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, can become less informative in the long-run.
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 such scores—particularly in marketplace contexts—have proliferated, as has the number of individuals and businesses subject to “reputation systems” (Levin, 2011; Hall and Krueger, Forthcoming; Katz and Krueger, 2016). Designing effective reputation systems has become a first-order question in the digital economy.

In online marketplaces, “reputations” are typically calculated from numerical feedback scores left by past trading partners. As many have noted, the distribution of feedback scores in various online marketplaces seems implausibly rosy.\(^1\) For example, the median seller on eBay has a score of 100% 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 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. Two distinct—but not mutually exclusive—reasons can cause rising feedback scores: (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.\(^2\) This “reputation inflation” erodes the comparability of feedback scores over time, and reduces the informativeness of reputation system.

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 the overwhelming majority of workers receiving perfect feedback.\(^3\) However, the distribution has not always been this

---
\(^1\)A PEW research center survey finds that while more than 80% of U.S. adults read online reviews before purchasing an item, almost 50% find the truthfulness of these reviews hard to assess (see http://www.pewinternet.org/2016/12/19/online-reviews).

\(^2\)This kind of inflation is similar to the conjecture about the increase in college grades—namely that students and work is not getting better but rather the same quality of work now earns a higher grade (Babcock, 2010; Butcher et al., 2014).

\(^3\)We use the terms “employer” and “worker” for consistency with the literature, and not as a comment
skewed—the fraction of workers receiving a perfect 5-star rating grew from 33% to 85% in just 6 years. Feedback scores in four other online marketplaces for which we obtained longitudinal data exhibit a similar increase over time.

To examine whether the increase is caused by raters becoming more satisfied or by raters lowering their 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 employers. At the same time, raters were still asked to give the status quo “public” feedback. We find that, while private feedback scores were decreasing, 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.

As a second alternative measure, we use the sentiment raters express in their written feedback that accompanies numerical scores. To capture this sentiment, we fit a model that predicts numerical feedback from 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 the training feedback was created. Using the predictions, we 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, and so on) that increased rater satisfaction and is reflected in feedback text, and the residual component that is due to inflation. Our estimates suggest that more than 50% of the increase in scores is due to inflation, and are robust across several specifications. Furthermore, insofar written feedback is also subject to inflation, our approach understates the extent of reputation inflation.

We next turn to understanding the cause of reputation inflation. We 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: 28.4% of employers who privately report that they would not hire the same worker again, publicly assign them 4 or more stars out of 5. The likely reason is that bad public feedback is costly in a way that bad private feedback is not. For example, employers may fear that workers will retaliate by complaining, bad-mouthing the rater, or withholding future cooperation. However, rated workers on this platform cannot engage in “tit-for-tat” rating behavior, as on the legal relationship of the transacting parties.
feedback scores are simultaneously revealed (Bolton et al., 2013). Furthermore, employers may not want to harm the ratee’s future prospects, as public feedback is consequential. These “reflected” costs make giving “bad” feedback costlier for raters than giving “good” feedback.

The cost of giving “bad” feedback could explain why feedback scores are higher than they would be if employers reported truthfully. However, it cannot by itself explain the dynamics of ever-higher scores we observe. Inflation requires the cost of leaving “bad” scores to increase over time. We hypothesize this is precisely what happens: the same feedback score (e.g., 2 “stars”) becomes 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 rated workers—is endogenous, depending on the current distribution of feedback scores, which in turn determines what inferences future buyers make from a score.4

We report evidence from a platform intervention that allows us to test our hypothesis directly. After collecting private feedback for 10 months, the platform began releasing batched aggregates of this private feedback to would-be employers. With this aggregation, private feedback remained quasi-anonymous, as the worker would not know which employer gave which feedback. However, private feedback became consequential to workers because it was publicly reported. Insofar that employers do not want to harm workers’ future prospects giving “bad” feedback suddenly had a cost, whereas before it had none.

