When Sheep Shop: Measuring Herding Effects in Product Ratings with Natural Experiments

Gael Lederrey
EPFL
gael.lederrey@epfl.ch

Robert West
EPFL
robert.west@epfl.ch

ABSTRACT

As online shopping becomes ever more prevalent, customers rely increasingly on product rating websites for making purchase decisions. The reliability of online ratings, however, is potentially compromised by the so-called *herding effect:* when rating a product, customers may be biased to follow other customers' previous ratings of the same product. This is problematic because it skews long-term customer perception through haphazard early ratings. The study of herding poses methodological challenges. In particular, observational studies are impeded by the lack of counterfactuals: simply correlating early with subsequent ratings is insufficient because we cannot know what the subsequent ratings would have looked like had the first ratings been different. The methodology introduced here exploits a setting that comes close to an experiment, although it is purely observational—a *natural experiment.* Our key methodological device consists in studying the same product on two separate rating sites, focusing on products that received a high first rating on one site, and a low first rating on the other. This largely controls for confounds such as a product's inherent quality, advertising, and producer identity, and lets us isolate the effect of the first rating on subsequent ratings. In a case study, we focus on beers as products and jointly study two beer rating sites, but our method applies to any pair of sites across which products can be matched. We find clear evidence of herding in beer ratings. For instance, if a beer receives a very high first rating, its second rating is on average half a standard deviation higher, compared to a situation where the identical beer receives a very low first rating. Moreover, herding effects tend to last a long time and are noticeable even after 20 or more ratings. Our results have important implications for the design of better rating systems.

1 INTRODUCTION

With every purchase but one click away, online shopping is extremely convenient and is accounting for an ever larger share of
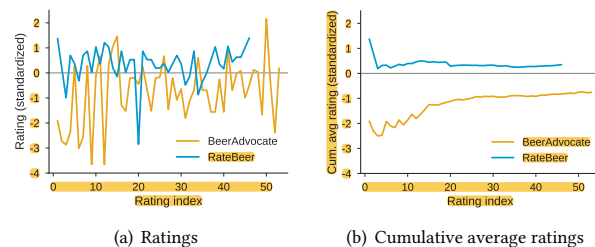
Figure 1: *Lost Rhino Ice Breaker IPA,* an example of a beer with low (high) ratings on BeerAdvocate (RateBeer) (scores standardized to be comparable across sites; *cf.* Sec. 3.2).

the retail market. The downside of online shopping is that it provides a less direct experience than going to an offline, brick-and-mortar store, where customers can taste, smell, touch, and feel a product before deciding whether to buy it. Online, we must rely on ratings provided by previous customers instead.

Online rating systems, however, suffer from the known problem of social influence, also termed **herding,** which expresses the fact that raters tend to be biased by the opinions of previous raters [2, 6, 8, 10, 13, 17] and which can make online rating systems fickle and sensitive to small variations in early ratings: if the first few reviews of a product happen to swing a certain way (or are purposefully engineered that way in an act of review spamming [9]), this can unduly skew subsequent reviews.

This can have severe implications for both customers, who may end up with unsatisfying products, and producers, whose high-quality products may end up being bought less than they deserve. Herding is therefore a behavior of great interest both sociologically as well as economically.

Studying the herding effect is difficult, however. Although **randomized experiments** have been successfully deployed to quantify social influence in rating behavior [6, 13], experiments can be risky from a business and ethical perspective, for either they wilfully subject products to random treatments with potentially harmful effects, or they are restricted to small laboratory settings [14], which reduces the generality of findings. **Observational studies** of herding [2, 3, 10], on the other hand, are less delicate in this regard than randomized experiments, but they are more delicate methodologically, as they require more care during the analysis.

To illustrate this point, consider a **naïve observational study of herding,** which simply measures if products that receive high early ratings also tend to receive high later ratings, and if at the same time products that receive low early ratings tend to receive low later ratings. The problem, obviously, with this hypothetical study is that both early and later ratings might be caused by a hidden correlate—the inherent quality of the product—rather than one

by the other. To credibly claim a causal relation, we would need to show that, *for a fixed product,* later ratings follow early ratings regardless of whether the first ratings were high or low. If, on the contrary, inherent quality were the real cause, then a good product would get good later ratings even if, for some reason, the first ratings were low. In other words, the naïve observational study suffers from the lack of counterfactuals: we cannot tell whether there has been herding in the ratings for a product if we do not know what would have happened if the first ratings had been different.

We address this challenge by studying not a single product rating website, but two of them in parallel. In particular, we will observe how the same product is rated independently on the two sites; if the product happens to receive a vastly different first-rating on site 1, compared to site 2, we can measure how the first-rating affects subsequent ones.

To explain the basic idea behind our method, we start with an example. For the sake of concreteness, and to set the stage for our case study, consider the case of beers as products, rated on the two major beer-rating websites, BeerAdvocate and RateBeer. Consider a beer B that has become available just recently. The RateBeer user to first rate beer B on RateBeer happens to love it, so she gives it a very high score. The BeerAdvocate user to first rate beer B on BeerAdvocate, on the contrary, happens to hate it, so he gives it a very low score. Of course, the inherent quality of beer B was exactly the same for both users, and each user was the first to rate beer B on the respective site, so they were not influenced by previous opinions. It is only due to chance that the first RateBeer rating was high, and the first BeerAdvocate rating low, instead of *vice versa.*

More generally, whenever the same beer gets a high first-rating on one site, and a low first-rating on the other site, it seems likely that it is haphazard whether the high first-rating happens on Rate-Beer or on BeerAdvocate. Whether this is indeed true needs to be established, but if it can be established, this means that nature has created a situation akin to an experiment for us: she has flipped a coin to decide on which of the two sites beer B was to be exposed to the "experimental" condition *high first-rating,* and on which to the "experimental" condition *low first-rating*—a so-called **natural experiment**. We may then analyze which effect the two conditions have on later ratings of beer B. In the presence of herding, later ratings on the site with the high first-rating should on average be higher than on the site with the low first-rating. Conversely, in the absence of herding, later ratings should on average be similar regardless of the first-rating.

