Reputation Inflation*

Apostolos Filippas† John J. Horton‡ Joseph M. Golden§

June 4, 2019

Abstract

A solution to marketplace information asymmetries is to have trading partners publicly rate each other post-transaction. Many have shown that these ratings are effective; we show that their effectiveness deteriorates over time. The problem is that ratings are prone to inflation, with raters feeling pressure to leave "above average" ratings, which in turn pushes the average higher. This pressure stems from raters' desire to not harm the rated seller. As the potential to harm is what makes ratings effective, reputation systems, as currently designed, sow the seeds of their own irrelevance.

*Author contact information and code are currently or will be available at http://www.john-joseph-horton.com. Thanks to Richard Zeckhauser, Foster Provost, Andrey Fradkin, David Holtz, Ramesh Johari, Nico Lacetra, Xiao Ma, and Aaron Sojourner for very helpful comments and suggestions.

†Fordham University, Gabelli School of Business

‡MIT Sloan School of Management

§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).[1] 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.[2] 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 "should" look like at a moment in time. However, if we look at how the distribution of scores changes over time, we can potentially learn more about what the reputation system is measuring. If feedback scores are rising, there are at least two distinct—but not mutually exclusive—reasons: (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.[3] This "reputation inflation" erodes the

[1]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 among 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.

[2]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).

[3]This kind of inflation is similar to the conjecture about the increase in college grades—namely that stu-

comparability of feedback scores over time and reduces the informativeness of reputation system—potentially making it completely uninformative.

In this paper, we examine the reputation system of a large online labor market, focusing on the evolution of average feedback scores over time and the causes for the dynamics we observe. Mirroring findings from other marketplaces, we find that the distribution of recent employer feedback for workers is highly top-censored, with an overwhelming majority receiving perfect feedback.[4] However, the distribution has not always been this skewed—the fraction of workers receiving the highest possible rating of 5 stars went from 33% to 85% in just 6 years. Feedback scores in four other online marketplaces for which we could obtain longitudinal data exhibit a similar increase over time.

To disentangle whether the increase is caused by raters becoming more satisfied or by raters lowering standards, we use longitudinal data that include both the feedback scores and an alternative measure of rater satisfaction. The idea is that if rater satisfaction is also captured by an alternative channel that does not inflate—or inflates at a lower rate—then we can exploit this difference to produce an estimate of inflation in the measure of interest.

As a first alternative measure of rater satisfaction, we use information obtained by the introduction of a parallel and experimental 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 platform's 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, and/or because a bad report would not harm the worker. We find that average private feedback scores were decreasing at the same time that average public feedback scores for the *same* transactions were increasing. This difference is evidence that raters were lowering their standards for public feedback rather than becoming more satisfied.

The private feedback feature provides us with evidence that inflation occurs, but it does not span the entire life time of the platform's operations—we observe both ratings for 10 months. As a second alternative measure of rater satisfaction, we use the sentiment raters express in the written feedback that accompanies numerical scores. 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, allowing us to learn the relationship that prevailed between written sentiment and numerical score when

dents and work is not getting better but rather the same quality of work now earns a higher grade (Babcock, 2010; Butcher et al., 2014).

[4] 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.

the training feedback was created. Using the predictions, we can then decompose the growth in average feedback scores into a component due to improvements in market "fundamentals" (e.g., improved marketplace features, better cohorts of workers, lower prices, less picky employers, and so on) that increased rater satisfaction and is reflected in feedback text, and the residual component that cannot be explained by improvements in fundamentals, and is hence due to inflation.

We find that although predicted feedback scores based on written feedback have increased over time, presumably due to improvements in fundamentals, they have not increased nearly as much as the actual, numerical feedback scores. Our estimates are robust across different specifications and training sets, and suggest that more than 50% 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 scores.

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 association between the same sentences and rater satisfaction 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. We will argue that the key to understanding reputation inflation is appreciating the role of costs, and how these costs affect raters. A starting point is the divergence between public and private feedback scores: 28.4% of those employers that *privately* report that they would definitely not hire the same worker in the future, *publicly* assign them 4 or more stars out of 5. The reverse essentially never happens—raters giving good private feedback and bad public feedback. The likely reason is that bad public feedback is costly in a way that bad private feedback is not. In surveys conducted by the platform, some employers report they fear retaliation, while others claim to not want to harm the rated individual, as public feedback is consequential. Although rated workers on this platform cannot retaliate by giving the employer bad feedback because it is simultaneously revealed (Bolton et al., 2013), 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 requires the cost of leaving a given "bad" score to increase over time. We hypothesize this is precisely what happens, with the same nominal feedback score (e.g., 2 "stars") becoming 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, which in turn determines what inferences future buyers 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 unique, 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.[5] While our observed data is consistent with our model, we report evidence from a platform intervention that allows us to test the predictions of our model directly.

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 employer gave which feedback, unless every rater in the batch gave the lowest or highest possible rating. However, with this private feedback now being publicly reported, the private scores became consequential to workers, who now had incentives to try to encourage good private feedback from employers. 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, whereas before it had none.

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. Of course, introducing a new, potentially more informative feedback measure may have caused improvements in fundamentals, and hence we still need to disentangle the effect of improvements in fundamentals and inflation on the observed changes.

We find that when the platform suddenly made private feedback scores public, private feedback scores began increasing immediately but there was limited change in the sentiment of written feedback—the no-longer-private feedback became inflated, mirroring what we ob-

[5] 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; ?).

served with public feedback. 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 that numerical feedback. Further, the results implicate 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. Projecting forwards, if the current rate of inflation persists, the average feedback would be the highest possible score on the private feedback scale in about 7 years after the switch to publicly revealing feedback.

Our paper makes several contributions. Our key contribution is documenting the extent of reputation inflation in a large online marketplace and identifying the likely cause of that inflation. Our long-run, whole-system perspective is possible because we use data spanning over a decade of the operations of the marketplace. We suspect the reputation inflation problem is widespread, given that many online marketplaces share the same features as the one we study in depth, and nearly all have those features that we show lead to inflation. Our collection of longitudinal data from other platforms supports this view—in every marketplace for which we could obtain data, we observe average feedback scores increasing, even though each of these marketplaces do not allow "tit-for-tat" rating behavior (Bolton et al., 2013).

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 in the extreme, put the reputation system on an inexorable path towards uninformativeness. Our analysis of the effects of public revelation of private feedback shows the key role of costs to raters, but also shows that this cost is not so easily diminished. The quasi-experimental revelation of private feedback demonstrates that simply making feedback anonymous is not sufficient to counteract inflation. In our case, as long as the feedback was consequential to workers, and employers know this, they find it costly to give "bad" feedback, even if the rated individual cannot retaliate. This suggests an inherent tension between ratings being consequential and ratings being informative. Whether reputation systems can be designed that are less prone to inflation remains an open research question.

Our paper also makes a methodological contribution, showing how to quantify reputation inflation in any marketplace where there are multiple measures of rater satisfaction. This methodology may also complement recent approaches to measuring monetary inflation using online data (Cavallo and Rigobon, 2016), particularly when prices are paired with product reviews—disentangling fundamentals and inflation in our context is conceptually similar to addressing the quality bias in monetary inflation measures (Council et al., 2002; Diewert, 1998; Cavallo et al., 2018).

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 Empirical context

The primary 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; Barach and 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 the 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 online marketplace literature has been reputation systems. Cabral and Hortaçsu (2010) also find that eBay sellers condition their behavior on their current reputations. Moreno and Terwiesch (2014) show how employers use reputation information—and subsequently how workers adjust their bidding strategies in light of this employer conditioning. The same authors 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, presumably due to the greater ease of monitoring for hourly contracts (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.

2.1 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.[6] 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 to interested parties for inspection. 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-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

[6]We use the present tense here to describe the reputation system before the introduction of private feedback.

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 seen 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, 81.8% of employers eligible to leave feedback have chosen to do so.

2.2 Feedback now and in the past

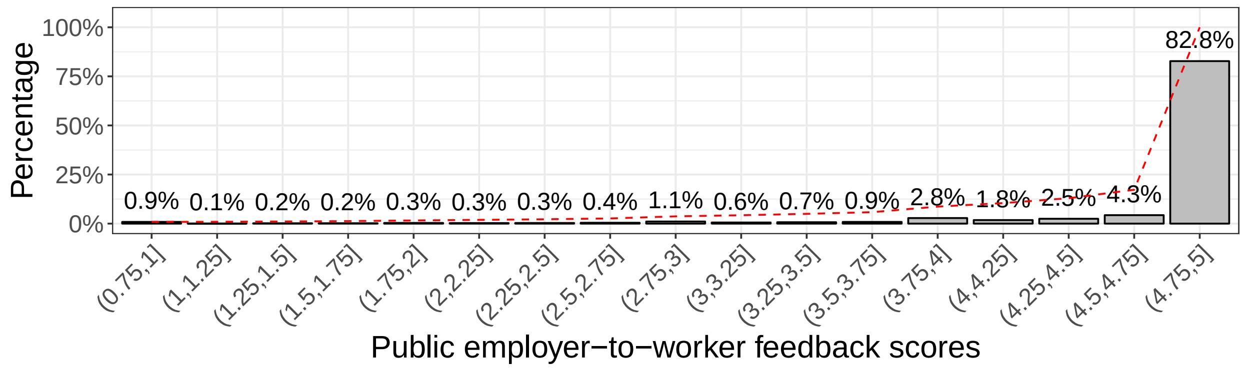
The distribution of employer-on-worker feedback scores in the market is highly right-skewed, but has not always been that way—scores have increased sharply over time. Most of the increase is explained by an increasing share of contracts receiving perfect feedback. These features of the data can be seen in the three panels of Figure 1.

Figure 1a depicts the histogram of public feedback scores for completed assignments that received a feedback score from the employer, from January 1, 2014 to May 11, 2016, for contracts worth more than $10.[7] 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.

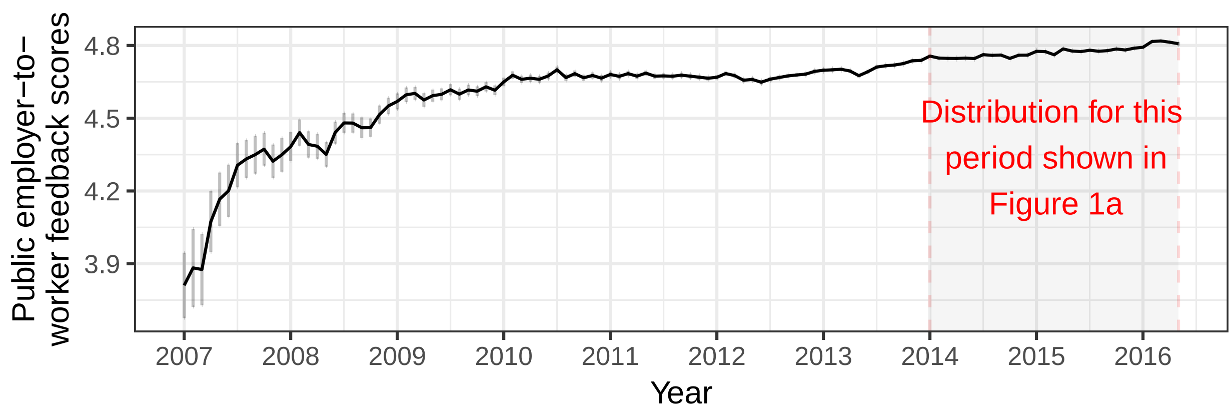
[7] We use this $10 restriction throughout the paper to remove mistaken, trial, and erroneous transactions.

