UP OR OUT—
THE DYNAMICS OF STAR RATING SCORES ON AIRBNB

Research paper

Teubner, Timm, TU Berlin, Germany, teubner@tu-berlin.de

Glaser, Florian, Karlsruhe Institute of Technology (KIT), Germany, florian.glaser@kit.edu

Abstract

This paper investigates the impact of dynamic processes, including survivorship, as a possible reason for the distribution skewness of star ratings in C2C platforms. We draw on actual Airbnb data covering a time frame of 19 months from October 2015 to May 2017, comprising information on the listings' number of ratings and average rating scores. Building on research approaches from empirical finance, we find that rating distributions vary markedly when differentiated by underlying volume. While for few ratings, basically the entire bandwidth of possible scores is represented, the distribution becomes narrower for larger numbers of ratings. Interestingly, this is not associated with changes in average rating scores. Also, we observe higher churn rates for listings with lower rating scores. The market is growing and exhibits high turnover rates of about 7% per month. Overall, we find that Airbnb's rating score skewness is caused by a multiplicity of influences, including survivorship and the constantly high market share of new arrivals. We discuss our findings in view of the important role of star ratings as a popular design element within the digital platform economy.

Keywords: *Airbnb, Star ratings, Survivorship bias, Reputation, Trust*

1 Introduction

With more than 3 million listings and an evaluation of $30bn, Airbnb represents the single most important and successful platform for peer-based accommodation sharing (Airbnb 2017; Forbes 2017), presumably for peer-to-peer based sharing in general. While it represents an interesting case as a business model and a field of application for many kinds of information systems in and by itself, it also serves as a template for various other ventures, reflected in many startups' claim to be "the Airbnb of …" (Horton and Zeckhauser 2016). Its function as a role model makes Airbnb worth studying all the more. This said, one of Airbnb's central challenges and strengths is the maintenance of trust between its users, between hosts and guests, that is (Gebbia 2016; Möhlmann 2015). One of the most salient and disputed means to achieve this is Airbnb's *star rating system*, where after a completed transaction, guests rate their hosts on a scale between one and five stars.[1] On each listing's page, an averaged score of all transactions is displayed, rounded to the half unit (Airbnb 2014; Zervas et al. 2015). Such ratings play a central role for the formation of trust in electronic C2C commerce and hence the realization of transactions altogether (Ert et al. 2016; Zervas et al. 2015).

The literature on Airbnb reports highly skewed distributions of rating scores, where the great majority of ratings is equal to or higher than 4.5 stars (Teubner et al. 2017; Zervas et al. 2015). This has raised some doubts about the functionality and effectiveness of Airbnb's reputation system, and it has almost become a fashion to deride it as dysfunctional and ineffective (Wolff-Mann 2016). While this observed skewness mostly remains a black box to date, recent literature has brought forward several possible explanations, including customer self-selection, herding behavior, and non- or under-reporting of negative experiences. The latter may, in turn, be due to several reasons such as personal contact, reciprocity, fear of retaliation, or publicity (Bridges and Vásquez 2016; Fradkin et al. 2017; Zervas et al. 2015). While such candidate explanations are increasingly being discussed, little have they been examined empirically. We suggest that reputation systems represent a fundamental component in the machine room of today's electronic commerce – in particular for peer-to-peer platforms with heterogeneous users, goods, and services. In this paper, we hence investigate the dynamics of Airbnb rating scores, presumably the most successful and influential peer-to-peer platform. Specifically, this paper's research objective is to explore how *processes of survivorship* shape the rating score distribution. The paper's main idea is that rating score distributions in peer-to-peer platforms are subject to survivorship bias, driven by the assumption that better-rated providers exhibit lower probabilities to drop out of the market.

To do so, the paper draws on actual Airbnb data covering a time frame of 19 months from October 2015 to May 2017. The data comprises a total of 43,288 listings and their associated characteristics, particularly including number of ratings and average rating scores. We find that rating distributions differ markedly when differentiated by their underlying volume, that is, on how many ratings they are based on. While for few ratings, basically the entire bandwidth of possible scores is represented, the distribution becomes narrower for higher numbers of ratings. Interestingly, this does not go along with changes in average rating scores. Also, we observe higher churn rates for listings with lower rating scores. Overall, the market is growing and exhibits high turnover rates. On average, 6.8% of all listings disappear while at the same time, new listings in the amount of 8.5% of the population are entering the market – every month(!). Overall, we find that Airbnb's rating score skewness is caused by a multiplicity of influences, including *survivorship* and the constantly high market share of *new arrivals*. The paper contributes to the literature by shedding light on an important pillar of electronic commerce, in particular for C2C platforms, by means of empirical analysis. In doing so, we foster the understanding of the intricacies of star ratings, reputation, and trust in electronic C2C commerce. This paper is among the first to leverage actual Airbnb data over an extended period of time to consider provider reputation.

The remainder of this paper is organized as follows. In Section 2, we outline the background of Airbnb, its reputation and star rating system, as well as a theoretical conception of how listings and ratings behave over time. Next, in Section 3, we describe our data model, our research approach, and our main

[1] Note that hosts also rate guests. Numerical guest ratings, however, are not displayed on the platform and it is not entirely clear what these ratings are used for (Airbnb 2016b).

results. Moreover, we consider alternative explanations and present a first approach to predict churn. Section 4 concludes this paper by a discussion of our findings in view of implications, limitations, and potential for future work.

2 Theoretical Background and Related Work

Many papers on Airbnb report highly skewed rating score distributions, where the great majority of ratings is equal to or higher than 4.5 out of 5.0 stars. Recent literature has brought forward several possible explanations for this remarkable skewness. First, non- or under-reporting of negative experiences may cause an overly positive distribution (Fradkin et al. 2017). Under-reporting itself may have various reasons. One suggestion is *reciprocity* which may lead guests and hosts to exchange utmost positive reviews, regardless of their actual experiences (e.g., based on a tit-for-tat logic or on fear of retaliation). Since 2014, however, Airbnb has made an important design change in this regard, where ever since reviews are only revealed once both parties have bindingly submitted their respective assessments (Airbnb 2014). Reciprocity can thus be ruled out as a driver of distribution skewness.

