

# Reputation in the Long-Run\*

Apostolos Filippas<sup>†</sup>    John J. Horton<sup>†</sup>    Joseph M. Golden<sup>‡</sup>

February 8, 2018

## Abstract

We present evidence from five online marketplaces showing that feedback scores have risen sharply over time, leading to substantial top-censoring. We use alternative measures of rater satisfaction—numerical and textual—to show that although some of the increase is explained by more satisfied raters, at least 48-50% is attributable to raters applying lower standards. We show that this “reputation inflation” is the equilibrium outcome in a market where (a) inferences made by future trading partners determine what constitutes “bad” feedback and (b) giving “bad” feedback is costly to raters. We exploit a natural experiment from the introduction of a new feedback system: raters were candid when the new feedback was private, but when it suddenly became public, reputations began inflating.

---

\*Author contact information and code are currently or will be available at <http://www.john-joseph-horton.com/>. Thanks to Richard Zeckhauser, Andrey Fradkin, David Holtz, Ramesh Johari, Nico Lacetra, Xiao Ma, and Aaron Sojourner for very helpful comments and suggestions. Helpful feedback was received at the Crowdsourcing Seminar at Carnegie Mellon University at the School of Computer Science, the NBER Summer Institute on the Economics of Digitization, and the MIT Conference on Digital Experimentation.

<sup>†</sup>NYU Stern School of Business

<sup>‡</sup>Collage.com

# 1 Introduction

Scores of various kinds—credit scores, school grades, restaurant and film “star” reviews, restaurant hygiene scores, Better Business Bureau ratings—have long been important sources of information for market participants. A large literature documents the economic importance of such scores (Resnick et al., 2000; Jin and Leslie, 2003; Resnick et al., 2006; Mayzlin et al., 2014; Ghose et al., 2014; Luca, 2016; Luca and Zervas, 2016). As more of economic and social life has become computer-mediated, opportunities to generate and apply new kinds of scores—particularly in marketplace contexts—have proliferated (Dellarocas, 2003), as has the number of individuals and businesses subject to these “reputation systems” (Levin, 2011; Farrell and Greig, 2016; Hall and Krueger, Forthcoming; Katz and Krueger, 2016).<sup>1</sup> Designing effective reputation systems and understanding their implications for online marketplaces has become a first-order question in the digital economy.

In online marketplaces, “reputations” are typically calculated from the numerical feedback scores left by past trading partners. As many have noted, the distribution of feedback scores in various online marketplaces seems implausibly rosy.<sup>2</sup> For example, the median seller on eBay has a score of 100% positive feedback ratings, and the tenth percentile is 98.21% positive feedback ratings (Nosko and Tadelis, 2015). On Uber and Lyft, it is widely known that anything less than 5 stars is considered “bad” feedback: Athey et al. (2018) find that nearly 90% of UberX Chicago trips in early 2017 had a perfect 5 star rating. Of course, there is no ground truth that tells us what the distribution of scores in some market

---

<sup>1</sup>These kinds of online marketplaces are becoming increasingly consequential as they grow rapidly, making their shortcomings consequential as well. Katz and Krueger (2016) find that the share of workers whose main job is an alternative work arrangement—defined to include independent contractors and freelancers—has increased from 10.7 percent in 2005, to almost 16 percent in 2015. A large part of this increase can be attributed to employment in online platforms; Farrell and Greig (2016) show that participation in the online platform economy amongst adults has risen from less than 0.2 percent in 2012, to 4.3 percent in 2016. The same authors also find that the annual growth rate of workers receiving income from online platforms exceeds 100 percent. Hall and Krueger (Forthcoming) report that more than 460,000 drivers were actively participating on the Uber platform by the end of 2015.

<sup>2</sup>A survey conducted by the PEW research center finds that while more than 80% of U.S. adults read online reviews before purchasing an item, almost 50% believe that it is hard to assess the truthfulness of these reviews (see <http://www.pewinternet.org/2016/12/19/online-reviews>).

“should” look like. However, to the extent that there is meaningful variation in latent transaction quality, a high degree of top-censoring—every transaction being assigned the same “5 star” label—suggests that the function of the reputation system is undermined.

In this paper we examine the reputation system of a large online labor market. We find that the distribution of recent employer feedback for workers is highly top-censored, with an overwhelming majority receiving perfect feedback.<sup>3</sup> However, the distribution has not always been this skewed—the fraction of workers receiving the highest possible rating of 5 stars went from 32% to 85% in just 6 years. Feedback scores in four other online marketplaces for which we obtain longitudinal data exhibit a similar increase over time, suggesting that this phenomenon is commonplace. Importantly, this increasing trend manifests itself despite the mechanisms that these marketplaces employ to prevent “tit-for-tat” rating behavior (Bolton et al., 2013).

Our main focus is understanding the cause behind the rise in feedback scores. There are two distinct—but not mutually exclusive—reasons feedback scores could be rising: (1) raters are becoming more satisfied, or (2) raters are lowering their standards. This second possibility—giving higher scores despite not being more satisfied—can be thought of as a kind of inflation.<sup>4</sup> Disentangling these two reasons requires longitudinal data that include both the feedback scores and an alternative measure of rater satisfaction.

As a first alternative measure of rater satisfaction, we use information obtained by the introduction of a new, parallel reputation system that asked employers to rate workers “privately.” The private feedback was not conveyed to the rated workers, nor made public to future would-be employers. At the same time, raters were still asked to give the status quo “public” feedback, both written and numerical. The conjecture motivating the private feedback feature was that raters would be more candid in private, willing to give “bad” feedback if not exposed to the reflected cost from angry workers. We find that private feedback scores were decreasing, while public feedback scores for the *same* transactions

---

<sup>3</sup>We use the terms “employer” and “worker” for consistency with the literature, and not as a comment on the legal relationship of the transacting parties.

<sup>4</sup>This kind of inflation is similar to the conjecture about college grade inflation (Babcock, 2010; Butcher et al., 2014).

were increasing. This difference suggests that raters lowered their standards, and hence this increase in public feedback scores is due to inflation rather than more satisfied raters. Importantly, this trend is not because of

The private feedback feature provides us with strong evidence that inflation occurs, but it does not span the entire life time of the platform’s operations—we observe both ratings for only 10 months. Further, it is important for platforms that do not collect private information to be able to assess the degree that their reputation systems are subject to inflation. As a second alternative measure of rater satisfaction, we use the sentiment raters express in the written feedback that accompanies numerical scores. We can compute this measure throughout the entire operations of the platform. To capture this sentiment, we fit a model that predicts numerical feedback from the text of written feedback. Critically, the model is fit using feedback from a narrow window of time early in our data set. With our method, we can learn the relationship that prevailed between written sentiment and score when the training feedback was created. Using the predictions, we can then decompose the growth in average feedback score into a component due to improvements in market “fundamentals” (e.g., improved marketplace features, better cohorts of workers, less picky employers, and so on) that increased rater satisfaction (as reflected in feedback text), and the residual component that cannot be explained by changes in fundamentals.

We find that although predicted feedback scores have increased over time, they have not increased nearly as much as the actual, numerical feedback scores. Our estimates suggest that about 48.6-50.2% of the increase in scores during a 6 year period was due to inflation, i.e., raters lowering their standards. Further, to the extent that written feedback is also subject to inflation, our approach *understates* the role of lower standards in explaining the rise in average feedback.

As a less model-dependent approach to quantifying inflation from textual feedback, we also compare the numerical feedback scores associated with the same common sentences appearing in written feedback in two different time periods. We show that the same sentences systematically have higher associated numerical feedback scores in the latter period. To the extent that the meaning of the same sentences has not changed between these time periods, reputation inflation is the culprit.

We next turn to understanding the cause of reputation inflation, or why raters lower standards over time. A starting point is noting the asymmetry in costs to giving different kinds of feedback, as it manifests in the divergence between public and private feedback scores: 28.4% of those employers that privately report that they would definitely not hire the same employee in the future, publicly assign them 4 or more stars. Some report they fear retaliation, while others claim to not want to harm the rated individual, as public feedback is highly consequential. Although rated workers can not retaliate by giving the employer bad feedback, they can still complain, bad-mouth the rater, withhold future cooperation, and so on. These “reflected” costs make giving “bad” feedback costlier to the rater than giving “good” feedback.

The cost of giving “bad” feedback could provide an explanation for why feedback scores are higher than they would be if more employers reported truthfully. However, it cannot by itself explain the dynamics of ever-higher scores we observe: inflation would require the cost of leaving a given “bad” score to increase over time. We hypothesize this is precisely what happens; the same nominal feedback score (e.g., 3 “stars”) can become costlier to the rated worker over time, and hence costlier for the rater to give. In other words, the cause of inflation is that what constitutes “bad” feedback—feedback that causes worse market outcome for raters receiving that score—is endogenous, depending on the current distribution of feedback scores, and what inferences future employers make from a score.

To formally illustrate how reputations can inflate, we present a simple model of a marketplace with a reputation system. We show that there exists a stable equilibrium in which sellers only report “good” feedback, regardless of actual performance, and reputations are universally inflated, even when raters derive some benefit from telling the truth.<sup>5</sup> While our observed data is consistent with our model, we report evidence from a platform intervention that allows us to test the assumptions and predictions of our model.

---

<sup>5</sup>In our model, the degree of reputation inflation depends on how much cost the rated entity can impose on the rater for “bad” feedback. This could explain why in less personal settings—such as consumers rating products on Amazon or restaurants on Yelp—ratings are more spread out. In contrast, inflation is likely more acute in highly “personal” settings, such as on peer-to-peer platforms (Sundararajan, 2013; Horton and Zeckhauser, 2017).

The platform intervention was that, after collecting private feedback for 10 months, the platform began releasing batched *aggregates* of this private feedback score to would-be employers. With this aggregation private feedback remained quasi-anonymous, as the worker would not know ex post which particular rating employer gave which feedback, unless every rater in the batch gave the same rating. However, with this private feedback now being reported, it became consequential to workers, who now had incentives to try to encourage good private feedback. Further, to the extent that employers care about the fate of workers and do not want to harm their future prospects—or believe that it could get “back” to them—giving “bad” feedback suddenly had a cost. Hence, the switch to public revelation was a positive shock.

This private feedback quasi-experiment is useful for our purposes because it allows us to test our model’s predictions, and assess its assumptions about the causes of reputation inflation, namely that (1) the rater’s choice of what score to give is “strategic”—in the sense that employers consider the likely costs and benefits to what they report—and as such, are more candid in private because there are no reflected costs, and (2) when costs are introduced by the switch to public revelation, inflation occurs.

When the platform suddenly made private feedback scores public, private feedback scores began increasing immediately but there was no corresponding change in the sentiment of written feedback—the no-longer-private feedback became inflated, mirroring what we observed with public feedback, albeit in a much shorter time window. Importantly, the fact that the sentiment of written feedback remained more or less constant provides us with direct evidence that written feedback sentiment is an estimate of satisfaction that is less prone to inflation than numerical feedback. Further, this change implicates the role of the reflected cost of “bad” feedback: when getting “bad” feedback became costly to workers, it also became costly to give, and there was less “bad” feedback. As “bad” feedback became scarce, what was mildly negative before became very negative, starting the inflation process described in our model.

Our paper makes several contributions. Our key contribution is documenting the extent and practical importance of reputation inflation in a large online marketplace with a state-of-the-art reputation system. We suspect the reputation

inflation problem is widespread, given that many marketplaces share the same features as the one we study, and nearly all have those features that we show lead to inflation. We also elucidate the root cause of that inflation, and our analysis reveals the dynamics that lead to this outcome. Our long-run, whole-system perspective is possible because we use a data set spanning over a decade of the operations of the marketplace. While our paper is not the first to explain how reputations can be biased (Dellarocas and Wood, 2008; Li and Hitt, 2008; Hu et al., 2017), we believe it is the first to show how individually rational choices about what feedback to leave can push the market towards a less informative equilibrium, and hence put the reputation system on an inexorable path towards uselessness. Further, to the best of our knowledge, this paper makes a methodological contribution that utilizes predictive modeling on longitudinal data to quantify the degree of inflation in online reputation systems.

Whether reputation systems can be designed that are less prone to inflation remains an important open research question. The quasi-experimental revelation of private feedback demonstrates that simply making feedback anonymous is not sufficient to counteract inflation. Instead, our findings suggest that the main reason behind inflation is rating individuals' unwillingness to harm the rated individuals' future prospects. This suggests an inherent tension between being consequential and being informative.

The rest of the paper is organized as follows. Section 2 describes our empirical setting, and documents that average feedback scores increase over time across a number of online platforms. Section 3 shows that this increase is largely due to inflation, by employing private and textual feedback information as alternative measures of rater satisfaction. A model for reputation inflation is presented in Section 4, and its predictions are tested in Section 5 by employing the quasi-experimental revelation of private feedback data information. We conclude in Section 6.

