Reputation Inflation in an Online Marketplace

John J. Horton
Leonard N. Stern School of Business
New York University

Joseph M. Golden
Elance-oDesk and University of Michigan*

February 16, 2015

Abstract

Average public feedback scores given to sellers have increased strongly over time in an online marketplace. Changes in marketplace composition or improved seller performance cannot fully explain this trend. We propose that two factors inflated reputations: (1) it costs more to give bad feedback than good feedback when feedback is *public* because buyers fear retaliation and (2) this cost is increasing in the market's average feedback score. Together, (1) & (2) push the market towards an equilibrium where feedback is always positive, regardless of performance. To address this problem, the marketplace allowed and encouraged buyers to additionally give *private* feedback. This private feedback was more candid and more predictive of future worker performance. The marketplace experimentally revealed aggregate private feedback scores which influenced employers' hiring decisions.

JEL J01, J24, J3

1 Introduction

Market outcomes—such as who trades with whom and on what terms—are explained in part by marketplace reputations. In the early reputation literature, reputations were modeled as the private beliefs market participants had about each other (Kreps and Wilson, 1982; Greif, 1993). With the advent of electronic commerce, platforms needed to create trust among strangers and so they reified "reputation," often simply by collecting and then showing the average feedback ratings made by prior trading partners (Dellarocas, 2003). A large and growing literature has conclusively demonstrated the importance of these reputations within online marketplaces (Resnick et al., 2000).

Online reputations always exist within the context of some designed *reputation system* that consists of rules about how feedback is collected, aggregated and shown. Designs differ, but all systems generally have the same aim: to reduce adverse selection. The reputation system can also reduce moral hazard by motivating behavior that will lead to "good" feedback, such as high effort and trustworthy behavior. However, if online reputations matter enough to motivate good behavior, they also matter enough to motivate less welcome behaviors, such as begging, bribes and threats.

*Author contact information, datasets and code are currently or will be available at http://www.john-joseph-horton.com/.
Thanks to Richard Zeckhauser, Ramesh Johari, Aaron Sojourner for very helpful comments and suggestions. Thanks to the Elance-oDesk corporation—and Samir Lakhani in particular—for their assistance with this project. Helpful (and hopefully not inflated) feedback was received at the Crowdsourcing Seminar at Carnegie Mellon University at the School of Computer Science and the MIT Conference on Digital Experimentation.

The nature of the threatened retaliation depends on the context, but it is generally some cost placed on the rating party.[1] In online marketplaces with bilateral reputation systems, the platform itself gives the rated party a ready-made threat/bribe, namely to match whatever feedback they receive with the same feedback. When retaliation becomes possible, the link between past performance and online reputation can grow tenuous, and if incentives for truth-telling are weak, the likely outcome is universally positive feedback scores and hence useless, highly inflated reputations.

In many marketplaces the observed distribution of reputation scores seems implausibly rosy. For example, the median seller on eBay has a score of 100% positive feedback ratings and the *tenth* percentile is 98.21% positive feedback reports (Nosko and Tadelis, 2014). In any actual marketplace it is difficult to say definitively whether reputations are inflated. There is no "ground truth" that tells us what the distribution reputations "should" look like. While we cannot say much about bias in a static setting, we can say much more in a dynamic setting. If average feedback is growing more positive—but there has been no change in marketplace composition or the attributes of transactions—then it seems likely that the informativeness of the reputation system is eroding.

In this paper, we document substantial "reputation inflation" in an online labor marketplace, oDesk, with average seller (worker) feedback scores increasing strongly over time: from the start of 2007 to mid-2014, average feedback scores on completed contracts increased by about one "star" (on a 1-5 star scale). To put this increase in perspective, in 2007, 28% of contracts ended with a feedback score of less than 4, whereas in 2014 this percentage had dropped to 9%. One strong piece of evidence that this is not merely an oDesk phenomenon is that we find a nearly identical pattern in the monthly feedback from Elance, a similar online labor market.[2]

We show that reputation inflation is not wholly explained by changes in marketplace composition, even though we would expect bad sellers (or hard-to-please buyers) to exit the marketplace: only half of this increase can be explained by composition. Instead, we believe that two factors are inflating reputations. First, giving negative feedback is more costly to the rater than giving positive feedback in part because the rated party can retaliate.[3] Second, what is considered "bad" feedback (and hence what prompts retaliation) depends upon the market penalty associated with that bad feedback. Together, these factors can create a ratchet-like dynamic of ever-increasing reputations. We formalize this argument in a simple adverse selection model and show that there exists a stable equilibrium in which sellers only report good feedback and reputations are thus universally inflated.

Theoretical arguments about inflation aside, oDesk believed it had a problem with inflated reputations. In response, oDesk instituted a new experimental "private feedback" system in which buyers and sellers privately reported on their experiences in addition to their status quo public feedback. oDesk began collecting private feedback from employers on May 9th, 2013. This private feedback was collected at the same time as the status quo public feedback, giving us a dataset of public and private feedback for the same completed job. For the new private feedback, employers were asked "Would you hire this

[1] As customers have had more opportunities to express and disseminate commercial opinions, there seems to be an increase in lawsuits designed to muzzle negative opinions. The strategy is commonplace enough that it has an acronym in the legal community, SLAPP, or "strategic litigation against public participation," to describe the practice of companies trying to deter bad feedback. For example, see "Venting Online, Consumers Can Find Themselves in Court", New York Times, May 31, 2010.

[2] Elance merged with oDesk in early 2014 but was previously operated on its own. The merger gave us access to the detailed micro-data on feedback scores needed to make the comparison. Recent work by Zervas et al. (2015) in the Airbnb context—which also has a kind of bilateral feedback system—documents that reputations are surprisingly high, though their data is cross-sectional rather than longitudinal.

[3] The existence of the expression "don't shoot the messenger" is some evidence for our claim that giving bad feedback is more costly than giving good feedback. Further, no student in the history of higher education has scheduled office hours to complain about receiving an A+.

freelancer [worker] again, if you had a similar project?" with multiple choice responses of "Definitely yes", "Probably yes", "Probably not" and "Definitely not." Unsurprisingly, buyers that claimed a good experience privately overwhelmingly claimed the same publicly. However, buyers that claimed a bad experience privately still gave the highest possible public feedback nearly 20% of the time.

Comparing the public and private feedback reveals a number of facts. First, the two scores are highly correlated, but the private score shows less top-censoring compression and far more examples of mildly negative sentiment. The private feedback scores contains more performance-relevant information than public feedback alone: a worker's initial private feedback is more predictive than their initial public feedback of both their next public and private feedback by some other employer. Using a sample of textual feedback comments and numerical public scores, we fit a model that predicts public feedback scores based on the text of the feedback comments. When the predicted public score is lower than the actual feedback, the private feedback was considerably more negative. This indicates that at least some of the negative sentiment captured by private feedback "shows up" in the text of the public feedback.

On April 29th, 2014, oDesk experimentally introduced a feature that revealed aggregated private feedback about applicants to some employers posting new openings. oDesk revealed aggregated private feedback as the percentage of previous employers who left private feedback that selected the top two private feedback tiers ("Definitely yes" and "Probably yes"). For example, a treated employer reviewing an applicant would see a notice on the applicants's profile that "85% [of past employers] would hire again," if that applicant had had enough past private feedback scores to aggregated them anonymously. Control employers received the status quo experience that made no mention of private feedback.

We find strong evidence that employers use this revealed information when deciding whose applications to review, whom to interview and whom to hire. The main effect seems to be that employers avoided hiring workers with bad private feedback scores, with no overall increase in hiring. There is no strong evidence that the intervention improved contract outcomes, but given that the overall change in the private feedback score of hired applicants was small, "downstream" effects on match outcomes were unlikely to be detected due to low power. There is evidence that applicants that were not eligible to receive a private feedback score on their profile (they had too few prior private feedback ratings) were less likely to be hired when employers could see private feedback scores of other applicants. This kind of crowd-out is troubling, as it could further raise the barriers for entry level workers, which Pallais (2013) shows are already high in this setting.