The private feedback quasi-experiment allows us to test our hypothesis that (1) the choice of what score to give is “strategic”—in the sense that employers consider the likely costs and benefits to what they report—and hence 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. Importantly, the fact that the sentiment of written feedback remained

4We formally illustrate how reputations inflate through a simple model of a marketplace with a reputation system (see Appendix D). We find a unique, stable equilibrium in which reputations are universally inflated, because employers are more likely to report “good” feedback regardless of actual worker performance. The degree of inflation depends on how much cost the rated party that receives “bad” feedback can impose on the rater who gives it; this could explain why ratings are more spread out in less “personal” settings—such as consumers rating products on Amazon or restaurants on Yelp—while reputation reflation is likely more acute in highly “personal” settings—such as peer-to-peer platforms (Sundararajan, 2016; Filippas et al., forthcoming).
more or less constant provides us with direct evidence that written feedback is less prone to inflation than numerical feedback. This result implicates the role of the “reflected” cost of “bad” feedback: when getting “bad” feedback became costly to workers, it also became costly to give, and there was less “bad” feedback. As “bad” feedback became scarce, what was mildly negative before became very negative, starting the inflation process.

Our paper makes several contributions. Our key contribution is documenting the extent of reputation inflation in a large online marketplace, and identifying its root cause. Our long-run, whole-system perspective is possible because we use data spanning over a decade of the operations of the marketplace. The reputation inflation problem is likely 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 data from other platforms supports this view—average feedback scores are increasing, despite none of these marketplaces allowing “tit-for-tat” rating behavior.

While our paper is not the first to exhibit that reputations can be biased (Dellarocas and Wood, 2008; Tadelis and Zettelmeyer, 2015; Tadelis, 2016; Hu et al., 2017), we believe it is the first to take a longitudinal approach and show how individually rational choices about what feedback to leave, push the market towards a less informative equilibrium. Our analysis of the public revelation quasi-experiment highlights the key role of costs to raters, and that these costs are hard to diminish: making feedback anonymous failed to counteract inflation. Insofar that employers know “bad” feedback is consequential to workers, employers find it costly to assign “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 less prone to inflation can be designed remains an open research question.

The rest of the paper is organized as follows. Section 2 describes our empirical setting, and documents increasing feedback scores over time. Section 3 shows that this increase is largely due to inflation by employing alternative measures of rater satisfaction. Section 4 examines the quasi-experimental revelation of private feedback information to illuminate the causes of reputation inflation. Section 5 concludes.

2 Empirical context

The setting for our study is a large online labor market. In online labor markets, employers hire workers to perform remote tasks, such as computer programming, graphic design, and data entry. Markets differ in their scope and focus, but common services provided by the platform include maintaining job listings, arbitrating disputes, certifying worker skills and, importantly, maintaining reputation systems (Horton, 2010).
Online labor markets offer a convenient setting for research due to the excellent measurement afforded in an online setting, and the ease of conducting field experiments (Horton et al., 2011). Much of the previous research has examined the role of information in employer decision-making (Pallais, 2013; Stanton and Thomas, 2015; Agrawal et al., 2016; Horton, 2017; Barach and Horton, 2017). One particular focus of this literature has been reputation systems. Cabral and Hortacsu (2010) also find that eBay sellers condition their behavior on their current reputations. Moreno and Terwiesch (2014) show how employers use reputation information, and how workers subsequently adjust their bidding strategies. Tadelis (2016) offers an excellent survey of this literature.

2.1 Status quo reputation system

On the focal platform, when one party ends a contract both parties are prompted to give feedback, both numerical and written. Numerical feedback is given on a 1-5 scale on several weighted dimensions, from which an aggregate score is calculated.\(^5\) A worker’s reputation score is the average of her scores, weighted by each project’s dollar value. 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 the platform’s response to the opportunism that becomes possible once an employer or worker has obtained a high, hard-to-lower reputation (Aperjis and Johari, 2010). The entire feedback “history” is publicly available.