## 2 Feedback over time

### 2.1 Empirical context

The setting for our study is a large online labor market. In online labor markets, firms and individuals hire workers to perform tasks that can be done remotely, such as computer programming, graphic design, data entry, and writing. Markets differ in their scope and focus, but common services provided by the platform include maintaining job listings, hosting user profile pages, arbitrating disputes, certifying worker skills and, importantly, maintaining reputation systems (Horton, 2010).

Online labor markets have offered a convenient setting for research, due to the excellent measurement afforded in an online setting, and the ease with which field experiments can be conducted (Horton et al., 2011). Much of the research has focused on the role of information in employer decision-making (Pallais, 2013; Stanton and Thomas, 2015; Agrawal et al., 2016; Chan and Wang, Forthcoming; Horton, 2017). There is also a growing literature on online labor markets as a phenomenon and as a domain to study online marketplaces more generally. This literature explores topics such as the nature of economic relationship created (Chen and Horton, 2016), the role of preference signaling (Horton and Johari, 2015; Horton, Forthcoming), and the bidding process (Zheng et al., 2016).

One particular focus of the literature has been reputation systems. Moreno and Terwiesch (2014) show how employers use reputation information—and subsequently how workers adjust their bidding strategies in light of this employer conditioning. Cabral and Hortacsu (2010) also find that eBay sellers condition their behavior on their current reputations. Moreno and Terwiesch (2014) also use written feedback as an alternative measure of rater satisfaction, though they extract the sentiment using unsupervised learning techniques, in contrast to our supervised learning where the label is the associated numerical feedback score. Dimoka et al. (2012) show that reputation is more important in labor than in product markets, as a bad seller may by chance offer a great product, but a bad worker almost certainly produces bad work. Reputation matters when hiring workers for fixed-price contracts, while its role is diminished for contracts



with hourly payments (Lin et al., Forthcoming). Kokkodis and Ipeirotis (2015) explore the “transferability” of reputations, showing that reputation scores can become more predictive of future performance when job category information is incorporated.

## 2.2 Status quo reputation system

On the platform used in our study, when one party ends a contract—typically the employer—both parties are prompted to give feedback.<sup>6</sup> Employers are asked to give both written feedback, e.g., “Paul did excellent work—I’d work with him again” or “Ada is a great person to work for—her instructions were always very clear,” and numerical feedback. The numerical feedback is given on several weighted dimensions: “Skills” (20%), “Quality of Work” (20%), “Availability” (15%), “Adherence to Schedule” (15%), “Communication” (15%) and “Cooperation” (15%). On each dimension, the rater gives a score on a 1-5 scale. The scores are aggregated according to the dimension weights. A worker’s reputation at a moment in time is the average of her scores on completed projects, weighted by the dollar value of each project. On the worker profile, a lifetime score is shown as well as a “last 6 months” score, which is more prominently displayed. Showing recent feedback is presumably the platform’s response to the opportunism that becomes possible once a employer or worker has obtained a high, hard-to-lower reputation (Aperjis and Johari, 2010; Liu, 2011). Despite the aggregation of individual scores into a reputation, the entire feedback “history” is available for inspection to interested parties. Workers can view the feedback given to previous workers rated by that employer and the feedback received by an employer from that same worker.

The reputation system could be characterized as state-of-the-art for a bilateral system, in the sense that direct tit-for-tat conditioning is not possible (Dellarocas, 2005; Bolton et al., 2013; Fradkin et al., 2015). Both the employer and the worker have an initial 14 day “feedback period” in which to leave feedback. The platform does not reveal public feedback immediately. Rather, the platform uses a “double-

---

<sup>6</sup>We use the present tense here to describe the reputation system before the introduction of private feedback.

blind” process. If both parties leave feedback during the feedback period, then the platform reveals both sets of feedback simultaneously. If, instead, only one party leaves feedback, then the platform reveals it at the end of the feedback period. Thus, neither party learns its own rating before leaving a rating for the other party.

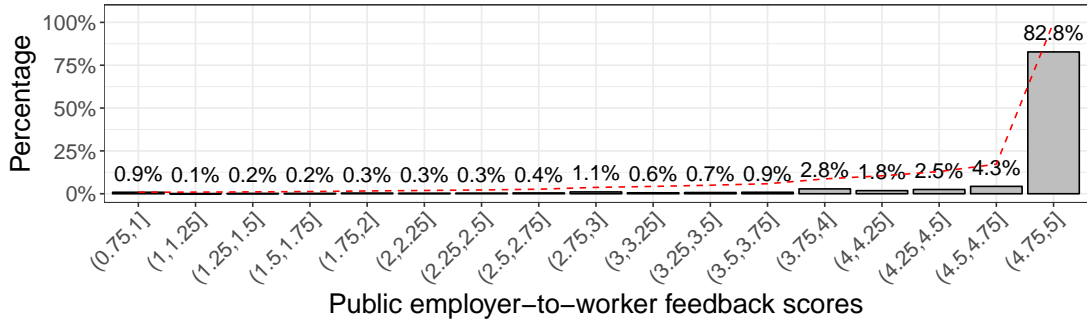
Despite the reputation system features designed to prevent tit-for-tat feedback, there is nothing to stop parties from engaging in “pre-play” communication about their intentions. We have some evidence that feedback manipulation occurs, from forum and blog postings, communication between employers and workers, and complaints directly to the platform. It is generally difficult to directly assess the severity of this problem, partially because communication about manipulation between the two parties may occur entirely in private, such as via email. However, a survey of platform employers found that 20% had felt pressure to leave more positive public feedback. Leaving feedback is not compulsory, though it is strongly encouraged. These encouragements seem effective, in that over the history of the platform, 68% of employers eligible to leave feedback have chosen to do so.

### 2.3 Feedback now and in the past

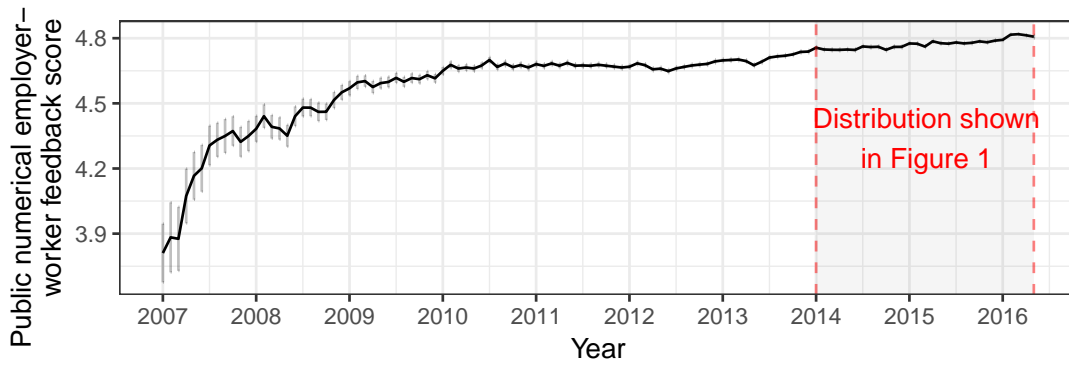
The distribution of employer-on-worker feedback scores in the market is highly right-skewed. Figure 1a depicts the histogram of completed assignments that received a feedback score from the employer, from January 1, 2014 to May 11, 2016. Public feedback scores are between 1 and 5 stars, inclusive, and with increments of 0.25 stars. Each bar is labeled with the percentage of total observations falling in that bin, and the red dashed line shows the cumulative number of assignments with feedback less than or equal to the right limit of the bin it is above. We observe that more than 80% of the evaluations fall in the 4.75 to 5.00 star bin (1,339,071 observations). The ratings distribution is *slightly* J-shaped (Hu et al., 2009), with some weight in the lowest bin of observations rated with exactly 1 star. The average feedback pooled over the whole sample shown in Figure 1a is 4.77.

Figure 1: Characteristics of employer-on-worker feedback in an online marketplace

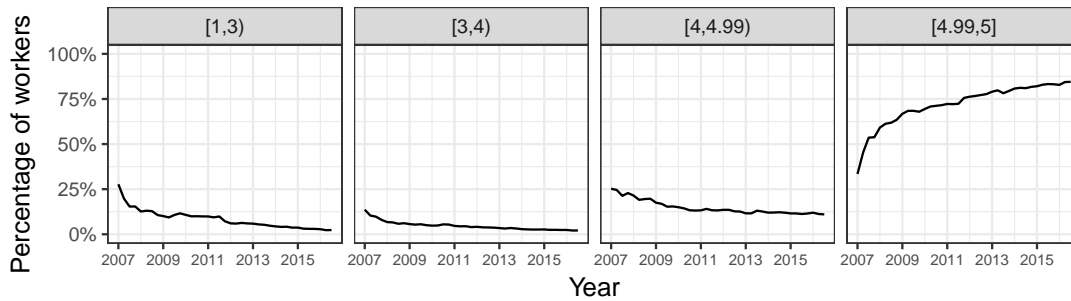
(a) Distribution of feedback scores for the period January 1, 2014 to May 11, 2016.



(b) Monthly average public feedback scores assigned on completed projects.



(c) Percentage of completed projects receiving different star ratings over time.



*Notes:* The top panel shows the histogram of public numerical ratings assigned by employers to workers, discretized by 0.25 star interval bins. The scale for feedback is 1 to 5 stars. The value of each bin is shown above it, and the red line depicts the empirical cumulative density function. The sample we use consists of all completed contracts from January 1, 2014 to May 11, 2016, for which the employer provided feedback. The middle panel plots the average public feedback scores assigned by employers to workers on completed contracts by month. The average scores are computed for every month, and a 95% interval is depicted for every point estimate. The shaded area denotes the data that was used in Figure 1a. This bottom panel plots the fraction of public feedback scores assigned in a given month into four bins, [1, 3), [3, 4), [4, 4.99), and 5 stars, over time.

In Figure 1b we plot the average monthly feedback over time. This measure is for contracts ending within that month, and hence approximately the month when that feedback was given. There is a clear increase in the feedback scores awarded on the platform: the numerical feedback score average has increased by more than one star over the ten years of operation of the platform, from 3.74 in the beginning of 2007, to 4.81 in May 2016. The strongest period of increase was 2007, when average feedback scores increased by about 0.53 stars.

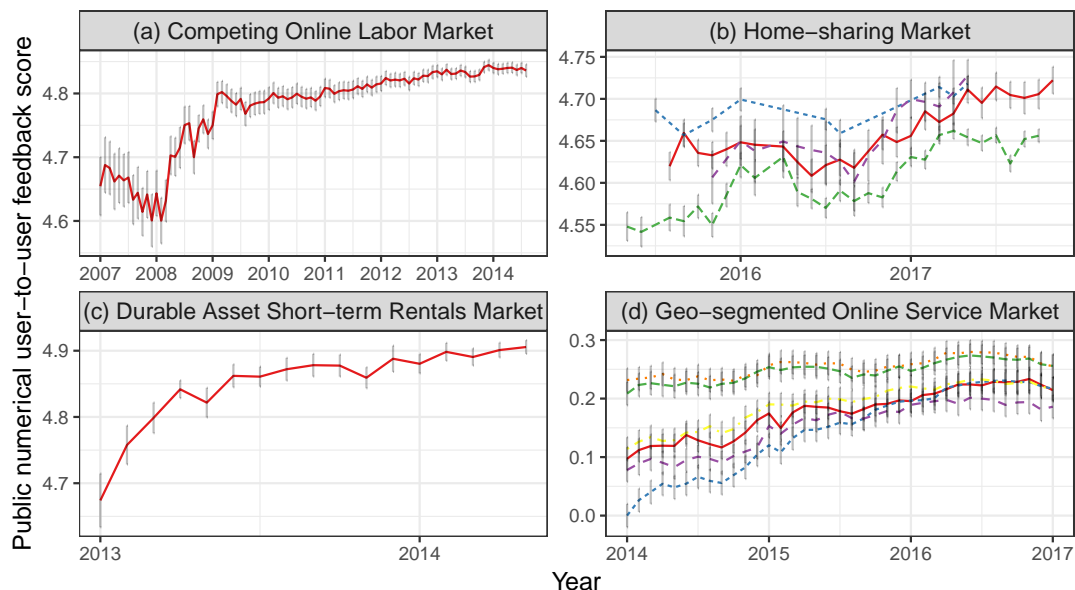
The increase in average feedback could be the outcome of raters giving less “bad” feedback, more “good” feedback, or some combination thereof. For example, we could see a decrease in the proportion of workers that receive 1-star ratings, and an increase in the percentage of workers receiving 3 stars. Figure 1c shows the fraction of contracts having a rating within different ranges, over time. We can see in the leftmost panel of Figure 1c that completed contracts regularly received ratings in the  $(0, 3]$  range in the early days of the platform. Further, early on the ratings assignments were reasonably dispersed, with every bin containing at least 15% of the employers’ ratings. Near the end of our data, completed contracts essentially never receive a rating in the  $(0, 3]$  star bin, despite nearly 30% of such contracts getting this rating originally. Instead, there has been a dramatic increase in the fraction of contracts getting exactly 5 stars: 32% of contracts received a 5-star rating at the start of sample, compared to 85% at the end of the sample.