More promising, the high degree of personal contact and interaction associated with staying at someone's apartment may prevent all too critical assessments. Submitting a negative rating to someone with whom one has spent time may simply feel socially awkward, like a "face-threatening act", and hence let users withhold their complaints (Bridges and Vásquez 2016; Ikkala and Lampinen 2015). As we demonstrate in Section 4, prior research and our data suggest that average rating scores and their distributions differ only marginally between transactions with and without extended personal contact between host and guest (4.707 vs. 4.690), rendering also this potential explanation rather unlikely.

Next, the public availability of reviews may impede guests from submitting negative reviews as they would implicitly admit poor judgement with regard to their accommodation choice (Bridges and Vásquez 2016). Also from a strategic perspective, all too critical reviews may impair their chances of being accepted by future hosts; after all, who would want to rely on the judgement of a pedant (Mulshine 2015)? Note, however, that only *text* reviews were displayed publicly and can be associated with the author. A listing's *star* rating, in contrast, represented an aggregated (i.e., averaged and rounded) value and it was not retrievable which guest submitted which rating exactly. Apparently, Airbnb has been experimenting with this where for some users, individual star rating scores are now displayed.

Another possible explanation refers to Airbnb as a rather novel mode of consumption with mostly first-time transactions (Teubner 2017). Compared to a hotel stay, guests on Airbnb may simply not exactly know what to expect, and hence not have high expectations or standards which could be disappointed. This may shift ratings toward the positive end of the scale. Systematic comparison of apartments that are listed both on Airbnb and TripAdvisor in fact reveals a stronger positive bias in Airbnb reviews than for the same apartments on TripAdvisor (Fradkin 2015). Similarly, comparing the platforms Airbnb and Booking.com for several European cities, there occurs a consistent gap of 20% in the average score in favor of Airbnb listings (Ert et al. 2016).

In this paper, we consider an additional potential explanation for the dynamics of rating scores. We suggest that distinct survivorship processes may facilitate the large proportions of top-rated listings even though the "initial" distribution may be much less skewed. That is, the large share of top-rated listings may be partially attributable to increased survivorship of well-rated listings. We outline the theoretical basis underlying this phenomenon in the following section.

2.1 Survivorship, regression toward the mean, and the law of large numbers

The theoretical conception of the survivorship bias is straightforward. Its main premise is that there occurs a logical error when concentrating on objects or observations that emerged as a result of some selection process and overlooking those that did not – typically due to their lack of visibility (Brown et al. 1992). Given that most markets do not represent static constructs but are in a steady process of change, survivorship bias applies when considering the prevalence of market players and player types. In this sense, listings on Airbnb are subject to an ongoing selection process, where new listings enter the market while others drop out. Prior research on reputation dynamics on Taobao.com (China's largest e-

commerce platform) has found that seller reputation has a positive effect on established sellers, whereas it does not for new sellers in the sense that "at any point of time, better-reputed established sellers are more likely to survive for another six months" (Fan et al. 2013, p. 4). Likewise, Cabral et al. (2010) considered time series of eBay seller reputation and found that when a seller first receives negative feedback, weekly sales dropped distinctly. Moreover, they found higher market exit rates for sellers with low than for those with good reputation. This common-sense observation is theoretically underpinned by *signaling theory*, linking the availability of reputation (i.e., signals) to the formation of trust (Riegelsberger et al. 2005; Spence 2002). The theory assumes two parties with conflicting interests and information asymmetry. It posits that sellers (or, more generally speaking, providers) can reduce information asymmetry and the associated uncertainty by providing signals, for instance, in the form of ratings by third parties (Basoglu and Hess 2014; Donath 2007; Dunham 2011). These will contribute to building consumer trust and hence to a listing's economic success.

For the dynamics of listings and rating scores on Airbnb, we assume that the principles of signaling and survivorship are at work. In fact, high reputation was found to translate into tangible economic value, where providers of well-rated listings charge higher prices for their offers (Ikkala and Lampinen 2015; Teubner et al. 2017). Such findings are consistent with earlier work on seller reputation and auction end prices on eBay, where better seller reputation is usually associated with higher end prices (Resnick et al. 2006). Moreover, Airbnb listings with better ratings are found to attract more demand (i.e., .169 customers/month per star; Ke 2017). In contrast, lower scores force owners to set lower prices as they are regarded as less trustworthy (Ert et al. 2016; Teubner et al. 2017). Providers with low reputation may hence face lower chances of being booked and not be able to charge sufficiently high prices to cover their cost or generate satisfying revenues. Moreover, Fradkin et al. (2017) report that Airbnb's ranking algorithm explicitly ranks down poorly rated hosts. As a consequence, we suggest that high reputation (in the form of high star ratings) will increase a listing's chances for economic success and hence its likelihood to remain in the market. As a result, favorable (self-) selection – as opposed to adverse selection (Akerlof 1970) – will create a tendency toward a prevalence of well-rated listings.

Other disciplines have been facing similar phenomena. In empirical finance, for instance, survivorship bias has long been studied (Brown et al. 1992). In particular, the performance of fund managers is a common subject of study and often analyzed over time to assess performance persistence. Usually the goal is to make a deduction regarding future performance of a fund based on its past performance (Grinblatt and Titman 1992). A crucial problem with such analyses results from the fact that investment funds are taken out of business when their performance remains poor for several periods. Statistical inference based upon samples which included funds that were present in earlier periods but ceased to exist before the last period of the sample were found to bias the results. Survivorship bias was studied intensively and databases widely used for investment fund performance analyses were adjusted as a counter measure (Carpenter and Lynch 1999). We argue that previous findings about the survivorship bias in research on investment funds might provide helpful insights and guidance for the analysis of digital platform user characteristics and their performance. On platforms such as Airbnb, from the customers' perspective, "performance" can be approximated by rating scores. Although the underlying business, associated risks, and variability are different, the dynamics that induce a survivorship bias in data samples exhibit vital similarities. For example, low performing funds are less frequently displayed at crucial points of sale and public rankings just like low-rated users on digital platforms are likely to be disadvantaged by filtering and sorting mechanisms (Fradkin et al. 2017). Eventually, in analogy to low performing funds being taken out of business, hosts that exhibit low ratings may over time leave the platform or become inactive due to lack of economic success.