Our experimental results support our hypothesis that costs drive inflation: when negative feedback costs are reduced—namely by allowing buyers to give feedback quasi-anonymously—we get more of it. Through the experimental validation, we show that buyers act upon the information contained in aggregated private feedback, which implies that they (correctly) believe that aggregated private feedback is informative, even conditioned on the publicly available signals. Interestingly, this result implies that employers already appreciate the biased nature of public feedback.

This paper is the first to directly document strategic rating behavior in an online market by showing that privately observed experiences frequently do not match publicly shared statements. By collecting both private and public statements about the same transaction, we can cut right to the heart of the matter. Our paper is not the first to explain how reputations can be biased (Dellarocas and Wood, 2008), but we believe it is the first to explain how individually rational choices about what feedback to leave can push the market towards an entirely uninformative equilibrium that cannot be fixed through statistical correction. However, the paper also has a silver lining of sorts in that the solution we present—aggregating private feedback into public scores—has potentially wide application, as it reduces the underlying reason why negative feedback is scarce.

The paper is organized as follows: Section 2 provides the empirical context, describing the current oDesk marketplace. Continuing this institutional focus, Section 3 delves into the status quo reputation system on oDesk and its quantitative characteristics. Special attention is paid to documenting the reputation inflation on oDesk and ruling out the possibility that other explanations completely describe the phenomena. Section 4 introduces an adverse selection model of buyer rating behavior in a labor market. Section 5 describes the relevant existing literature on reputation systems. In Section 6, the private feedback collection intervention is described and its informativeness—i.e., its ability to predict future market outcomes—is compared to the status quo public feedback. Section 7 describes the experimental intervention in which future employers in the treatment group were shown the aggregate private feedback scores of their applicants. Section 8 concludes.

2 Empirical context

During the last ten years, a number of online labor markets have emerged. In these 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 platforms include maintaining job listings, hosting user profile pages, arbitrating disputes, certifying worker skills and maintaining reputation systems. On oDesk, would-be employers write job descriptions, self-categorize the nature of the work and required skills and then post the vacancies to the oDesk website. Workers learn about vacancies via electronic searches or email notifications.

Workers submit applications, which generally include a wage bid (for hourly jobs) or a total project bid (for fixed-price jobs) and a cover letter. In addition to worker-initiated applications, employers can also search worker profiles and invite workers to apply. After a worker submits an application, the employer can interview and hire the applicant on the terms proposed by the worker or make a counteroffer, which the worker can counter, and so on. The process is not an auction and neither the employer nor worker are bound to accept an offer.

To work on hourly oDesk contracts, workers must install custom tracking software on their computers. The tracking software, or "Work Diary," essentially serves as a digital punch clock that allows for remote monitoring of employees. When the worker is working, the software logs the count of keystrokes and mouse movements; at random intervals, the software also captures an image of the worker's computer screen. All of this captured data is sent to the oDesk servers and then made available to the employer for inspection. This monitoring makes hourly contracts and hence employment relationships possible, which in turn makes the oDesk marketplace more like a traditional labor market than project-based online marketplaces where contracts are usually arm's-length and fixed price.

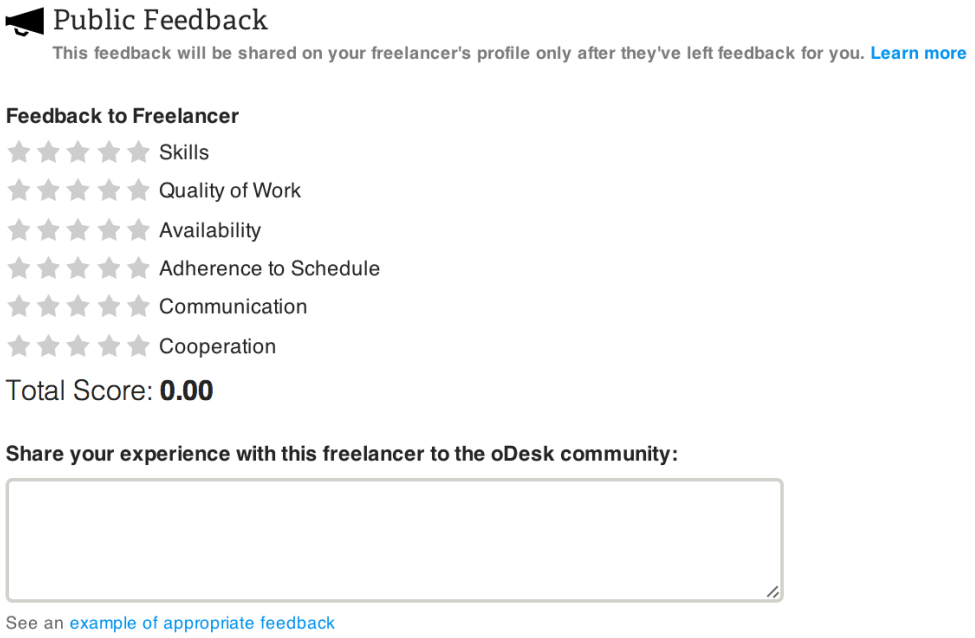
In the first quarter of 2012, $78 million was spent on oDesk. The 2011 wage bill was $225 million, representing 90% year-on-year growth from 2010. As of October 2012, more than 495,000 employers and 2.5 million workers have created profiles (though a considerably smaller fraction are active on the site). Approximately 790,000 vacancies were posted in the first half of 2012. See Agrawal et al. (2013a) for additional descriptive statistics on oDesk.

There has been some research which focuses on the oDesk marketplace. Pallais (2013) shows via a field experiment that past worker experience on oDesk is an excellent predictor of being hired for subsequent work on the platform. Stanton and Thomas (2012) use oDesk data to show that agencies (which act as quasi-firms) help workers find jobs and break into the marketplace. Agrawal et al. (2013b) investigate what factors matter to employers in making selections from an applicant pool and present some evidence of statistical discrimination, which can be ameliorated by better information.

3 Status quo reputation system on oDesk

On oDesk, when one party ends a contract—customarily the employer—both parties are prompted to give feedback. Feedback includes both a short written portion, 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 quantitative feedback. The employer-on-freelancer quantitative feedback is given on several weighted dimensions—"Skills" (20%), "Quality of Work" (20%), "Availability" (15%), "Adherence to Schedule" (15%), "Communication" (15%) and "Cooperation" (15%). Figure 1 shows the current public feedback interface used on oDesk for employers rating workers. The weighted mean of the various dimensions generates an overall feedback score for the completed project. These project-level feedback scores are then aggregated to give a total feedback score, which is a dollar-weighted mean. On the worker profile, a lifetime score is shown as well as a "last 6 months" score. Showing recent feedback is presumably the platform's response to the opportunism that becomes possible once a buyer or seller has obtained a high, hard-to-lower reputation (Aperjis and Johari, 2010; Liu, 2011).

Figure 1: Public feedback interface



Notes: This is the interface presented to employers for giving public feedback at the conclusion of a contract.