## 2.4 Evidence from other online marketplaces

An important question about the strong increase in feedback scores observed in Figure 1b is whether the same pattern can be found in other online platforms. Unfortunately, longitudinal data is rare: evidence presented in previous work mostly pertains to the distribution of feedback scores at a moment in time, and hence it does not allow us to examine whether high ratings are the outcome of an over time increase. To answer this question, we collected data from four other online marketplaces.

We first obtained longitudinal ratings data from a competing online labor market. The average monthly feedback given by employers to workers is plotted

Figure 2: Longitudinal buyer-on-seller feedback scores for a collection of online marketplaces



*Notes:* This figure plots the average public feedback scores assigned in four online peer-to-peer marketplaces. Scores are assigned by employers to workers in panels (a), by guests (users renting properties) to hosts (users renting out properties) in panel (b), by renters (those renting the durable asset) to providers (those renting out the durable asset) in panel (c), and by customers to providers of a service in panel (d). The panel (d) scores are demeaned by the grand mean for all observations. Scores are assigned upon the completion of each transaction, and the scale for feedback is 1 to 5 stars. For each observation, average scores are computed for every time period, and a 95% interval is depicted for every point estimate.

in the top panel of Figure 2. Although the average feedback on this marketplace starts from a higher point—likely reflecting the greater age of this platform (launched in 1999)—an increase is also present over time.

We also obtained longitudinal ratings data from a large home-sharing platform. Home-sharing platforms are peer-to-peer marketplaces that facilitate short-term rentals for lodging (Filippas and Horton, 2017). The average monthly feedback given by guests (those who are renting properties) to hosts (those who are renting out properties) is plotted in the second panel of Figure 2 for 4 large cities located in Europe and in the United States. Though our data do not cover the entire operations of the platform in each city, we observe a similar increase in the numerical ratings over time.

We next obtained over time numerical feedback data from an online marketplace that facilitates the short-term rental of a durable asset. The third panel of Figure 2 plots the average monthly feedback given by users (renters of the asset) to users (providers of the asset) after the transaction has taken place. We notice that average feedback scores exhibit a similar pattern of strong increase over time.

Finally, we obtained longitudinal ratings data from a large online marketplace for services. The average monthly feedback assigned to providers by users of the platform is plotted in the bottom panel of Figure 2 for 6 large cities in the United States. We observe a similar pattern of strong over time increase in our data.

Increasing feedback scores are common in other online marketplaces with similar reputation systems for very different goods and services. We next focus on the first online labor market, and we utilize our longitudinal transaction-level data to examine the reasons that lead to this increase.

### 3 Reputation Inflation

The previous section documents a substantial increase in feedback scores over time across several online platforms. There are two broad sets of reasons that may have led to this increase: (1) rater satisfaction has increased, and (2) raters are lowering their standards. We cannot disentangle these two sets of reasons by only using the numerical feedback scores, as these scores potentially reflect changes in both rater satisfaction and standards. Instead, we turn to alternative measures of rater satisfaction. The main idea here is that if an alternative measure of rater satisfaction is more accurate than numerical scores, inflating slower or not at all, then comparing relative changes over time between the two will allow us to estimate a lower bound on the degree to which feedback scores inflate.

As a first alternative measure of rater satisfaction, we use private feedback information collected by the platform. Unlike public feedback, private feedback information was not accessible by either future employers or workers, and hence “bad” private feedback was costless for workers to receive. As a second alternative measure of rater satisfaction, we use the public written feedback that employers generate. We will argue that written feedback is an attractive alternative measure

of rater satisfaction because changes in sentiment are more likely to reflect changes in rater satisfaction, rather than a change in standards.

### 3.1 Private feedback as an alternative measure

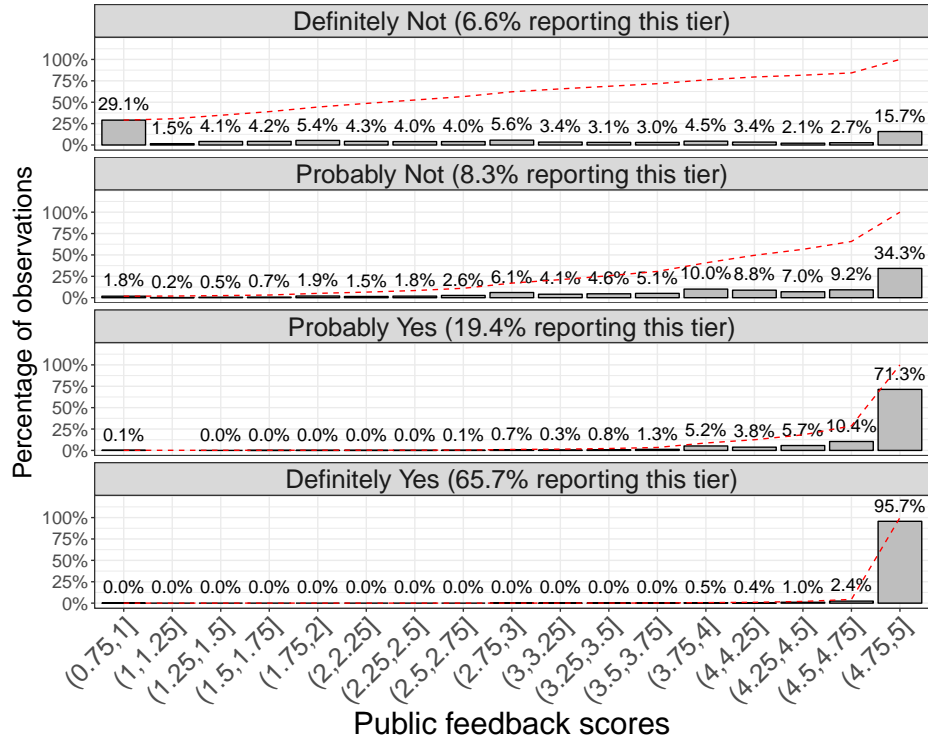
Our first alternative measure comes from a platform experiment that asked for an additional “private” feedback measure of feedback. The platform introduced this new experimental “private feedback” feature in April, 2013. With this system, on the completion of a contract, employers were asked to generate private feedback in addition to public feedback. Critically, the platform let the employers know that private feedback would not be shared with the rated worker or future would-be employers, and would only be collected by the platform for internal evaluation purposes. Because of this private structure, private feedback offers an alternative measure that is potentially not subject to inflation—we elaborate on the reasons why in Sections 4 and 5.

Employers were initially asked the private feedback question, “Would you hire this freelancer [worker] again, if you had a similar project?” There were four response options: “Definitely yes,” “Probably yes,” “Probably not,” and “Definitely not.” Starting on the beginning of May 2014, the employers were asked to rate the worker on a numerical scale of 0 to 10.

#### 3.1.1 Comparing private and public feedback

As employers assigned both public and private feedback for the same contract, we can compare the two scores assigned by the same employer on the same contract. In Figure 3, each of the four possible responses to the private feedback question are shown as facets. The percentage of employers giving that feedback score is shown in parenthesis. Underneath, the distribution of public scores for that private feedback is shown. Although the most common response was “Definitely Yes,” about 15% of the employers gave unambiguously bad private feedback (“Definitely Not” and “Probably Not”). In contrast, during the same period less than 4% of the employers gave a numerical score of 3 stars or less. Given this gap, we might suspect that some employers expressing a negative private sentiment are less candid in public.

Figure 3: Distribution of publicly given feedback to workers, by private feedback.



Notes: This figure plots the distribution of public feedback scores, computed separately for every set of users that gave the same answer to the private feedback question. The red dotted line plots the cumulative distribution function.

Turning to the conditional public feedback, we can see that employers that selected “Definitely yes,” also left very positive public feedback, with more than 95% of these observations falling into the highest bin. Employers leaving more negative private feedback, also assign lower public feedback scores. At the extremes, among those employers that selected the “Definitely No” answer to the private feedback question, 29.1% assigned a 1-star rating publicly. Surprisingly, however, the second most common choice for these employers at 15.7% was in the 4.75 to 5.00 bin, while 28.4% publicly assigned more than 4 stars. In short, many privately dissatisfied employers publicly claimed to be satisfied.



### 3.1.2 Private feedback over time

The platform began eliciting numerical private feedback by employers on workers in May 2014. Similarly to the private feedback question, the platform informed employers that numerical private feedback would not be released to either other employers or workers, but instead would only be used internally. As a consequence of this new platform feature we now possess two separate numerical measures; we can use private feedback as an alternative measure of employer satisfaction, and examine whether there is a difference in their behavior over time.

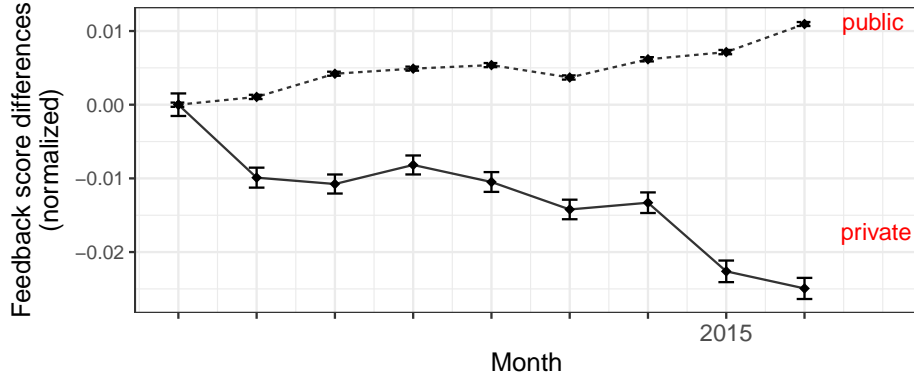
In Figure 4 we plot the average monthly feedback trends over time, for the numerical public and private feedback during the period that both were collected. Both scores are normalized by their respective average during the first period for which both were collected. Public feedback scores exhibit a small increase during the period of interest, as we noted in Figure 2. For the same period of time, however, private feedback scores exhibit a strong decreasing trend. This divergence between public and private feedback scores strongly suggests that public numerical scores are subject to considerable inflation. As the decrease in private feedback scores suggests a decline in rater satisfaction, we can infer that the increase due to inflation accounts for *at least* the observed increase in public feedback scores.

In Appendix A we verify that the process through which employers assign private feedback does not change over time. We should note that other assumptions about the private feedback assignment behavior of employers, such as that private standards are getting harsher even though the quality of transactions is increasing are improbable.

## 3.2 Written feedback as an alternative measure

Private feedback scores offer an alternative measure of rater satisfaction during the period where both types of feedback were elicited by the platform. However, it is important to examine whether feedback scores were also subject to inflation earlier in the lifetime of the platform—feedback scores sharply increased during 2007. Further, a different alternative measure is useful for platforms that do not elicit private feedback, but aim to assess the extent that their reputation systems

Figure 4: Monthly average public and private feedback scores assigned to workers by employers.



*Notes:* This figure plots the evolution of the average public feedback scores (dotted line) versus the average private feedback scores (solid line) assigned by employers to workers. The average scores are computed for every month, and are normalized by the value of their respective first observation (May 2014). A 95% confidence interval is shown for each mean.

are subject to inflation. Toward that end, we employ written feedback left by employers as an alternative measure of rater satisfaction to conduct a similar analysis.

To make the two kinds of feedback comparable, we fit a predictive model that predicts numerical feedback scores from the feedback text. The predictive model is fit on a narrow time window, and the fitted model is then used to estimate out-of-sample feedback scores of the written feedback for the entire sample.

Words used in written feedback can certainly become “inflated,” with work that would have elicited a “good” now garnering a “great.”<sup>7</sup> However, some words found in written feedback, such as “unresponsive,” are less subjective and are likely to retain the same meaning, more or less, over time. We also suspect that written feedback is inherently less subject to inflationary pressures. The reason is that the nature of written feedback and how it is used on the platform makes it less costly for raters to be candid, and these costs are central to explaining reputation inflation (which we discuss later in the paper).

The reason that the costs to the rater for giving “negative” written feedback

<sup>7</sup>For example, an employer’s written feedback in our dataset reads: “This is the most impressive piece of coding in the history of software development!”

are lower than for numerical feedback is that it is harder for workers to complain about textual “tone” than it is to complain about a non-perfect star rating. Workers also likely care less about bad written feedback because of how it is presented, and are thus less likely to try to compel better feedback: written feedback is not aggregated or put on a scale, and hence cannot as easily be used for cross-worker comparisons. These comparisons are precisely what makes the feedback consequential for workers. Relatedly, many would-be employers would likely not bother to read the often voluminous collection of written feedback, making any particular written feedback less consequential.