Fig. 1 shows that beers such as the anonymous B really exist, in this case an India pale ale named *Lost Rhino Ice Breaker*. The time series of this beer's ratings is plotted in Fig. 1(a), with the cumulative average displayed in Fig. 1(b). The curves show that *Lost Rhino Ice Breaker* received a much lower first-rating on BeerAdvocate than on RateBeer, and that it never managed to recover from its dismal start on BeerAdvocate, while it thrived on RateBeer—a difference that may be due to herding.[1] This example is but anecdotal, of course, and the goal of this paper is to move from such examples to reliable causal statements. Our main contributions are twofold: First, we introduce an observational methodology for quantifying how consistently early ratings influence later ones via the natural

---

[1]The irony of illustrating herding with a lost rhino is incidental.
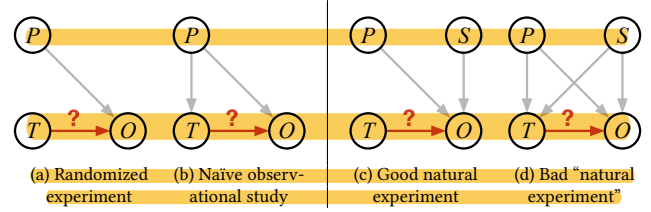

Figure 2: Variable dependencies for different kinds of study (Sec. 2). T: treatment (first-rating); O: outcome (subsequent ratings); P: rated product; S: rating site. Only (a) and (c) allow us to draw conclusions about a causal link between T and O.

experiment sketched above (Sec. 2). Second, we apply our method to a real dataset of product ratings (Sec. 3), after carefully ruling out confounds and thus confirming that indeed we have identified a natural experiment (Sec. 4). Our results provide strong evidence of substantial herding effects (Sec. 5): the second rating for a product is on average half a standard deviation higher when a beer received a very high first-rating, compared to when it received a very low first-rating. Furthermore, herding effects are tenacious and can be noticed even after 20 or more reviews. We conclude the paper by discussing implications of our findings and prior as well as future work (Sec. 6).

2 METHODOLOGY: NATURAL EXPERIMENT

To illustrate the difference between a proper randomized experiment and the naïve observational study delineated in the introduction, we represent both scenarios as Bayesian networks in Fig. 2(a–b). The diagrams contain three random variables: product (P), treatment (T), and outcome (O). The treatment T captures whether the first-rating for product P is high or low; the outcome O captures whether subsequent ratings for P are high or low.

In a randomized experiment (Fig. 2(a)), treatment assignment is decided by a coin flip, independent of any product properties, so different treatment groups are indistinguishable with respect to product properties and are therefore directly comparable. Hence, if we observe significantly different outcomes for different treatments, this difference is likely to be caused by the treatment.

In a naïve observational study (Fig. 2(b)), on the contrary, treatment assignment may depend on product properties, so the latter may influence both treatment and outcome. In this case, a correlation between treatment and outcome may not be causal, but both may instead be caused separately by a *confound,* a latent property of the product. As mentioned, an obvious confound could be the inherent quality of the product, as good products tend to be rated highly both by the first reviewer (treatment) and by subsequent reviewers (outcome), even in the absence of herding.

Our methodology circumvents the problem of product-induced confounds by studying not one single rating website (as would be done in a naïve observational study), but rather two separate rating websites with overlapping sets of rated products. This still constitutes an observational study, but one that naturally controls for confounds and thus comes close to a randomized experiment in spirit—a situation commonly known as a *natural experiment.*

Let us call the two sites S_1 and S_2, and consider a product P rated on both sites. If P received a high first-rating (*i.e.*, treatment) on S_1 and a low first-rating on S_2 (or *vice versa*), then—under conditions

to be discussed below (Sec. 2.2)—this situation may be seen as emulating two "parallel universes": we can observe what happens to the *same* product under each possible treatment. In other words, we now have counterfactuals, the lack of which is the major shortcoming of the naïve, single-site observational study.

2.1 Step-by-step description

We now describe our methodology in detail. It encompasses 5 steps.

Step 1: Match products across the two rating websites. The fundamental device exploited by our natural experiment consists in tracking the same product on two separate websites, so we work only with products that we can identify on both sites and discard all others. Sec. 4.1 describes the matching algorithm we used on a pair of beer rating websites; it is based on string similarities between product names, and we believe it is general enough to be adapted to other datasets as well.

Step 2: Define the paired-treatment groups. On each site, label each product as "high" (H), "medium" (M), or "low" (L), depending on the first-rating (treatment) it received on the site. A first-rating is defined as H if it is in the top p percent of the first-ratings on the respective site, as L if it is in the bottom p percent, and as M otherwise. (In our specific case study, we use $p = 15$.) After this step, each product falls into one of the nine *paired-treatment groups* defined by the cross product $\{T_1 T_2 : T_1, T_2 \in \{H, M, L\}\}$, where the two letters capture the treatments the product received on sites S_1 and S_2 via the respective first-ratings; *e.g.*, HH means that the product received high first-ratings on both sites, HL that it received a high first-rating on S_1 and a low one on S_2, LH that it received a low first-rating on S_1 and a high first-rating on S_2, *etc.*

Step 3: Balance the paired-treatment groups, making sure that, for each $(T_1, T_2) \in \{H, M, L\} \times \{H, M, L\}$, we have the same number of products in the paired-treatment group $T_1 T_2$ as in the group $T_2 T_1$. We may achieve this simply by randomly subsampling from the larger of the two groups. This step ensures that, for each site S and each paired-treatment group $T_1 T_2$ (with $T_1 \neq T_2$), the probabilities of the two treatments T_1 and T_2 are 50% each.

Step 4: Aggregate paired-treatment groups containing the same set $\{T_1, T_2\}$ of treatments. This reduces the number of paired-treatment groups to six: HH, HM, HL, MM, ML, LL; *e.g.*, after aggregation, the group HL contains both products with H on S_1 and L on S_2 and products with L on S_1 and H on S_2. This is done to have more data points per group, but we advise to also perform a separate analysis on the non-aggregated data as a sanity check.

Step 5: Compare the outcomes for different treatments within the same paired-treatment group. We consider the groups HL, HM, and ML, where the same product received different first-ratings on the two sites. By comparing subsequent ratings across the two sites, we can estimate the treatment effect in isolation from any product-related confounds (such as inherent quality), which are controlled for by fixing the product. In particular, for a given product P and a given rating index i, we compare P's i-th rating on the site where it received the higher first-rating with its i-th rating on the site where it received the lower first-rating. If the difference is positive, this supports the hypothesis of a causal link between treatment (first-rating) and outcome (i-th rating), *i.e.*, herding. Tracking the difference as a function of the rating index i also lets us study if, and how fast, herding attenuates with time.

The most interesting paired-treatment group is HL, as it corresponds to the starkest difference in treatments. Since, however, it will generally occur less frequently in practice than the less extreme groups (HM and ML), we recommend to also study the latter. Finally, as a sanity check, it is also advisable to include the symmetric groups HH, MM, and LL in the analysis.

2.2 Assumptions

The above methodology allows us to estimate the causal effect of the first-rating (treatment) on subsequent ratings (outcome) if the two following assumptions hold.

The first and most crucial assumption, and in fact the defining property of a natural experiment, is that **treatment assignment is haphazard:** whether a given product P with different first-ratings receives its higher first-rating on S_1 and its lower first-rating on S_2 or *vice versa* must not depend on any properties of P, S_1, and S_2. In other words, the treatment assignment T must be independent of the product P and of the site S (Fig. 2(c)). If this is the case, and if variations in T are correlated with variations in O, then the link between T and O is likely to be causal. Otherwise (Fig. 2(d)), properties of the product or of the site, or a combination of the two, could explain both the treatment assignment and the outcome—we might have mere correlation without causation, which would defeat the very purpose of considering a matched, rather than a naïve, single-site observational study.