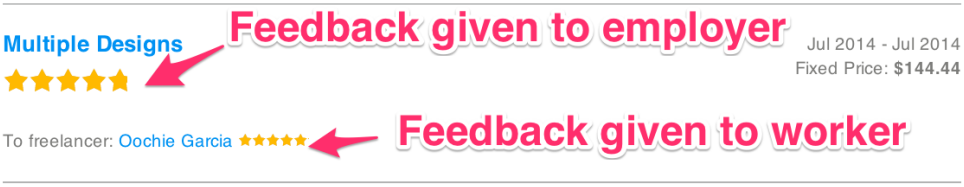
Both buyer and seller have an initial 14 day "feedback period" in which to leave feedback. oDesk does not reveal public feedback immediately. Rather oDesk uses the following "double-blind" process. If both parties leave feedback during the feedback period, then oDesk reveals both sets of feedback simultaneously. If instead, only one party leaves feedback, then oDesk reveals it at the end of the feedback period. Thus, neither party learns its own rating before leaving a rating for the other party. Once either both parties have left each other feedback, or the feedback period has elapsed, neither party can enter or revise the feedback they have left without permission of the other party. Despite this seeming "fix" to prevent tit-for-tat feedback, there is nothing to stop parties from engaging in "pre-play" communication

about their intentions. [4]

We have some evidence that feedback manipulation occurs, from forum and blog postings, communication between buyers and sellers, and complaints directly to oDesk, but it is 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 oDesk employers found that 20% had felt pressure to leave more positive public feedback. Feedback is not compulsory, though it is strongly encouraged. Of employers eligible to leave feedback, 16% do not leave feedback, while 8% of workers do not leave feedback on employers.

Average numerical feedback—weighted by the dollars spent on the related contract—is shown on worker profiles. The entire feedback history is also available to interested parties. Workers can also view the feedback given to previous workers evaluated by that employer and the feedback received by an employer from that same worker. Figure 2 shows an example of what a freelancer [worker] can view about an employer's past relationships.

Figure 2: Public view of the feedback given to and given by an employer



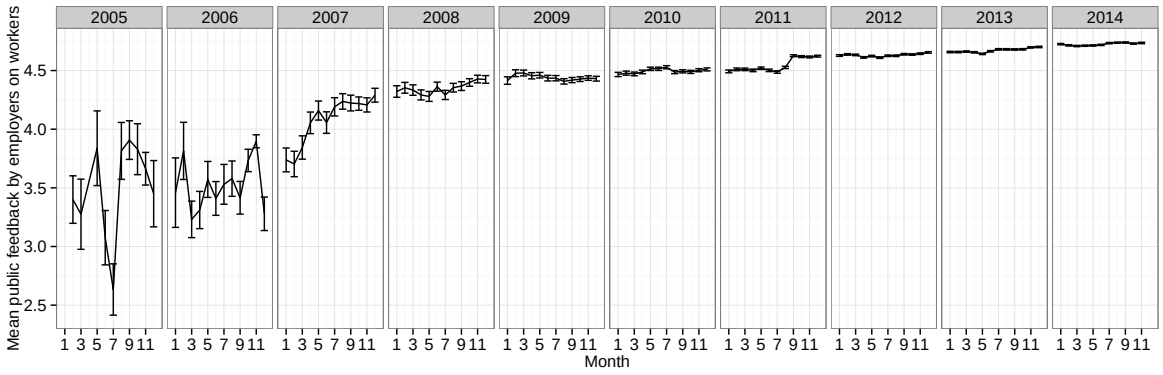
Notes: This image is screen-shot of an employer's feedback history, which is publicly viewable. An employer would have an entry like this one for every contract. It shows not only the feedback they gave to the evaluated worker, but also the feedback they received.

3.1 Dynamics and current distribution of employer-on-worker feedback scores

On oDesk, the average feedback scores given to workers have risen over time. Figure 3 shows the historical monthly average feedback of employers on workers. For each month, the mean is shown, as well as a 95% confidence interval. We can see that in the early years of the platform, the average fluctuated a great deal, as the total number of completed contracts was small. However, over time, the number of feedback scores per month increased and average feedback grew more stable month to month. There is a strong positive trend over time, with the greatest period of growth occurring in 2007. Elance is a similar online labor market, with a similar public reputation system, which merged with oDesk in 2014. Figure 4 shows that historical public feedback given to workers on Elance follows all of the same patterns as on oDesk, though the period of rapid inflation occurred later.

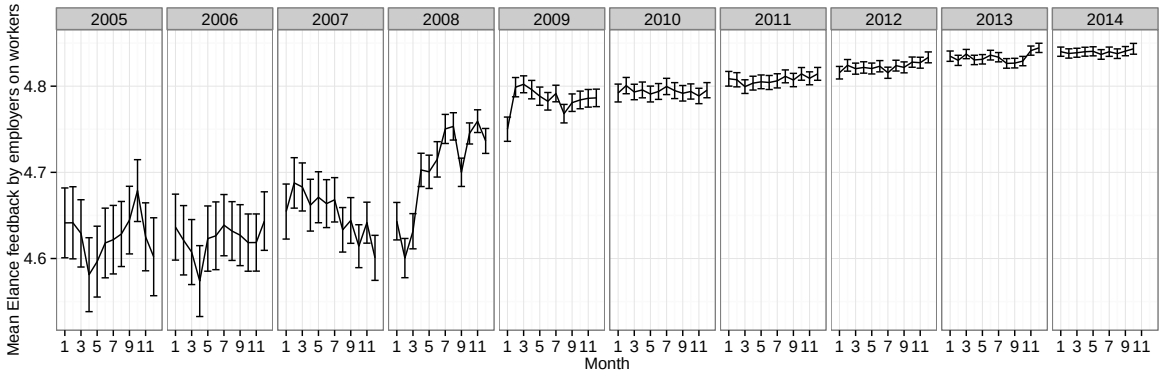
[4] Each party can independently grant the other party permission to change their feedback once per job, once the "feedback period" ends. The feedback changing process's primary purpose is to give buyers and sellers the opportunity to directly work out their own problems. However, the feedback changing process is not "double-blind", so buyers and sellers have more opportunity to strategically manipulate the feedback changing process than the initial feedback leaving process. In particular, one party could coerce the other into improving their feedback, mildly by, for example, asking or begging for good feedback, or aggressively by, for example, withholding work until the buyer leaves good feedback. Additionally, one party may offer to raise the feedback they have left, but only if the other party does so first. There are two ways that feedback can be removed. First, oDesk can remove feedback ratings that violate site policies, such as feedback that is manipulated. Second, if a seller issues a complete refund for a job, oDesk removes the feedback.

Figure 3: Monthly average public feedback scores on oDesk to workers over time



Notes: This figure shows the average monthly feedback given by employers to workers for contracts on oDesk ending that month. For each point observation, a 95% confidence interval is shown. Contracts for which no feedback was left are excluded.

Figure 4: Monthly average public feedback scores on Elance to workers over time



Notes: This figure shows the average monthly feedback given by employers to workers for contracts on Elance ending that month. For each point observation, a 95% confidence interval is shown. Contracts for which no feedback was left are excluded.

An upward trend in feedback is only *consistent* with reputation inflation. It is possible that the pool of workers is getting better over time as poor-performing workers exit.

3.2 Is the positive trend in feedback score caused by worker and employer compositional changes?

Over time, the composition of workers and employers in the marketplace could change: we would expect worker quality to improve if "bad" workers exit the marketplace. We could also imagine that employers that are relatively hard to please would also exit. Together, these two trends could yield a "good" workforce and easy-to-please employers, the upshot of which would be legitimately high feedback.

We can test this "composition" hypothesis for reputation inflation by testing whether we still see a positive time trend in feedback when controlling for the identity of the hired worker and hiring firm.

As we know the entire hiring and feedback history of everyone in the marketplace, we can use the same econometric techniques pioneered by Abowd et al. (1999) for working with matched employer-employee datasets. In Table 1, in Column (1) we first report an OLS estimate of

$$\text{PubFB}_{ijt} = \beta_0 + \beta_1 t + \epsilon \tag{1}$$

where PubFB_{ijt} is the public feedback received by worker i on contract j at time t. The sample is restricted to workers with at least 10 completed contracts on oDesk. In Column (2), we estimate

$$\text{PubFB}_{ijt} = \beta_0 + \beta_1 t + c_i + c_j + \epsilon \tag{2}$$

where c_i and c_j are employer- and worker-specific fixed effects and t is the relative year in which the feedback rating was made.

As expected, in Column (1) of Table 1, the coefficient on the evaluating year is positive and highly significant. When we add the worker- and firm-specific effects in Column (2), the coefficient on t goes down, but it is still positive and highly significant.[5] There is still a substantial increase in reputation scores even when controlling for composition. However, this conclusion depends on the assumption that workers are not improving over time.