To the extent that written feedback offers a a more or less unchanging measure of rater satisfaction, it is useful for disentangling changes in rater satisfaction from changes in rater standards. Importantly, to the extent that written feedback is also subject to inflation, our approach will underestimate the magnitude of the inflation in scores.

### **3.2.1 Predicting numerical feedback from written feedback**

To extract the sentiment of the written feedback, we employ a standard machine learning approach. We use a sample of our written feedback corpus as the training set, with the associated numerical scores as the set of labels. We fit a model that predicts the public feedback score, given the text of the written feedback. We use a standard natural language preprocessing pipeline: our data is stripped of accents and special characters, is lowercased, stopwords are removed, a matrix of token counts (up to 3-grams) is created, and is weighed using the TFIDF method. We then perform an extensive grid search over a set of different learning algorithms and their corresponding hyperparameters, evaluating each configuration using a 5-fold cross validation.<sup>8</sup> We keep the best performing model in terms of average squared error.<sup>9</sup> An important step in our method is deciding the time period where the training corpus for our model will come from.

---

<sup>8</sup>The algorithms we use are linear regression, lasso regression, ridge regression, gradient boosting regression, and random forest regression.

<sup>9</sup>Our analysis is performed using the Python scikit-learn package implementation. The package’s webpage provides a detailed description of the implementations of each model (see [http://www.scikit-learn.org/stable/user\\_guide.html](http://www.scikit-learn.org/stable/user_guide.html)).

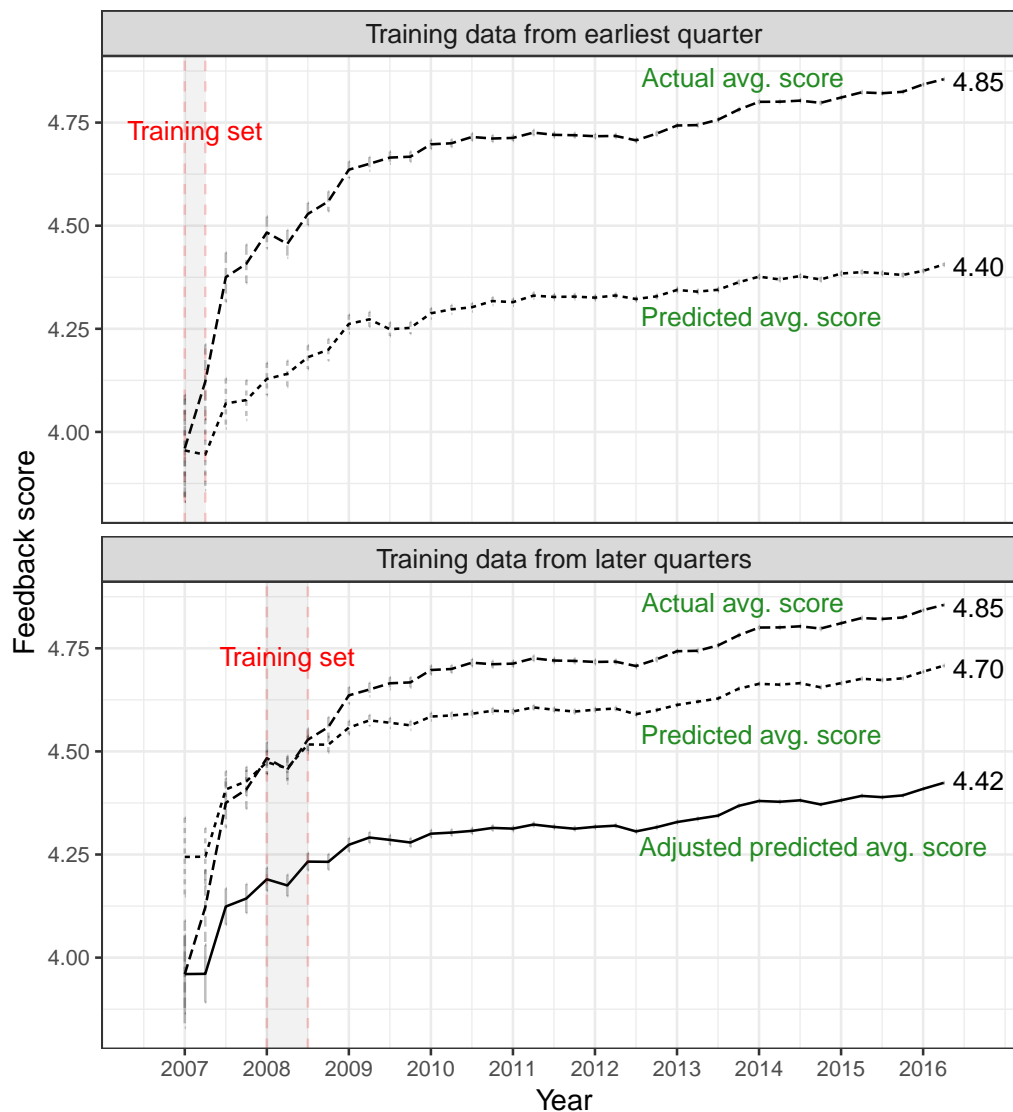
The average quarterly feedback over time, for both the numerical public feedback, and the feedback predicted from the written feedback, are plotted in the top panel of Figure 5. The predictive model is trained with a written feedback corpus from the earliest quarter in our data (indicated with a dashed red line), and consists of 1,492 feedback samples. As expected, the predicted and actual scores match up during the training period. Going forward, both scores increase, but the predicted feedback score increases at a much slower rate. On average, numerical feedback goes from 3.96 in the beginning of 2006 to 4.85 stars at the beginning of 2016. In contrast, the average score predicted from the written feedback only goes to 4.40 stars.

In the bottom panel of Figure 5 we again plot the average quarterly feedback over time, for both the numerical public feedback and the feedback predicted from the written feedback. However, our training sample now comes from a longer time period indicated by the two vertical red lines, and is larger, consisting of 10,555 feedback samples. As expected, the predicted and actual scores closely match up during the training period. However, in the period before, the predicted score is higher than the numerical score, and vice versa post the training period. We adjust the second score by a constant, so that the predicted score matches the actual feedback score in the beginning of our dataset. With this adjustment, the average predicted feedback score at the end of the data set “should” have only been 4.42 stars.

The divergence between written sentiment and numerical feedback implies that a substantial amount of the increase in numerical feedback scores is due to lower rater standards. Our approach also allows us to quantify the contribution of lower rater standards to the increase. Using the first quarter sample, the point estimate is that 50.8% of the increase in feedback scores is due to inflation, whereas the larger sample from the middle of the data implies 48.6% of the increase is due to inflation.

It is essential to note that to the extent written feedback is also subject to inflation (“good” work now garners a “great”), our method understates the extent of reputation inflation taking place on the platform, and so we view our approach as providing a lower bound estimate. The robustness tests we perform in Appendix B provide additional evidence that our approach provides a lower

Figure 5: Numerical public feedback and predicted score from textual feedback using the first quarter as the training period.



*Notes:* This figure plots the evolution of average public feedback scores (dashed line) versus the average predicted score of textual feedback (dotted line) assigned by employers to workers. A 95% interval is depicted for every point estimate. The shaded area indicates the quarters from which training data was obtained for the corresponding predictive model. The training sets consist of 1,492 samples (top panel) and 10,555 samples (bottom panel). Adjusted predicted scores (dashed line in the bottom panel) are calculated by subtracting the constant from the predicted scores that allows the left endpoints of the adjusted and actual score lines to coincide.

bound estimate.

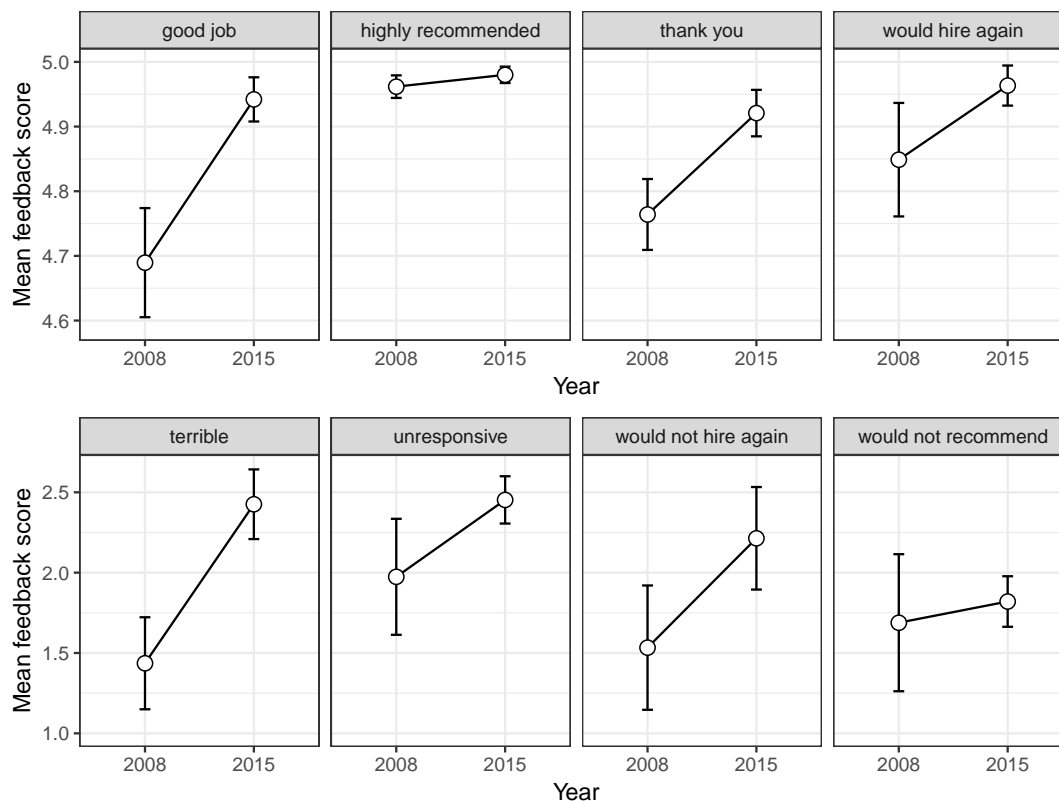
### 3.2.2 Average feedback scores of sentences over time

A potential shortcoming with the approach of Section 3.2.1 is that the lexical composition of reviews could presumably change over time. While we have no evidence that supports this hypothesis, in what follows we take an alternative approach: as a more direct measure of inflation, we examine whether the same sentences found in written feedback correspond to different feedback scores at different points in time. We select written feedback from 2008 and 2015. We scan through the written feedback generated in these periods for lexically similar sentences, and then compare average feedback by phrase, across the two periods. To find candidate sentences, we split reviews into sentences, detect named entities, such as names and locations, and replace them by tokens, and keep pairs of identical sentences used across the two years, and the corresponding assigned numerical rating. Our procedure allows us to extract 7,300 pairs of lexically identical sentences from the two periods.

To illustrate our approach, Figure 6 shows the average numerical feedback scores for a set of example short sentences that are commonly used on employer written feedback, by period. We selected sentences spanning both “good” and “bad” feedback, and which most frequently occurred in the corresponding written feedback in our data, such as “highly recommended,” “would hire again,” “terrible,” and “would not recommend.” We can see across terms that the numerical feedback scores associated with identical sentences have increased considerably over time, and this increase has affected both positive and negative sentences.

Using the whole collection of identified sentences, we find that the average difference in numerical feedback scores is 0.11 stars, with a 95% confidence interval of [0.096, 0.124]. As the increase in average numerical feedback scores between the 2008 and 2015 period is 0.35 stars, a back-of-the-envelope calculation suggests about 31.5% of the overall increase is due to inflation, which is close to our lower bound estimate from our predictive modeling approach. Note that we are excluding data from year 2007 when the majority of the overall increase occurred.

Figure 6: Difference over time in the average employer feedback score associated with a set of example sentences.



Notes: This figure shows the average numerical feedback associated with exact sentences found in the text of numerical reviews, in 2008 and 2015. A 95% confidence interval is shown for each mean.

## 4 A model of reputation dynamics

To help explain the implications of our empirical findings and understand why inflation occurs, we develop a model of employers leaving public feedback to workers in a competitive online labor market. Though our framing here is a labor market, the same framework can be applied to the more general case of buyers and sellers giving public feedback.

The assumptions of our model are motivated by the differences between public and private feedback scores revealed by the divergence we observed in Figure 4. Private feedback scores are not observable by other employers when they are

making hiring decisions, and hence receiving “bad” private feedback is less costly for workers. As a result, employers are more truthful in private, i.e., they assign lower private feedback scores, suggesting a link between the cost of receiving “bad” feedback and the cost of assigning “bad” feedback. Further, private feedback scores assigned to workers in the past are not observable by employers when they are assigning private feedback scores. As a result, employers can not condition their own private feedback rating on the ratings that were assigned by other employers.