Note that, by construction, treatment assignment is independent of the product alone: each product is included twice in the matched dataset, once per site, on one of them with a higher, and on one with a lower, first-rating, such that we have a 50/50 distribution over the two possible treatments for a given product. Similarly, again by construction, treatment assignment is independent of the site alone: after balancing paired-treatment groups (step 3 in Sec. 2.1), each site has as many products with a higher as with a lower first-rating, resulting in a 50/50 distribution over the two possible treatments in each paired-treatment group for a given site.

This does not imply, though, that treatment assignment is independent of the combination of product and site;[2] *e.g.*, users on site S_1 might like a certain kind of product more than users on site S_2, which could result in an increased probability of a higher first-rating (treatment) for that kind of product on S_1, compared to S_2.

In our setup, showing that treatment is indeed independent of the combination of product and site establishes the *internal validity* of the study. How to show that a specific study is internally valid depends on the datasets being used. (See Sec. 4.3 for how we proceed in the case of beer ratings.) Although we cannot speak of a natural experiment if this independence does not immediately hold in the matched dataset, one might still achieve it by explicitly balancing the dataset, *e.g.*, via propensity-score matching [15].

The second assumption is that the **matched dataset accurately reflects the full dataset.** In general, not all products are present on both rating sites, so matching will select a subset of all products. If the matched sample is biased, *i.e.*, systematically different from the full dataset before matching, this might preclude us from generalizing our findings from the natural experiment to the set of all

---

[2] While $X \perp\!\!\!\perp (Y, Z)$ ("X is independent of (Y, Z)") implies $X \perp\!\!\!\perp Y$ and $X \perp\!\!\!\perp Z$, the statement does not hold in the opposite direction.

Table 1: Dataset size.

	BeerAdvocate	RateBeer
Breweries	16,758	24,189
Beers	280,823	442,081
Beers ($\geq$ 5 ratings)	96,156	166,043
Beers ($\geq$ 10 ratings)	61,193	104,062
Beers ($\geq$ 20 ratings)	38,533	60,451
Users	153,704	70,174
Users ($\geq$ 10 ratings)	48,595	17,744
Users ($\geq$ 100 ratings)	14,488	6,419
Ratings	8,393,032	7,122,074

products rated on the two sites. For instance, it is conceivable that particularly good products are more likely to be present on both sites, which would make our findings specific to good, rather than average, products. By showing that the matched sample is unbiased, we establish the so-called *external validity* of the study. (See Sec. 4.3 for how we proceed in the case of beers as products.)

3 DATA: TWO BEER RATING WEBSITES

We apply our methodology to the specific scenario of beer ratings. This setting is well suited for several reasons: the market is dominated by two large websites dedicated to the rating and reviewing of beers—BeerAdvocate and RateBeer—, each with a long history reaching back nearly 20 years, with very similar site designs, and with a large overlap of rated products.

An older version of the data was made available by McAuley *et al.* [11, 12], but as that version was produced in 2012, we re-crawled it; the data now extends from 2001 to August 2017.[3]

Although we focus on one case study, our method applies equally to other pairs of rating websites, as long as the intersection of the sets of rated products is large (*cf.* our discussion in Sec. 6).

3.1 Description of beer rating websites

BeerAdvocate and RateBeer are the two largest online beer-related websites. Although they provide a general space for beer aficionados, with articles, discussion forums, and trading platforms, their main purpose is to collect and curate beer ratings provided by users. On both sites, beers are rated with respect to five aspects (look, smell/aroma, taste, feel/palate, overall), which are then combined via a weighted sum into a rating score between 1 and 5. The sites are also similar with respect to layout and visual appearance. They both prominently show the most recent ratings as well as the current cumulative average on the page of each beer (RateBeer also shows the rank of the beer among all beers), so we can assume that users about to rate a beer become aware of this information.

3.2 Basic analysis of rating datasets

Here we discuss some properties of the two beer rating datasets that are relevant for our study of herding. We start by summarizing the size of the datasets in Table 1, which shows that each site contains ratings for hundreds of thousands of beers from tens of thousands of breweries, rated by tens of thousands of users, totaling millions of ratings.

The two sites attract rather different user populations. In particular, BeerAdvocate is mostly frequented by users from the U.S. (74% of users), followed by Canada (2%), with less than 1% of users from any other single country. RateBeer's user base, though also predominantly from the U.S. (38%), is more balanced, with 5% of users from Canada, 4% from England, 2% Poland, 2% from Australia, *etc.*

BeerAdvocate's more U.S.-centric user base is also reflected in the breweries whose beers are rated on the sites: 44% of all breweries represented on BeerAdvocate are from the U.S., while the fraction is only 29% on RateBeer. Other countries have similar percentages across sites (*cf.* "Unmatched" in Table 4). These differences imply that selecting a matched sample of beers rated on both websites cannot possibly reflect the overall distribution of products on both sites equally well, an issue we address in Sec. 4.3.

Fig. 3(a) plots the histograms of ratings for both websites. We clearly see that users on BeerAdvocate tend to give higher ratings than users on RateBeer.[4] As a side note, we also point out that the rating distributions of Fig. 3(a) differ vastly from the bimodal distributions with mostly extremely high or extremely low values that have frequently been observed on other rating websites [8] and that have been attributed to a "brag-and-moan" effect. On beer rating websites, ratings seem to be less affected by selection bias due to disappointment or positive surprise.

The histograms of Fig. 3(a) pool all ratings from 2001 through August 2017. Next, we group ratings by year and plot the annual mean (Fig. 3(b)) and standard deviation (Fig. 3(c)). We observe that neither quantity stays constant over time: the mean increases, while the standard deviation decreases, from year to year. Assuming that the inherent quality of beers being rated stays roughly constant, the rising mean may be interpreted as score inflation, while the sinking standard deviation could indicate a consolidating consensus about what should constitute the score of an average beer.

This implies that, in order to compare ratings across sites and time periods, we must account for biases stemming from site conventions (shifted rating histograms) and from a temporal drift in these conventions (rising means and sinking standard deviations). Instead of raw ratings, we therefore consider *standardized ratings* (also known as *z*-scores): for each site and each year, we compute the mean and standard deviation over all ratings. We then subtract the mean of year t from all ratings submitted in year t and divide them by the standard deviation of year t, such that each year's set of ratings has mean 0 and standard deviation 1.