Although we cannot add a worker-specific time regressor (as it would be co-linear with the overall time regressor), we can include a term for worker experience, exper_{it}, which is the count of previous projects by worker i at time t. As this variable is highly skewed—the max in the sample is more than 800 completed projects—we take the log of previous projects by worker i and add it as a regressor. In Column (3), Table 1 we report an estimate of

$$\text{PubFB}_{ijt} = \beta_0 + \beta_1 t + \beta_2 \log\left(\text{exper}_{it}\right) + c_i + c_j + \epsilon. \tag{3}$$

We can see that the coefficient on experience is actually *negative*, suggesting that the reduction in the time trend from Column (1) to Column (2) actually over-stated the reduction in feedback due to compositional effects.

The within-worker negative relationship between experience and feedback might seem counter-intuitive, but at least two mechanisms could be at play. First, a long-held concern with reputation systems is that once a reputation is established, the holder of that reputation has an incentive to "defect," doing less work or bearing fewer costs since the reputational penalty from doing so is not so great. Cabral and Hortaçsu (2010), examining the dynamics of reputation on eBay from a constructed panel of sellers, find that a worker's current reputation score seemingly affects their behavior. Sellers that know they are about to leave the marketplace seem to provide worse service. While this is slightly different than free-riding on a good reputation, it does suggest that effort is endogenous with respect to feedback score. Second, workers are not exogenously given jobs with set wages and so we might expect that as workers improve or gain experience, they might try more demanding jobs and/or higher-paying jobs in which pleasing the employer is more difficult.

3.3 Current distribution of feedback scores

The positive trend in average feedback to sellers on oDesk has resulted in a distribution of feedback scores with a heavy right tail. Figure 5 shows the distribution of public feedback by employers on workers from May 9th, 2013 to June 2th, 2014. The distribution is highly skewed, with slightly more than 80% of evaluations being in the 4.75 to 5.00 star bin. However, there is some weight in the lowest bin, which contains the fraction of observations with exactly 1.00 star, the lowest possible rating.

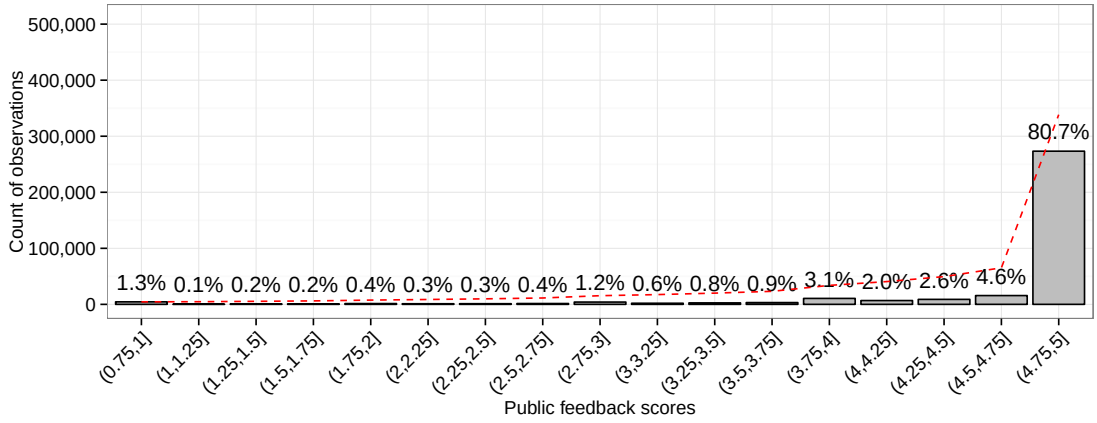
[5] We estimate the double-fixed effects model using the software developed by Gaure (2013).

Table 1: Estimates of the effect of relative evaluation date on oDesk public feedback score (1 to 5 star scale) on workers, with and without controls for marketplace participant composition

	Dependent variable:		
	Employer feedback on worker (5-star scale):		
	(1)	(2)	(3)
Years from start of oDesk	0.056^{***}	0.020^{***}	0.042^{***}
	(0.0004)	(0.001)	(0.002)
Log worker experience			-0.031^{***}
			(0.001)
Constant	4.203^{***}		
	(0.004)		
Worker and Firm FE?	No	Yes	Yes
Observations	1,707,647	1,707,647	1,707,647
R^2	0.010	0.422	0.422

Notes: The dependent variable in both regressions is the public numerical feedback score given to a worker by the employer that hired them, at the conclusion of the contract. The sample consists of all completed contracts on oDesk, as of December 1st, 2014, by workers with 10 or more contracts. The key independent variable in both regressions is the year of the evaluation, relative to the start of the oDesk marketplace. Column (1) reports an ordinary least squares regression, whereas Column (2) has worker and employer fixed effects. As such, the Column (2) regression is intended to control for changes in marketplace composition. For Column (2), standard errors are clustered at the employer level, whereas in Column (1) conventional standard errors are shown. Significance indicators: $p \leq 0.05 : *$, $.p \leq 0.01 : **$ and $p \leq .001 : ***$.

Figure 5: Distribution of public feedback to workers



Notes: This figure shows—for each bin—the count of completed assignments receiving a public feedback score from the employer falling in the bin. Public feedback scores are between 1 and 5 stars, inclusive. 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. Each bar is labeled with the percentage of total observations in that bin. This sample consists of all completed contracts from May 9th, 2013 to June 2th, 2014 for which the employer provided feedback.

4 An adverse selection model of reputation

To explore the incentives created by a reputation system, we develop a simple model of an employer giving public feedback to workers in a competitive labor market. Workers have a type q. They are matched at random with employers, after which they produce output $y \in \{0,1\}$, with $\Pr(y=1|q)=q$ which they can sell in the product market for \$1. The firm observes y and then gives a signal, i.e., gives feedback, to the marketplace, $s \in \{0,1\}$. The firm gets a benefit $b>0$ whenever $s=y$, that is when they tell the truth. However, when $y=0$ and the firm reports $s=0$ it pays a cost, $c(\Delta w)$, where Δw is the difference in a worker's expected productivity—as inferred by some future employer—when they receive bad feedback versus good feedback. Firms choose a strategy p, which is their probability of truthfully reporting $s=0$ when $y=0$. Note that the firm never has an incentive to report $s=0$ when $y=1$. Let p_e be the equilibrium choice by all other firms.

There are two worker types: q_H and q_L, with $q_H > q_L$. The fraction of high-types in the market is θ, which is common knowledge. Future employers only observe the most recent feedback from a worker. Both sides are price-takers and so workers are just paid their expected marginal product, which is simply

$$w = \Pr(q=q_H|s)q_H + \left(1-\Pr(q=q_L|s)\right)q_L. \tag{4}$$

If $s=0$, then a Bayesian firm infers that

$$\Pr(q=q_H|s=0) = \frac{\theta(1-q_H)}{\theta(1-q_H)+(1-\theta)(1-q_L)} \tag{5}$$

(note that the p_e term divides out) and when $s=1$,

$$\Pr(q=q_H|s=1) = \frac{\theta q_H}{\theta q_H + (1-\theta)q_L + (1-p_e)\theta(1-q_H) + (1-p_e)\theta(1-q_L)}. \tag{6}$$

The expected wage difference from good and bad feedback at the equilibrium level of truth-telling, p_e, is thus

$$\Delta w = \mathbf{E}[w|s=1] - \mathbf{E}[w|s=0] \tag{7}$$

$$= \frac{(q_H - q_L)^2(1-\theta)\theta}{\big(1-p_e(1-\bar{y})\big)\big(1-\bar{y}\big)}. \tag{8}$$

where $\bar{y} = \theta q_H + (1-\theta)q_L$. We can see that $\Delta w > 0$ for all p_e, implying that there is always a cost to the firm of giving bad feedback, which they must compare to b, the benefit.