## 4.1 Setup

Consider an online labor market composed of workers and employers. Workers are matched at random with employers, after which workers produce output  $y \in \{0, 1\}$ . The worker produces output  $y = 1$  with probability  $\Pr(y = 1|q) = q$ , from which the employer obtains utility equal to 1, by selling the output on some product market. The employer obtains zero utility in the case that output  $y = 0$  is produced.

Workers are characterized by their quality  $q \in \{q_L, q_H\}$ , with  $q_L < q_H$ . Employers know the fraction of high quality workers in the marketplace, which we denote by  $\theta$ . After the employer observes the worker’s realized output  $y$ , she generates a signal to the marketplace in the form of feedback  $s \in \{0, 1\}$ , where  $s = 1$  denotes “good” feedback, and  $s = 0$  “bad” feedback. In the next “round,” employers observe the most recent feedback assigned to the worker, and form Bayesian beliefs about the worker’s quality. We assume that both sides are price-takers, and hence workers are paid their expected marginal product, which is

$$w_s = \Pr(q = q_H|s)q_H + (1 - \Pr(q = q_H|s))q_L.$$

The worker’s cost of bad feedback, realized in the next round, is the difference in compensation between receiving good feedback,  $w_{s=1}$ , and bad feedback,  $w_{s=0}$ , that is

$$\Delta w = w_{s=1} - w_{s=0}.$$

Whenever the employer tells the truth, that is when  $s = y$ , she obtains a benefit



$b > 0$ .<sup>10</sup> If the worker’s output is good ( $y = 1$ ), then the employer has no incentive to lie and always assigns good feedback ( $s = 1$ ) to the worker. However, in the case that the worker produces no output ( $y = 0$ ) and the employer truthfully reports  $s = 0$ , the worker incurs a cost  $\Delta w$ , which is the wage penalty in the next round. We assume that some fraction of this cost is “reflected” back on the employer.<sup>11</sup> Employers differ in how much of this cost is reflected: let  $c_i$  be the employer-specific fraction of this cost that is reflected back on the rating employer. The employer thus incurs a cost of  $c_i \Delta w$ , where  $c_i$  is drawn from some distribution  $F : [\underline{c}, \bar{c}] \rightarrow [0, 1]$ , with  $\underline{c} \geq 0$ .

In light of these reflected costs, some employers might give positive feedback even if the worker’s output was bad, thereby avoiding the cost of giving bad feedback. This decision will depend on  $c_i$ , and so employer  $i$  will not report truthful feedback if

$$b \leq c_i \Delta w. \tag{1}$$

Let  $p$  denote the fraction of employers that generate truthful feedback in the most recent round, and assume that  $p$  is common knowledge. When considering a particular worker that received bad feedback in the previous round, i.e.,  $s = 0$ , the Bayesian employer infers that

$$\begin{aligned} \Pr(q = q_H | s = 0; p) &= \frac{\Pr(s = 0 | q = q_H; p) \Pr(q = q_H)}{\Pr(s = 0; p)} \\ &= \frac{(1 - q_H)\theta}{(1 - q_H)\theta + (1 - q_L)(1 - \theta)}. \end{aligned}$$

Note that the  $p$  term divides out as  $s = 0$  always implies truthful reporting. In contrast, if the worker received good feedback, i.e.,  $s = 1$ , the Bayesian employer

---

<sup>10</sup>This benefit includes idiosyncratic reasons to report truthfully as well as platform-specific benefits, such as awards by other users for being an accurate reviewer (e.g. user review ratings on Yelp). [Abeler et al. \(2016\)](#) find strong evidence about individuals’ preferences for truth-telling, both in 72 previous studies and in their experiments. Surprisingly, the propensity for truth-telling persists even in one-shot games.

<sup>11</sup>These costs include the cost of harming the worker’s future prospects, the cost of the worker complaining or withholding future cooperation, and even the cost from other workers being unwilling to work for the employer in the future if the employer has a reputation as a “strict” rater.

infers that

$$\begin{aligned}\Pr(q = q_H | s = 1; p) &= \frac{\Pr(s = 1 | q = q_H; p) \Pr(q = q_H)}{\Pr(s = 1; p)} \\ &= \frac{(q_H + (1 - q_H)(1 - p))\theta}{(q_H + (1 - q_H)(1 - p))\theta + (q_L + (1 - q_L)(1 - p))(1 - \theta)}.\end{aligned}$$

The cost of bad feedback to a worker is then

$$\Delta w(p) = w_{s=1;p} - w_{s=0;p} = \frac{\theta(1 - \theta)(q_H - q_L)^2}{k - pk^2}, \quad (2)$$

where  $k = \theta(1 - q_H) + (1 - \theta)(1 - q_L)$ , which is the probability that a randomly chosen worker will produce bad output.

We see from Equation 2 that  $\Delta w(p) > 0$  for all  $p$ , implying that as long as  $c_i > 0$ , there is always a cost to the employer of giving bad feedback, which they must compare to their benefit  $b$  from telling the truth. Further, when  $p$  is large, i.e., when most of the employers truthfully report, feedback is a more accurate measure of quality, and hence the value of positive feedback increases, along with the wage penalty  $\Delta w(p)$ . In contrast, when the majority of firms lie, the signal from good feedback is less informative, and the wage penalty narrows. We note that this makes which feedback is “good” and “bad” endogenous in our model: the market effect of a feedback score depends on what is considered “good” and “bad,” which in turn depends on the choices of all other employers.

We now consider what an equilibrium of this market would be. Let  $p_E$  denote the fraction of firms that truthfully assign negative feedback when the market equilibrium has been attained. The equilibrium fraction is found by solving the equation

$$p_E = F\left(\frac{b}{\Delta w(p_E)}\right), \quad (3)$$

to which a solution always exists for any continuous distribution function, and is unique for distributions with increasing hazard rate. Importantly, the two

extreme cases where

$$p_E = \begin{cases} 1, & \text{if } b \geq \bar{c}\Delta w(1) \\ 0, & \text{if } b \leq \underline{c}\Delta w(0) \end{cases}$$

correspond to an all-truthful and an all-lying equilibrium. If the benefit to assigning truthful feedback is higher than the cost for every employer, then no employer has incentive to lie ( $p_E = 1$ ), while if the costs are too high, all employers lie ( $p_E = 0$ ).<sup>12</sup> To the extent that we think of employers as both strategic and narrowly self-interested, the all-lying equilibrium is the likely equilibrium, as the benefit  $b$  is likely small or sometimes even zero, while the employer-specific costs  $c_i$  could be substantial.

## 4.2 Convergence and the evolution of average feedback

We now consider the marketplace's convergence to the equilibrium prediction. Consider a marketplace where every employer starts off truthfully reporting feedback, that is,  $p_0 = 1$ . To avoid cases where the convergence process is trivial, we also assume that the equilibrium truth-telling fraction is not the all-truthful equilibrium.

In every period, employers randomly match with workers, workers produce outputs, and employers subsequently report feedback. Among the employers, a fraction  $\theta_B = (1 - \theta)(1 - q_L) + \theta(1 - q_H)$  receives a bad output, i.e.,  $y = 0$ . These employers then compare their benefit from truth-telling with the cost of truthfully reporting bad feedback. Employers whose cost from truth-telling is lower than the benefit give bad feedback to the workers. Therefore, a fraction  $l_0 = \theta_B[1 - F(\frac{b}{\Delta w(p_0)})]$  begins to lie after the first period, and hence  $p_1 = p_0 - l_0$ .

We now examine the convergence of this process. Let  $T(x) = F(b/x)$  be the proportion of sellers that are better off truthfully reporting if the cost of bad feedback is  $x$ . From Equation 3 we obtain  $T(p_E) = p_E$ . Since  $F$  is a cumulative distribution function, and  $\Delta w$  is convex and decreasing in its argument,  $T$  is a decreasing but non-negative function. As a result,  $p_2 < p_1$ , but  $l_1 < l_0$ , and hence

---

<sup>12</sup>In the case where all employers have the same cost,  $p_E$  can be interpreted as the probability of truthfully generating public negative feedback in the resulting mixed strategy equilibrium.

$p_1 - p_2 < p_0 - p_1$ . Following the same argument, we can inductively show that the dynamics of the marketplace result in convergence to the equilibrium truth-telling fraction  $p_E$ , and that the rate of convergence decreases as the market approaches the equilibrium point. This is precisely the pattern we observed empirically in Figure 2: reputation initially inflates fast, but then flattens out as the equilibrium fraction is approached.

### 4.3 Model discussion and implications

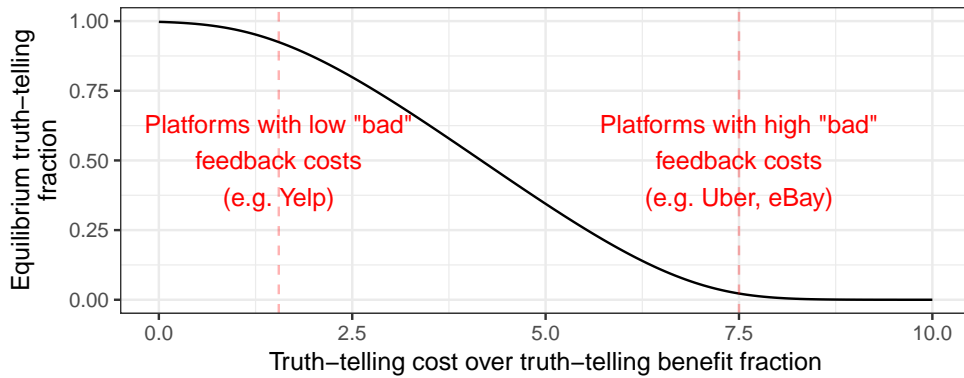
In online labor markets and other peer-to-peer marketplaces, it seems likely that both wage penalties for workers and employers’ reflected cost coefficients are high. Reasons for the high worker cost of bad feedback include that workers are often highly substitutable, but each worker has few transactions, and hence each rating is more consequential. Further, feedback scores are often the only signal of quality. On the employer side, as transactions are more personal, the reflected costs are likely higher. In contrast, when individuals assign feedback to products (e.g. movie reviews) there is likely no reflected cost, and our model predicts that there will be no inflation. Indeed, numerical scores on such platforms exhibit no inflation, but are rather characterized by lower averages, a much higher spread, and, in some cases, a decreasing temporal pattern (Li and Hitt, 2008; Cabral and Hortacsu, 2010; Moe and Schweidel, 2012; Godes and Silva, 2012; Hu et al., 2017).

However, the underlying importance of rater costs seems to hold even when reviews are much more impersonal. In a recent study, Proserpio and Zervas (Forthcoming) find that after hotels started replying to bad reviews on a travel review website, the number of users who leave bad feedback decreases, despite no change in hotel quality. It seems probable that raters might find the hotel’s response embarrassing, and/or it becomes clear that an actual person is “harmed” by the negative review.

To provide a graphical depiction of this intuition, we plot in Figure 7 the equilibrium truth-telling percentage for different truth-telling cost to truth-telling benefit ratios. To increase truth-telling costs, we increase the mean of the distribution of “reflected” cost coefficients, keeping everything else constant. When

the cost-to-benefit ratio is low, we see that most of the employers truthfully report their feedback in equilibrium. This is the case for platforms such as Yelp or movie rating websites, where the raters are giving feedback to businesses, and transactions are less personal. Furthermore, reviewers in these platforms likely view themselves as performing a service for fellow consumers, and being known for good, honest reviews is at least part of the incentive people have for participating. In the language of our model, these sites have a higher  $b$ . As the cost-to-benefit ratio increases, the equilibrium truth-telling percentage approaches zero, and the average feedback scores are hence inflated: this is the case for platforms such as Uber or eBay, where the “reflected” costs are high.

Figure 7: Truth-telling equilibrium fraction as a function of the ratio of the mean cost over the benefit of generating truthful feedback.



*Notes:* For this figure, the parameters used in computing the equilibrium truth-telling fractions are  $q_H = 0.8, q_L = 0.2, \theta = 0.5, b = 1, F \sim N(\mu_C, 1)$ . Alternative distribution and parameter choices of parameters yield qualitatively similar results.

The employers’ cost of assigning bad feedback is proportional to the workers’ cost from receiving bad feedback in our model. This assumption is made because workers are more likely to retaliate when the cost of negative public feedback is higher, thereby reflecting a higher cost back on the employer. Modern reputation systems attempt to decrease the probability and the cost of retaliation.