4 MATCHING PRODUCTS ACROSS WEBSITES

Our method hinges on a set of products rated on two separate websites. An alignment of products across sites (*e.g.*, via consistent unique identifiers) is typically not given explicitly; rather, one usually needs to perform the matching oneself heuristically. In this section, we describe our algorithm for achieving a high-quality alignment between beers from BeerAdvocate and RateBeer (Sec. 4.1), report basic statistics of the matched sample (Sec. 4.2), and discuss its external and internal validity (Sec. 4.3).

---

[3] Data available upon request. Code: https://github.com/epfl-dlab/when_sheep_shop.

[4] We found that the distinct spikes in BeerAdvocate's histogram are caused by reviews that gave the same score to all five aspects (*cf.* Sec. 3.1) and that tend to be very short or even empty, which seems to indicate that these reviews were entered in a hurry.

(a) Histograms of ratings   (b) Mean of ratings per year   (c) Standard deviation of ratings per year
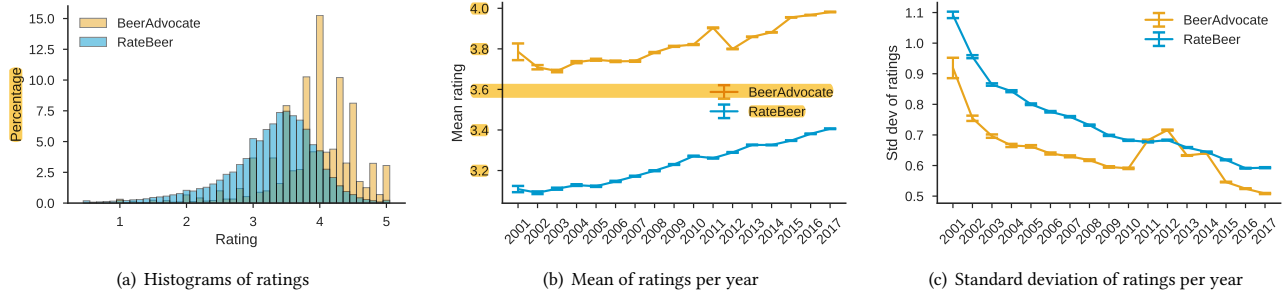
Figure 3: Statistics of the two beer-rating datasets. Error bars in (b–c) capture 95% confidence intervals.

4.1 Matching algorithm

When matching products across websites, we should favor precision over recall: not matching all potentially matchable products simply decreases the sample size and is therefore acceptable (as long as it does not bias the dataset, *cf.* Sec. 4.3), whereas matching non-identical products to each other introduces noise into the data. With this in mind, we designed a matching algorithm geared toward precision, potentially at the expense of recall.

We proceed in two phases. First, we align breweries across sites, then we align beers only within breweries; *i.e.*, we only consider pairs of beers as potential matches if their respective breweries were matched to each other before. The same basic procedure is used both for matching breweries and for matching beers, in each case operating on name strings.

The algorithm starts by representing names as TF-IDF vectors and computes all pairwise cosine similarities. Inverse document frequency (IDF) weighting serves to down-weight common terms such as "brewery", "company", "beer", "ale", *etc.* When aligning names from two sets, the number of matches is upper-bounded by the size of the smaller set, so we iterate over the smaller set and, for each name, find the optimal match in the larger set. We only keep a match if its cosine similarity is above a threshold θ and if the best match has a much higher similarity than the second best match (to rule out ambiguities), by requiring a gap in similarities of at least δ. If this greedy procedure pairs the same element from the larger set with more than one element from the smaller set, we discard all these pairs.[5]

As mentioned, we use the same procedure for matching breweries and then for matching beers within breweries. When matching breweries, we additionally require an exact match of locations (states for U.S. breweries, and countries for others). When matching beers, we additionally require an exact match in alcohol by volume; also, before computing cosine similarities, we first remove from each beer name all tokens that also appear in the brewery name, as sometimes the name of the beer contains the brewery name in one dataset, but not in the other (*e.g.*, "Ingobräu Meistersud" *vs.* just "Meistersud").

4.2 Matched dataset

The above algorithm has two parameters, θ and δ. We find $\theta = 0.8$ and $\delta = 0.3$ to work well in practice, as shown by an evaluation in

---

[5] Alternatively, we could run a proper matching algorithm, such as the Hungarian algorithm, but as we aim to maximize precision, we opted for first being fully greedy and then generously discarding all potentially bad matches.

Table 2: Dataset size after matching.

	Minimum number of ratings per beer			
	0	5	10	20
Breweries	6,084	2,561	1,711	1,079
Beers	45,640	12,890	7,424	4,051
Ratings on BA	955,968	873,944	812,070	732,165
Ratings on RB	1,020,638	761,496	650,642	542,961

Table 3: Number of beers per paired-treatment group (Sec. 2.1, step 2) after matching, before balancing (Sec. 2.1, step 3).

		BA		
		H	M	L
	H	585	1,213	116
RB	M	1,210	6,593	1,242
	L	138	1,225	568

which we inspected 500 matched brewery pairs and 500 matched beer pairs and, based on this ground truth, estimate precision as 99.6% for matched breweries (2 of the 500 inspected matches were wrong) and 100% for matched beers.

The size of the matched dataset is summarized in Table 2. In our result analysis (Sec. 5), we restrict ourselves to beers with at least a minimum number of ratings, so the table lists sizes for various values of this threshold. Although matching reduces the dataset by a lot, we are still left with tens of thousands of beers from thousands of breweries, with close to a million ratings on each site.

Matching ensures that each remaining product has been rated on each of the two websites. As explained in Sec. 2.1 (step 2), each beer falls into one of nine paired-treatment groups (HH, HM, HL, MH, *etc.*). Table 3 displays the size of all groups (for beers with at least five ratings on each site, as the bulk of our analysis will be conducted on this set). We observe that the groups are rather well balanced "out of the box" (*e.g.*, 1,210 beers in HM *vs.* 1,213 in MH, *etc.*), even before balancing them explicitly (step 3 of Sec. 2.1).

We recomputed and inspected the rating histograms (*cf.* Fig. 3(a)) on the subset of beers in the matched sample only and found them essentially indistinguishable from the full sample shown in Fig. 3(a). This implies that the vastly different rating distributions of the two sites are not caused by users on one of the sites having a systematic preference for rating inherently better or worse beers (*e.g.*, certain rating sites might see themselves as "bashing sites"); rather, the difference must stem from different scoring standards.

To make scores comparable across sites, we thus standardize all ratings as described in Sec. 3.2, by subtracting the annual mean and dividing by the annual standard deviation. As seen in Fig. 4(a),

the two sites' rating histograms are entirely overlapping after standardization, so ratings may now be compared across sites.