An evaluating firm chooses a best response function $\phi(p_e)$, which is their probability of reporting $s = 0$ when $y = 0$, i.e., telling the truth:

$$\phi(p_e) = \begin{cases} 1 & : b > c(\Delta w(p_e)) \\ 0 & : b < c(\Delta w(p_e)) \\ p' & : b = c(\Delta w(p_e)) \end{cases} \tag{9}$$

where p' is any p in $[0,1]$. There are three equilibria: $p_e = 0$, $p_e = 1$ and $p_e = p|b = c(\Delta w(p))$. The first two equilibria are stable. The $p_e = 0$ equilibrium is the all-lying equilibrium in which all firms report positive feedback regardless of performance. The $p_e = 1$ equilibrium is the all-truthful equilibrium in which all firms tell the truth. In the all-lying equilibrium, if a small number of firms begin telling the truth, so long as $p_e < p'$, the individual firm's best response does not change. Similarly, in the all-truth-telling equilibrium, if a small number of firms begin lying, so long as $p_e > p'$, the individual firm's best response is still to tell the truth. The mixed strategy equilibrium, $p_e = p|b = c(\Delta w(p))$, is unstable, as a small change in another firm's strategy would tip the individual firm towards always telling the truth or always lying, depending on the direction of the perturbation.

Among the two pure strategy equilibria, is there any reason to believe one or another is more likely? Both are clearly possible and the size of the benefits and costs matter, but it is interesting to note that in the all-truth-telling $p = 1$ equilibrium, the cost of bad feedback—and hence the incentive for workers to impose costs following bad feedback on employers—is higher than in the all-lying equilibrium. Let $\Delta w(p)$ be the equilibrium loss when all firms choose p_e. We can see that

$$\frac{\partial \Delta w}{\partial p} = \frac{(q_H - q_L)^2\theta(1-\theta)}{(1-p(1-\bar{y}))^2} > 0, \tag{10}$$

which suggests that the more common truthful feedback is in the marketplace, the stronger pressure firms face to lie when receiving bad performance.

5 Prior work

Dellarocas (2003) provides an overview of how online marketplaces use feedback-based reputations. Although the examples in this paper are dated, the basic framework is not. These reputations are presented as "scores" which are potentially informative to future buyers and sellers, who must consider the context of some reputation system—which sets the rules for how and when feedback is collected and how that collected feedback is transformed into a reputation score.

As reputation systems and electronic commerce proliferated, two related literatures emerged: one that focused on measuring the importance of reputation in such markets (Luca, 2011; Resnick et al., 2006) and another focused on their biases and limitations. One source of bias is non-response, with

participants with "extreme" views being more likely to rate. Dellarocas and Wood (2008) shows that those that are mildly disappointed are far more likely to stay silent.[6] Another more invidious source is outright fraud—Mayzlin et al. (2014) and Luca and Zervas (2013) provide evidence of fake reviews and the economic conditions that encourage fake reviews. Another kind of "fake" review occurs when the rating party gives a feedback score for some reason other than actual performance, for example, because they were threatened or bribed.

Despite the concern about fake reviews, these reputation systems have proven critical in solving informational problems in online markets: a large empirical literature documents the importance.[7] In matching markets—or in product markets where buyers and sellers care about who precisely they trade with—reputation is particularly important. As more of economic life becomes computer-mediated, online or "algorithmic" reputation will presumably grow in importance (Varian, 2010). While the importance of reputation is well-documented in the literature, comparatively less has been written about the economic incentives parties have when leaving feedback, and how these incentives in turn affect the dynamics and functioning of the reputation system. Leaving feedback has been characterized as a public goods problem, but the focus has been on the incentive to leave any feedback at all—not the strategic nature of that feedback (Bolton et al., 2005).

An exception to this characterization is Bolton et al. (2013), who present theoretical and empirical evidence that the original design of eBay's public reputation system, which was developed between 1996 and 2007, made it easy for parties to engage in reputation extortion and/or generate low-information tit-for-tat evaluations and that these problems were likely substantial in practice. Later, eBay substantially reduced the importance of bilateral feedback and the scope for tit-for-tat evaluations.[8]

The eBay solution of eliminating or at least substantially weakening bi-lateral feedback is not universally applicable. eBay could make the changes described above in part because improved online payment technology made eBay sellers more or less indifferent over their precise counter-party. However, a unilateral feedback system is not a general solution: in some markets, sellers inherently have preferences over buyers, particularly when the seller uses (or abuses) the owner's capital (e.g., Airbnb, RelayRides) or when what is being produced is a collaborative effort, such as in online labor markets (e.g., Elance-oDesk). Soujourner et al. (2014) make this point convincingly, showing experimentally that workers on Amazon Mechanical Turk, an online labor market, rely on the feedback ratings of buyers to decide whom to work for. In these kinds of marketplaces bilateral feedback systems are commonplace.[9] Even Uber, which entered an industry where buyers were formerly anonymous and had no reputations, uses a bilateral system, as buyer/rider attributes such as promptness and sobriety matter.

[6] Interestingly, Nosko and Tadelis (2014) essentially turn the insights from Dellarocas and Wood (2008) into a new feature on eBay and then validate the basic conclusion: silence is associated with a bad experience and the extant reviews are positively biased.

[7] See, for example, Resnick et al. (2000), Resnick and Zeckhauser (2002), Resnick et al. (2006)

[8] This original system allowed both buyers and sellers to rate each other and see the other's evaluation before offering their own, and starting in 2005, allowed the two parties in a transaction to mutually agree to withdraw feedback for each other. Recognizing that on eBay, moral hazard is a much bigger problem on the seller side, in 2007 eBay implemented an additional, private, one-way feedback system called Detailed Seller Ratings, where buyers rate sellers after a transaction, but these ratings are anonymous and only presented in aggregate. Additionally, in 2008, eBay partially eliminated bilateral feedback in the public feedback system by allowing sellers to only either rate buyers positively, or not at all. Despite removing the ability for sellers to leave bad feedback for buyers, Nosko and Tadelis (2014) demonstrate in the eBay context that many buyers still fear giving bad feedback: many use an "if you can't say something nice, don't say anything at all" strategy, with transactions where the buyer left no feedback for the seller strongly indicating a negative buyer experience.

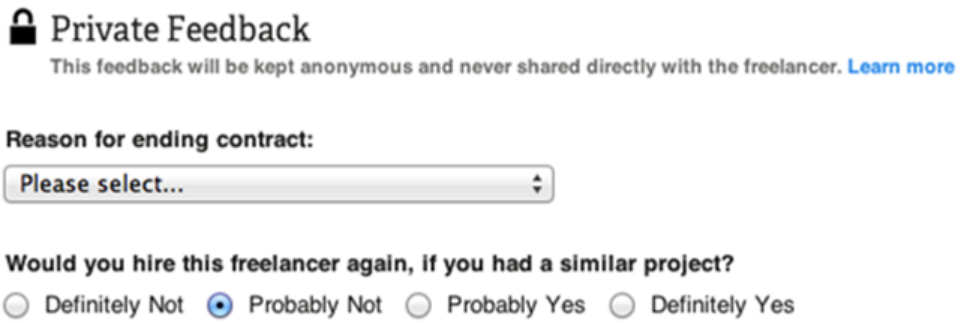
[9] "For Uber, Airbnb and Other Companies, Customer Ratings Go Both Ways", New York Times, December 2nd, 2014. Accessed online at `http://www.nytimes.com/2014/12/02/business/for-uber-airbnb-and-other-companies-customer-ratings-go-both-ways.html`

There is a fundamental asymmetry between leaving good feedback and leaving negative feedback: negative feedback is costlier to give than positive feedback. Although some buyers may decide to exercise their "voice" and complain after a bad experience (Hirschman, 1970), many other similarly situated buyers decide, perhaps through gritted teeth, to give positive feedback to a terrible seller: they do not want the hassle of retaliation; they worry about being sued for libel; they feel bad for the seller; they don't want to pay a premium to future trading partners (who would consider their proclivity to give bad feedback) and so on. The asymmetry in positive and negative feedback manifests itself in offline domains as well: Hirschman (1970) discussed this in the context of consumers facing the decline of service from traditional firms, and while firms would prefer "voice," consumers often choose "exit" instead. We also see this in job references, with prior employers often unwilling to give candid opinions about past employees. The penalty paid by laid-off employees (Gibbons and Katz, 1991) could presumably be remedied if firms thought they could get candid appraisals from past employers.