Figure 1: Employer-on-worker feedback characteristics in an online marketplace

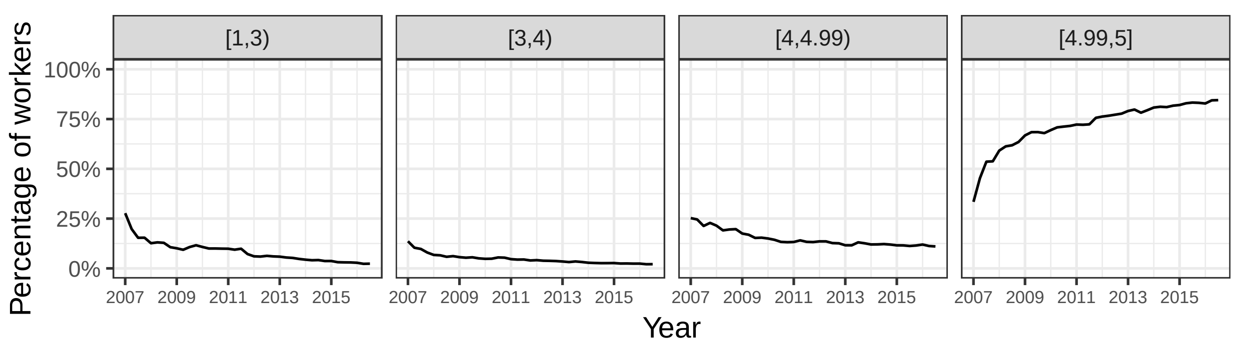
(a) Distribution of feedback scores for the period January 1, 2014 to May 11, 2016.



(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 contracts from January 1, 2014 to May 11, 2016, for which the employer provided feedback. See Section 2.2 for the description of the sample. 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, 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 4.85 in May 2016. The strongest period of increase was 2007, when average feedback scores increased by about 0.53 stars.

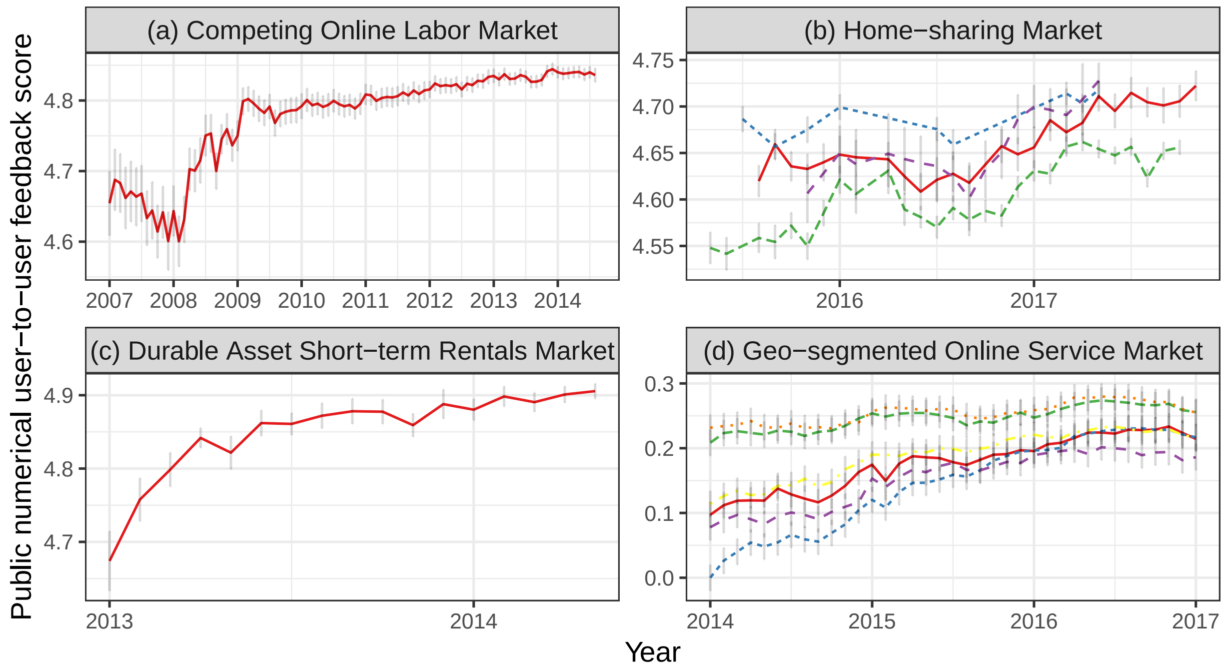
The increase in average feedback could be the outcome of raters giving less "bad" feedback, more "good" feedback, or some combination thereof. For example, we could see a decrease in the proportion of workers that receive 1-star ratings, and an increase in the percentage of workers receiving 3 stars. 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: 33% of contracts received a 5-star rating at the start of sample, compared to 85% at the end of the sample.

2.3 Evidence from other online marketplaces

Before exploring the causes for the pattern observed in Figure 1b, we turn to the question of whether the observed increase is something specific to this particular platform or a general feature of reputation systems. Although there is substantial evidence of right-skewed distributions of ratings at a moment in time (Nosko and Tadelis, 2015; Athey et al., 2018), we are unaware of other research showing that this right-skewness arises over time rather than being present at launch. For this reason, we obtained data from four other online marketplaces, two of which contain city-level data.

The average feedback scores for the various marketplaces are shown in Figure 2. Panel (a) shows longitudinal data in a competing online labor market. Panel (b) plots longitudinal ratings data from four major cities in the United States and Europe in a large home-sharing platform. Home-sharing platforms are peer-to-peer marketplaces that facilitate short-term rentals for lodging (Filippas and Horton, 2017). Panel (c) plots numerical feedback data from an online marketplace that facilitates the short-term rental of a durable asset (Sundararajan, 2013; ?). Panel (d) plots longitudinal ratings data from six major cities in the United States in a large online marketplace for services.

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.

The goods and services that are transacted in these marketplaces differ. In panel (a) ratings are assigned by employers to workers, in panel (b) by guests (those who are renting properties) to hosts (those who are renting out properties), in panel (c) by users (renters of the asset) to users (providers of the asset) after the transaction has taken place, and in panel (d) by consumers of the service to providers of the service. Further, these platforms greatly differ in the marketplace mechanisms they employ. For example, in the home-sharing marketplace renters choose the provider, but in the service marketplace the platform assigns a provider to the consumer. However, the reputation system mechanisms of interest are more or less identical: transactions are personal (peer-to-peer), ratings are given after the transaction has taken place and are consequential for the rated party, and the platforms all utilize mechanisms that prevent “tit-for-tat” rating behavior.

Despite the differences in what is being transacted, we observe an increase in ratings over time that mirrors the pattern that we found in our focal marketplace. This provides us with evidence that increasing feedback scores are likely common in online marketplaces with similar reputation system characteristics, even when other marketplace characteristics vary, irrespective of the transacted goods and services. We now turn our attention to the first online marketplace, where we have access to longitudinal transaction-level data, which enables a richer analysis.

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.

There are many reasons that rater satisfaction could change across time, even if standards remained fixed. The platform may have implemented better search and recommendation features that can lead to more effective matching of employers with sellers, or improvements in the way tasks are posted on the platform may have resulted in better transmission of information and setting of expectations. Further, workers may be getting better at performing the task, better at conducting online work, or even exerting more effort. Cohorts of workers and employers of systematically higher quality may be joining the platform over time, or fewer alternative options for the same task may be becoming available at a lower price. Other endogenous reasons include employers identifying and continuously transacting with a subset of desirable workers, thereby minimizing screening, transaction, and uncertainty costs. While all these reasons imply that transaction quality is improving, we can think of similar reasons that may have led to lower quality.

We cannot hope to account for all of the potential fundamental changes—and thereby detect the increase due to changes in standards.[8] Instead, we can side-step this issue by using an alternative measure of rater satisfaction. Suppose during some period of time, employers form matches and complete projects, getting some value at some price. Let a given employer i obtain utility u_i from some transaction. This utility is unobservable, and may be affected by both fundamentals-related and idiosyncratic factors, such as the worker's quality, platform features, match-specific characteristics, environmental factors, the employer's past experience, the existence of alternative options, or even the employer's mood. Such factors are also unobservable and may vary over time.

The employer leaves feedback $s_i = s(u_i) + \sigma_i$, where $s(\cdot)$ is common among employers, while σ_i is idiosyncratic. In words, the component σ_i captures differences in how employers translate their utility to the measure of satisfaction. The employer also leaves alternative

[8]That said, in Appendix A, we directly show that the most plausible selection and/or composition stories cannot explain the rise of average feedback scores.

feedback a_i. Similarly, we assume that this alternative measure of satisfaction has a common component and an idiosyncratic component, that is, $a_i = a(u_i) + \alpha_i$. Let U denote a collection of data points, such that we observe $s(u)$ and $a(u)$ for every $u \in U$. We assume that we can estimate a function $\hat{s}$ on U such that

$$\hat{s}(a_i(u_i)) = s(u_i) + \epsilon + \eta,$$

where we decompose the error introduced from estimating function $\hat{s}$ to an idiosyncratic error ϵ, and an employer-specific error η, with $\mathbb{E}[\epsilon] = 0$ and $\mathbb{E}[\eta] = 0$. As such, we get $\mathbb{E}[\hat{s}(a_i)|U] = \mathbb{E}[s|U]$. Intuitively, the function $\hat{s}$ maps the alternative measure, a, to the same scale as the primary measure, s.

Now suppose we have a new collection of observations U' at some later date, and we suspect that the primary satisfaction measure has changed to $s(u_i) + g(u_i)$, while the alternative measure, a, remains constant. We can then apply the function $\hat{s}$ on the new set of observations U', and compute the difference in expectation as

$$\begin{aligned} \Delta s &= \mathbb{E}[s_i - \hat{s}(a_i)|U'] \\ &= \mathbb{E}[g(u_i)|U'] - \mathbb{E}[\epsilon|U'] - \mathbb{E}[\eta|U']. \end{aligned} \tag{1}$$

The three terms Δs is decomposed to in Equation 1 are important. The term $\mathbb{E}[g(u_i)|U']$ is what we label the inflation component: the change in the primary feedback that cannot be explained by changes in fundamentals. The term $\mathbb{E}[\epsilon|U']$ is the component due to any differential selection of latent utilities for which $\hat{s}$ over—or under—predicts in expectation. The $\mathbb{E}[\eta|U']$ term is the component due to any differential selection of raters, e.g., raters who systematically differ in how they translate their utilities to the primary and alternative feedback measures.

It is critical to note that this method does not lead to bias simply because average transaction utility has gotten better or worse for all the potential reasons we identified in the beginning of Section 3. Essentially, this approach circumvents the problem of estimating the underlying utilities, and gives us a new, albeit more tractable problem of verifying that the estimated function $\hat{s}$ does not introduce systematic bias when applied on U', i.e., $\mathbb{E}[\epsilon|U'] = 0$ and $\mathbb{E}[\eta|U'] = 0$.