To increase truthful feedback, the platform could increase employers’ benefit from reporting it. For example, the platform could provide monetary incentives for users who generate feedback, as dissatisfied users often choose not to leave any feedback (Dellarocas and Wood, 2008; Nosko and Tadelis, 2015). Such types of

interventions generally have two problems. First, it is not clear how the platform can measure the truthfulness of feedback. Second, this solution is expensive, and hence not likely to be scalable (Fradkin et al., 2015). This implies that platforms are more likely to find success through decreasing reflected costs. We study one such intervention that decreased the costs of assigning bad feedback in the following section.

## 5 Private feedback revelation

Our model in Section 4 proposes a process by which reputation inflates in online marketplaces. A key feature of our model is Equation 1, which posits that an employer lies if  $b \leq c_i \Delta w$ . If the cost of bad feedback to the workers is zero, then employers should be truthful for any positive value of  $b$ , and thus would generate more “bad” feedback. If the cost of “bad” feedback to the workers changes, then the fraction of truthful employers should also change. Further, as the cost of “bad” feedback is endogenous, our model also predicts an over time convergence to new equilibria. As such, in the event that the costs of assigning “bad” feedback change, we do not expect to see a “jump” to the new equilibrium, but rather a convergence thereto.

To test the predictions of our model empirically, we would need a setting where the workers’ costs of bad feedback—and hence the employers’ reflected costs—are exogenously changed. Fortunately, the platform provided such a change.

### 5.1 Increasing workers’ cost of receiving “bad” private feedback

Our interpretation of private versus public feedback is that for “bad” public feedback, the cost to the worker,  $\Delta w$ , was positive, whereas for “bad” private feedback the same cost was zero. As a result, private feedback was more candid, i.e., more employers were more likely to report  $s = 0$  when  $y = 0$ , as the employers’ costs were increasing in the workers’ cost of negative feedback. We now consider what happened when the platform made a change that raised the cost  $\Delta w$  from zero to some positive amount, by changing how that private feedback was used on the

platform.

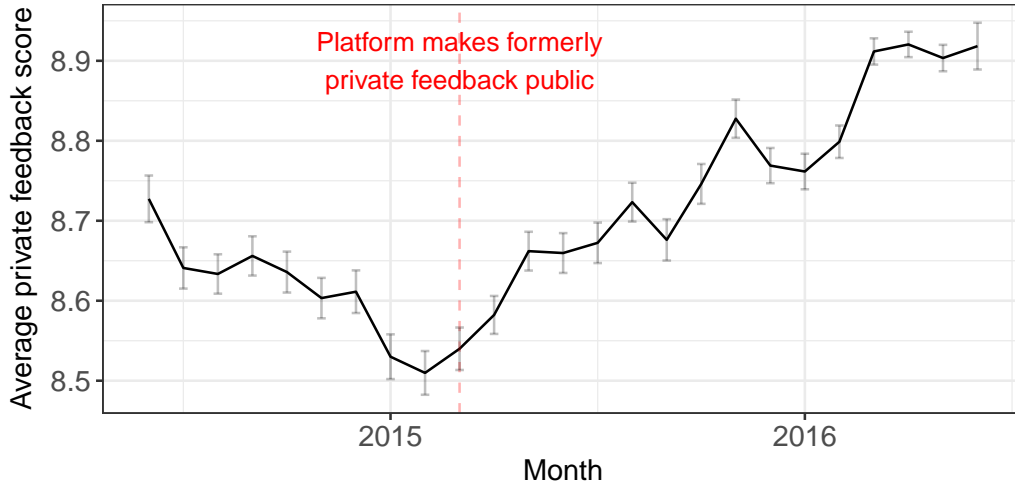
The change was the platform’s announcement in March 2015 that the private feedback ratings would be used to compute a new aggregate feedback score for workers. The aggregate score on a worker’s profile was only updated after the worker received five new feedback scores, to prevent workers from identifying which employer gave them which feedback. This score be shown on the profile of each worker and therefore be publicly available, but anonymous in the sense that one could not associate individual scores with employers.

To the extent that employers used this new score in their hiring decisions, the workers’ cost of private feedback went from zero to some positive number. In the logic of our model, the platform’s hope was that by not allowing workers to know which employer gave feedback, the distribution of  $c$  was pushed towards 0. However, if many employers do not want to hurt the worker or fear some other kind of generalized ex post retaliation, the private feedback measure could start to inflate after the policy change.

To see how the revelation of private feedback affected the private feedback scores, the time series of average scores by month are shown in Figure 9. Unlike public feedback, private feedback was initially relatively stable, exhibiting little fluctuation and no clear trend. The date of the platform’s private feedback policy change is indicated by the vertical red line. We can see that, immediately after the revelation, the formerly flat series started exhibiting an increasing trend in the average private feedback scores. In terms of our model, when the cost  $\Delta w$  went up, the cost of truth-telling went up and employers started lying. Importantly, there was no immediate jump to the new truth-telling equilibrium fraction, providing strong evidence in favor of our assumption about the endogenous nature of what constitutes “bad” feedback.

More formally, we can also place the results of Figure 9 in a difference-in-differences regression framework. Table 1 reports two regressions of the aggregate monthly private feedback on the interaction effect of the private feedback revelation on the aggregate private feedback. In Column (1) we employ a fixed effects regression, while in Column (2) we use a random effects regression, while controlling for month-specific effects in both. We observe that the private feedback revelation had a strong positive and highly statistically significant effect,

Figure 8: Monthly average “private” feedback given by employers to workers.



*Notes:* This figure shows the average monthly private feedback (on a 1 - 10 point scale) given by employers to workers. The left red dashed vertical line indicates the point in time when the platform switched from the private feedback question to eliciting private feedback on the 0-10 scale. The right red dashed vertical line indicates the point in time in which employer private feedback scores were aggregated and added to worker profiles. These aggregate scores were changed after the worker received five new feedback scores, to prevent workers from identifying which employer gave them which feedback. Prior to this point, scores were only collected by the platform and not used publicly in any way.

that is robust across the two specifications.

Of course, the increase in private feedback scores could be reflective of improvements in fundamentals. For example, the introduction of a less inflated and more informative measure of reputation may have led to better hiring decisions, or to workers exerting more effort. However, we can use the same “text as alternative measure of satisfaction” approach by using the written feedback given by employers. While the regressions provided in Table 1 partially control for such considerations, we can also employ the method developed in Section 3 to fit a model that predicts private feedback scores from public written feedback, using data obtained from employers that left both public written feedback and private feedback scores during a short time window. In Figure 10 we plot the average private feedback of contracts that also received textual feedback (solid line), versus the predicted private feedback from the corresponding textual feedback (dashed line). The predicted and actual private feedback scores closely match up until the

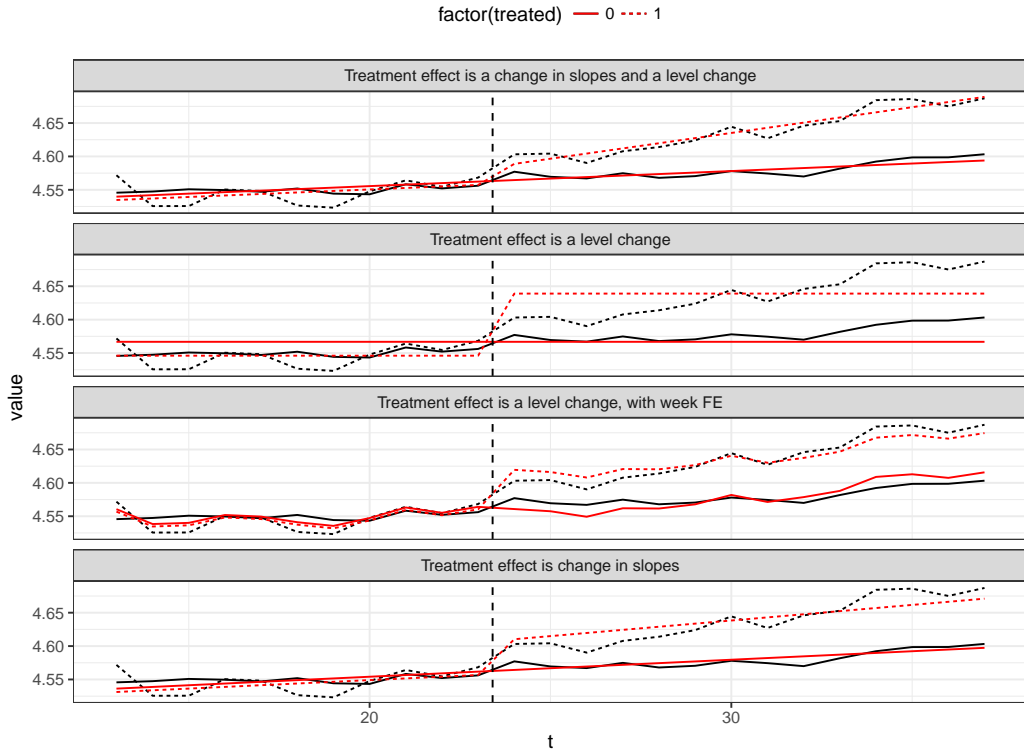


Table 1: Effects of “private feedback” public revelation on aggregate private feedback scores

	<i>Dependent variable:</i>	
	(1)	(2)
Post-Revelation $\times$ Aggregate Feedback	0.062*** (0.008)	0.070*** (0.007)
Constant		4.563*** (0.006)
Model Type	FE	RE
Month Effects	Y	Y
Observations	50	50
R <sup>2</sup>	0.945	
Adjusted R <sup>2</sup>	0.882	
Log Likelihood		111.863
Akaike Inf. Crit.		-213.725
Bayesian Inf. Crit.		-204.165

*Notes:* This table reports regressions where the outcome is the monthly aggregate feedback and the predictor is an indicator variable for the private feedback revelation. Both specifications control for month-specific effects, with Column (1) utilizing a fixed effects model, and Column (2) a random effects model. Significance indicators:  $p \leq 0.05$  : \*,  $p \leq 0.01$  : \*\*, and  $p \leq .001$  : \*\*\*.

Figure 9: Effects of public revelation on “private” feedback scores



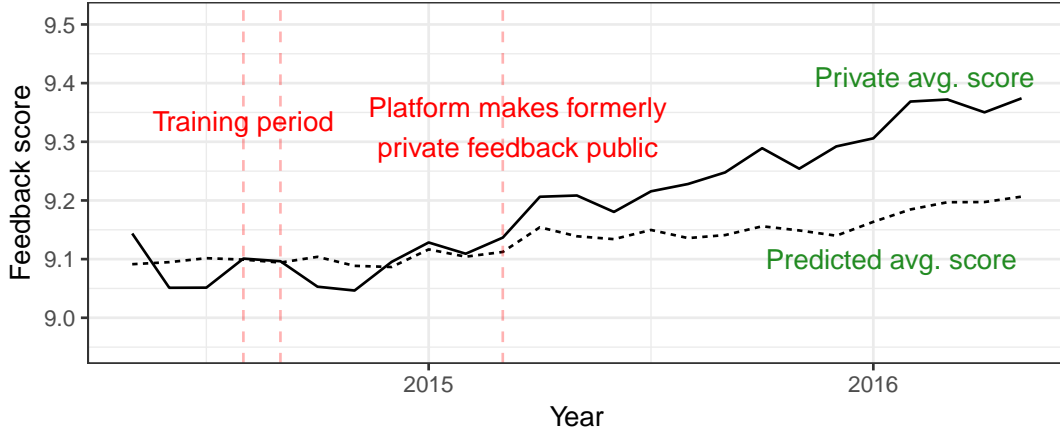
*Notes:*

time that the platform makes the private feedback information publicly available.

After the private feedback revelation change, both scores increase, but the actual private feedback scores clearly overstate that trend. After the private feedback revelation, the average private feedback score increased by 0.27 points, while the average predicted scores only increased by 0.10 points, indicating that at least 63% of the increase is due to inflation.

Increasing employer costs to leaving negative feedback results in inflation—which is our central claim. Further, this increase in scores does not happen instantly, leading to a one-off step change in feedback, but rather occurred over time, consistent with our modeling of the inflation process.

Figure 10: Private feedback score and predicted private score from textual feedback using the eight and ninth quarters of 2014 as the training period.



*Notes:* This figure plots the evolution of average private feedback scores (solid line) versus the average private score predicted from textual feedback (dashed line) assigned by employers to workers. The left solid red lines indicate the period from which training data was obtained for the predictive model. The training set consists of 61,316 samples. The right red dashed vertical line indicates the point in time in which employer private feedback scores were aggregated and added to worker profiles. We also note that the average private feedback score for contracts for which employers also left written feedback is higher (compare the y-axis scale to that in Figure 9); this indicates that when employers have had a bad experience, they avoid publicly stating their opinion via text.