Beyond processes of survivorship, we suggest that the reputational dynamics of Airbnb listings are subject to rather mundane statistical phenomena. In this regard, *regression toward the mean* refers to the observation that if a variable is extreme on its first measurement, it will tend to be closer to the average on the second measurement (Barnett et al. 2005; Bland and Altman 1994; Stigler 1997). Also, a high standing rating may elicit customer expectations of an exceptionally good experience, which may effectively increase the chances of *disappointment* and, in turn, facilitate lower customer ratings. Likewise,

if being booked at all, listings with low ratings may facilitate *unexpectedly* positive experiences. In either case, the self-regulating dynamics of expectations drive the overall rating score towards moderate values. Moreover, the *law of large numbers* suggests that when repeatedly performing the same experiment often enough (i.e., receiving a star rating), the average of these results converges towards the expected value. Applied to ratings of Airbnb listings, this suggests that listings with ever-increasing review counts will have a decreasing likelihood to exhibit the highest possible rating scores.

Hence, several statistical implications can be transferred from other domains to the case of Airbnb. First, spurious performance persistence might be revealed and indicate that previous performance predicts future performance to an unjustifiable degree. Second, performance of a survivorship biased sample displays a J-shape of performance if plotted against initial period performance. This shape falsely indicates higher performance in later periods if performance was high in the initial period, hence leading to an overestimation of persistence over time. Third, analyses of performance-conditioned samples might be biased due to survivorship effects and the dropping out of low performing listings. In the following, we hence scrutinize actual Airbnb panel data in view of how survivorship bias and related dynamic phenomena affect the appearance of rating scores.

3 Data, Methods, and Results

The data was retrieved from InsideAirbnb.com (Cox 2017; Wired 2017). For 44 cities, InsideAirbnb provides a detailed data set of all retrieved listings at that time, including various variables such as each listing's price, room type, rating score, and the number of ratings. For some of these cities, a monthly data history is available. In view of rating scores, note that although Airbnb displays ratings on a scale from 1 to 5 stars (rounded to the half unit), the underlying data is more fine-grained. It ranges on a scale from 20 to 100 in steps of 1, where the interval [20, 24] corresponds to 1.0 star, [25, 34] to 1.5 stars, and so on (see Teubner 2017 for details). In this paper, we focus on the Berlin data. Preprocessing yielded a data structure as exemplified in Table 1 with a total of 43,288 distinct listings. Note that many listings only started to occur in a later than the first time frame (e.g., first row). Others dropped out and did not show up again (e.g., second row). Also note that for some listings, there occurred *gaps*, that is, time frames with no data but with data before *and* after this time frame (e.g., third row). As a first step of data preprocessing, such gaps were filled in order not to overestimate arrival and churn rates. Specifically, a gap between month i and i+x was filled by repeating the rating score of month i.

Listing ID	t_1	t_2	t_3	t_4	t_5	…	t_{19}
1	—	—	96	96	97	96	96
2	80	80	80	75	65	65	—
3	75	75	75	—	75	80	80
4	60	88	95	95	97	98	98
…							

Table 1. Data structure with indication of missing values, i.e., rating scores [20, 100]

As a next step, we consider the number of available listings in each period as well as the number of appearing and disappearing listings, see Figure 1. Overall, the available number of listings is growing by 1.6% per month on average. As indicated in Figure 1, there occurred some turmoil in 2016 due to the commencement of a law in Berlin, restricting short-term rental (BATO 2016). Until May 2017, the market has settled at just above 20,000 listings. Of these, 15,895 have a rating score, that is, have received at least one rating.

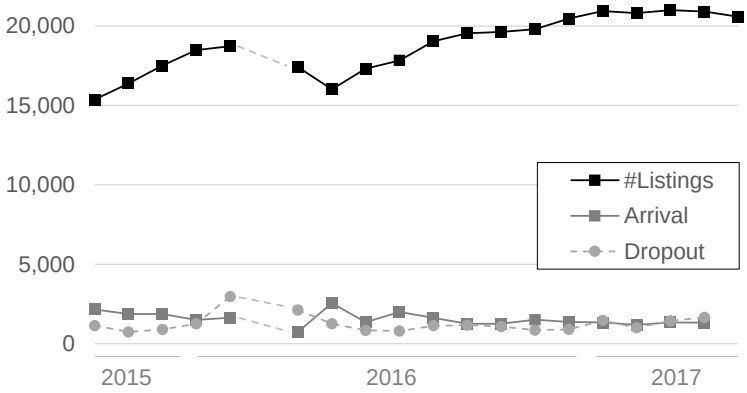

Figure 1. Overall number of Airbnb listings, arrivals, and market dropouts for Berlin.

Figures 2 depicts the distributions of (a) *star rating scores* and (b) *number of ratings* of the 20,576 listings available in the most recent data set (May 2017). It is important to note that there are listings with no ratings and hence no rating score. The most common number of reviews is in fact "0". The overall mean is 16.69 with a standard deviation of 29.39 and a median of 6. With regard to star ratings, the most common value is 5.0 stars (mean = 4.696, sd = .425, median = 5.0). The distribution is as follows: .55 (5.0 stars), .33 (4.5 stars), .09 (4.0 stars), and .03 (≤3.5 stars). For the underlying detailed rating score value, the mean is 93.39 with a standard deviation of 8.05 and a median of 95.