A second important observation worth noting is that if the alternative feedback measure inflates as well, then our method provides a lower bound for the inflation term. More specifically, if written feedback has also changed to $a(u_i) + h(u_i)$, then we get from Equation 1 that $\Delta s \leq \mathbb{E}[g(u_i)|U']$. Similarly, our method underestimates the degree of inflation if $\hat{s}$ introduces increasing systematic bias over time, that is, if the terms $\mathbb{E}[\epsilon|U']$ and $\mathbb{E}[\eta|U']$ are positive.[9]

For our analysis of the focal market, we will employ two alternative measures of employer satisfaction. One is a "private" feedback measure collected on completed contracts; the other is the written feedback left by raters, from which we will extract the sentiment. For reasons we will explain, we believe both measures are less prone to inflationary pressures. For both measures, we will find a substantial gap when we compute Equation 1, and we will present a variety of evidence that the gap is not due to selection with respect to either $\mathbb{E}[\epsilon|U']$ and $\mathbb{E}[\eta|U']$.

Our decomposition task is conceptually similar to estimating monetary inflation in the presence of quality changes (Sidrauski, 1967; Lucas Jr and Rapping, 1969; Friedman, 1977; Galı and Gertler, 1999; Mishkin, 2000; Berentsen et al., 2011). Although quality changes are acknowledged, they are typically sidestepped by a "basket-of-goods" approach (Diewert, 1998). The implicit assumption underlying such methods is that consumers derive the same satisfaction from some "basic" goods and services, irrespective of the time period.[10] In the context of online marketplaces, however, there is no "basic" or "standard" transaction, that is, a transaction with an immutable associated rater satisfaction.

3.1 Private feedback as an alternative measure

Our first alternative measure of rater satisfaction comes from a platform experiment that asked for an additional "private" feedback measure of feedback. The platform introduced this new experimental 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. 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 at the beginning of June 2014, the employers were instead asked to rate the worker on a numerical scale of 0 to 10. 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.

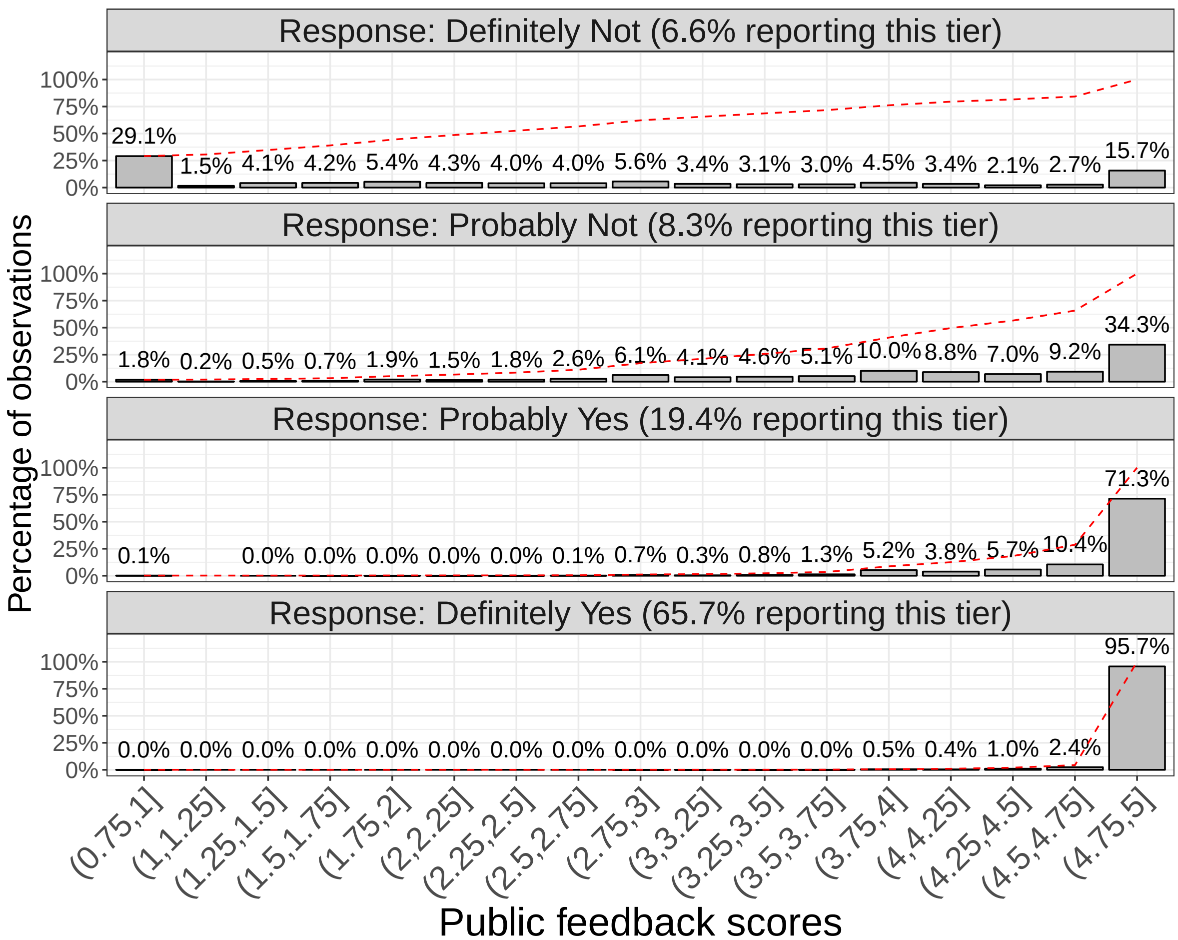
[9]The results of this section hold under the more general assumption $\mathbb{E}[\epsilon|U]+\mathbb{E}[\eta|U] = c$, for some arbitrary constant c, and $\mathbb{E}[\epsilon|U'] + \mathbb{E}[\eta|U'] \geq \mathbb{E}[\epsilon|U] + \mathbb{E}[\eta|U]$.

[10]A consumer derives equal utility from a loaf of bread in 2000, as she will in 2020 (approximately, and, of course, not from the same loaf of bread). Issues with such measures of monetary inflation mostly arise when aggregate consumer utility from "basic" goods and services changes and is hard to measure, such as for goods including phones, computers, and even cars.

3.1.1 Comparing private and public feedback

Employers assigned both public and private feedback for the same contract. Figure 3 shows the distribution of public feedback, conditioned on the private feedback. The percentage of employers giving that feedback score is shown in parenthesis in each panel. 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 response to the private feedback question: "Would you hire this freelancer [worker] again, if you had a similar project?"



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.

Employers who leave more negative private feedback do assign lower public feedback scores: 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, and 28.4% publicly assigned more than 4 stars. In short, many privately dissatisfied employers publicly claimed to be satisfied. We can see that the reverse—privately satisfied employers giving bad public feedback—essentially never happens. Employers who selected "Definitely yes" left very positive public feedback, with more than 95% of these observations falling into the highest bin.

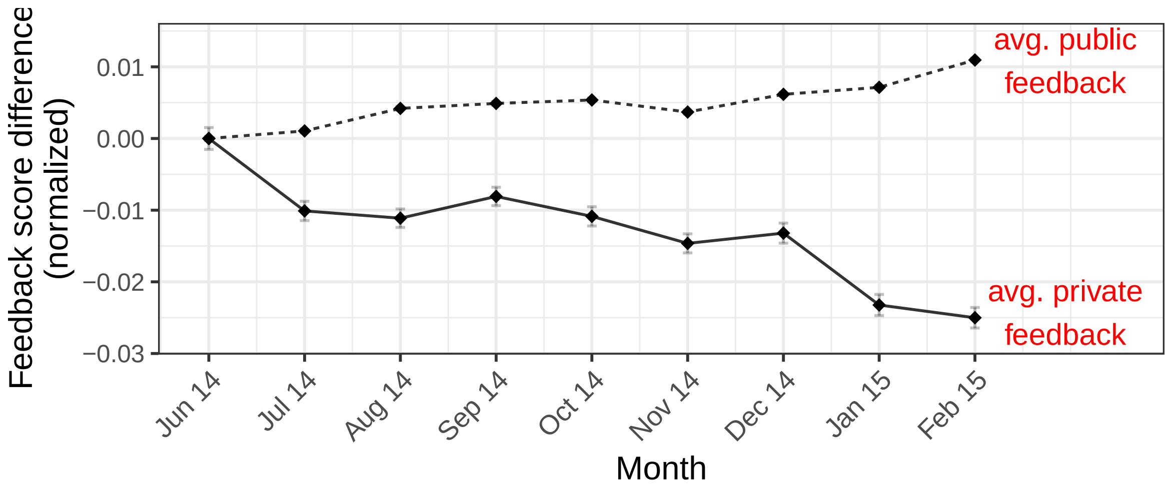
3.1.2 Private feedback over time

The platform began eliciting numerical private feedback by employers on workers in June 2014. In Figure 4 we plot the average monthly feedback over time, for the numerical public and private feedback during the period that both were collected. To make the two scores comparable, we normalize them by the first observed mean. In the language of Section 3, we use $\hat{s}(a_t) = (a_t - a_0)/a_0$, which has the advantage of simplicity.[11] A 95% interval is plotted around each observation. Public feedback scores exhibit a small increase during the period of interest (as we saw in Figure 1b). For the same period of time, Figure 4 shows that private feedback scores exhibit a strong decreasing trend.

It is critical to note that the average feedback scores shown in Figure 4 are being assigned by the same employers on the same contracts. The decrease in private feedback scores suggests a decline in rater satisfaction, and yet public feedback increased. In short, it is hard to rationalize some change in fundamentals alone that could generate this pattern. What seems more probable is that public feedback scores are subject to inflation, whereas the private scores are not because of their private nature. And at least during this period of time, the private feedback suggests that there was some reduction in rater satisfaction. In Appendix B we conduct a series of robustness checks for the above result. We rule out 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—that could rationalize the divergent trends. Later, we will use a change in how the public feedback score works to more directly explore its tendency towards inflation, or lack thereof.

[11] We opt for a simple $\hat{s}$ in this section for exposition purposes. We use a more intricate mapping function in Section 3.2.

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 (June 2014). A 95% confidence interval is shown for each mean.

3.2 Written feedback as an alternative measure

Although private feedback scores offer an alternative measure of rater satisfaction during the period where both types of feedback were elicited by the platform, they only cover a fairly short period of time. Recall from Figure 1b that most of the increase happened back in 2007. In contrast to the private feedback, we have written feedback over the entire platform history. Using written feedback as an alternative measure has the advantage that many other platforms also collect this data, and hence our method could be used in more cases.

To make the two kinds of feedback comparable, we fit a predictive model, $\hat{s}(\cdot)$ 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."[12] However, some words found in written feedback, such as "unresponsive," are less subjective and are likely to be associated with the same degree of satisfaction, 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

[12]For example, an employer's written feedback in our data reads: "This is the most impressive piece of coding in the history of software development!"

these costs are central to explaining reputation inflation (as we will show in Section 4).

There are several reasons that the costs to the rater for giving "negative" written feedback are lower than for numerical feedback. First, it is harder for workers to complain about textual "tone" than it is to complain about a non-perfect star rating. Another reason is that the platform does not aggregate written feedback or put it on a scale. Furthermore, the platform does not present the written feedback to the employer during the initial worker screening phase—only average numerical feedback scores are presented, and written feedback is harder to access. Many would-be employers would also likely not bother to read the often voluminous collection of written feedback, making any particular written feedback less consequential. As such, written feedback cannot be used for cross-worker comparisons by future employers as easily as numerical feedback does; these comparisons are precisely what makes the feedback consequential for workers.