The reputation system could be characterized as a state-of-the-art bilateral system, in the sense that “tit-for-tat” conditioning is not possible (Bolton et al., 2013). The platform does not reveal public feedback immediately, but rather uses a “double-blind” process. If both parties leave feedback during an initial 14 day “feedback period,” then both sets of feedback are revealed simultaneously. If, instead, only one party leaves feedback, then the platform reveals it at the end of the feedback period. Thus, neither party learns its own rating before leaving a rating for the other party. Leaving feedback is strongly encouraged, but not compulsory: 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. Figure 1a depicts the histogram of public feedback scores from January 1, 2014 to May 11, \(^6\)

\[^5\] The dimensions and weights are: “Skills” (20%), “Quality of Work” (20%), “Availability” (15%), “Adherence to Schedule” (15%), “Communication” (15%), and “Cooperation” (15%).
2016, for contracts worth more than $10.\textsuperscript{6} 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. More than 80% of the ratings fall in the 4.75 to 5.00 star bin (1,339,071 observations). The average feedback for the sample of Figure 1a is 4.77.

Scores have not always been highly right-skewed. Figure 1b shows the average monthly feedback over time. There is a clear increase in the feedback scores awarded on the platform: the feedback score average has increased 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. Figure 1c shows the fraction of contracts having a rating within different ranges, over time. In the early days of the platform rating assignments were reasonably dispersed, with completed contracts regularly receiving ratings in the (0,3] range. Near the end of our data, completed contracts essentially never receive a rating in the (0,3] range. 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.

\textsuperscript{6}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.
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 in feedback scores is specific to the focal 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; Filippas et al., forthcoming). Panel (d) plots longitudinal ratings data from six major cities in the United States in a large online marketplace for services.

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. Furthermore, 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 nearby available provider to the consumer upon request of the service. However, the fundamental features of the reputation system mechanisms these marketplaces employ are more or less identical to the ones of the focal marketplace: 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 the 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, and irrespective of the transacted goods and 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.
3 Decomposing changes in reputation

Two broad sets of reasons may have led to substantially higher feedback scores: (1) rater satisfaction has increased, and (2) raters have lowered their standards.

Rater satisfaction can change over time, even if standards remain fixed. Reasons include improvements in search and recommendation features resulting in better employer-worker matches, workers becoming more experienced or exerting more effort, higher-quality cohorts of workers and employers joining the platform, employers incurring lower transaction and uncertainty costs by continuously transacting with a subset of desirable workers, and lower prices. Improvements in these “fundamentals” would result in higher average transaction quality, and vice versa. We cannot hope to account for all potential “fundamental” changes—and detect the increase in scores due to changes in standards. Instead, we can sidestep this issue by employing alternative measures of rater satisfaction.

Let $u_i$ denote the utility employer $i$ obtains after some transaction. Utilities are unobservable, and potentially affected by both “fundamentals”-related and idiosyncratic reasons, which are also unobservable and can change over time. Employer $i$ leaves primary feedback

$$s_i = s(u_i) + \sigma_i,$$

where $s$ is common amongst employers and monotonically increasing in the latent utility, and $\sigma_i$ captures idiosyncratic differences in employers’ rating behavior. The employer also leaves an alternative feedback measure $a_i$, that we can similarly decompose into

$$a_i = a(u_i) + \alpha_i.$$

Assume that, at time $t$, utilities follow some latent distribution $F$. Let $U$ denote a collection of observations such that for all $u \in U$ we observe $s(u)$ and $a(u)$—but not $u$. We can use $U$ to estimate a function $\hat{s}$ from the alternative feedback measure to the primary feedback measure, i.e.,

$$\hat{s}(a(\cdot)) = s(\cdot) + \epsilon,$$

where $\epsilon$ is the estimation error. Essentially, $\hat{s}$ estimates the conditional expectation $\mathbb{E}[s|a]$.