6 Collecting private feedback

In response to the ever-increasing public feedback and the presumed reduction in information from this inflation, in April, 2014, oDesk introduced a new experimental "private feedback" feature in which buyers/employers—in addition to giving public feedback—could also give private feedback. This private feedback was not shared with the evaluated worker, and was, at first, just collected by oDesk. Figure 6 shows the private feedback interface presented to employers. Employers answered the private feedback question, "Would you hire this freelancer [worker] again, if you had a similar project?", with four response options: "Definitely yes"; "Probably yes"; "Probably not" and "Definitely not".

Figure 6: Private feedback interface

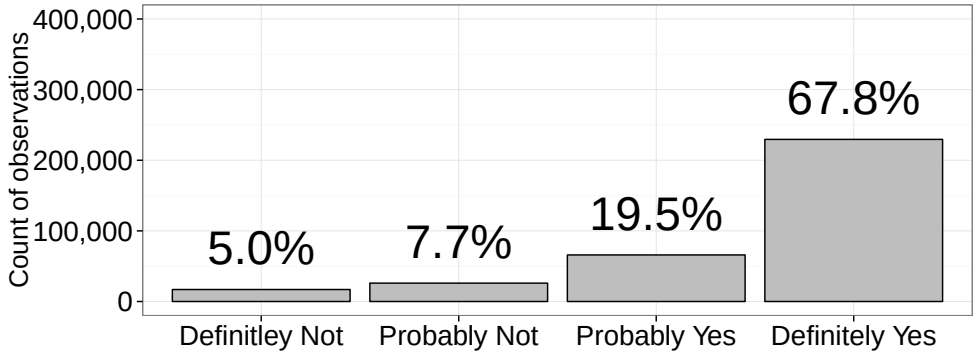


Notes: This is the private feedback interface presented to employers.

Figure 7 shows the distribution of responses to the private feedback question. Although the most common response was "Definitely Yes", there is a substantial fraction giving some form of negative feedback ("Definitely Not" and "Probably Not").

As employers gave both public and private feedback on the same contract, we can compare scores within an employer. Figure 8 shows the distribution of public feedback, by each of the four private feedback categories. As we would expect, those employers selecting "Definitely yes" also left very positive public feedback. For employers selecting "Definitely No", 23.4% did give only 1 star, but the second most common choice at 17.8% was in the 4.75 to 5.00 bin.

Figure 7: Distribution of private feedback—answers from employers to the question"Would you hire this freelancer [worker] again, if you had a similar project?"



Notes: This figure shows the count of buyers choosing each of the four private feedback options following the completion of a contract. The question the employer was asked was the first version of the private feedback question, "Would you work with this freelancer [worker] again, if you had a similar project?" For each bar, the percentage of responses is also shown.

6.1 Informational content of public and private scores

A natural question is whether this private feedback score just captures information already contained in the public feedback. One way to test this is to look at a chronologically ordered pair of a worker's contracts and see whether the private feedback on the first contract predicts their public or private feedback on the second contract. Doing this with the first and second contracts of workers with private and public feedback scores for both, we estimate

$$s_{i2}^{a} = \beta_0 + \beta_1 s_{i1}^{b} + \epsilon \tag{11}$$

where i indexes the worker and s_{i1} is the feedback on the first contract and s_{i2} is feedback on the second and a and b indicate the type of feedback (public or private, depending on the regression). We find that private feedback—rather than public feedback—is more predictive of subsequent private *and* public feedback.

Table 2 shows a regression of a worker's next *public* feedback score based on either their previous public or private score. We normalize all scores to have zero mean and unit standard deviation. In Column (1), we regress public-feedback on public-feedback. Unsurprisingly, prior feedback is strongly positively correlated with subsequent feedback. Given the distribution of realized feedback scores, the R^2 is unsurprisingly low.

In Column (2), the public score is still the outcome, but the regressor is the normalized private feedback score. To normalize, we assigned scores of 1, 2, 3 and 4 to the "Definitely No" to "Definitely Yes" scale. This public-on-private feedback has a higher R^2 and the coefficient on the private feedback score is larger than in Column (1). As public and private feedback scores are both normalized, coefficients are directly comparable. In Column (3), both scores are added as levels, while in Column (4) their interaction is added. The Column (4) interaction effect is positive and significant, suggesting that the public and private scores each contain information not fully captured by the other score.

In Table 3, the outcome measure is the worker's private feedback score on the next contract. In Column (1), the regressor is the public feedback score, while in Column (2) it is the private feedback score.

Table 2: Predicting a worker's next public FB from previous FB

	Dependent variable:			
	Worker's 2nd Public Feedback (FB), z-score			
	(1)	(2)	(3)	(4)
1st Public FB	0.135^{***}		0.062^{***}	0.129^{***}
	(0.009)		(0.012)	(0.022)
1st Private FB		0.150^{***}	0.110^{***}	0.108^{***}
		(0.009)	(0.012)	(0.012)
1st Private FB × 1st Public FB				0.032^{***}
				(0.009)
Constant	-0.111^{***}	-0.115^{***}	-0.114^{***}	-0.135^{***}
	(0.009)	(0.009)	(0.009)	(0.011)
Observations	16,910	16,910	16,910	16,910
R^2	0.013	0.016	0.017	0.018

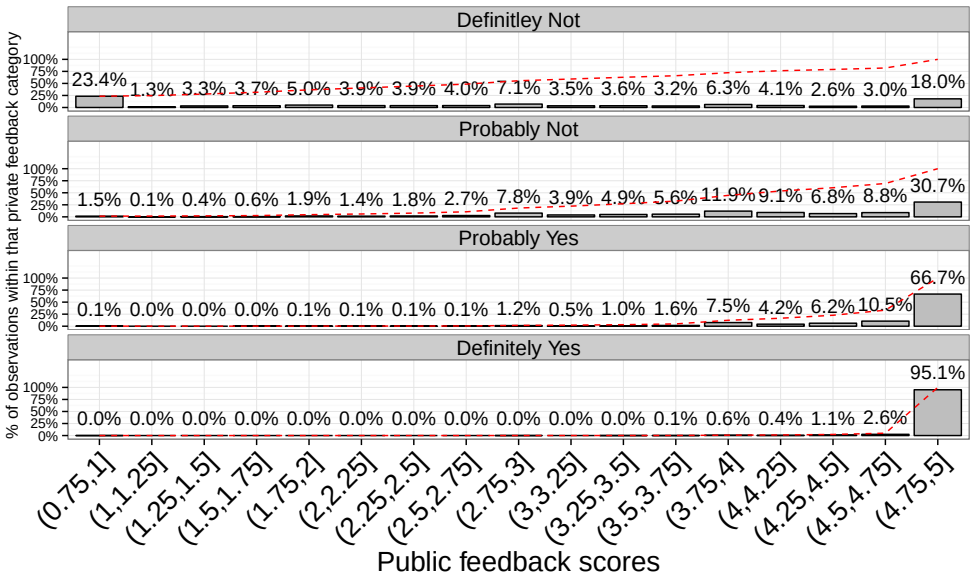
Notes: The dependent variable in each regression is a worker's second public feedback z-score. The dependent variables are either the 1st public feedback—in Column (1)—, 1st private feedback—in Column (2)—or both kinds of feedback as well as the interaction. The sample consists of all workers receiving at least two complete feedback scores since the introduction of. the private feedback system. All regressions are estimated using OLS. Significance indicators: $p \leq 0.05 : *$, $.p \leq 0.01 : **$ and $p \leq .001 : ***$.