## 6 Conclusion

This paper documents that the reputation system in an online marketplace was subject to inflation—we observe systematically higher scores over time, which cannot be fully explained by overall improvements in fundamentals. Data from three other marketplaces with similar characteristics exhibit the same trend, suggesting that the inflation problem is widespread. A theoretical model is developed which hypothesizes that the root of inflation is the costs that raters incur when leaving negative feedback. A market intervention where the costs to negative feedback were increased—namely by making previously private feedback public—yielded data consistent with the predictions of our model.

Though our findings have direct implications for online platforms that are affected by this inflation, reputation inflation is likely not an online phenomenon, but is seemingly also present in the non-digital world. For example, there is

widespread concern about grade inflation, and some schools have taken steps to counter it (Butcher et al., 2014). The debate found in this literature stream mirrors many of the issues we examine in this paper, namely whether the increase in grades is due to fundamentals, such as better student cohorts, or due to different standards, and whether information is lost. Babcock (2010) finds evidence that efficiency concerns are important, with inflated grades seemingly reducing student effort. Assessing the degree to which inflation is consequential for market actors is an interesting next step; Appendix C provides evidence the information the system conveyed decreased over time.

This paper illustrates a core market design problem, and elucidates its root cause. Whether there are effective platform design responses to this phenomenon is an open question; simultaneously-revealed ratings and anonymizing ratings through aggregation did not seem to prevent inflation from occurring in our data set. An important next step is to propose and evaluate alternative reputation mechanisms that may overcome reputation inflation. Introducing additional dimensions to a reputation system may slow down the eventual erosion of its informativeness (Chen et al., Forthcoming). Our model suggests some approaches, such as raising the benefit  $b$  to truthfully reporting feedback, or lowering the reflected cost coefficient  $c$ . One could interpret the mandatory grading curves often found in non-digital reputation systems as a policy that simultaneously affects both parameters.<sup>13</sup>

---

<sup>13</sup>Officer evaluation reports in the US Army limit senior raters to indicating only 50% or less of the officers they rate as “most qualified.”

## References

- Abeler, Johannes, Daniele Nosenzo, and Collin Raymond**, “Preferences for truth-telling,” 2016.
- Agrawal, Ajay, Nicola Lacetera, and Elizabeth Lyons**, “Does standardized information in online markets disproportionately benefit job applicants from less developed countries?,” *Journal of International Economics*, 2016, *103*, 1–12.
- Aperjis, Christina and Ramesh Johari**, “Optimal windows for aggregating ratings in electronic marketplaces,” *Management Science*, 2010, *56* (5), 864–880.
- Athey, Susan, Juan Camilo Castillo, and Dan Knoepfle**, “Service Quality in the Gig Economy,” 2018.
- Babcock, Philip**, “Real costs of nominal grade inflation? New evidence from student course evaluations,” *Economic Inquiry*, 2010, *48* (4), 983–996.
- Bolton, Gary, Ben Greiner, and Axel Ockenfels**, “Engineering trust: Reciprocity in the production of reputation information,” *Management Science*, 2013, *59* (2), 265–285.
- Butcher, Kristin F, Patrick J McEwan, and Akila Weerapana**, “The effects of an anti-grade-inflation policy at Wellesley College,” *The Journal of Economic Perspectives*, 2014, *28* (3), 189–204.
- Cabral, Luis and Ali Hortacsu**, “The dynamics of seller reputation: Evidence from eBay,” *The Journal of Industrial Economics*, 2010, *58* (1), 54–78.
- Chan, Jason and Jing Wang**, “Hiring preferences in online labor markets: Evidence of a female hiring bias,” *Management Science*, Forthcoming.
- Chen, Daniel L and John J Horton**, “Are online labor markets spot markets for tasks? A field experiment on the behavioral response to wage cuts,” *Information Systems Research*, 2016, *27* (2), 403–423.

- Chen, Pei-Yu, Yili Hong, and Ying Liu**, “The value of multi-dimensional rating systems: Evidence from a natural experiment and randomized experiments,” *Management Science*, Forthcoming.
- Dellarocas, Chrysanthos**, “The digitization of word of mouth: Promise and challenges of online feedback mechanisms,” *Management Science*, 2003, *49* (10), 1407–1424.
- , “Reputation mechanism design in online trading environments with pure moral hazard,” *Information Systems Research*, 2005, *16* (2), 209–230.
- **and Charles A Wood**, “The sound of silence in online feedback: Estimating trading risks in the presence of reporting bias,” *Management Science*, 2008, *54* (3), 460–476.
- Dimoka, Angelika, Yili Hong, and Paul A Pavlou**, “On product uncertainty in online markets: Theory and evidence,” *MIS Quarterly*, 2012, *36* (2), 395–426.
- Farrell, Diana and Fiona Greig**, “The online platform economy: Has growth peaked?,” 2016.
- Filippas, Apostolos and John J Horton**, “The Tragedy of your Upstairs Neighbors: When is the Home-Sharing Externality Internalized?,” 2017.
- Fradkin, Andrey, Elena Grewal, Dave Holtz, and Matthew Pearson**, “Bias and reciprocity in online reviews: Evidence from field experiments on Airbnb,” in “Proceedings of the Sixteenth ACM Conference on Economics and Computation” ACM 2015, pp. 641–641.
- Gelman, Andrew, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin**, *Bayesian data analysis*, Vol. 2, CRC press Boca Raton, FL, 2014.
- Ghose, Anindya, Panagiotis G Ipeirotis, and Beibei Li**, “Examining the impact of ranking on consumer behavior and search engine revenue,” *Management Science*, 2014, *60* (7), 1632–1654.

- Godes, David and José C Silva**, “Sequential and temporal dynamics of online opinion,” *Marketing Science*, 2012, *31* (3), 448–473.
- Hall, Jonathan V and Alan B Krueger**, “An analysis of the labor market for Ubers driver-partners in the United States,” *ILR Review*, Forthcoming.
- Horton, John J**, “Online labor markets,” *Internet and network economics*, 2010, pp. 515–522.
- , “The effects of algorithmic labor market recommendations: evidence from a field experiment,” *Journal of Labor Economics*, 2017, *35* (2), 345–385.
- , “Buyer uncertainty about seller capacity; Causes, Consequences and a Partial Solution,” *Management Science*, Forthcoming.
- **and Ramesh Johari**, “At what quality and what price?: Eliciting buyer preferences as a market design problem,” in “Proceedings of the Sixteenth ACM Conference on Economics and Computation” ACM 2015, pp. 507–507.
- **and Richard J Zeckhauser**, “Owning, using and renting: Some simple economics of the “Sharing Economy”,” *Working paper*, 2017.
- , **David G Rand**, **and Richard J Zeckhauser**, “The online laboratory: Conducting experiments in a real labor market,” *Experimental Economics*, 2011, *14* (3), 399–425.
- Hu, Nan, Jie Zhang, and Paul A Pavlou**, “Overcoming the J-shaped distribution of product reviews,” *Communications of the ACM*, 2009, *52* (10), 144–147.
- , **Paul A Pavlou**, **and Jie Zhang**, “On self-selection biases in online product reviews,” *MIS Quarterly*, 2017, *41* (2).
- Jin, Ginger Zhe and Phillip Leslie**, “The effect of information on product quality: Evidence from restaurant hygiene grade cards,” *The Quarterly Journal of Economics*, 2003, *118* (2), 409–451.

- Katz, Lawrence F and Alan B Krueger**, “The rise and nature of alternative work arrangements in the United States, 1995-2015,” 2016.
- Kokkodis, Marios and Panagiotis G Ipeirotis**, “Reputation transferability in online labor markets,” *Management Science*, 2015, *62* (6), 1687–1706.
- Levin, Jonathan D**, “The Economics of Internet Markets,” Technical Report, National Bureau of Economic Research 2011.
- Li, Xinxin and Lorin M Hitt**, “Self-selection and information role of online product reviews,” *Information Systems Research*, 2008, *19* (4), 456–474.
- Lin, Mingfeng, Yong Liu, and Siva Viswanathan**, “Effectiveness of reputation in contracting for customized production: Evidence from online labor markets,” *Management Science*, Forthcoming.
- Liu, Qingmin**, “Information acquisition and reputation dynamics,” *The Review of Economic Studies*, 2011, *78* (4), 1400–1425.
- Luca, Michael**, “Reviews, reputation, and revenue: The case of Yelp.com,” *Working Paper*, 2016.
- **and Georgios Zervas**, “Fake it till you make it: Reputation, competition, and Yelp review fraud,” *Management Science*, 2016, *62* (12), 3412–3427.
- Mayzlin, Dina, Yaniv Dover, and Judith Chevalier**, “Promotional reviews: An empirical investigation of online review manipulation,” *The American Economic Review*, 2014, *104* (8), 2421–55.
- Moe, Wendy W and David A Schweidel**, “Online product opinions: Incidence, evaluation, and evolution,” *Marketing Science*, 2012, *31* (3), 372–386.
- Moreno, Antonio and Christian Terwiesch**, “Doing business with strangers: Reputation in online service marketplaces,” *Information Systems Research*, 2014, *25* (4), 865–886.
- Nosko, Chris and Steven Tadelis**, “The limits of reputation in platform markets: An empirical analysis and field experiment,” 2015.

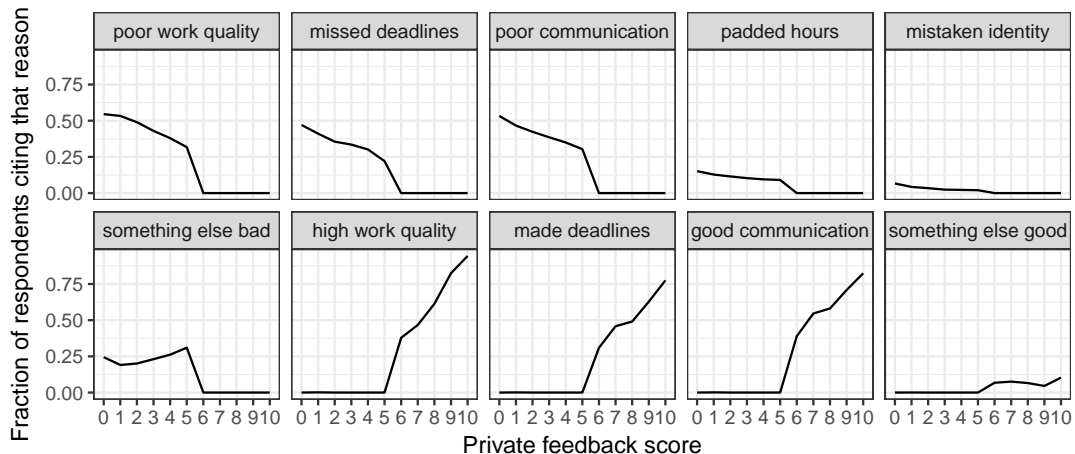


- Pallais, Amanda**, “Inefficient Hiring in Entry-level Labor Markets,” *The American Economic Review*, March 2013, 104 (11).
- Proserpio, Davide and Georgios Zervas**, “Online reputation management: Estimating the impact of management responses on consumer reviews,” *Marketing Science*, Forthcoming.
- Resnick, Paul, Ko Kuwabara, Richard Zeckhauser, and Eric Friedman**, “Reputation systems,” *Communications of the ACM*, 2000, 43 (12), 45–48.
- , **Richard Zeckhauser, John Swanson, and Kate Lockwood**, “The value of reputation on eBay: A controlled experiment,” *Experimental Economics*, 2006, 9 (2), 79–101.
- Stanton, Christopher T and Catherine Thomas**, “Landing the first job: The value of intermediaries in online hiring,” *The Review of Economic Studies*, 2015, 83 (2), 810–854.
- Sundararajan, Arun**, “From Zipcar to the sharing economy,” *Harvard Business Review*, 2013, 1.
- Zheng, Alvin, Yili Hong, and Paul A Pavlou**, “Matching in two-sided platforms for IT services: Evidence from online labor markets,” *Working paper*, 2016.

## A Robustness checks on using private feedback as an alternative measure of user satisfaction

One concern with any new feedback feature is that raters might simply not understand the new ratings. However, we have evidence that employers, at least collectively, understood quite well what the scale meant. When asked for private feedback, the platform also displayed a set of reasons that the employer could optionally select to indicate the reason for their score. Positive reasons were shown when the assigned feedback was above 5, while negative reasons were shown otherwise (during the 0 to 10 scale period). We use this “reason” information to verify that employers did not misrepresent the private feedback question. The fractions of private feedback reports citing these different reasons against the assigned private feedback score (0 to 10 scale) are plotted in Figure 11. We can see that there is a clear trend in the “correct” direction for both scores, indicating that private feedback scores were correctly assigned, at least on average.

Figure 11: Fraction of users citing a given reason when giving private feedback, by score.