Assume now that at some latter period $t'$, employer satisfaction has shifted, i.e., the latent utility distribution has changed to $F'$, and rater standards have shifted, i.e., the primary feedback measure has changed to $s'$. Although we no longer observe $s$, the function $\hat{s}$ allows

---

7That said, in Appendix A, we directly show that the most plausible selection and composition stories cannot explain the observed increase in feedback scores.
us to recover what the average value of the feedback measure $s$ “would” have been had rater standards not shifted. We get

$$\Delta_s = \int s'(u) - \hat{s}(a(u)) dF'(u) = \underbrace{\mathbb{E}_{F'}[s'] - \mathbb{E}_{F'}[s]}_{\text{reputation inflation}} - \mathbb{E}_{F'}[\epsilon].$$

Insofar $\mathbb{E}_{F'}[\epsilon] = 0$, we can estimate the extent of reputation inflation under the utility distribution $F'$. Crucially, if the alternative measure $a$ is also subject to inflation, then the estimate $\Delta_s$ provides a lower bound for the inflation term.

Using an alternative feedback measures to estimate reputation inflation circumvents the problem of estimating latent utilities, but presents us with a different—albeit more tractable—problem of verifying that $\mathbb{E}_{F'}[\epsilon] = 0$. The $\mathbb{E}_{F'}[\epsilon]$ term can be non-zero due to (i) differential selection of latent utilities of which $\hat{s}$ over- or under-estimates in expectation, and (ii) because of systematic changes in how raters translate their utilities to the primary and alternative feedback measures.

In what follows, we employ “private” and written feedback as two alternative measures of employer satisfaction. We will directly verify that both measures are not affected by systematic biases, and are subject to less inflationary pressure.

### 3.1 Private feedback as an alternative measure

Our first alternative measure of rater satisfaction comes from a platform experiment that elicited an additional “private” feedback measure. At the completion of a contract, employers were asked to generate private feedback—in addition to public feedback. Critically, the platform let the employers know that private feedback would not be shared with other workers or employers, and would only be collected by the platform for internal evaluation purposes. As it was less costly for workers, this private feedback offers an alternative measure that is potentially not subject to inflation—we elaborate on this argument in Section 4. Employers were initially asked the private feedback question, “Would you hire this freelancer[worker] again, if you had a similar project?” Appendix B.1 shows that employers were substantially

---

8Our decomposition task is conceptually similar to estimating monetary inflation in the presence of quality changes (Sidrauski, 1967; Friedman, 1977; 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—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. In the context of online markets, however, there is no “basic” or “standard” transaction, i.e, a transaction with an immutable associated rater satisfaction.
more candid in their private feedback. Starting on June 2014, employers were instead asked to rate workers on a numerical scale of 0 to 10.

Figure 3a reports 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. Public feedback scores exhibit a small increase during the period of interest (as we saw in Figure 1b). However, for the same period of time, private feedback scores exhibit a strong decreasing trend.

It is critical to stress that the average feedback scores shown in Figure 3a are being assigned by the same employers on the same contracts. The decreasing private feedback scores hence suggest a decline in rater satisfaction, and yet public feedback increased. As such, 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.\(^9\)

### 3.2 Written feedback as an alternative measure

Our second alternative measure of rater satisfaction comes from written feedback. In contrast to private feedback, we have written feedback over the entire platform history. Similar to the private feedback, 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. Furthermore, the platform does not aggregate written feedback or put it on a scale, and 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. 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 feedback consequential for workers.

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, using the written feedback corpus as the training set and the associated numerical scores as the set of labels.\(^10\) We then use the fitted model to estimate out-of-sample

\(^9\)Appendix B presents a series of robustness checks, ruling out other assumptions that could rationalize the divergent trends—e.g., that private standards are getting harsher although transaction quality is increasing.

\(^10\)We use a standard natural language processing pipeline: text 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. To find the best performing algorithm, we conduct an extensive grid search, evaluating each configuration of hyperparameters using a 5-fold cross validation in terms of average squared
Figure 3: Using alternative measures of rater satisfaction to quantify reputation inflation

(a) Numerical public feedback score and private feedback score.

(b) Numerical public feedback score and predicted score from textual feedback

Notes: This figure presents evidence of reputation inflation in public feedback scores by employing alternative measures of rater satisfaction. The top panel shows the evolution of the average public feedback scores (solid line) versus the average private feedback scores (dashed line) assigned by employers to workers, for the same contracts. The average scores are computed for every month, and are normalized by the value of their respective first observation (June 2014). The bottom panel shows 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 reviews.
feedback scores of the written feedback for the entire sample.