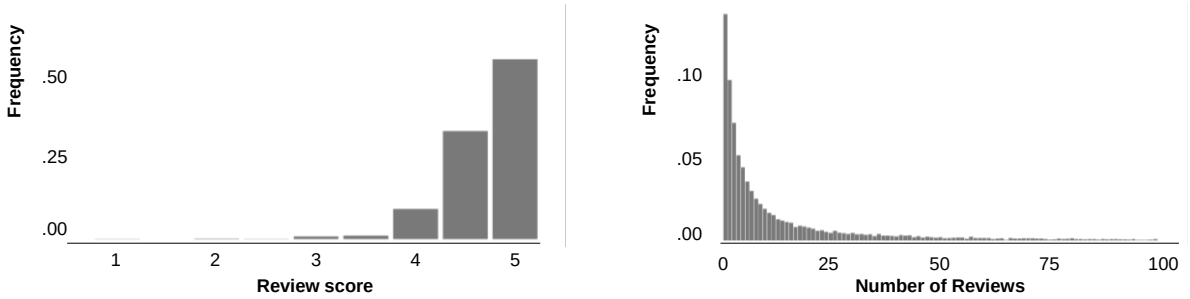

Figure 2. (a) Distribution of star rating scores; (b) Distribution of number of reviews.

Interestingly, listings receive .77 new ratings each month on average (sd = 1.61, max = 47). Given that about half of all requests are accepted by Airbnb hosts and that only about half of all stays are reflected in reviews (estimates vary between 30% and 70%), this yields 3 incoming booking requests per month on average (Cox 2017; Edelman et al. 2017; Fradkin 2015).

3.1 Differentiation of rating scores

We now take a closer look at the distribution of rating scores when the distribution is differentiated by the underlying number of ratings. Figure 3 depicts these distributions with the rating score flipped to the y-axis. In Figure 3(b), from left to right, the number of ratings increases where each bin has twice the size of the preceding one. All corresponding values are provided in Table 1. As can be seen from this, the rating scores' range decreases from 80 (=100–20) to 24 (=99–75). Similarly, the interquartile range (IQR) decreases from 20 (=100–80; 1 rating) to 5 (=96–91; ≥64 ratings). At the same time, the correlation between rating score and number of ratings is insignificant, that is, a listing's rating score does not linearly depend on the underlying number of ratings (r = –.013; n.s.)

Result 1: *For increasing numbers of ratings, the distribution of rating becomes narrower (i.e., smaller inter-quartile range) and transitions from multi- to unimodality while retaining its mean (see Figure 3 and Table 2).*

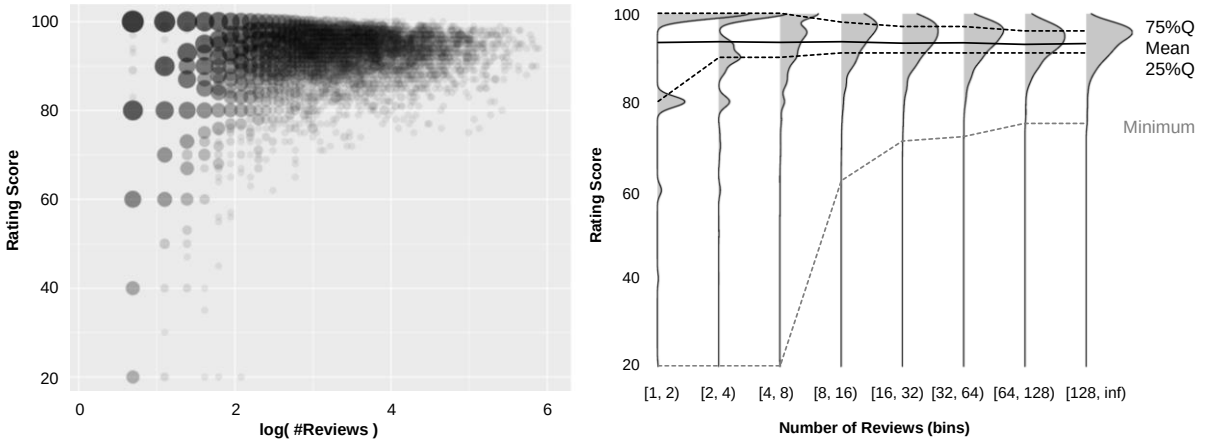

Figure 3. *(a) Scatter plot of listings (x = log(#reviews), y = ratings score [20, 100]). Bubble size indicates number of listings on particular position; (b) Rating score distributions, differentiated by underlying number of reviews (bins), including indications of mean, interquartile range, and minimum values.*

#Reviews	N	Mean	SD	IQR	Min	25%	Median	75%	Max
[1, 2)	2,464	93.35	12.70	20	20	80	100	100	100
[2, 4)	3,298	93.54	9.30	10	20	90	100	100	100
[4, 8)	3,193	93.40	7.30	10	20	90	95	100	100
[8, 16)	2,659	93.55	5.68	7	62	91	95	98	100
[16, 32)	1,958	93.21	4.93	6	71	91	94	97	100
[32, 64)	1,365	93.29	4.43	6	72	91	94	97	100
[64, 128)	715	92.92	4.14	5	75	91	94	96	100
[128, ∞)	243	93.17	4.19	5	75	91	94	96	99
All	**15,895**	**93.39**	**8.05**	**10**	**20**	**90**	**95**	**100**	**100**

Table 2. *Key indicators of rating score distribution for different review counts (May 2017).*

3.2 Dynamics of rating scores

As a next step, we consider transitions between the possible rating score "states" of a listing. Here, a state refers to a listing's star rating score, where all scores of 3.5 stars or less are aggregated into a joint category. We can now compute the state transition probabilities based on the observed data. These values indicate, for instance, how likely a listing with 4.5 stars is to increase/decrease its rating to 5.0/4.0 stars in the subsequent month. Also, this yields a differentiated set of churn rates, that is, the probabilities of a listing dropping out of the market, depending on its rating score. Moreover, we obtain an assessment of the probabilities for different star ratings of newly arriving listings. A summary of these state transitions is depicted in Figure 4(a). As can be seen there, the highest values pertain to listings retaining their rating scores (values between 80.5% and 89.1%). A new listing entering the market has a probability of 60.9% to be rated 5.0 stars, 19.9% for 4.5 stars, 14.6% for 4.0, and 4.6% for 3.5 stars or less (values printed in blue). With regard to churn, lower rating scores are associated with higher churn rates (7.1% < 7.4% < 8.2% < 11.6%). Thus, listings with 3.5 stars or less exhibit 63% higher churn rates than those with 5.0-star ratings.