To the extent that written feedback offers a more or less unchanging measure of rater satisfaction, that is, the rater satisfaction measure a remains constant, 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. We have some evidence that written feedback does inflate, in that the private feedback score (pre-public revelation) was declining while the sentiment of written feedback was increasing.

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. The algorithms we use are linear regression, lasso regression, ridge regression, gradient boosting regression, and random forest regression. We then keep the best performing model in terms of average squared error.[13]

The average quarterly feedback scores over time, for both the numerical public feedback,

[13]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).

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 4.86 stars at the beginning of 2016. In contrast, the average score predicted from the written feedback only goes to 4.25 stars.

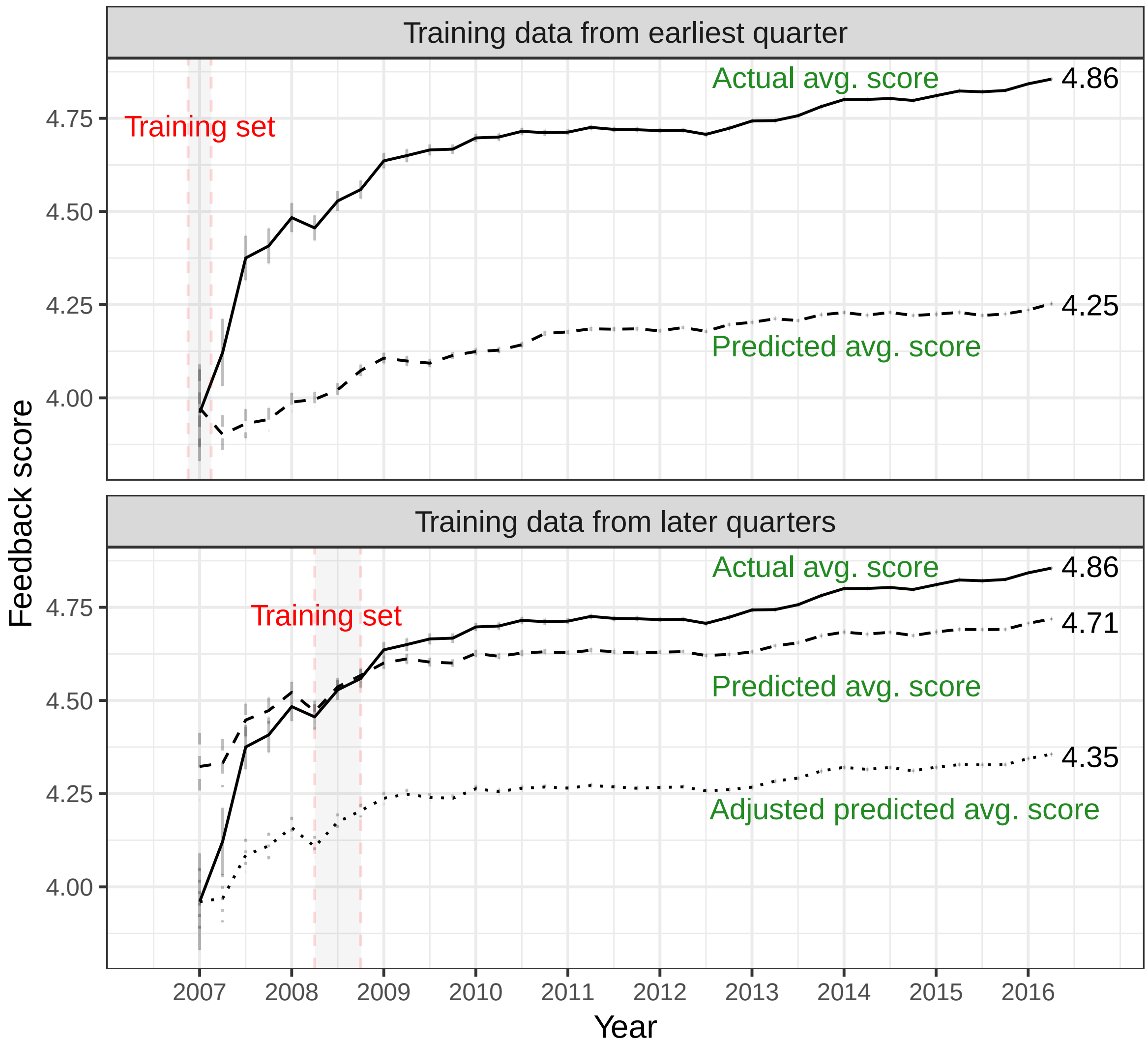
In the bottom panel of Figure 5 we again plot the average quarterly feedback over time, for both the numerical public feedback and the feedback predicted from the written feedback. However, our training sample now comes from a longer time period indicated by the two vertical red lines, and is larger, 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 the opposite holds after 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 data. With this adjustment, the average predicted feedback score at the end of the data "should" have only been 4.35 stars. Reassuringly, the two corpuses give similar results.

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 allows us to quantify the contribution of lower rater standards to the increase. Using the first quarter sample, the point estimate is that 67.7% of the increase in feedback scores is due to inflation, whereas the larger sample from the middle of the data implies 56.6% of the increase is due to inflation. However, this approach does require the assumption that there is no selection with respect to bias in the model or the rater, i.e., that $\mathbb{E}[\epsilon|U']$ and $\mathbb{E}[\eta|U']$ are constant. Although this assumption is not directly testable, in Appendix C we report a number of tests looking for evidence of selection bias with respect to the written measure, finding no evidence against our assumption. 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.

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. In the language of our model, $a(u_i)$ has shifted to $a(u_i) + h(u_i)$. While we have no evidence that supports this hypothesis, in what follows we

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

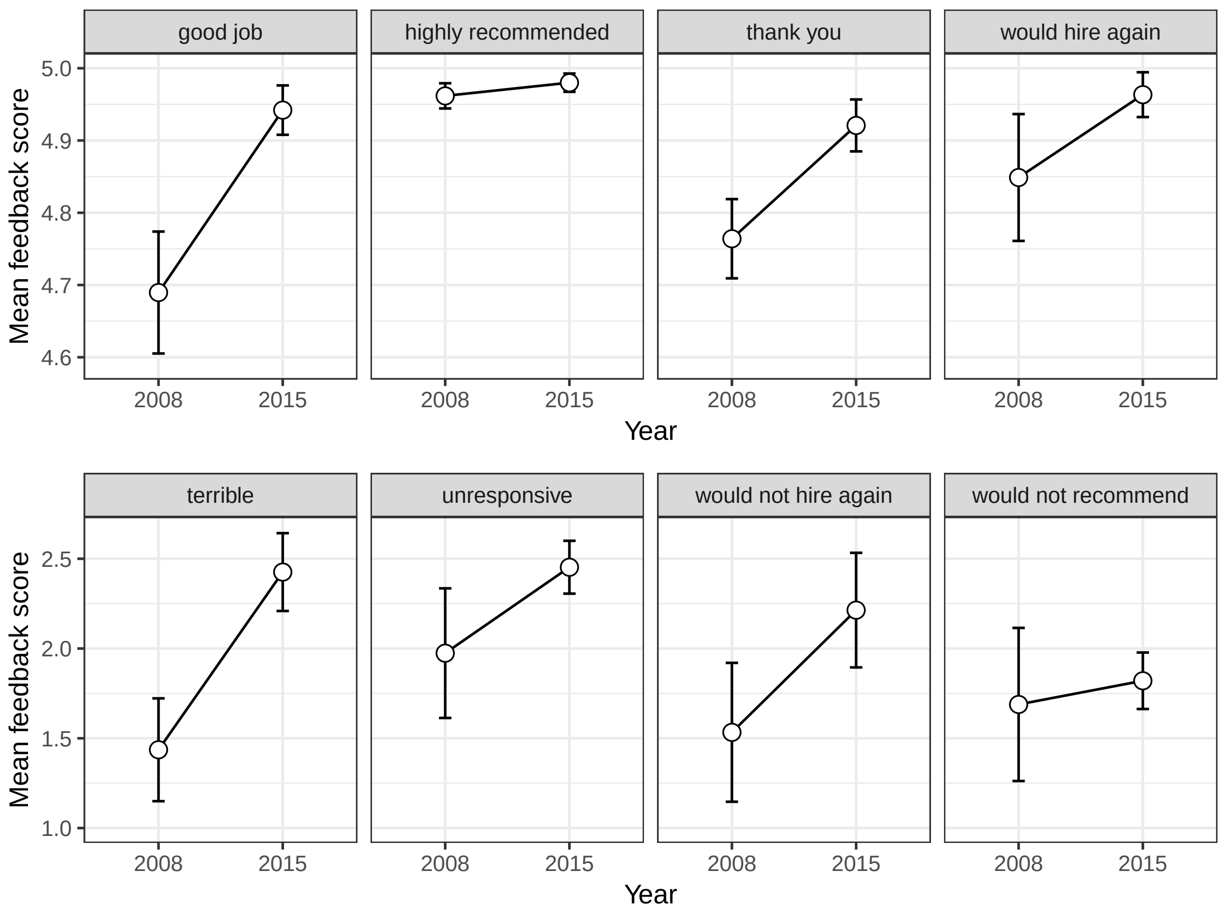


Notes: This figure plots the evolution of average public feedback scores (solid line) versus the average predicted score of textual feedback (dashed line) assigned by employers to workers. A 95% interval is depicted for every point estimate. The shaded area indicates the quarters from which training 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 (dotted 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.

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 find all lexically identical sentences generated in these periods, and then compare average feedback by sentence, 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.

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.

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 0.21 stars, a back-of-the-envelope calculation suggests about 52.3% of the overall increase is due to inflation, which is close to our lower bound estimate from our predictive modeling approach from Section 3.2.1.

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 a reputation system in a competitive 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 feedback.

Motivated by the strategic reporting we observed in Section 3.1—reviewers choosing to leave "good" public feedback despite an unsatisfactory experience (at least as stated in private)—we develop a model where raters decide whether to be candid following a bad experience. As we never observe strategic misreporting of good private experience—i.e., good private feedback but bad public feedback—we assume that the choice in our model is restricted to whether a bad private experience should be publicly reported.

Employers in our model have an incentive to truthfully assign "bad" feedback after a bad experience, captured as a positive benefit from truth-telling. 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.[14] At the same time, employers incur a cost when they assign "bad" feedback, which is increasing in the cost of the workers from receiving this bad feedback. This "reflected" disutility includes the employer's aversion to 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.

Our model gives reflected costs a large role. This feature was motivated by the differences we observed 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

[14] 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.

of assigning "bad" feedback.

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 θ. 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 $b > 0$. 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 bad output ($y = 0$) and the employer truthfully reports $s = 0$, the worker incurs a cost Δw, which is the wage penalty in the next round. We assume that some fraction of this cost is "reflected" back on the employer. 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 $F : [\underline{c}, \bar{c}] \to [0, 1]$, 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

$$b \leq c_i \Delta w. \tag{2}$$

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 infers that

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

The cost of bad feedback to a worker is then

$$\Delta w(p) = w_{s=1;p} - w_{s=0;p} = \frac{\theta(1 - \theta)(q_H - q_L)^2}{k - pk^2}, \tag{3}$$

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 3 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, as many workers receiving "good" feedback actually did not produce the output. We note that this relationship between the wage penalty p makes which feedback is "good" and "bad" endogenous in our model—the characterization depends on p, which in turn depends on the choices of all other employers, who are reacting to that wage penalty.