The average quarterly feedback scores over time, for both the numerical public feedback, and the feedback predicted from the written feedback, are plotted in Figure 3b. As expected, the two 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.

The divergence between written sentiment and numerical feedback implies that a substantial amount of the increase in numerical feedback scores is due to lower rater standards. Our approach also allows us to quantify the degree of inflation: the point estimate is that 67.7% of the increase in feedback scores is due to inflation.\(^\text{11}\)

Importantly, written feedback can certainly become “inflated,” with work that would have elicited a “good” now garnering a “great.”\(^\text{12}\) We have some evidence that written feedback does inflate, in that the private feedback scores were declining while the sentiment of written feedback was increasing (see Figure 3). Regardless, to the extent that written feedback is also subject to inflation, our approach will underestimate the magnitude of the inflation in scores, and hence our estimates can be interpreted as lower bounds.

Our approach requires 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.

A potential threat to our approach 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 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, find all lexically identical sentences generated in these periods, and compare average feedback by sentence, across the two periods. To illustrate our approach, Figure 4 shows the average numerical feedback scores for a set of commonly used short sentences, by period. Across terms, we see that the numerical feedback scores associated with identical sentences have increased considerably over time, and this increase has affected both positive and negative sentences.

\(^{11}\)Employing different training periods and/or predictive algorithms yields similar results (see Appendix C).

\(^{12}\)One written feedback in our data reads: “This is the most impressive piece of coding in the history of software development!”
Figure 4: Difference over time in the feedback scores associated with identical sentences.

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

4 Causes of reputation inflation

We now examine the causes of reputation inflation. Recall from Section 3.1 that employers were more candid about bad performance in private than in public. A compelling explanation is that raters incur a cost for assigning “bad” feedback, which increases in the cost of the rated party from receiving this “bad” feedback. In the focal market, this “reflected” disutility includes the employer’s aversion to harming the worker’s future prospects, the cost of the worker complaining, and even the cost from other workers avoiding an employer that has a “strict” rater reputation. In addition to “reflected” costs, if what feedback is “good” and “bad” is endogenous, i.e., depends on the current market feedback average, this “reflected” disutility can lead to reputation inflation.13 This would explain why public and private

13We present a model of reputation inflation based on these two primitives in Appendix D.
feedback scores diverged: “bad” public feedback had a positive cost to workers, and hence was costly for employers to assign, but “bad” private feedback had zero cost for workers, and hence zero cost for employers.

To test this hypothesis, we examine a platform change that raised the workers’ cost of “bad” private feedback from zero to some positive amount. In March 2015, the platform started using private feedback ratings to compute a new aggregate feedback score for workers. This score was publicly available, shown on workers’ profiles, but anonymous in that workers could not associate individual scores with employers: the aggregate score on a worker’s profile was updated after the worker received five new feedback scores, to prevent workers from identifying which employer gave them which feedback.

Insofar that employers used this new score in their hiring decisions, the workers’ cost of “bad” private feedback went from zero to some positive number. Crucially, if employers simply do not want to harm the worker’s future prospects, then even the batched release of private feedback would cause private feedback scores to inflate. Of course, an over time private feedback increase does not prove inflation; for example, if employers could form better matches because of their access to the private feedback score, then we would expect higher future private feedback scores because of improvements in “fundamentals.” As before, we address this concern by using written comments to construct an alternative measure of employer utility.

Figure 5 plots the monthly average private feedback, and the monthly average predicted private feedback, using the approach described 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 shaded region indicates the training period for the model used to predicted private feedback from written feedback. The vertical dashed line indicates the day the revelation took place. Prior to public revelation, actual and predicted private feedback are quite similar, but after the revelation, the actual numerical rating increases while the predicted rating does not (this effect is highly significant—see Appendix E). This finding is robust across several specifications, providing empirical support for the “reflected” costs hypothesis.
Figure 5: 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. These specifications are described in each facet’s title, and more details can be found in Appendix E.
5 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 improvements in “fundamentals.” Data from four other marketplaces exhibit the same trend, suggesting that reputation inflation is widespread. We develop an approach to quantify inflation based on using alternative measures of rater satisfaction. A market intervention that increased the costs of negative feedback by making previously private feedback public, yielded data supporting the role of “reflected” costs as the root cause of inflation.