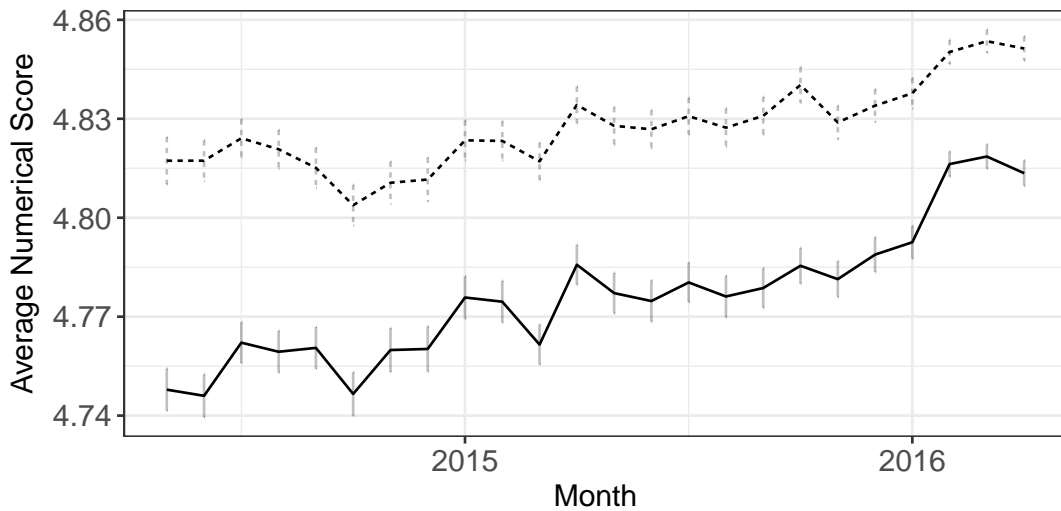


*Notes:* This figure plots the fraction of feedback reports that cited each reason as the basis of the feedback being positive or negative, against the private feedback score given. Across every case, we notice that employers that assigned more extreme feedback scores tend to cite reasons of the same sentiment more frequently.

Another plausible concern is that employers could be self-selecting into when

they will leave private feedback, and that changes in private feedback scores reflect changes in the selection process. Figure 12 plots the evolution of numerical public feedback for all contracts (solid line), and contracts for which private feedback was also assigned (dashed line). We observe that contracts in which private feedback is also assigned receive higher average public ratings, implying that it is the more satisfied employers that also assign public feedback. However, the degree to which this bias occurs does not change throughout our data set.

Figure 12: Average public numerical scores for all contracts, and for contracts to which private feedback was assigned.



*Notes:* This figure plots the monthly average feedback scores for all contracts (solid line), and the monthly average feedback scores for contracts for which private feedback was also assigned. A 95% interval is depicted for every observation. Scores are assigned upon the completion of each transaction, and the scale for numerical feedback is 1 to 5 stars.

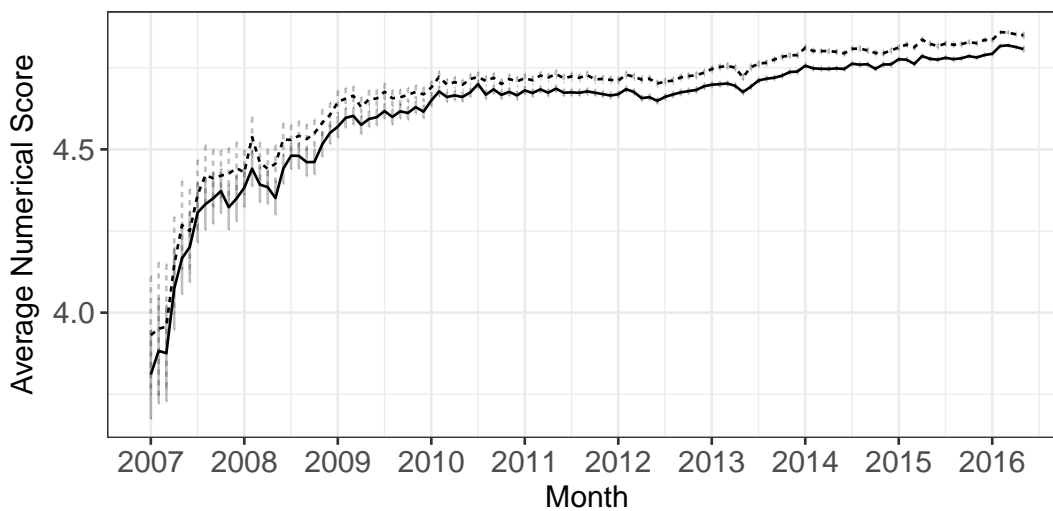
## B Robustness checks on using written feedback as an alternative measure of user satisfaction

A concern about the use of written feedback as an alternative measure of rater satisfaction is that employers' assignment behavior changes over time. In what follows we conduct robustness tests to identify potential sources of bias for our

analysis.

As with private feedback, a plausible concern is that employers may be more or less satisfied when deciding to assign written feedback in addition to numerical feedback. Figure 13 plots the evolution of numerical feedback for all contracts (solid line), and all contracts for which written feedback was also assigned (dashed line). We observe that contracts in which written feedback is also assigned receive higher ratings, implying that more satisfied employers assign written feedback. However, the degree to which this bias occurs does not change throughout our data set. Further, since written feedback is positively biased, comparing the predicted scores from text versus the evolution of all scores gives us a lower bound for the degree of inflation.

Figure 13: Monthly average numerical scores, and monthly average numerical scores when written feedback was assigned.



*Notes:* This figure plots the monthly average feedback scores for all contracts (solid line), and the monthly average feedback scores for contracts to which written feedback was also assigned. Scores are assigned upon the completion of each transaction, and the scale for feedback is 1 to 5 stars. For each observation, average scores are computed for every time period, and a 95% interval is depicted for every point estimate.

## C Examining the informational implications of reputation inflation

The impact of reputation inflation could be minimal if market participants “know” about the rate of inflation and adjust accordingly; even if individuals are not well-informed, the platform could implement statistical adjustments in its design of the reputation system to uncover the “true” (non-inflated) scores. However, if the pooling in the highest feedback “bin” becomes acute, statistical corrections cannot recover the lost information. This is partially due to the fact that, by design, numerical scale systems are prone to top-censoring; for the question “rate on a scale from 1 to X,” the value of X must be pre-specified.<sup>14</sup> Changes in the reputation system, such as adding a higher ceiling in the feedback scores or additional dimensions of reputation, may temporarily mitigate—but do not solve—the problem.<sup>15</sup>

To see the problem created by top-censoring, consider the information conveyed by observing a binary variable  $X$ , as it is captured by the information-theoretic entropy  $H(X) = p \log(p) + (1 - p) \log(1 - p)$ , where  $p$  is the probability of one outcome. As  $p$  goes to either 1 or 0, the information conveyed by the variable—in our case, the observed feedback score—goes to zero. However, this binary characterization of the reputation system is a simplification that could elide an important way in which rising—and even more compressed scores—could convey just as much (or even more) information. Consider increasing all nominal scores by some fixed amount and then “shrinking” all scores toward some new higher mean. This transformation would have no informational implications. To assess informativeness, we need to take an empirical approach.

To assess the informativeness of the feedback scores about worker quality over time, we conduct a variance decomposition, showing how the fraction of unexplained variance in feedback scores changes over time. Suppose that the

---

<sup>14</sup>This is why reputation inflation differs from monetary inflation; a sandwich that used to cost \$0.50 and may now cost \$12. However, this could not happen if price was mechanically restricted to be below \$1.

<sup>15</sup>See also <https://www.youtube.com/watch?v=KOO5S4vxi0o>.

data generating process of a worker’s feedback is

$$\text{SCORE}_{it} = a_{it} + c_t + \epsilon_{it}, \tag{A1}$$

where  $a_{it}$  is the worker’s true quality,  $c_t$  is a baseline time effect, and  $\epsilon_{it}$  is some noise term such that  $E[\epsilon_{it}] = 0$ .<sup>16</sup> If, over time, more of the variation in feedback scores can be explained by the variation in the noise term rather than by variation in the quality of individuals, then a feedback score is becoming less informative of the worker’s true quality.

Consider a Bayesian employer trying to infer the quality of a worker from a score: the more the feedback score is attributable to noise, the lesser its impact on the employer’s posterior belief of worker’s quality after observing this score. To wit, let  $\Pr(a) \sim N(a_0, \sigma_0^2)$  be the employer’s prior distribution for worker quality, and let  $\epsilon \sim N(0, \sigma^2)$  be the noise term with known variance  $\sigma^2$ , and  $a$  be the worker’s true quality, which the employer forms a posterior about after observing a feedback score. After observing the worker’s feedback score SCORE, the employer’s posterior is

$$\Pr(a|\text{SCORE}) = N\left(\frac{\frac{1}{\sigma^2}\text{SCORE} + \frac{1}{\sigma_0^2}a_0}{\frac{1}{\sigma^2} + \frac{1}{\sigma_0^2}}, \frac{1}{\frac{1}{\sigma^2} + \frac{1}{\sigma_0^2}}\right).$$

From the above equation, as  $\sigma^2 \rightarrow \infty$ ,  $\Pr(a|\text{SCORE}) \rightarrow \Pr(a)$ , or in words, as the noise component of the score explains more of the variance, the observed feedback becomes less informative, and at the limit, has no effect on the employer’s beliefs.<sup>17</sup>

To explore the informativeness of feedback scores empirically, we make two assumptions. First, for a suitably small window of time (i.e., a quarter), we assume that the baseline time effect,  $c_t$ , is fixed. Second, we assume that the population distribution of  $a_{it}$  can have a changing mean, reflecting shifts in worker quality, but its variance is constant; workers could be getting systematically better or

---

<sup>16</sup>For simplicity, we are treating the feedback score as continuous. The logic is identical in the dichotomous case.

<sup>17</sup>Gelman et al. (2014) provide a derivation of this result, which can also be found in most standard Bayesian analysis textbooks.

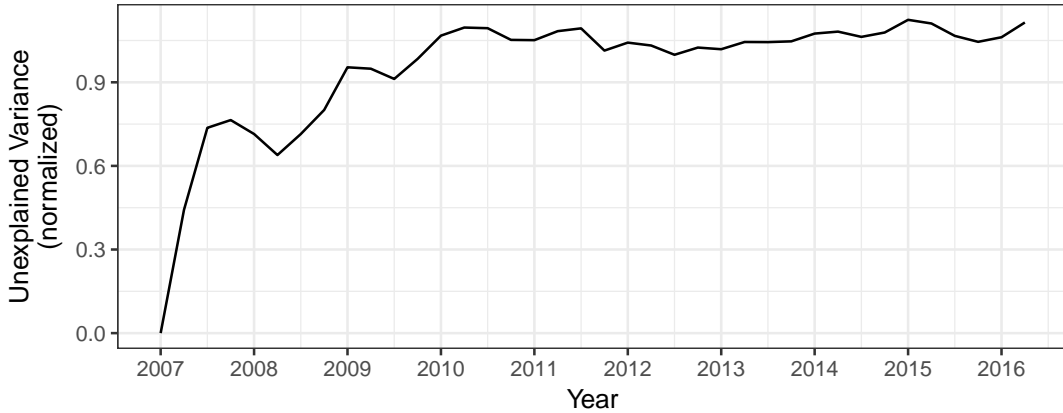
worse, but their abilities are not getting more or less spread out.

The fraction of variance due to noise is the quantity

$$\frac{\text{Var}(\epsilon)}{\text{Var}(\text{SCORE})} = 1 - \frac{\text{Var}(a)}{\text{Var}(\text{SCORE})}. \quad (\text{A2})$$

If this ratio increases over time, feedback scores are becoming less informative. We can easily compute this fraction for a time period  $t$  by performing the regression implied by Equation A1—the quantity of Equation A2 is  $1 - R_t^2$ , where  $R_t^2$  is the coefficient of determination from the period  $t$  regression.

Figure 14: Feedback score variance not explained by worker quality over time. Scores are reported as percentage differences with respect to the minimum unexplained variance.



*Notes:* Unexplained variance is reported the percentage difference with respect to the minimum unexplained variance of the time series, which is attained at the first period of this figure. The data of each quarter consists of workers with at least 2 jobs in that quarter, as otherwise the fixed effect  $a_{it}$  would perfectly predict their feedback score. Utilizing different cutoffs does not quantitatively change our results.

We fit the regression described in Equation A1 on the feedback scores generated in every quarter of our data separately. On each of these regressions, we are using fixed worker effects to estimate  $a_{it}$ , thereby allowing worker quality to evolve in time, even “within” a worker. Figure 14 plots the percentage difference of  $1 - R_t^2$  from the minimum unexplained variance, which is found at the first period in our data. The increase in unexplained variance from 2007 to 2016 is about 118% (from 0.32 to 0.70). This strong positive trend in the explained variance

implies that the relative importance of noise in explaining feedback grows over time, which in turn implies that the informativeness of feedback about worker quality has deteriorated.