Result 2: *Listings with lower star rating scores exhibit higher churn rates, that is, a higher probability to drop out of the market in the subsequent month.*

Now, this network of states and state transitions can effectively be understood as a Markov chain. Markov chains enable to model and quantify the described dynamics (Gagniuc 2017). Here, the population of listings is assumed to be distributed across different states, which reflect rating scores (i.e., 5.0, 4.5,

and so on). It is then assumed that, depending on a listing's state, it has certain probabilities of transitioning into another state (e.g., from 4.5 to 5.0 stars, or from ≤3.5 stars to "out of business"). Applying a Markov chain model, the steady state distribution will then depend on the transition probabilities between states, churn and arrival rates, as well as the overall magnitude of supply growth, stagnation, or decrease. Assuming a constant rate of new arrivals, the Markov process converges against a "steady state" which represents a certain distribution of star rating scores.

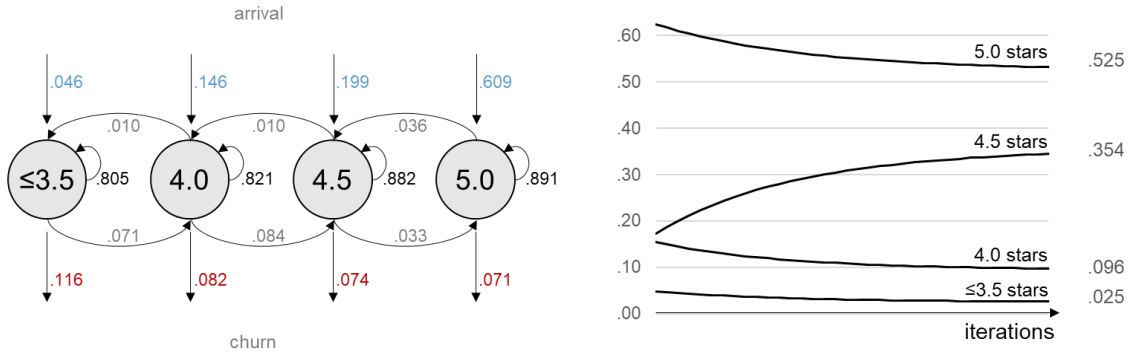

Figure 4. (a) State transition network (partial); (b) Approximation of steady state distribution.

As depicted in Figure 4(b), the steady state proportions are ≤3.5: 2.5%, 4.0: 9.6%, 4.5: 35.4%, 5.0: 52.5%. These values (obtained through the Markov process) almost perfectly replicate the observed distribution of rating scores. As an assessment of distribution skewness, we employ Fisher-Pearson coefficient of skewness G_1 (Doane and Seward 2011), which is $G_1^{steady} = 41.29$. Note that this steady state distribution is much less extreme than that of new arrivals ($G_1^{new} = 93.93$). In particular, note that the proportion of 5.0-star ratings is *lower* than in the initial proportion. Thus, although natural selection processes have an impact on the development of the overall rating score distribution (see Results 1 and 2), it cannot be concluded that this (alone) governs the observed distribution's skewness.

While a detailed assessment of all involved processes in beyond the scope of this paper, we briefly outline two additional relevant aspects in this regard. First, in every period, a significant set of new arrivals enters the market (8.5%). The specific distribution of these new listings hence represents a constant and formative factor for the overall distribution. Second, processes of mean regression and the law of large numbers most certainly have a mitigating effect on the distribution. This is reflected in the state transition probabilities depicted in Figure 4(a). Here, the 3.6% transitioning from 5.0 to 4.5 star ratings has a much larger impact than the 3.3% moving from 4.5 to 5.0 stars.

Result 3: *Survivorship bias alone does not explain Airbnb's rating score skewness since a) the distribution of first ratings is already highly skewed and b) additional processes overlay the effects of survivorship on rating distribution.*

3.3 Predicting churn

We now take a closer look at churn rates. As described above, churn refers to a listing which drops out of the market (for good). The analysis revealed that lower rating scores imply higher churn rates. Beyond rating scores, we take the control variables price, number of ratings, and room type (i.e., entire apartment or private room) into account. These variables are used as independent variables within a logistic regression to predict churn, where the churn indicator is "1" if in time step t to t+1 the listing drops out of the market, and "0" otherwise. As summarized in Table 4, this analysis confirms and corroborates the robustness of the findings outlined above. Specifically, higher rating score values have a significant and negative effect on churn ($b = –.019$, $p < .01$). Also, churn rates appear not to depend on price or apartment type. Furthermore, the number of ratings has a significant effect on churn, where larger numbers imply lower churn rates ($b = –.003$, $p < .05$). This analysis corroborates this paper's finding that lower ratings are associated with churn – also when controlling for number of ratings, room type, and price.

3.4 Other explanations

The literature suggests the high degree of social interaction between hosts and guests as a potential driver of high ratings and hence rating distribution skewness (e.g., Bridges and Vásquez 2016). To address this possible alternative explanation, we leverage the fact that listings on Airbnb can be differentiated by type. While "entire homes/apartments" refer to listings that are rented out entirely and hence host and guest only meet for key exchange (if at all), "private rooms" refer to guest rooms within the host's apartment (typically with shared bathroom, kitchen, etc.), rendering the personal and social dimension much more prevalent (Ikkala and Lampinen 2015). Note that Airbnb's third category ("shared rooms") plays a negligible role. We compare the rating distributions for listings from the "entire homes/apartments" category (51.6%) with those from the "private room" category (47.2%). Systematic rating differences due to social interaction during the stay should be reflected in the differentiation of these accommodation categories. A first visual assessment shows that both distributions look quite similar, where, in fact, private rooms exhibit a higher proportion of 5.0 star ratings (57.3% vs. 53.8%) and a lower proportion of 4.5 star ratings (31.2% vs. 34.8%) (see Figure 5). A two-sample Kolmogorov-Smirnov test suggests the distributions to be different (D = .040, p < .001). Considering average values for the two categories, we find the private room ratings to be slightly higher than those for entire apartments (4.707 vs. 4.690 stars). An independent sample t-test as well as a (non-parametric) Wilcoxon test both yield a significant difference (*delta* = .017 stars, T = 2.571, $p_{t\text{-}test}$ = .0104; W = 29846000, $p_{Wilcoxon}$ < .001), but the effect size is small (Cohen's d = .043).