It is important to note that, whereas the distribution of standardized ratings is identical for the two sites (Fig. 4(a)), there are numerous individual products with vastly different ratings on BeerAdvocate *vs.* RateBeer. To emphasize this point, Fig. 4(b) contains a scatter plot of ratings on BeerAdvocate *vs.* ratings on RateBeer, where each beer is summarized by its average standardized rating on each site. The fact that the point cloud disperses widely off the diagonal clearly shows that many beers are perceived differently on the two sites. Viewed this way, the purpose of our natural experiment is to determine how beers such as *Lost Rhino Ice Breaker* from the introduction wind up in the fringe of the point cloud of Fig. 4(b) (where *Lost Rhino Ice Breaker* is marked as a red cross)—by virtue of herding or by sheer good or bad luck.

4.3 Validity of matched sample

Drawing correct conclusions from our observational study requires the assumptions laid out in Sec. 2.2. The purpose of this section is to show that these assumptions, in particular external and internal validity, are empirically met by the matched beer rating dataset.

External validity. Favoring precision over recall when matching (Sec. 4.1) comes at the expense of losing many matches in which we are less confident, which may introduce selection bias and can potentially impair the external validity of our results: if the sample we study is fundamentally different from the overall population, our conclusions might not generalize from the former to the latter.

First recall from Sec. 3.2 that there are some significant differences between the two sites: BeerAdvocate is more U.S.-centric in terms of products and users, and it is also smaller in terms of the number of beers rated (Table 1). Since the number of matched beers is upper-bounded by the number of beers in the smaller dataset, the best we could hope to do is match all beers in BeerAdvocate to their corresponding beers in RateBeer. This would preserve the original data distribution in BeerAdvocate and skew RateBeer's distribution to match it.

Fig. 5 and Table 4 show that, even though the smaller BeerAdvocate is not a subset of RateBeer, matching still produces a dataset very similar to BeerAdvocate. Fig. 5 inspects three exemplary properties (mean average beer rating, number of ratings per beer, and number of beers per brewery) of the data before (left box in each pair) *vs.* after matching (right box in each pair), for both BeerAdvocate (left panel in each figure) and RateBeer (right panel in each figure). We observe that matching does not noticeably alter BeerAdvocate's distributions, whereas RateBeer's do change.

Table 4 lists the most common countries of origin for breweries present in each dataset before matching, as well as in the matched dataset. We make two observations. First, the distribution over countries is similar in both datasets even before matching, with the exception that BeerAdvocate contains a much larger fraction of U.S. breweries. Second, matching mimics the distribution of the smaller dataset, BeerAdvocate, more closely than that of RateBeer. We also compared the style distributions before and after matching, with the same result that the matched dataset mirrors BeerAdvocate's distribution closely (in decreasing order of frequency:

Table 4: Brewery locations before and after matching (Sec. 2.1, step 1), ordered by percentage after matching.

	Unmatched		Matched
	BA	RB	
United States	44.4%	28.6%	47.8%
Germany	8.5%	8.3%	6.4%
England	6.1%	8.8%	5.8%
Canada	5.1%	3.7%	4.9%
Italy	2.2%	4.3%	2.7%
Belgium	2.0%	1.9%	2.5%
France	2.4%	3.5%	2.3%
Spain	1.9%	3.2%	2.3%
Australia	2.4%	2.3%	2.2%
Netherlands	1.5%	2.1%	2.1%

American IPA 10.8% before, *vs.* 12.1% after, matching; American Pale Ale 6.2% *vs.* 6.8%; Saison/Farmhouse Ale 5.0% *vs.* 5.9%, *etc.*).

We conclude that our matched sample is unbiased with respect to BeerAdvocate (as explained above, the best we could hope for), such that the conclusions we draw can at the very least be generalized to all of BeerAdvocate. Note that this is a conservative statement; we have seen no explicit indications why our conclusions should not also hold on all of RateBeer.

Internal validity. As argued in Sec. 2.2, we need to show that the treatment assignment T (the first-rating a beer receives) is independent of the rating site S and the rated product P. Although, as discussed there, we have $T \perp\!\!\!\perp S$ and $T \perp\!\!\!\perp P$ by construction, these do not automatically imply $T \perp\!\!\!\perp (S, P)$. For instance, it is in principle possible (though not likely) that users on site S_1 love all pale beers and hate all dark beers, while users on S_2 love all dark beers and hate all pale beers. This would entail that all pale beers would see both a high treatment (first-rating) and a high outcome (subsequent reviews) on S_1; and that all dark beers would see both a high treatment and a high outcome on S_2. Here, a correlation between treatment and outcome would not be causal, but due to the confound of site-specific preferences. We therefore need to check empirically that the distribution of treatment assignments (*i.e.*, the probability of receiving a higher first-rating) is approximately equal for all combinations of site and product properties. Notice that only properties available *before* treatment should be taken into account here, as all other properties might be consequences, rather than causes, of the treatment. This precludes us, *e.g.*, from considering ratings received by the respective beer.

Inspecting treatment probabilities for all beer properties on each site would be elusive, especially given our limited dataset size. But we argue that the most likely confounds would be captured by beer style and producer country: controlling for style also roughly fixes the most salient properties of a beer, such as bitterness, color, alcohol content, *etc.*; and controlling for producer country accounts for the fact that BeerAdvocate users are more likely American and might therefore be biased toward (or against) American beers.

The numbers are presented in Table 5 for the most frequent styles and countries. There is one table per group of interest (HM, ML, HL). For each combination of beer property (style or country) and site, we list the number of beers with the higher treatment on

(a) Histograms of ratings (std.) after matching
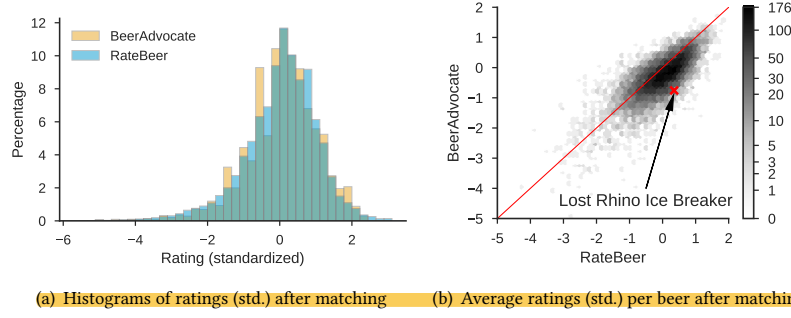(b) Average ratings (std.) per beer after matching