Table 3: Predicting a worker's next private FB from previous FB

	Dependent variable:			
	Worker's 2nd Private Feedback (FB), z-score			
	(1)	(2)	(3)	(4)
1st Public FB	0.116^{***}		0.025^{*}	0.090^{***}
	(0.008)		(0.011)	(0.020)
1st Private FB		0.154^{***}	0.138^{***}	0.135^{***}
		(0.008)	(0.011)	(0.011)
1st Private FB × 1st Public FB				0.031^{***}
				(0.008)
Constant	-0.071^{***}	-0.075^{***}	-0.075^{***}	-0.096^{***}
	(0.008)	(0.008)	(0.008)	(0.010)
Observations	16,910	16,910	16,910	16,910
R^2	0.011	0.020	0.020	0.021

Notes: The dependent variable in each regression is a worker's second public feedback z-score. The dependent variables are either the 1st public feedback—in Column (1)—, 1st private feedback—in Column (2)—or both kinds of feedback as well as the interaction. The sample consists of all workers receiving at least two complete feedback. scores since the introduction of the private feedback system. All regressions are estimated using OLS. Significance indicators: $p \leq 0.05 : *$, $.p \leq 0.01 : **$ and $p \leq .001 : ***$.

Figure 8: Distribution of publicly given feedback to workers, by private feedback



Notes: This figure shows the distribution of public feedback scores, conditional upon the employer's private feedback score.

Comparing the two, we see that the private feedback score explains considerably more of the variation: the coefficient is about 40% larger in the private feedback case and the R^2 for the regression is nearly doubled. In Columns (3) and (4), we see again that the private feedback score tends to complement the public feedback score.

6.2 Does public written feedback convey private feedback?

As noted earlier, in addition to leaving numerical public feedback, employers can also leave written feedback. Obviously this written feedback is not placed on a scale, as is numerical feedback. However, we can fit a model that predicts the public feedback score, given the text of the written feedback. We can then see whether—when controlling for the public feedback score—workers that received relatively poor private feedback also received more negative written feedback.

Suppose we have a collection of written feedback texts. From these texts, we can construct a document term matrix: the rows correspond to the texts and the columns to the "terms." The indicated terms are those that are common across texts but are not "stop words" such as "I", "we", "am", "you" and so on which are removed because they are not informative about sentiment. The matrix entries are simply indicators for whether job i contains term j.

For example, suppose the terms were "excellent", "good" and "poor" and the three example texts were

- Job 1: "Robin did **excellent** work".
- Job 2: "He did an **good** job and his communication was excellent as well. I would work with Paul again."
- Job 3: "A **poor** performance all around."

The resultant document term matrix (with the ordering "excellent", "good" and "poor" for the columns)

would be

$$\mathbf{X} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}. \tag{12}$$

Using this matrix, we can write the simple linear regression

$$s_i = \beta_0 + \beta \cdot \mathbf{X} + \epsilon. \tag{13}$$

However, estimating this model in practice would yield a predictive model with poor performance due to over-fitting, as document term matrices are usually very large. For this reason, some kind of regularization is needed.

What we did empirically was to first create create a sample of 25,000 written feedback texts. The pre-regularization document term matrix contained 690 terms after the removal of stop-words and an initial pruning of sparse terms. We then used the Lasso model for regularization, picking tuning parameters with cross-validation (Friedman et al., 2009). A total of 277 terms survived the regularization. See Appendix C for estimation details.

Using this fitted model, we can measure to what extent the private information contained in the private feedback was already "there" in the public textual comments. To formalize this, using the full sample of contracts with private, public and written feedback, we can estimate the regression

$$s_z^{PRI} = \beta_0 + \beta_1 s_z^{PUB} + \beta_2 \hat{s}_z^{TXT} + \beta_3 \left(s_z^{PUB} \times \hat{s}_z^{PRI} \right) + \epsilon \tag{14}$$

where s_z^{PRI} is the private feedback z-score, s_z^{PUB} is the public feedback z-score and $\hat{s}_z^{TXT}$ is the *estimated* public feedback score, given the textual comments, i.e., the predictions from the regularized estimate from Equation 13. Table 4 contains estimates of this equation: in Column (1) the constraints $\beta_2 = 0$ and $\beta_3 = 0$ are applied, in Column (2) the constraint $\beta_3 = 0$ is applied and in Column (3), the full estimate is shown.

From Column (1), we can see that, as expected, the public feedback and private feedback scores are highly correlated. In Column (2), the *predicted* public feedback—based on the feedback text and fitted model—is added as a regressor. The positive and significant coefficient on the predicted public feedback, $\hat{\beta}_2$, shows that, conditional upon the public score given, some of the information conveyed by the private score is conveyed by the language of the textual feedback. In Column (3), the interaction between the public and predicted public feedback is added. The interaction term between the public and the predicted public feedback is positive and highly significant. This is what we would expect if employers gave candid feedback on all channels when the worker did well but at least some employers gave dishonest public feedback but honest private and textual feedback when the worker performed poorly.

6.3 Characteristics of employers with a candor gap

Employers presumably differ in their incentives to leave overly-positive feedback. For one-off employers not planning to use the marketplace again, leaving damaging negative feedback to a poorly performing worker is less costly. Given that negative feedback is damaging to a worker, they would avoid employers that seem prone to give negative feedback. On oDesk, a worker can see the whole employment history of an employer, including the public feedback scores given.

Table 4: Predicting private feedback from the qualitative, textual feedback

	Dependent variable:		
	Worker's Private Feedback (z-score)		
	(1)	(2)	(3)
Public FB (PF)	0.550^{***}	0.503^{***}	0.540^{***}
	(0.001)	(0.001)	(0.001)
Predicted Public FB (PPF)		0.083^{***}	0.111^{***}
		(0.001)	(0.001)
PF × PPF			0.020^{***}
			(0.0004)
Constant	0.000	0.000	-0.011^{***}
	(0.001)	(0.001)	(0.001)
Observations	717,432	717,432	717,432
R^2	0.303	0.307	0.309
Residual Std. Error	0.835 (df = 717430)	0.832 (df = 717429)	0.831 (df = 717428)

Notes: In this table, the observations are a random sample of completed oDesk contracts with public and private employer-on-worker feedback scores. In each regression, the dependent variable is the private feedback (z-score) received by the worker. In Column (1), the regressor is the 1-5 star public feedback for that contract. In Column (2), the same regression is run as in Column (1), but with the inclusion of the predicted public feedback score, with the prediction based on the text of the written feedback for that contract: the predictions come from a model trained on out-of-sample data. For details on the prediction mode, see Appendix C. Significance indicators: $p \leq 0.05 : *$, $.p \leq 0.01 : **$ and $p \leq .001 : ***$.

While we cannot exogenously manipulate whether an employer plans to be a long- or short-term use of oDesk, we can at least see if employer *future* experience is correlated with leaving overly-positive feedback. Table 5 shows how an employer's feedback on their *first* reviewed assignment is related to their total number of assignments. The outcomes are public feedback in Column (1), private feedback in Column (2).

We can see that public feedback is strongly increasing in the employer's future number of jobs posted. This could be mechanical in that employers with more positive initial experiences are more likely to post future jobs. Yet in Column (2) we can see that these employers give substantially more negative private feedback on the first evaluation.

Table 5: Feedback given by the employer on first job, by future experience

	Dependent variable:	
	Public Feedback (z-score)	Private Feedback (z-score)
	(1)	(2)
Num. future employer assignments (log)	0.054^{***}	-0.040^{***}
	(0.005)	(0.005)
Intercept	-0.122^{***}	0.099^{***}
	(0.007)	(0.006)
Observations	112,916	112,916
R^2	0.001	0.001

Notes: This table reports regressions of the employer-given public feedback (Column (1)) and private feedback, (Column (2)) for the first observation by an employer. These outcomes are regressed on the future assignments in. the marketplace by that same employer. Significance indicators: $p \leq 0.05 : *$, $.p \leq 0.01 : **$ and $p \leq .001 : ***$.