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

$$p_E = F\left(\frac{b}{\Delta w(p_E)}\right), \tag{4}$$

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

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

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 ($p_E = 0$).[15] 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 4 we obtain $T(p_E) = p_E$. Since F is a cumulative distribution function, and Δw is 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 $p_1 - p_2 < p_0 - p_1$. Following the same argument, we can inductively show that the dynamics of the marketplace result in convergence to the equilibrium truth-

[15] 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.

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 all marketplaces we have data on spanning their entire operations, i.e., in Figure 1b and in panels (a) and (c) of Figure 2: reputation initially inflates fast, but then flattens out as the equilibrium fraction is approached.

5 Effects of making feedback consequential

Our model in Section 4 proposes a process by which reputation inflates. A key feature of our model is Equation 2, which posits that an employer misreports following a bad experience 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 a convergence to new equilibrium following a change in costs. 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 gradual convergence to some new equilibrium.

Recall from Section 3.1 that employers were more candidate about bad performance in private than they were in public. Our interpretation of private versus public feedback is that for "bad" public feedback, the cost to the worker, Δw, 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 Δw 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 would 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 bad 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 would remain unaffected and close to zero, even though Δw increased from zero to some positive value. However, if many employers simply do not want to hurt the worker or fear some other kind of generalized ex post retaliation, then even the batched release of private feedback scores would keep the weight of the distribution of c above 0, which should cause the private feedback measure to inflate from the all-truthful equilibrium to some new equilibrium.

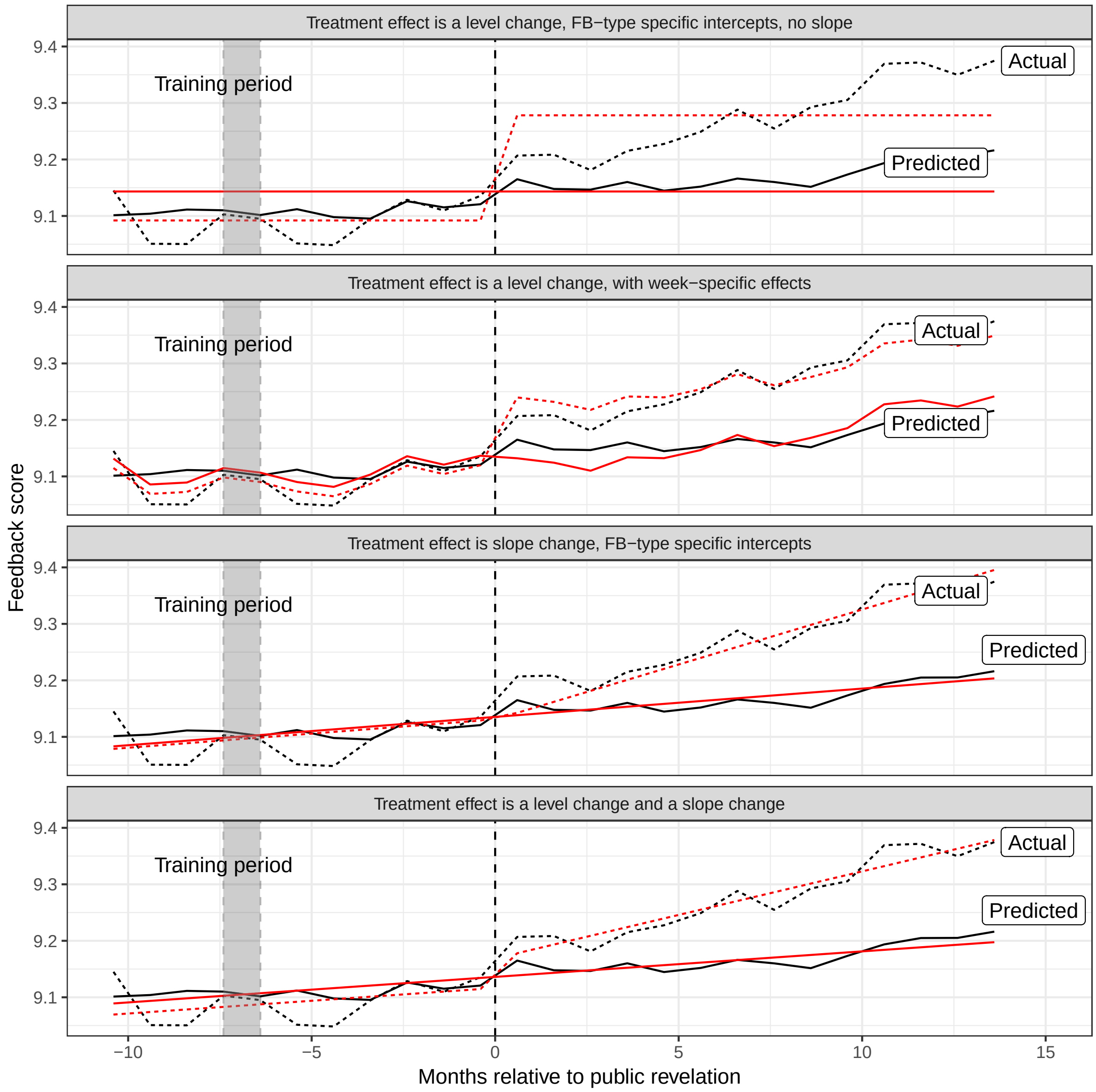
Of course, simply observing that the time series of numerical private feedback rises after the platform change does not prove inflation. The new feedback system was intended to improve matches, and so the same concern from our earlier analysis applies—namely that any increase in the private feedback score following revelation reflects changes in fundamentals. For example, if employers could now form better matches because of their access to the private feedback score measure, then we would expect higher future private feedback scores. As before, we address that concern by using written comments to construct an alternative measure of employer utility.

Figure 7 plots the monthly average private feedback and the monthly average *predicted* private feedback, using the same modeling method we used earlier in Section 3.2. In each panel, the monthly averages are shown by type, as well as the fitted values under different regression specifications. The predicted private feedback is the prediction of a model trained during a period before the revelation, which is indicated with a shaded region. The day the platform switched to batched public revelation is indicated with vertical dashed line. The figure shows that prior to public revelation, the actual and predicted private feedback are quite similar, but that after public revelation, the actual numerical rating increases while the predicted rating does not.

To quantify the effect of revelation, we switch to a regression framework. However, as we have some choice over the regression specification, the different panels of Figure 7 show various alternatives. In the top panel, we report the simplest specification, which is for the treatment to simply have a level effect and to allow the two feedback-types to differ by a fixed amount before the change. We can see that this specification clearly fails to capture the underlying time trend in both series, and especially for the numerical feedback in the post period. In the next panel down, the specification maintains the assumption of a level feedback, but includes a week-specific effect. This specification better captures the underlying trend in both measures that caused the previous specification to perform poorly, but it still performs inconsistently in the post-period, over-estimating the actual feedback early in the period, and then under-estimating it later, and vice versa for the predicted feedback. This is consistent with the simple level-change specification not capturing some of the dynamics of the effects of the treatment e.g., a change in slopes.

In the third panel from the top, we give both types of feedback a common linear time trend, but then allow that trend to change in the post-period for the actual feedback. With

Figure 7: Monthly average private feedback scores and average predicted private feedback scores



Notes: This figure shows the average monthly private feedback (on a 1 - 10 point scale) given by employers to workers, both actual and predicted. Predicted scores are derived from the employer's written textual public feedback, with the predictive model fit using data from the shaded region. The 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. The red lines in the lower panels correspond to predictions from various difference-in-differences model specifications.

a common slope, the fit in the pre-period is much better than when we forced the two types to only differ by a level (in the top panel). However, we can see that earlier in the pre-period, only allowing a linear change in slopes under-predicts the actual feedback score, suggesting some immediate effect and not just a change in slopes.

In the bottom panel, the specification allows for both a level treatment effect and a change in slopes. This specification seems to work the best, with the predicted series closely matching the realized value. We will make use of this insight when we switch to estimating the effects of public revelation at the level of the individual contract rather than at the level of monthly averages. This has the advantage of allowing us to directly control for employer-specific effects and thus directly control for some of the potential sources of bias described by Equation 1.

As our interest is in the divergence between the public and private feedback scores, we switch our outcome to Δs_i, which is the numerical private feedback rating minus the predicted private rating based on the sentiment of the written text. By taking this difference, we eliminate the need (or the possibility) of including time-based fixed effects.

Table 1, Column (1) reports an estimate of the effects of public revelation on the gap between the actual and predicted feedback scores. We can see that after the switch to revelation, the gap increased. The effect size of 0.13 is about 8% of the population standard deviation in Δs. All standard errors in this table are clustered at the level of the individual employer.

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

	Dependent variable:			
	Δs, (Actual - Predicted) Private FB Ratings			
	(1)	(2)	(3)	(4)
Post-revelation	0.133***	0.167***	0.138***	0.067***
	(0.007)	(0.007)	(0.013)	(0.018)
Post × Month				0.011***
				(0.002)
Constant	−0.013*	−0.105***		
	(0.007)	(0.006)		
Less than 25 contracts for C and E	N	Y	Y	Y
Employer FE	N	N	Y	Y
Observations	899,842	537,640	537,640	537,640
R^2	0.001	0.002	0.498	0.498
Adjusted R^2	0.001	0.002	0.184	0.184

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 $p \leq .001 : ***$.

One limitation of including all assignments as the unit of analysis is that it over-weights employers and workers with many contracts. In Column (2), we restrict the sample to employers and workers with fewer than 25 completed contracts in total. We can see that the effect size is somewhat larger with this restricted sample, but is broadly similar in magnitude.

As we noted in Section 3, one reason why a gap might emerge between some measure and an alternative measure is that in the post-period is selection of raters with idiosyncratically large or small gaps—recall the $\mathbb{E}[\eta|U']$ term from Equation 1. To assess this possibility, in Column (3) we add an employer-specific fixed effect to the regression. The effect size is somewhat smaller when the fixed effects are included, but the coefficient implies that the increase in the gap between the private numerical rating *within* employers is quite close to the average effect.

As we saw in Figure 7, there was visual evidence for a change in the trend and not just a level difference. As such, for our preferred specification, in Column (4), we include both a post-indicator and a linear time trend for the post period and continue using the restricted sample and the employer-specific fixed effect. We can see that some of the treatment effect detected in Columns (1)-(3) was the accumulation of a trend of an increasing gap in the post-period.

Column (3) shows that average scores kept rising after public revelation. Although the trend appears to be linear, if private feedback follows the same pattern as the public feedback, we might expect the growth to slow, particularly as it nears the top value. If we project the Column (3) estimated trend of a 0.011 per-month increase in the numerical score into the future, the average numerical feedback would be equal to the top value in the scale $(10 - 9.1)/(0.011)/12 \approx 7$ years. In short, in the long-run of about 7 years, we would expect very little information to be left in the new reputation system.

6 Discussion and 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 four other marketplaces with similar characteristics exhibit the same trend, suggesting that the reputation inflation problem is widespread. We propose and evaluate an approach to quantify inflation based on using alternative measures of rater satisfaction. 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 of negative feedback were increased—namely by making previously private feedback public—yielded data consistent with the predictions of our model.