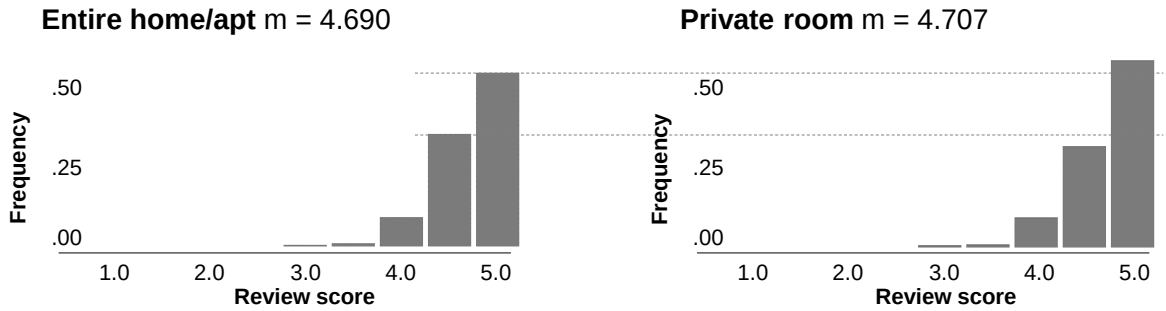

Figure 5. Comparison of rating distribution by room type.

To control for potential confounding effects (entire apartments may be systematically more expensive, exhibit more/less reviews, etc.), we conducted an additional linear regression (Table 4). This revealed that when controlling for price and number of reviews, room type does not have a significant effect on rating score. Thus, overall, social interactions during the stay do not appear to be the reason for rating skewness. These results echo prior findings in this regard. Specifically, based on 2.3 million Airbnb listings, Ke (2017) report an average rating of 4.67 stars with little to no differences in the star rating distribution when differentiated by room type. Also Teubner et al. (2017) report insignificant correlation between room type and rating scores.

		DV: Churn (0/1)		DV: Rating score	
Independent variable	**Scale**	**Coef.**	**Sig.**	**Coef.**	**Sig.**
Rating score	[20, 100]	**–.019**	**	–	–
Number of ratings	>= 0	**–.003**	*	–.003	
Room type	(entire apt = 1, private room = 0)	–.134		–.248	
Price	(EUR)	–.001		**–.002**	**
Intercept		**–1.383**	***	**93.486**	***

Table 4. Regression summaries (*** p < .001; ** p < .01; * p < .05)

4 Discussion and Conclusion

Akerlof (1970)'s "market of lemons" exemplifies that markets with information asymmetry between providers and consumers may fail if providers cannot reliably signal quality. Since markets such as Airbnb exist, function, and flourish, the reputation of a provider gained through prior transactions does evidently represent a mechanism by which information about the provider's behavior can be credibly communicated to consumers. The provider's reputation as a signal reduces information asymmetries and thereby allow the market to function (Melnik and Alm 2002). Given this distinct role of reputation, it is important to understand how the systems and mechanisms employed to build and manage reputation work in general, and for specific applications in particular. Star rating systems (usually based on scales from 1 to 5 stars; Airbnb 2016a) represent one such system with extensive adoption in practice.

For providers, consumers, and platform operators, it is hence crucial to understand the meaning rating scores actually carry and how they emerge dynamically over time – potentially explaining much of what is casually referred to as the ratings' *skewness*. This paper hence follows a longitudinal approach which, to the best of our knowledge, is the first to do so for Airbnb. A growing body of recent work has leveraged "snapshots" of Airbnb, for instance, of listings (e.g, Edelman and Luca 2014; Ke 2017; Teubner et al. 2017) or reviews (e.g., Bridges and Vásquez 2016; Brochado et al. 2017). Time-series-based approaches such as the present one allow to investigate phenomena such as survivorship processes. Understanding the intricacies and implications of market survivorship processes is thereby practically important in at least two ways. First, an all too naïve assessment of immediate market data will yield spurious analyses and biased conclusions as it tends to focus on a special subset of the data (e.g., successful funds or listings). Second, customer churn represents a threat to basically all businesses. For Airbnb alike, it is hence essential to understand when, how, and why providers drop out of the market as these do not simply represent "customers" but the very pillar of Airbnb's (platform-) business model.

In this regard, we find a remarkable turnover rate of 7% per month and observe that churn rates are higher for low rated listings. This association of churn and rating scores is robust against controlling for other explanatory factors such as price, room type, and number of listings. Importantly, our findings may actually *underestimate* actual churn rates as we conceptualized churn as the permanent disappearance of a listing from the market. Many listings, however, may simply remain "dead" on the platform and may not be removed or deactivated (since there is no incentive for hosts to actually do so). Future work may attempt to identify and control for such non-active listings. Moreover, anecdotal evidence suggests that hosts who received low ratings may abandon their present account and open a new account for the same apartment, thus enabling a fresh start. This may, to some extent, also explain the high rates of churn and arrivals. Future work should attempt to identify and re-connect these "follow-up" listings.

Recent research sets out to predict how many new transactions a listing generates based on its properties (e.g., star rating, number of past transactions), suggesting that a difference of one star is associated with an additional .16 transactions per month on average (Ke 2017). Given that an average listing exhibits .77 transactions per month, having a five- rather than a four-star rating yields a difference of over 20%.