Figure 4: Standardized ratings after matching. Red cross in (b) marks example from Fig. 1; gray tones capture number of beers per bin. Although overall rating distributions are identical, many individual beers are rated differently across sites.
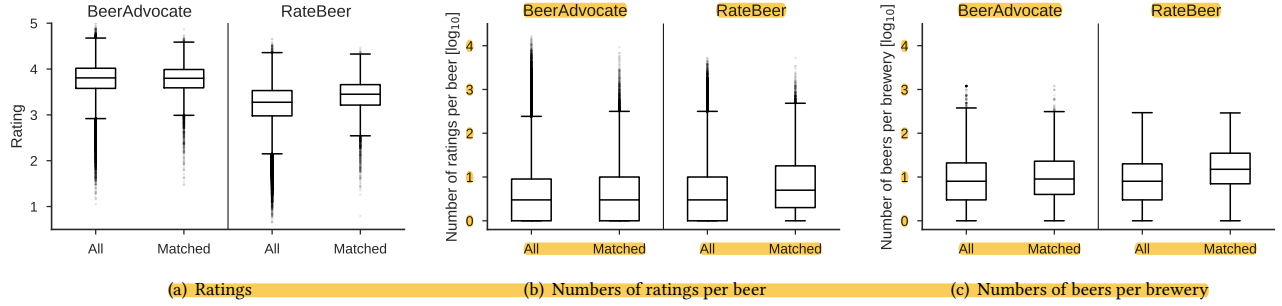

(a) Ratings
(b) Numbers of ratings per beer
(c) Numbers of beers per brewery

Figure 5: Comparison of three dataset properties before *vs.* after matching.

that site and the implied probability of receiving the higher treatment. As there are two treatments in each group, perfect independence would yield probabilities of 50% everywhere.

While achieving such an exact balance is infeasible with our limited dataset, Table 5 shows that we come rather close; in particular, we achieve treatment probabilities close to 50% for the most frequent styles and countries (which have the biggest impact).

These numbers mean that our cross-site product matching results in a treatment assignment that is (approximately) independent of site and product properties (*cf.* Fig. 2(c)). Though not randomized by us as the researchers, treatment assignment is mostly haphazard, such that we may indeed speak of a natural experiment.

5 RESULTS

Now that we have introduced our matched dataset and established that correlations between first and later ratings are highly likely to be causal, we proceed to analyzing the dataset with respect to such correlations, to estimate the effect of first on later ratings.

After processing the data as described in Sec. 2.1 (steps 1–4), there are 3 paired-treatment groups of interest: HL, HM, ML. We estimate the effect of a higher (H for HL/HM; M for ML) *vs.* a lower (L for HL/ML; M for HM) first-rating separately for each of them.

Recall that each group contains the same beer twice, once on BeerAdvocate, once on RateBeer; and once with a higher, once with a lower, first-rating. As established in Sec. 4.3, on which site a beer receives the higher first-rating is essentially haphazard. Therefore, in the presence of herding, a beer will on average receive higher subsequent ratings on the site on which it happened to receive the higher first-rating, and lower subsequent ratings on the other site.

In the absence of herding, subsequent ratings will be indistinguishable between the two sites on average. In other words, we have herding if and only if first and subsequent ratings are correlated.

In this light, Fig. 6(a) provides a clear indication of herding. The dark curves in the center correspond to the paired-treatment group HL. The dark red (green) curve summarizes ratings received on the site with a first-rating of H (L); the horizontal axis shows the rating index $i = 1, \ldots, 5$, and the vertical axis, the i-th standardized rating on the respective site, averaged over all beers in the HL group. The plot includes only beers with at least five ratings, so the same set of beers contributes to all indices i. We also emphasize that the dark red and green curves are computed on exactly the same set of beers, just with a different treatment per curve. Thus, we conclude from Fig. 6(a) that the same beer's second rating is about half a standard deviation higher if the first-rating was H, *vs.* if it was L.

The same effect can be observed for the less extreme paired-treatment groups (HM and ML; Fig. 6(b)), but with less extreme differences, as expected: here, a different first-rating translates into a second-rating difference of about a quarter standard deviation. Fig. 6(a–b) also show that the impact of herding extends beyond the second rating; the effect size is roughly constant on any of the first five ratings. It is therefore interesting to ask whether the herding effect lingers indefinitely or is ultimately overridden by the inherent quality of the respective beer. To address this question, we consider only beers with a substantial number of ratings (at least 20) and compare their long-term averages (based on all ratings received up until the datasets were crawled) for different first-ratings. The results of this analysis, plotted in Fig. 6(c), are clear: even after 20 or more ratings, a high first-rating (H for HL) entails a rating on average 0.28 standard deviations higher than what we would see

Table 5: Counts (#) and probabilities (Pr) of higher treatment for three paired-treatment groups, both sites (BA, RB), and top beer styles and brewery countries. Most values being similar for BA and RB supports internal validity of our study (Sec. 4.3).

	HM					ML					HL			
	#(H)		Pr(H)			#(M)		Pr(M)			#(H)		Pr(H)	
Style	BA	RB	BA	RB	Style	BA	RB	BA	RB	Style	BA	RB	BA	RB
Amer. IPA	105	88	0.54	0.46	Amer. IPA	64	64	0.50	0.50	Amer. IPA	7	5	0.58	0.42
Amer. Double/Imp. IPA	100	81	0.55	0.45	Amer. Pale Ale	36	44	0.45	0.55	Amer. Double/Imp. IPA	8	2	0.80	0.20
Amer. Pale Ale	32	32	0.50	0.50	Fruit/Vegetable Beer	30	25	0.55	0.45	Amer. Pale Ale	3	3	0.50	0.50
Amer. Wild Ale	27	36	0.43	0.57	Amer. Amber/Red Ale	32	20	0.61	0.39	Russian Imperial Stout	0	5	0.00	1.00
Saison/Farmhouse Ale	33	30	0.52	0.48	Amer. Double/Imp. IPA	28	18	0.61	0.39	Amer. Amber/Red Ale	3	2	0.60	0.40
Amer. Double/Imp. Stout	22	36	0.38	0.62	Saison/Farmhouse Ale	20	26	0.44	0.56	Amer. Wild Ale	4	1	0.80	0.20
Amer. Black Ale	15	20	0.43	0.57	Amer. Blonde Ale	28	13	0.68	0.32	Amer. Barleywine	2	3	0.40	0.60
Amer. Porter	14	18	0.44	0.56	Amer. Porter	14	24	0.37	0.63	Amer. Blonde Ale	1	3	0.25	0.75
Russian Imperial Stout	13	13	0.50	0.50	Amer. Pale Wheat Ale	18	19	0.49	0.51	Amer. Porter	2	2	0.50	0.50
American Stout	10	14	0.42	0.58	German Pilsener	14	21	0.40	0.60	Amer. Black Ale	0	4	0.00	1.00
Country	BA	RB	BA	RB	Country	BA	RB	BA	RB	Country	BA	RB	BA	RB
United States	544	493	0.52	0.48	United States	445	449	0.50	0.50	United States	46	38	0.55	0.45
Canada	24	37	0.39	0.61	Canada	84	53	0.61	0.39	Canada	7	7	0.50	0.50
Belgium	30	31	0.49	0.51	Belgium	28	36	0.44	0.56	Germany	3	6	0.33	0.67
England	7	13	0.35	0.65	Germany	20	21	0.49	0.51	Belgium	5	2	0.71	0.29
Australia	4	14	0.22	0.78	Australia	20	19	0.51	0.49	Australia	2	4	0.33	0.67
Germany	10	7	0.59	0.41	England	9	19	0.32	0.68	Switzerland	1	2	0.33	0.67
Sweden	10	6	0.62	0.38	Netherlands	7	12	0.37	0.63	Denmark	0	2	0.00	1.00
Italy	3	8	0.27	0.73	Italy	4	11	0.27	0.73	Austria	1	1	0.50	0.50
Denmark	4	7	0.36	0.64	Scotland	5	6	0.46	0.54	Netherlands	1	1	0.50	0.50
New Zealand	0	9	0.00	1.00	New Zealand	3	5	0.38	0.62	New Zealand	1	1	0.50	0.50