Reputation inflation is likely most acute in peer-to-peer platforms, such as online labor markets and sharing economy marketplaces, where 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 Hortaçsu, 2010; Moe and Schweidel, 2012; Godes and Silva, 2012; Hu et al., 2017). However, the underlying importance of rater costs seems to persist 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.

Reputation inflation 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 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.

The grade inflation literature considers the effects of this inflation, with Babcock (2010) finding evidence that inflated grades seemingly reduce student effort. In this paper, we do not explore the consequences of reputation inflation, though doing so would be an interesting next step. We do present some evidence that reputation scores have become less informative over time in our focal marketplace in Appendix D.

For would-be marketplace designers, our 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. Our model suggests some approaches, such as raising the benefit b to truthfully reporting feedback or lowering the reflected cost coefficient c. Lowering the penalty from bad feedback also would "work," though doing so would likely undermine the purpose of the reputation system.

To raise b, 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). This approach would be costly to implement, as the total compensation would have to be analogous to the size of the platform. Platforms could also emphasize reviewers as performing a service for fellow consumers. Our evidence suggests that raters have somewhat altruistic motivations in giving ratings, which the platform could try to tap, emphasizing the assistance reviews provide to other parties or the necessity of negative feedback to workers in learning how to improve. Platforms could also give other incentives for having a reputation for good, "honest" reviews. Local business reviewing website Yelp employs mechanisms such as badges for top reviewers, as well as making the feedback score distribution of each reviewer publicly accessible, and casual inspection suggests far more centered ratings distributions. Mandatory grading curves often found in non-digital reputation systems are a policy that raises b, though for settings where buyers evaluate sellers as a "flow" forcing a distribution is challenging.[16]

Platforms already take steps to lower c. These steps, such as simultaneously-revealed ratings (in place since the start of the platform) and anonymizing ratings through aggregation (as was the case with the private feedback change), did not prevent inflation from occurring in our data. Recent research from other domains suggests that the aggregate effect of retaliation concerns is not the primary source of reflected costs; Fradkin et al. (2017) find that introducing a simultaneous reveal feature in Airbnb reviews only reduced the percentage of 5-star ratings by 1.5 percentage points. An alternative explanation is that reflected costs are due to raters incurring a greater personal cost—or guilt—the greater the harm they impose on the rated worker. Our analysis of the private feedback revelation supports this "guilt" view, as we find that feedback began to inflate following public revelation, even though it would be hard for workers to retaliate.

[16] 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.

Barach, Moshe and John J Horton, "How do employers use compensation history?: Evidence from a field experiment," 2017.

Berentsen, Aleksander, Guido Menzio, and Randall Wright, "Inflation and unemployment in the long run," *American Economic Review*, 2011, *101* (1), 371–98.

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

Cavallo, Alberto and Roberto Rigobon, "The billion prices project: Using online prices for measurement and research," *Journal of Economic Perspectives*, 2016, *30* (2), 151–78.

_ , W Erwin Diewert, Robert C Feenstra, Robert Inklaar, and Marcel P Timmer, "Using online prices for measuring real consumption across countries," Technical Report, National Bureau of Economic Research 2018.

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.

Council, National Research et al., *At what price?: conceptualizing and measuring cost-of-living and price indexes*, National Academies Press, 2002.

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.

Diewert, W Erwin, "Index number issues in the consumer price index," *Journal of Economic Perspectives*, 1998, *12* (1), 47–58.

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, and David Holtz, "The determinants of online review informativeness: Evidence from field experiments on Airbnb," *Working paper*, 2017.

_ , _ , 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.

Friedman, Milton, "Nobel lecture: inflation and unemployment," *Journal of political economy*, 1977, *85* (3), 451–472.

Galı, Jordi and Mark Gertler, "Inflation dynamics: A structural econometric analysis," *Journal of Monetary Economics*, 1999, *44* (2), 195–222.

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 Uber's 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.

_ **, 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.

Jr, Robert E Lucas and Leonard A Rapping, "Real wages, employment, and inflation," *Journal of Political Economy*, 1969, *77* (5), 721–754.

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.

Mishkin, Frederic S, "Inflation targeting in emerging-market countries," *American Economic Review*, 2000, *90* (2), 105–109.

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.

Sidrauski, Miguel, "Inflation and economic growth," *Journal of Political Economy*, 1967, *75* (6), 796–810.

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 Other reasons for the feedback score increase

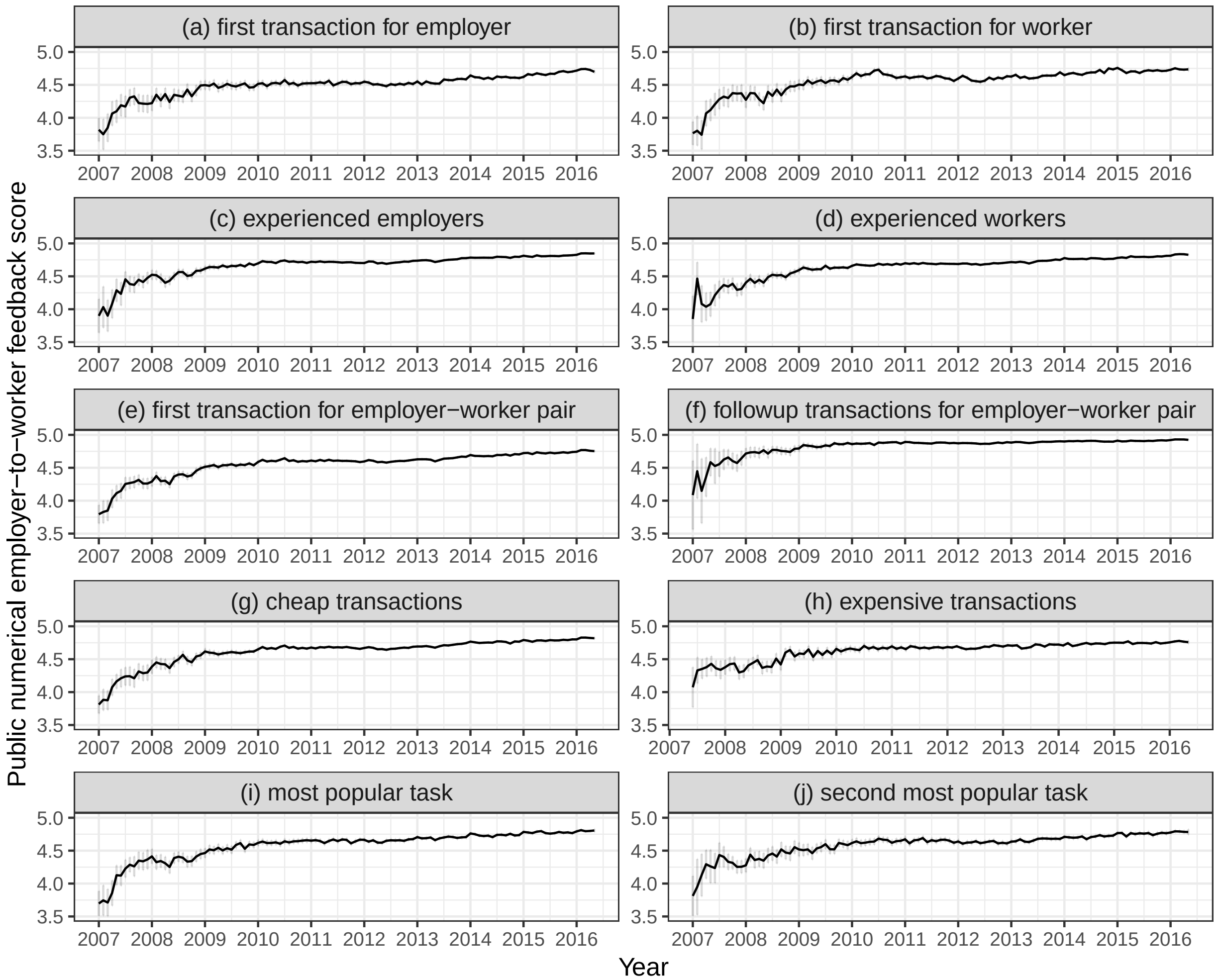
The increase in numerical scores we observed in Figure 1b and 2 may be the outcome of changes related to employer/worker composition, experience, and selection, as well as changes in the composition of work types transacted on the platform. Though we address such concerns through the alternative measures approach in Section 3, we believe it is useful to also offer direct supplementary evidence that the observed pattern persists regardless of these factors.

Toward that end, we plot average feedback scores for various subsets of the focal marketplace data in Figure 8. In panel (a) and (b) we plot the average employer feedback scores for the first transactions of employers and workers respectively. The same increase in feedback scores holds for both of these cases, suggesting that the observed increase is not a function of experience: inexperienced employers give higher ratings over time, and inexperienced workers also receive higher ratings. Panels (c) and (d) depict the average feedback scores only for transactions in which employers and workers respectively are experienced—workers and employers with more than four previous transactions in the platform. Average scores are only slightly higher than those of first-time users, but exhibit the same over-time increase. Panel (e) plots only feedback scores from first time transactions between a pair of employer and worker, where we observe the same pattern. In panel (f) we plot average feedback scores for transactions where the employer and worker have transacted in the past. As the employer chose to transact with the same worker, we expect that the employer was more satisfied. Indeed, we observe that average scores start higher, yet we see a similar over-time increase in average numerical scores. Together, panels (e) and (f) indicate that selection can not explain the observed increase.[17] Panel (g) plots average scores for transactions worth less than 100 US dollars, panel (h) plots average scores for transactions worth more than 1000 dollars, and panels (i) and (j) plot average scores for the two most frequent types of tasks. We see that the composition of types of transaction in the platform also can not explain the observed increase.

Reasons related to selection and composition do not explain the observed trend: in Section 3 we develop an alternative measures approach that accounts for such reasons, and show that the observed increase is largely inflationary.

[17]Further, this reconfirms the findings of panels (b) and (d) in Figure 2. In the home-sharing marketplace, it is unlikely that selection is a major factor since users are unlikely to repeatedly travel to the same destination for leisure, while in the service marketplace the platform matches providers with consumers. Further, supplier capacity is highly constrained in these platforms, and it is hence unlikely that the same provider will be available in the future.

Figure 8: Monthly average feedback scores assigned by employers to workers for various subsets of the online labor market data.