In view of survivorship processes, we find that such processes do play a role for star ratings on Airbnb, reflected by increasing churn rates for listings with lower ratings. Other processes, however, play a countervailing role and appear to dominate the picture insofar as that the distribution of new arrivals has higher skewness than the overall distribution. Survivorship bias alone does hence not explain Airbnb rating score skewness. When differentiated by their *number* of ratings, the picture is unambiguous (see Figure 3). Rating distributions become narrower (i.e., lower IQRs) and unimodal for higher numbers of ratings. For Airbnb listings, roughly speaking, it is hence either up or out. Moreover, while of course Airbnb's rating score distribution is far from uniform, the popular narrative that "all ratings are 5 stars anyway" must be rejected. After all, the scale allows for (apparently sufficient) differentiation.

Our findings are consistent with the those from empirical finance. High star rating listings seem to exhibit a high performance over time, that is, performance persistence. This relationship, however, might be biased due to the survivorship processes in the sample. A large number of accounts might have

dropped out at the time the snapshot was taken and hence the sample might be biased by the large proportion of surviving units. Given these initial indications of survivorship bias being present, we argue that other analyses of digital platform data might be subject to the bias and hence exhibit spurious correlation findings. We call for future research on generalizable insights regarding potential survivorship sample biases on C2C platforms. A holistic, clear, and statistically sound picture of the implications for inferences based on survivorship-biased samples are crucial foundations of guidelines for future empirical work based on data from digital platforms.

Future work may, for instance, apply hidden Markov models, a variant of Markov models in which the modeled system is assumed to be a Markov process with *unobserved* states. This appears well-suited to capture the notion that we can only observe an indirect proxy for quality (i.e., rating score) whereas the listings' actual (underlying) quality cannot be directly observed. Moreover, we call for future work to use the available data on InsideAirbnb.com or proprietary data sets to compare, replicate, and put the findings presented in this paper into (a global) perspective. Specifically, InsideAirbnb provides continuous chronological data for several other cities (e.g., New York City, San Francisco, Vienna). This seems to be particularly appropriate as there does exist systematic structure within rating scores across cities and regions (Dann et al. 2017). Airbnb listings in US and Canadian cities exhibit particular high scores. Also *prices* in North America are systematically higher in comparison to other regions. Moreover, the cities' and regions' Airbnb penetration (listings per 1,000 inhabitants) varies considerably. Values range from 1.6 (Manchester) to above 25 for Copenhagen (26.9) and Paris (25.2).

Overall, this paper must be seen as a first approach and future work is needed to substantiate, qualify, and extend our findings and method. For instance, a two-dimensional conception of Markov states (rating score and volume) may provide further insights since – as we have shown in the brief regression analysis on churn (Table 4), dropout rates do not only depend on rating quality but on quantity too. Note, in this regard, that higher underlying numbers of ratings will make the overall score more robust against single low evaluations. In this sense, an average star rating of 5.0 based on 3 ratings will drop to 4.0 in reaction to a spiteful or accidental 1-star rating. In contrast, with a basis of 15 ratings, the displayed score remains at 5.0 (i.e., rounded up from 4.75). Hence, listings with larger numbers of rating build up immunity against negative outliers. Thus, there is virtue in having *many* ratings (outlier immunity). But of course, there is also value in *the very first* ratings. In this sense, Livingston (2005) found that eBay seller profit highly when they receive first positive evaluations but that marginal returns to additional reviews are severely decreasing. This paramount importance of first reviews may be instrumental in explaining why the distribution of ratings of new arrivals on Airbnb is as skewed as it is. In a way, the first visual assessment sets a decisive switch for the economic path of listings and hosts (see also Gutt and Herrmann 2015).

The debate around five-star rating systems has revealed a great amount of customer confusion in perceptions of what is and should be considered a good rating score at all. For the TIME, Ethan Wolff-Mann (2016) nicely put it by stating that "at some point, maybe around when orange became the new black, four stars became the new zero." For star ratings as a design element of two-sided platforms, it is hence crucial to develop a clear understanding of the meaning of ratings and scores since otherwise, they will degenerate into pixels with zero practical value. With this paper, we intend to contribute to the ongoing debate by leveraging empirical data and developing novel insights based on temporal dynamics, which we believe is instrumental for understanding not only the *whats* and *wheres* of star rating systems, but also its *hows* and *whys*.

References

Airbnb. 2014. "Building trust with a new review system," *Airbnb* (available at http://blog.atairbnb.com/building-trust-new-review-system/; retrieved January 19, 2017).

Airbnb. 2016a. "How do star ratings work?," (available at https://www.airbnb.com/help/article/1257/how-do-star-ratings-work; retrieved September 14, 2017).

Airbnb. 2016b. "Can guests see their star ratings?," (available at https://community.withairbnb.com/t5/Hosting/Can-guests-see-their-star-ratings/td-p/126861; retrieved October 2, 2017).

Airbnb. 2017. "About us," (available at www.airbnb.com/about/about-us; retrieved February 15, 2017).

Akerlof, G. A. 1970. "The market for 'lemons': Quality uncertainty and the market mechanism," *The Quarterly Journal of Economics*, (84:3), pp. 488–500.

Barnett, A. G., van der Pols, J. C., and Dobson, A. J. 2005. "Regression to the mean: What it is and how to deal with it," *International Journal of Epidemiology*, (34:1), pp. 215–220.

Basoglu, K. A., and Hess, T. J. 2014. "Online business reporting: A signaling theory perspective," *Journal of Information Systems*, (28:2), pp. 67–101.

BATO. 2016. "The ban on vacation apartments in Berlin," *BATO Newsletter*, (7:7), pp. 1–2.

Bland, J. M., and Altman, D. G. 1994. "Some examples of regression towards the mean," *Statistical Notes*, (309), p. 780.

Bridges, J., and Vásquez, C. 2016. "If nearly all Airbnb reviews are positive, does that make them meaningless?," *Current Issues in Tourism*, pp. 1–19.

Brochado, A., Troilo, M., and Shah, A. 2017. "Airbnb customer experience: Evidence of convergence across three countries," *Annals of Tourism Research*, (63), pp. 210–212.

Brown, S. J., Goetzmann, W. N., and Ross, S. A. 1992. "Survivorship bias in performance studies," *Review of Financial Studies*, (5:4), pp. 553–580.