Reputation inflation is likely most acute in peer-to-peer platforms, such as online labor and sharing economy markets, where wage penalties for workers and employers’ “reflected” costs are high. Reasons for the higher worker cost of negative feedback include that feedback scores are often the sole signal of quality, workers are typically highly substitutable and have few transactions, and hence each rating is more consequential. As transactions are more personal, the “reflected” costs for employers are also likely higher. In contrast, when individuals assign feedback to products (e.g. movie reviews) there is likely no reflected cost, and inflation will be less acute. Indeed, numerical scores on such platforms are characterized by lower averages, a higher spread, and, in some cases, a decreasing temporal pattern (Cabral and Hortaçsu, 2010; Moe and Schweidel, 2012; Godes and Silva, 2012).

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 examined, namely whether the increase in grades is due to “fundamentals,” such as better student cohorts, or due to lower standards. This literature also considers some negative effects of inflation, such as that inflated grades seemingly reduce student effort (Babcock, 2010). Exploring the consequences of reputation inflation is an interesting next step.

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. Platforms could provide monetary incentives for users who generate feedback, as dissatisfied users often leave no feedback (Nosko and Tadelis, 2015), but this approach could be costly to implement. Platforms could also emphasize reviewers as performing a service for fellow consumers, or provide other incentives for “honest” reviews: Yelp employs mechanisms such as badges for top reviewers, and makes the feedback score distribution of each reviewer publicly accessible. Mandatory grading curves are often employed in non-digital reputation systems, although it is challenging to force a distribution.
in settings where buyers evaluate sellers as a “flow.”

Platforms already take steps to lower “reflected” costs. 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 simultaneously revealed Airbnb reviews reduced the percentage of 5-star ratings by only 1.5%. Our analysis of the private feedback revelation quasi-experiment supports that “reflected” costs are due to raters incurring a greater personal cost—or guilt—the greater the harm they impose on the rated worker.

---

14Senior US Army raters are limited to rating fewer than 50% of officers as “most qualified.”
References


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 6. 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. 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.

Furthermore, 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. Furthermore, supplier capacity is highly constrained in these platforms, and it is hence unlikely that the same provider will be available in the future.
Figure 6: 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.
Figure 7: 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.
B Robustness tests for private feedback

B.1 Comparing private and public feedback

Employers assigned both public and private feedback for the same contract. Figure 8 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.

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.

B.2 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 8: 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.

### B.3 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 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.

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. Furthermore, 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.
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.

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.
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 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. Furthermore, since written feedback is positively biased, comparing the predicted scores from text versus the evolution of all scores gives us a lower bound for the degree of inflation.

Figure 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.

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.

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

\begin{figure}
\centering
\includegraphics[width=\textwidth]{figure13.png}
\caption{Percentage of employers leaving written feedback in addition to public numerical feedback.}
\end{figure}

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.

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 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 $\mathbb{E}[^{\epsilon}\mid U']$.

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 $r_T = \sum_{i \text{ left feedback in } T} r_i$. 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 \( r_t = \sum i \text{ left feedback in } T \land t r_i \), for any \( t \neq T \). If for \( t > T \) the quantities \( r_t \) 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 15. 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.

C.3 Alternative training periods

In the bottom panel of Figure 15 we perform the same empirical exercise as in Section 3.2, again plotting 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. 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. Reassuringly, the two corpora give similar results.

C.4 Predictive algorithm performance

We present more details about the performance of the algorithms used to extract the written feedback sentiment in Section 3.2.

