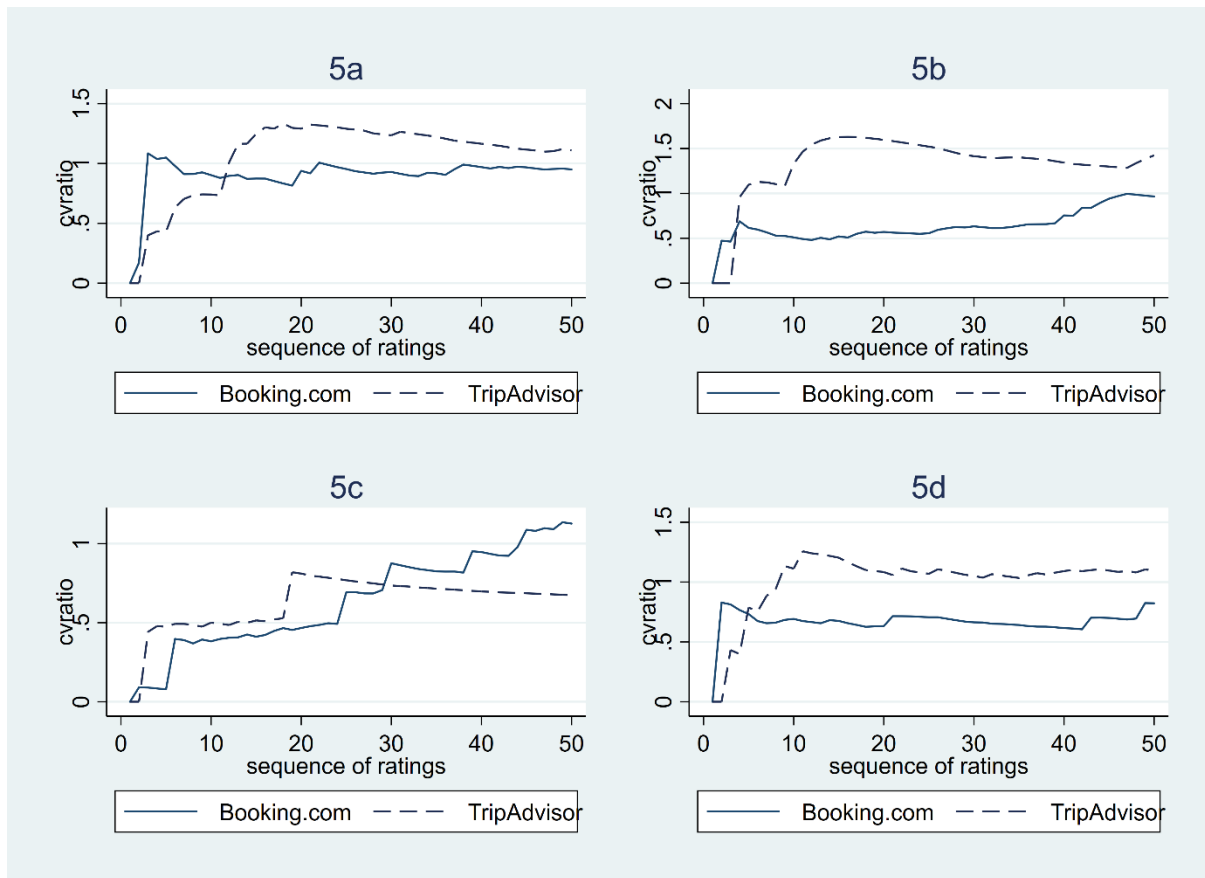
Notes: cvratio is the ratio between the cumulative coefficient of variation and its final value, computed for each hotel and each platform; ratings are sorted according to their review date and aggregated for each percentile of the distribution. Values of cvratio are the averages of the pool of hotels for each percentile of the distribution. The sub-samples of hotels for which cvratio is lower than 1 in the first decile of the distribution are reported in (3a) for Booking.com and (3b) for TripAdvisor. The sub-samples of hotels for which cvratio is higher than 1 in the first decile of the distribution are reported in (3c) for Booking.com and (3d) for TripAdvisor.

Figure 4 – 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 and normalized by percentile.

Figure 5 – 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.

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

Sample	Obs	Mean TA	Mean BC	KS test	WMW test	Prob mean TA > mean BC	Effect size (Cohen's d)	
Total	182	4.20	4.11	0.232***	52.55 ***	63.7%	0.252	SM
1 and 2 stars	7	4.30	4.27	-	-	-	0.118	S
3 stars	123	4.19	4.07	0.240***	47.84 ***	65.8%	0.311	SM
4 and 5 stars	52	4.21	4.17	0.240***	23.76 ***	61.5%	0.113	S
City Centre	85	4.17	4.15	0.192***	29.16 ***	52.9%	0.046	S
Seaside	97	4.22	4.07	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 Standard Deviation between TripAdvisor (TA) and Booking.com (BC)

Sample	Obs	St. dev. TA	St. dev. BC	KS test	WMW test	Prob sd TA > sd BC	Effect size (Cohen's d)	
Total	182	0.92	0.76	0.289***	76.80 ***	79.7%	0.819	L
1 and 2 stars	7	0.96	0.67	-	-	-	1.481	L
3 stars	123	0.92	0.77	0.276***	49.97 ***	76.4%	0.787	L
4 and 5 stars	52	0.92	0.76	0.324***	51.79 ***	86.5%	0.797	L
City Centre	85	0.93	0.75	0.357***	69.08 ***	90.6%	1.074	L
Seaside	97	0.91	0.78	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.

Table 3 – The difference in Mean Absolute Percentage Deviation between TripAdvisor (TA) and Booking.com (BC)

Sample	Obs	MAPD TA	MAPD BC	KS test	WMW test	Prob MAPD,TA > MAPD,BC	Effect size (Cohen's d)	
Total	182	17.8	15.2	0.228***	47.20 ***	73.1%	0.485	M
1 and 2 stars	7	16.7	12.3	-	-	-	0.975	L
3 stars	123	18.0	15.4	0.198***	25.25 ***	69.1%	0.447	M
4 and 5 stars	52	17.6	14.9	0.307***	37.43 ***	82.7%	0.530	M
City Centre	85	18.0	14.6	0.335***	51.89 ***	84.7%	0.712	ML
Seaside	97	17.7	15.6	0.162***	16.42 ***	62.9%	0.342	M

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.