Cabral, L. L., Hortacsu, A., and Hortaçsu, A. 2010. "The dynamics of seller reputation: Evidence from eBay," *The Journal of Industrial Economics*, (58:1), pp. 54–78.

Carpenter, J. N., and Lynch, A. W. 1999. "Survivorship bias and attrition effects in measures of performance persistence," *Journal of Financial Economics*, (54), pp. 337–374.

Cox, M. 2017. "Inside Airbnb: Adding data to the debate," (available at http://insideairbnb.com/; retrieved July 5, 2017).

Dann, D., Teubner, T., and Weinhardt, C. 2017. "Poster child and guinea pig—Insights from a structured literature review on Airbnb," *Working Paper*.

Doane, D. P., and Seward, L. E. 2011. "Measuring skewness: A forgotten statistic?," *Journal of Statistics Education*, (19:2), pp. 1–18.

Donath, J. 2007. "Signals in social supernets," *Journal of Computer-Mediated Communication*, (13:1), pp. 231–251.

Dunham, B. 2011. "The role for signaling theory and receiver psychology in marketing," in *Evolutionary Psychology in the Business Sciences*, G. Saad (ed.), (Vol. 53), Springer, Heidelberg, pp. 225–256.

Edelman, B. G., and Luca, M. 2014. "Digital discrimination: The case of Airbnb.com," *Working Paper*.

Edelman, B. G., Luca, M., and Svirsky, D. 2017. "Racial discrimination in the sharing economy: Evidence from a field experiment," *American Economic Journal: Applied Economics*, (9:2), pp. 1–22.

Ert, E., Fleischer, A., and Magen, N. 2016. "Trust and reputation in the sharing economy: The role of personal photos in Airbnb," *Tourism Management*, (55:1), pp. 62–73.

Fan, Y., J, J., and M, X. 2013. "Losing to win: Reputation management of online sellers," *Working Paper*.

Forbes. 2017. "Airbnb closes $1 billion round to keep delaying IPO," (available at https://www.forbes.com/sites/briansolomon/2017/03/09/airbnb-closes-1-billion-round-to-keep-delaying-ipo/; retrieved June 20, 2017).

Fradkin, A. 2015. "Search frictions and the design of online marketplaces," *Working Paper*.

Fradkin, A., Grewal, E., and Holtz, D. 2017. "The determinants of online review informativeness:

Evidence from field experiments on Airbnb," *Working Paper*.
Gagniuc, P. A. 2017. *Markov Chains: From Theory to Implementation and Experimentation*, John Wiley & Sons.
Gebbia, J. 2016. "How Airbnb Designs for Trust," *TED.com*.
Grinblatt, M., and Titman, S. 1992. "The persistence of mutual fund performance," *Journal of Finance*, (47:5), pp. 1977–1984.
Gutt, D., and Herrmann, P. 2015. "Sharing means caring? Hosts' price reaction to rating visibility," in *ECIS 2015 Proceedings*, pp. 1–13.
Horton, J. J., and Zeckhauser, R. J. 2016. "Owning, Using and Renting: Some Simple Economics of the 'Sharing Economy,'" *NBER Working Paper Series*.
Ikkala, T., and Lampinen, A. 2015. "Monetizing network hospitality: Hospitality and sociability in the context of Airbnb," in *CSCW'15 Proceedings*, pp. 1033–1044.
Ke, Q. 2017. "Sharing means renting? An entire-marketplace analysis of Airbnb," *Working Paper*.
Livingston, J. A. 2005. "How valuable is a good reputation? A sample selection model of Internet auctions," *Review of Economics and Statistics*, (87:3), pp. 453–465.
Melnik, M. I., and Alm, J. 2002. "Does a seller's ecommerce reputation matter? Evidence from eBay auctions," *The Journal of Industrial Economics*, (50:3), pp. 337–349.
Möhlmann, M. 2015. "Collaborative Consumption: Determinants of Satisfaction and the Likelihood of Using a Sharing Economy Option Again," *Journal of Consumer Behaviour*, (14:3), pp. 193–207.
Mulshine, M. 2015. "After a disappointing Airbnb stay, I realized there's a major flaw in the review system," *Business Insider* (available at http://www.businessinsider.com/why-airbnb-reviews-are-a-problem-for-the-site-2015-6?IR=T; retrieved January 19, 2017).
Resnick, P., Zeckhauser, R., Swanson, J., and Lockwood, K. 2006. "The value of reputation on eBay: A controlled experiment," *Experimental Economics*, (9:2), pp. 79–101.
Riegelsberger, J., Sasse, M. A., and McCarthy, J. D. 2005. "The mechanics of trust: A framework for research and design," *International Journal of Human Computer Studies*, (62:3), pp. 381–422.
Spence, M. 2002. "Signaling in retrospect and the informational structure of markets," *American Economic Review*, (92:3), pp. 434–459.
Stigler, S. M. 1997. "Regression towards the mean, historically considered," *Statistical Methods in Medical Research*, (6:2), pp. 103–114.
Teubner, T. 2017. "The web of host-guest connections on Airbnb—A social network perspective," *Working Paper*.
Teubner, T., Hawlitschek, F., and Dann, D. 2017. "Price determinants on Airbnb: How reputation pays off in the sharing economy," *Journal of Self-Governance and Management Economics*, (5:4), pp. 53–80.
Wired. 2017. "A lone data whiz is fighting Airbnb — and winning," (available at https://www.wired.com/2017/02/a-lone-data-whiz-is-fighting-airbnb-and-winning/).
Wolff-Mann, E. 2016. "Here's everything wrong with online reviews - and how to fix them," (available at http://time.com/money/page/online-reviews-trust-fix/; retrieved October 26, 2017).
Zervas, G., Proserpio, D., and Byers, J. 2015. "A first look at online reputation on Airbnb, where every stay is above average," *Working Paper*.

Acknowledgements

The authors would like to thank Christoph M. Flath and David Dann for their comments and feedback while hatching the idea for this project. Moreover, we would like to thank the three anonymous reviewers for their constructive feedback.