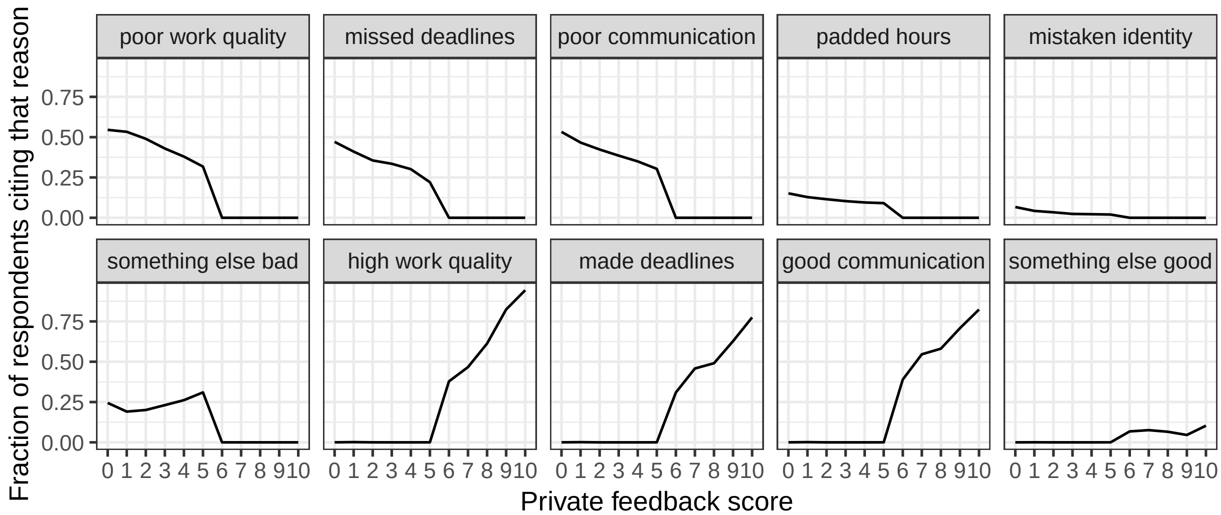
Notes: This figure plots the average public feedback scores assigned in the focal online labor market. Scores are assigned by employers to workers upon the completion of each transaction, and the scale for feedback is 1 to 5 stars. For each observation, average scores are computed for every time period, and a 95% interval is depicted for every point estimate. Panels (a) and (b) plot average scores for the first employers' and workers' transactions respectively. Panels (c) and (d) plot transactions where employers and workers respectively, had more than 4 previous transactions on the platform. Panels (e) and (f) plot average scores for each employer-worker pair's first, and followup transactions respectively. Panels (g) and (h) plot average scores for transactions with cost of less than 100 and more than 1000 U.S. dollars respectively. Panels (i) and (j) plot average scores for the two most common freelancing tasks.

B Robustness tests for private feedback

B.1 Misinterpreting private feedback

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 misinterpret the private feedback question. The fractions of private feedback reports citing these different reasons against the assigned private feedback score (1 to 10 scale) are plotted in Figure 9. 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 9: Fraction of users citing a given reason when giving private feedback, by score.

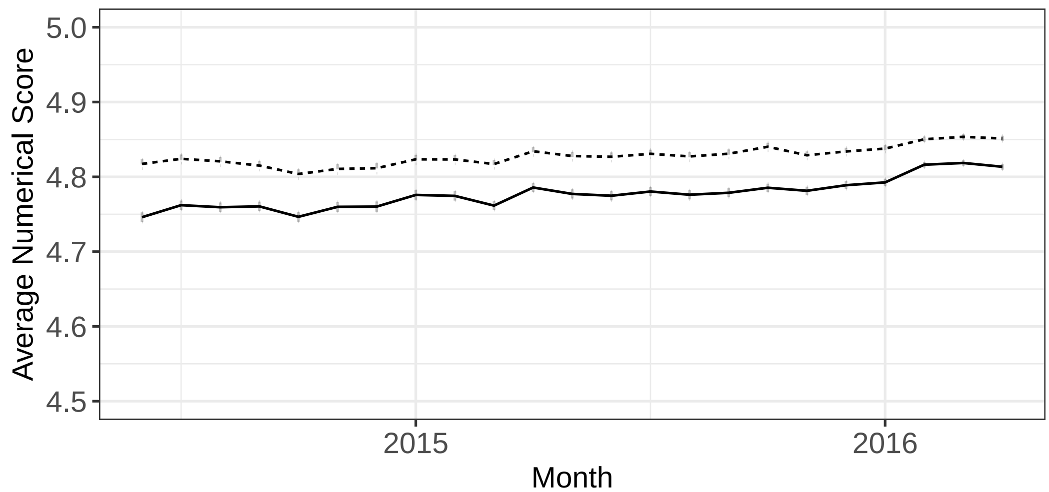


Notes: This figure plots the fraction of feedback reports that cited each reason as the basis of the feedback being positive or negative, against the private feedback score given. Across every case, we notice that employers that assigned more extreme feedback scores tend to cite reasons of the same sentiment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

B.2 Selection issues

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 10 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 employer who publicly indicate higher satisfaction are more likely to also assign private feedback. The two lines closely resemble each other throughout the period where we collect both types of feedback, indicating no systematic change over time.

Figure 10: 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.

Another concern is that employers decision to leave private feedback when they leave public feedback could change over time. Figure 11 plots the percentage of contracts that received private feedback amongst these contracts that received public feedback. We observe that there is no systematic change over time in employers' decisions to assign private feedback when they assign public feedback. Further, the percentage of employers that chooses to leave private feedback is high, with an average of 81.4% of employers deciding to also assign private feedback.

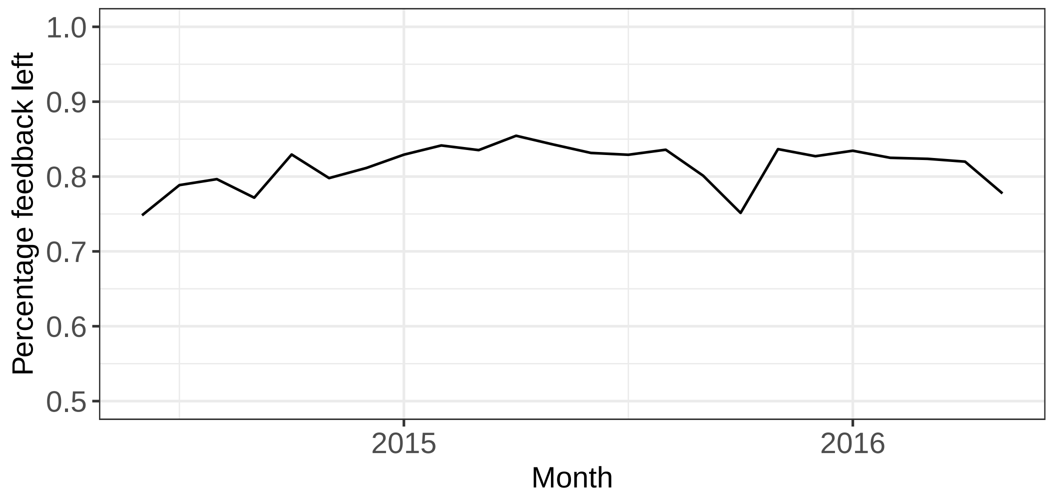
C Robustness tests for written feedback

C.1 Selection issues

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 12 plots the evolution of numerical feedback for all contracts (solid line), and all contracts for

Figure 11: Percentage of employers leaving private feedback in addition to public numerical feedback.



Notes: This figure plots the monthly percentage of contracts for which employers assigned private feedback, amongst those contracts for which employers also assigned numerical feedback.

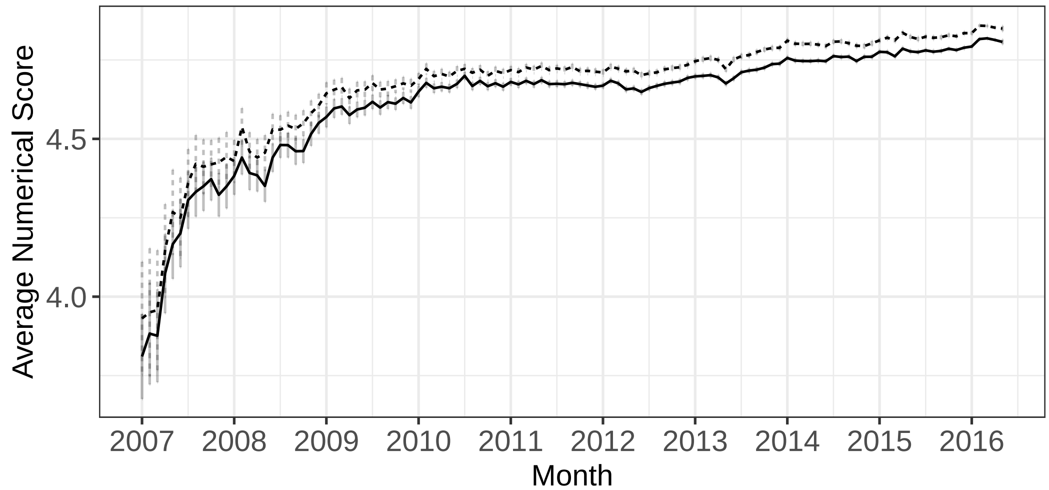
which written feedback was also assigned (dashed line). We observe that contracts in which written feedback is also assigned receive higher ratings, implying that more satisfied employers assign written feedback. However, the degree to which this bias occurs does not change throughout our data. Further, 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.

Similarly to private feedback, a concern is that employers decision to leave written feedback when they leave public feedback could change over time. Figure 13 plots the percentage of contracts that received written feedback for those contracts that also received public feedback. We observe that there is no systematic change over time in employers' decisions to assign private feedback when they assign public feedback. The percentage of employers that chooses to leave written feedback is also high, with an average of 79.2% of employers deciding to also assign written feedback.

C.2 Composition of raters

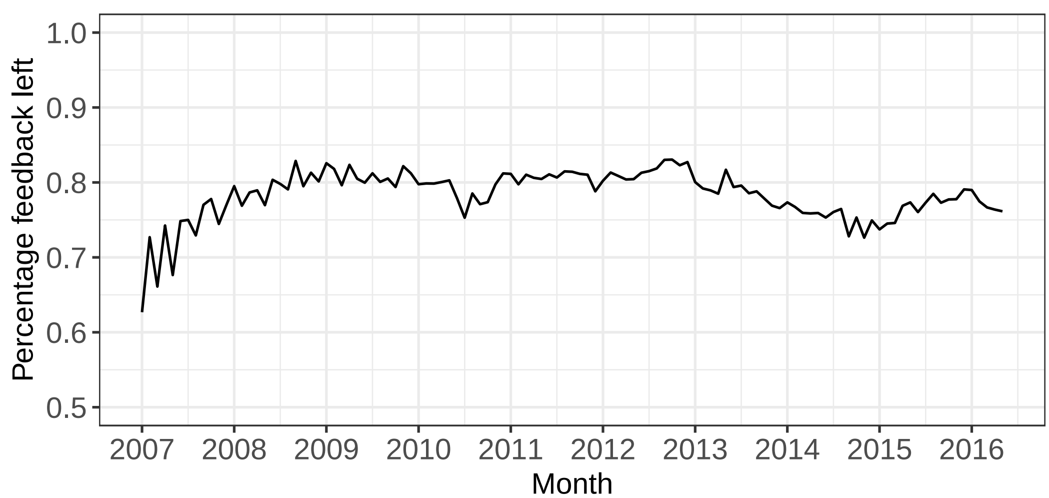
Shifts in the composition of raters could potentially introduce bias in using written feedback as an alternative measure of satisfaction. More specifically, the widening gap between numerical scores and scores predicted from written feedback could be the outcome of employers with this rating behavior—employers who assign higher scores for the same written

Figure 12: 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.

Figure 13: Percentage of employers leaving written feedback in addition to public numerical feedback.