Figure 16a plots the scatterplot of numerical scores versus predicted scores from written feedback for the algorithm trained on data coming from the earliest quarter. Figure 16b 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. Furthermore, 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 presented in Figure 4.
Figure 15: 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.
Figure 16: 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.
D 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. 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 3a. Private feedback scores are not observable by other employers when they are making hiring decisions, and hence receiving “bad” private feedback is less costly for workers. As a result, employers are more truthful in private, i.e., they assign lower private feedback scores, suggesting a link between the cost of receiving “bad” feedback and the cost of assigning “bad” feedback.

D.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

---

16Abeler 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.
utility equal to 1, by selling the output on some product market. The employer obtains zero utility in the case that output $y = 0$ is produced.

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

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

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

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

Whenever the employer tells the truth, that is when $s = y$, she obtains a benefit $b > 0$. 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 $\Delta w$, which is the wage penalty in the next round. We assume that some fraction of this cost is “reflected” back on the employer. 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 $F : [c, \bar{c}] \to [0, 1]$, with $\bar{c} \geq 0$.

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

$$b \leq c_i\Delta w.$$  \hspace{1cm} (A1)

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

\[
\Pr(q = q_H|s = 0; p) = \frac{\Pr(s = 0|q = q_H; p)\Pr(q = q_H)}{\Pr(s = 0; p)} \frac{(1 - q_H)\theta}{(1 - q_H)\theta + (1 - q_L)(1 - \theta)}.
\]

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

\[
\Pr(q = q_H|s = 1; p) = \frac{\Pr(s = 1|q = q_H; p)\Pr(q = q_H)}{\Pr(s = 1; p)} \frac{(q_H + (1 - q_H)(1 - p))\theta}{(q_H + (1 - q_H)(1 - p))\theta + (q_L + (1 - q_L)(1 - p))(1 - \theta)}.
\]

The cost of bad feedback to a worker is then

\[
\Delta w(p) = w_{s=1;p} - w_{s=0;p} = \frac{\theta(1 - \theta)(q_H - q_L)^2}{k - pk^2}, \tag{A2}
\]

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 A2 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. Furthermore, 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

\[
p_E = F\left(\frac{b}{\Delta w(p_E)}\right), \tag{A3}
\]

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

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

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 \((p_E = 0)\).\(^{17}\) 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.

**D.2 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 \(l_0 = \theta_B [1 - F(b/\Delta w(p_0))]\) 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 A3 we obtain \(T(p_E) = p_E\). Since \(F\) is a cumulative distribution function, and \(\Delta w\) 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-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.

\(^{17}\)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.
E Details on the econometric analysis of the quasi-experiment

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 5 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 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.

As our interest is in the divergence between the public and private feedback scores, we switch our outcome to $\Delta s_i$, 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 $\Delta s$. 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

<table>
<thead>
<tr>
<th></th>
<th>Dependent variable:</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>∆s, (Actual - Predicted) Private FB Ratings</td>
<td>(1)</td>
<td>(2)</td>
<td>(3)</td>
</tr>
<tr>
<td>Post-revelation</td>
<td>0.133***</td>
<td>0.167***</td>
<td>0.138***</td>
<td>0.067***</td>
</tr>
<tr>
<td></td>
<td>(0.007)</td>
<td>(0.007)</td>
<td>(0.013)</td>
<td>(0.018)</td>
</tr>
<tr>
<td>Post × Month</td>
<td>0.011***</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.002)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Constant</td>
<td>−0.013*</td>
<td>−0.105***</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.007)</td>
<td>(0.006)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Less than 25 contracts for C and E</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>Employer FE</td>
<td>N</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
</tr>
<tr>
<td>Observations</td>
<td>899,842</td>
<td>537,640</td>
<td>537,640</td>
<td>537,640</td>
</tr>
<tr>
<td>R²</td>
<td>0.001</td>
<td>0.002</td>
<td>0.498</td>
<td>0.498</td>
</tr>
<tr>
<td>Adjusted R²</td>
<td>0.001</td>
<td>0.002</td>
<td>0.184</td>
<td>0.184</td>
</tr>
</tbody>
</table>

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 \( p \leq 0.001 : *** \).

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. 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 5, 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 $(10 − 9.1)/(0.011)/12 ≈ 7$ 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.