(a) Extreme paired-treatment groups   (b) Less extreme paired-treatment groups   (c) Long-term average ratings
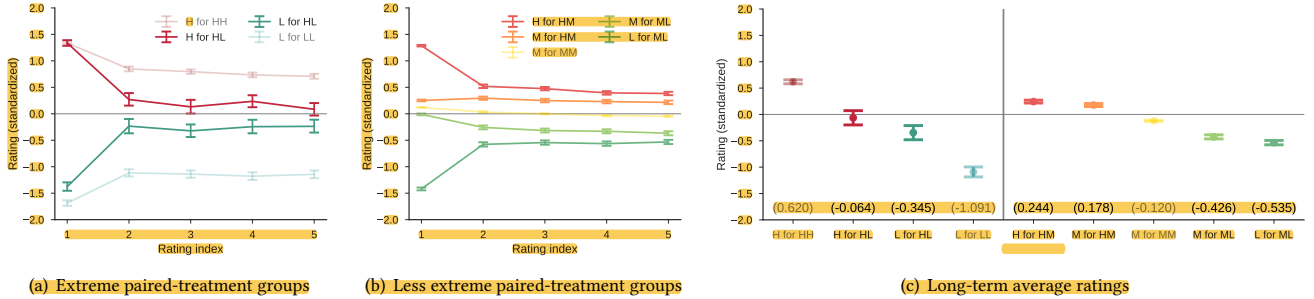
Figure 6: Herding effects in beer ratings; (a–b) identical beer receives significantly higher (lower) subsequent ratings if first-rating was higher (lower); (c) herding effects last long: even after 20 or more ratings, product average ratings are significantly higher (lower) for beers with higher (lower) first-ratings. Ratings are standardized; error bars show 95% confidence intervals.

after a low first-rating (L for HL). The effect is again less extreme for less extreme first-rating differences (HM, ML), but it is still noticeable (and with non-overlapping 95% confidence intervals).

The plots of Fig. 6 were computed after aggregating symmetric paired-treatment groups, such that, *e.g.*, beers that received H on BeerAdvocate and L on RateBeer were grouped together with beers that received L on BeerAdvocate and H on RateBeer (*cf.* step 4 in Sec. 2.1). As a sanity check, we also investigate treatment differences in the non-aggregated groups (Table 6), concluding that a higher first-rating entails higher subsequent ratings regardless of the site on which the higher first-rating occurred.

Fig. 6 and Table 6 directly capture (standardized) rating scores. Plain scores are, however, not all that matters; rating sites tend to also rely heavily on rankings, *e.g.*, when making recommendations to users. Therefore, we also repeat our analysis by measuring outcomes in terms of ranks (normalized to lie between 0 and 1), rather than scores, as follows (we explain our setup for the HL group; it is analogous for the other groups). For each rating index $i = 1, \ldots, 5$, we rank all beers in the group by their i-th rating, resulting in one ranking for each of the two sites. Then, we compute, for each beer, its mean rank on the site where it received H as the first-rating, as

well as on the site where it received L. Comparing the mean for H with the mean for L, we observe that normalized ranks with respect to second ratings are 61% on average on the H site, and only 43% on the L site. We consider this normalized-rank difference of 18% rather substantial. (Even when considering fifth, rather than second, ratings, we still measure a difference of 10%.)

6 DISCUSSION AND RELATED WORK

Summary of results. Our results show clear evidence of herding in the ratings collected by the two most prominent online beer-rating platforms. We find that a very high first-rating leads to the following five ratings being about half a standard deviation higher, compared to a situation in which the exact same beer receives a very low first-rating (Fig. 6(a–b)). These differences in absolute scores also translate into large differences in ranking positions (18% for the second, 10% for the fifth, rating), which are essential for product recommendations. The problem would be mitigated if the platform managed to swiftly "forget" the first reviews and to converge to the true, inherent quality of the beer being rated, but we find that this is not the case, with the effects of first-ratings lingering until after the beer has received 20 or more reviews (Fig. 6(c)).

Table 6: Standardized 5th rating (with 95% confidence intervals) for disaggregated paired-treatment groups, *i.e.*, without step 4 of Sec. 2.1; *e.g.*, cell (L, H on RB) contains mean 5th rating on site with first-rating L when first-rating H happened on RB.

	H on BA	H on RB		H on BA	H on RB		M on BA	M on RB
H	-0.032 [-0.193, 0.123]	0.226 [0.045, 0.393]	H	0.270 [0.225, 0.314]	0.498 [0.460, 0.542]	M	-0.506 [-0.563, -0.450]	-0.229 [-0.279, -0.176]
L	-0.392 [-0.572, -0.231]	-0.066 [-0.220, 0.099]	M	0.057 [0.006, 0.107]	0.376 [0.334, 0.423]	L	-0.657 [-0.721, -0.596]	-0.414 [-0.467, -0.363]

Whether the first review is positive or negative might come down to random factors such as if the sun was shining when the first reviewer tasted the beer, if they had a bad stomach, or if they had been in a fight with their husband, which may then kick off a domino effect with potentially severe consequences: since many users rely on rating sites to decide what to buy, randomness among the first reviews can tangibly affect the business of producers.

Implications for rating-site design. A simple idea to address this problem would be to hide all reviews of a product as long as it has received less than a minimum number of ratings. If, *e.g.*, ratings are hidden until there are at least ten of them, this will mean that, effectively, the ten first ratings are independent of one another and not affected by herding. Once the eleventh reviewer arrives, they will see an average rating that reflects the inherent quality of the product much more closely than any one single review. As a consequence, even if the eleventh reviewer is biased by previous ratings, they will be biased by something much less haphazard. Future work should verify this hypothesis in an A/B test.

Community overlap. We point out that our conclusions hold even when information flows between the two sites via users active on both sites (such users exist in practice): as we focus on products with divergent first-ratings, a surmised dependence between the two sites (with respect to the products we study) would have to be one of deliberate anti-herding, which would be hard to explain.