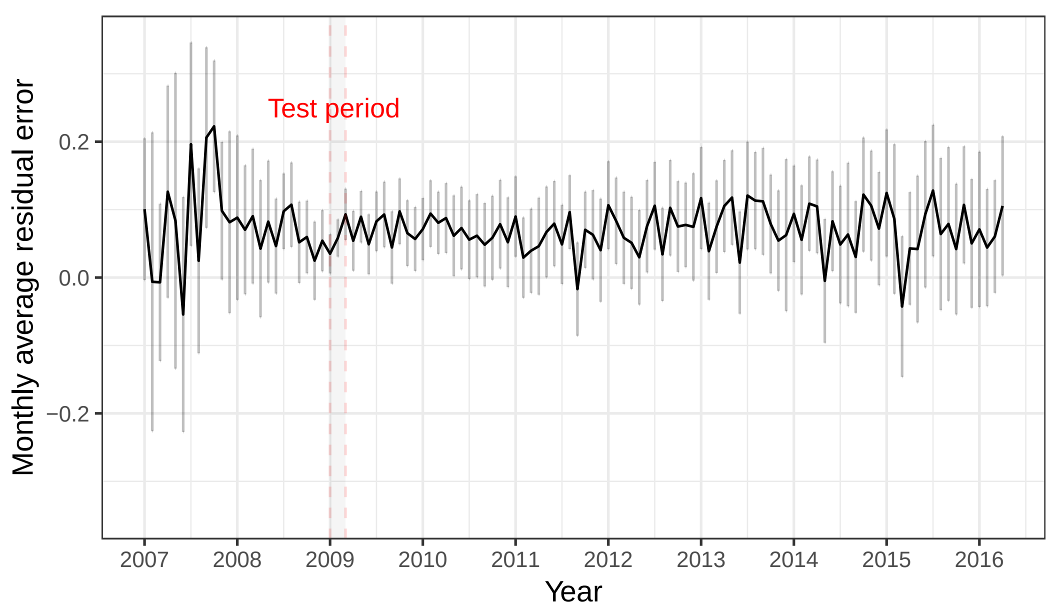
Notes: This figure plots the monthly percentage of contracts for which employers assigned written feedback, amongst those contracts for which employers also assigned numerical feedback.

feedback—joining the platform over time, or, equivalently, employers with the opposite rating behavior dropping out. In the language introduced in Section 3, this issue can be thought of as a systematic changes in $\mathbb{E}[\epsilon|U']$.

We test against this hypothesis as follows. For a period of time T, we compute the average residual error r_i for each employer i that left feedback during T, defined as the divergence between the numerical scores and the predicted scores from the associated written feedback employer i assigned. The employer average residual error is then $\bar{r}_T = \sum_{\text{i left feedback in T}} r_i$. We then test whether, amongst these employers, there is a systematic drop-out behavior that has led to employers with wider gaps remaining in the platform in the post period (and, respectively, whether only employers with narrower gaps were present in the pre-period). We can do so by simply computing $\bar{r}_t = \sum_{\text{i left feedback in T and t}} r_i$, for any $t \neq T$. If for $t > T$ the quantities $\bar{r}_t$ show a systematic increase, then this composition shift in rater types may bias our estimates.

Figure 14 carries out this analysis for employers who left feedback in January and February of 2009. For the predicted scores, we employ the predictions of the model in the lower panel of Figure 5. We find no evidence of a systematic trend in neither the pre-period, nor the post-period, suggesting that our inflation estimates are not subject to this source of bias. Conducting the analysis for other periods in our data or for other predictive models, yields qualitatively identical results.

Figure 14: Employer average residual error in for employers who left feedback during January and February 2009.



Notes: This figure plots the employer average residual error over time for the set of employers who left feedback during the period indicated by the shaded area. The average residual errors are computed for every month, and a 95% interval is depicted for every point estimate.

C.3 Predictive algorithm performance

We present more details about the performance of the algorithms used to extract the written feedback sentiment in Section 3.2.1.

Figure 15a plots the scatterplot of numerical scores versus predicted scores from written feedback for the algorithm trained on data coming from the earliest quarter. Figure 15b plots the same scatterplot for the algorithm trained on data coming from the later quarters. Since the training data is skewed towards higher scores in both cases, the algorithms are expected to over-predict, but both predictive models attain good performance, with the mass of their predictions being close to the 45 degree line. Further, note that this performance is attained despite the fact that we should expect somewhat large variance between scores and written feedback amongst different employers. The appropriateness and good performance of our models is further verified by the fact that the estimates we obtain closely match the performance of our model-free approach in Section 3.2.2.

D 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.[18] 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.[19]

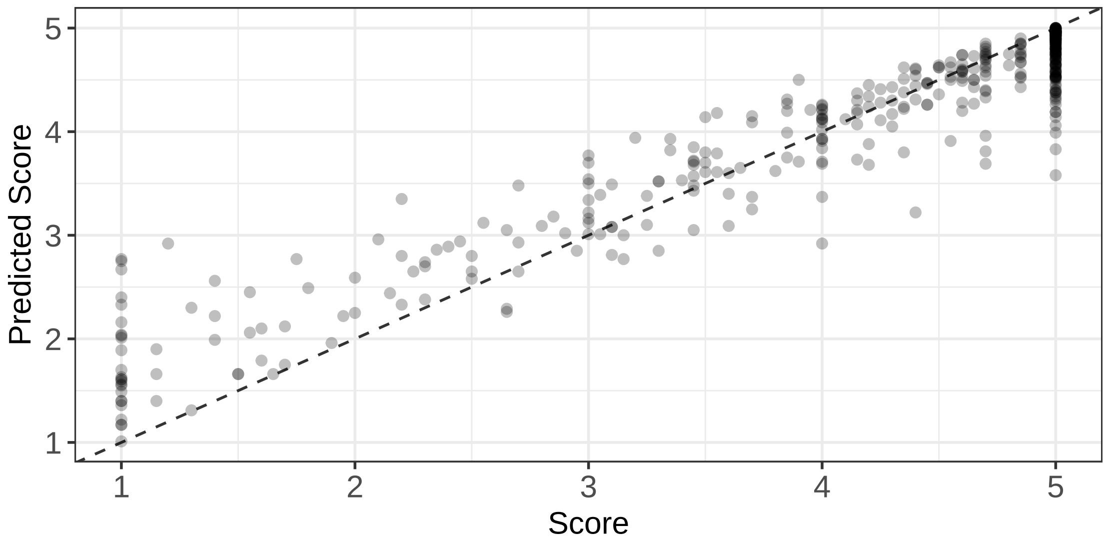
To see the problem created by top-censoring, consider the information conveyed by the observation of a binary variable X, as it is captured by the information-theoretic entropy $H(X) = p\log(p) + (1-p)\log(1-p)$, where p is the probability of one outcome. As p goes to either 1 or 0, the information conveyed 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

[18] 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.

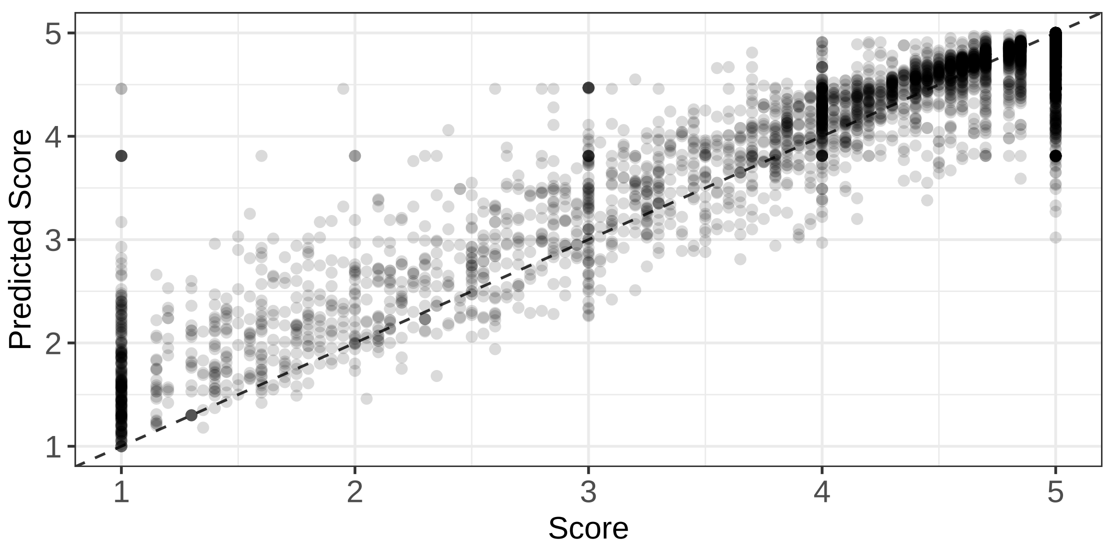
[19] See also https://www.youtube.com/watch?v=KOO5S4vxi0o.

Figure 15: Numerical score versus predicted score from text scatterplot.

(a) Performance on training set from earliest quarter.



(b) Performance on training set from later quarter



Notes: The top panel plots the scatterplot of numerical scores assigned to contracts versus numerical scores predicted from the associated written feedback for the algorithm trained on data from the earliest quarter, while the bottom panel plots the same scatterplot for the algorithm trained on data from the later quarter. The scale for feedback is 1 to 5 stars. The 45 degree line represent the performance of a "perfect" prediction algorithm.

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 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 ϵ_{it} is some noise term such that $E[\epsilon_{it}] = 0$.[20] 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 σ^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 $\sigma^2 \to \infty$, $\Pr(a|\text{SCORE}) \to \Pr(a)$, 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.[21]

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

The fraction of variance due to noise is the quantity

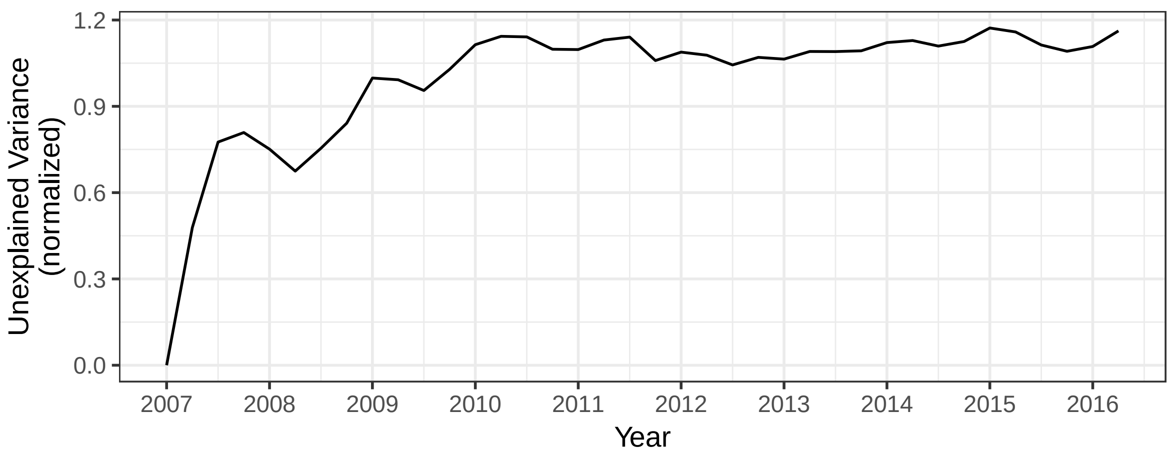
$$\frac{Var(\epsilon)}{Var(\text{SCORE})} = 1 - \frac{Var(a)}{Var(\text{SCORE})}. \tag{A2}$$

[20]For simplicity, we are treating the feedback score as continuous. The logic is identical in the dichotomous case.

[21]Gelman et al. (2014) provide a derivation of this result, which can also be found in most standard Bayesian analysis textbooks.

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 16: 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 16 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.