Prior work on herding. Early work on human herding behavior (primarily from marketing and economics) was inspired by work from biology on the behavior of animal herds [7], which possibly explains why some of the earliest empirical studies of human herding considered farmers [18] and investment bankers [16].

Due to data scarcity, much early work was theoretical [1], but with the rise of the Web, empirical studies have become more feasible. The strongest evidence, naturally, comes from experimental studies. Prominently, Muchnik *et al.* [13] inserted random first-ratings into a news-story website and studied how users reacted to these treatments. Interestingly, while they found that both up- and down-votes skew later votes, the effects of down-votes were offset by social correction in their case, *i.e.*, by benign users over-compensating for down-votes with subsequent up-votes. In our case, even if social correction should occur, it certainly does not override the negative influence of haphazard early ratings (Fig. 6). A follow-up experiment [6] yielded evidence for herding on the social bookmarking site Reddit, but without evidence for social correction.

Experiments are powerful tools, but they are expensive to run and involve random manipulations that raise ethical challenges, which, *e.g.*, kept Muchnik *et al.* [13] from disclosing the site on which they had operated. Simulations [17] and observational studies [2, 3, 10] can serve as an alternative, but circumventing the problems we have mentioned in the introduction tends to require complex modeling assumptions and ways to control for confounds.

We, on the contrary, propose a methodology based on natural experiments, which, although also observational, eliminates the need for explicitly controlling for confounds by leveraging a situation where treatment assignment is haphazard. Our approach is inspired by the method of *double pair comparison* [4, 5], which was first applied to study the effectiveness of car safety belts (a concise summary of the study is given by Rosenbaum [15, Sec. 1.4]).

Applicability of our method. Being cheap and not interfering with the systems being studied are clear advantages of observational studies, especially because different settings may be affected by herding in different ways (*cf.* the above case of social correction on news stories [13] *vs.* Reddit posts [6]), such that we should study a variety of cases. Luckily, our method applies very generally (Sec. 2). We simply require a set of products rated on two separate websites and alignable across the two sites. We emphasize again the importance of verifying the validity of each setting before analyzing results. In particular, we need to ascertain that the matched sample of products is unbiased with respect to the set of all samples (external validity), and that matching products across sites indeed results in first-ratings (treatment assignment) being independent of product and site properties (internal validity).

When assessing validity, we are limited to observed product features. In particular, we argued that the style and country of a beer are the primary potential confounds, as they capture most other conceivable confounds, observed or unobserved (Sec. 4.3). Despite this extrinsic argument, we stress that one can never fully rule out unobserved confounds, something researchers should be aware of when applying our method to other datasets. When one does not have overwhelming extrinsic arguments supporting the independence of treatment from product and site properties (internal validity), one may perform a sensitivity analysis [15] to quantify how strongly the treatment would have to depend on such properties before we would alter our conclusions.

Future work. We hope that researchers will adopt our method to study herding in further scenarios. We believe that product ratings on Amazon constitute a particularly interesting case, as Amazon has sites in multiple languages (*e.g.*, Amazon.com, Amazon.de, Amazon.fr), each with an independent rating system, yet covering overlapping subsets of a wide spectrum of products. Also, as each product has a unique Amazon-wide identifier, matching is trivial.

Our results raise several interesting questions: Are certain users (*e.g.*, newcomers) more susceptible to herding than others? Can exposure to haphazard ratings lastingly alter a user's later behavior (rather than only a product's later ratings)? And finally, given ratings for the same product from several websites, can we develop models for combining them into a more truthful aggregate score?

REFERENCES

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

[2] Judith A Chevalier and Dina Mayzlin. 2006. The Effect of Word of Mouth on Sales: Online Book Reviews. *Journal of Marketing Research* 43, 3 (2006), 345–354.

[3] Wenjing Duan, Bin Gu, and Andrew B Whinston. 2008. Do Online Reviews Matter? An Empirical Investigation of Panel Data. *Decision Support Systems* 45, 4 (2008), 1007–1016.

[4] Leonard Evans. 1986. Double Pair Comparison: A New Method to Determine How Occupant Characteristics Affect Fatality Risk in Traffic Crashes. *Accident Analysis & Prevention* 18, 3 (1986), 217–227.

[5] Leonard Evans. 1986. The Effectiveness of Safety Belts in Preventing Fatalities. *Accident Analysis & Prevention* 18, 3 (1986), 229–241.

[6] Maria Glenski and Tim Weninger. 2017. Rating Effects on Social News Posts and Comments. *ACM Transactions on Intelligent Systems and Technology* 8, 6 (2017), 78.

[7] William D Hamilton. 1971. Geometry for the Selfish Herd. *Journal of Theoretical Biology* 31, 2 (1971), 295–311.

[8] Nan Hu, Paul A Pavlou, and Jennifer Zhang. 2006. Can Online Reviews Reveal a Product's True Quality? Empirical Findings and Analytical Modeling of Online Word-of-Mouth Communication. In *Proceedings of the 7th ACM Conference on Electronic Commerce*. 324–330.

[9] Nitin Jindal and Bing Liu. 2007. Analyzing and Detecting Review Spam. In *Proceedings of the 7th IEEE International Conference on Data Mining*. 547–552.

[10] Young-Jin Lee, Kartik Hosanagar, and Yong Tan. 2015. Do I Follow my Friends or the Crowd? Information Cascades in Online Movie Ratings. *Management Science* 61, 9 (2015), 2241–2258.

[11] Jure Leskovec and Andrej Krevl. 2016. SNAP Datasets: Stanford Large Network Dataset Collection (2014). *http://snap.stanford.edu/data* (2016).

[12] Julian J McAuley, Jure Leskovec, and Daniel Jurafsky. 2012. Learning Attitudes and Attributes from Multi-Aspect Reviews. *Proceedings of the 12th IEEE International Conference on Data Mining* (2012), 1020–1025.

[13] Lev Muchnik, Sinan Aral, and Sean J Taylor. 2013. Social Influence Bias: A Randomized Experiment. *Science* 341, 6146 (2013), 647–651.

[14] Andreas Roider and Andrea Voskort. 2016. Reputational Herding in Financial Markets: A Laboratory Experiment. *Journal of Behavioral Finance* 17, 3 (2016), 244–266.

[15] Paul R Rosenbaum. 2010. *Design of Observational Studies*. Springer.

[16] David S Scharfstein and Jeremy C Stein. 1990. Herd Behavior and Investment. *The American Economic Review* 80, 3 (1990), 465–479.

[17] Ting Wang and Dashun Wang. 2014. Why Amazon's Ratings Might Mislead You: The Story of Herding Effects. *Big Data* 2, 4 (2014), 196–204.

[18] Kislev Yoav and Nira Shchori-Bachrach. 1973. The Process of an Innovation Cycle. *American Journal of Agricultural Economics* 55, 1 (1973), 28–37.