Another approach is to examine all contracts, but include a worker-specific fixed effect. In Table 6, Column (1), the dependent variable is public feedback. In Column (2) the dependent variable is private feedback. We can see that employers with greater future usage give more positive public feedback, more negative private feedback and, as a result, they give less honest feedback.

Table 6: Employer feedback with hired worker specific fixed effects

	Dependent variable:	
	Public FB (z-score)	Private FB (z-score)
	(1)	(2)
Num. future employer assignments (log)	0.078^{***}	-0.018^{***}
	(0.002)	(0.002)
Observations	338,456	338,456
R^2	0.395	0.381

Notes: This table reports regressions of the employer-given public feedback (Column (1)), private feedback, (Column (2)). These outcomes are regressed on the total number of future assignments by that employer relative to the date of the observation. Each regression includes worker-specific fixed effects. Significance indicators: $p \leq 0.05 : *$,$^{\cdot}$ $.p \leq 0.01 : **$ and $p \leq .001 : ***$.

7 Experimental revelation of aggregated private feedback

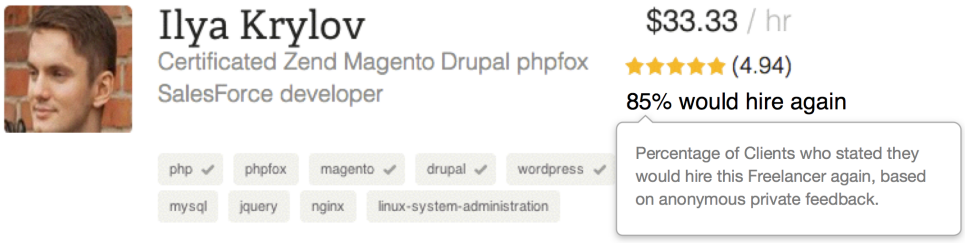
The reason oDesk collected private feedback was to potentially show aggregates of that feedback to future would-be buyers, with the goal of improving allocative efficiency. A very simple experiment was conducted: employers posting a job were randomized into one of two experimental groups:

- Control: Employers have the status quo hiring experience, in which they are shown the standard applicant characteristics, such as work experience, cover letter, hourly rate, skills and skills tests and public feedback.
- Treatment: In addition to all the information conveyed in the Control experience (including public feedback), for applicants that are eligible (to be defined below), aggregated *private* feedback is also shown.

An example worker profile as presented to employers in the treated group is shown in Figure 9. Near the worker's traditional 1-5 star public feedback, there is also a statement "85% would hire again" with the mouse-over elaboration "Percentage of Clients [employers] who stated they would hire this Freelancer [worker] again, based on anonymous private feedback." A worker was only eligible to have their "hire again" percentage shown if they had at least five completed contracts worth more than $50.00 and three distinct employers left private feedback.

Randomization was effective: Table 7 shows the means for a collection of pre-assignment job opening attributes and the p-value for a two-sided t-test comparing those means. For none of the compared attributes is there a significant difference by assignment. This is unsurprising as the same oDesk tool for randomization has been used repeatedly in numerous prior experiments. Appendix B contains information on the day-by-day assignment of new openings and also confirms that the randomization was effective.

Figure 9: Public display of aggregated private feedback



Notes: This figure shows how a worker's private feedback would be displayed. Note that this is the presentation for a fictional worker.

Although randomization was at the level of the employer, observations are at the level of the individual worker/employer pair. As we are primarily interested in the effect of the treatment on who gets hired—and more specifically on the interaction between a worker's private feedback score, their public feedback score and the treatment indicator—we want to keep the hierarchical structure of the data intact rather than collapse outcomes to employer-level outcomes alone. This hierarchical structure requires either a multi-level model or a clustering correction for the standard errors. Given that different jobs on oDesk receive very different numbers of applicants based on the nature of the job and employer, simply pooling observations and correcting standard errors is unattractive, as it weights some openings more

Table 7: Means of opening characteristics, by treatment assignment

	Treatment mean: $\bar{X}_{TRT}$	Control mean: $\bar{X}_{CTL}$	Difference in means: $\bar{X}_{TRT} - \bar{X}_{CTL}$	p-value
Observation Counts				
	97,631	98,335		
Type of work				
Technical (1 if yes, 0 otherwise)	0.283 (0.001)	0.282 (0.001)	0.002 (0.002)	0.449
Non-Technical	0.717 (0.001)	0.718 (0.001)	-0.002 (0.002)	0.449
Type of work—(more detailed)				
Admin	0.121 (0.001)	0.122 (0.001)	-0.001 (0.001)	0.434
Writing	0.127 (0.001)	0.126 (0.001)	0.001 (0.002)	0.682
Web	0.283 (0.001)	0.282 (0.001)	0.002 (0.002)	0.449
Design	0.177 (0.001)	0.176 (0.001)	0.001 (0.002)	0.626
Software	0.111 (0.001)	0.112 (0.001)	-0.000 (0.001)	0.750
Vacancy attributes				
Job description length > median	0.430 (0.002)	0.432 (0.002)	-0.003 (0.002)	0.253
Required prior oDesk experience	0.112 (0.001)	0.114 (0.001)	-0.002 (0.001)	0.157

Notes: This table contains the means for various pre-treatment assignment job opening attributes. The p-value for a two-sided means comparison t-test is shown.

heavily than others. For this reason, we analyze the experimental data using a multi-level model with opening-specific random effects.

In the experimental sample, 32.5% of all applications and 43.8% of all accepted applications, i.e., hires, were eligible for employers to view the applicant's aggregated private feedback in the treatment group. Figure 10 shows the distribution of aggregated private feedback scores of applicants.

Let i index individuals and j index employer job openings. Let y_{ij} be some outcome for individual i after applying to opening j. Outcomes include whether i had their profile evaluated by the employer, whether the individual was invited to interview, and whether the individual was ultimately hired. All of these outcomes are indicators of employer interest in a particular applicant. We estimate

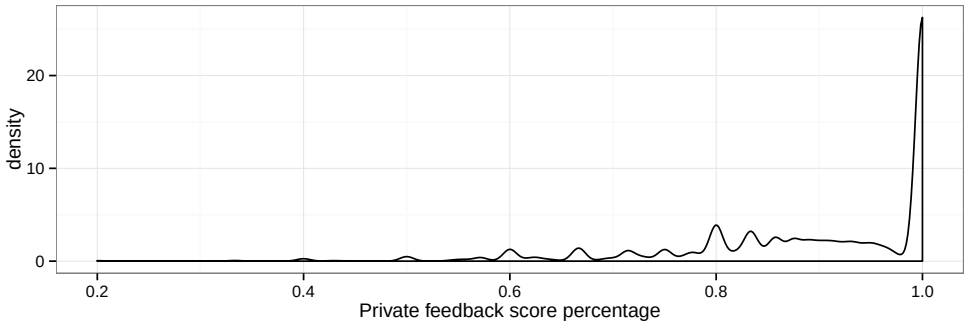
$$y_{ij} = \alpha_j + \beta_0 \text{ShwnPrvtFB}_{ij} + \beta_1 \text{PrvtFB}_{ij} + \beta_2 \left(\text{PrvtFB} \times \text{ShwnPrvtFB}_{ij}\right) + \epsilon_{ij} \tag{15}$$

where ShwnPrvtFB is an indicator for whether the private feedback score was shown (i.e., the employer was in the treatment group), PrvtFB was the private feedback score of the applicant at the time they applied and α_j is an opening-specific random effect.

The sample is all applicants *eligible* to have their aggregated private feedback shown. In each regression, the independent variables are the employer's treatment indicator, the worker's private feedback score and their interaction. Workers cannot condition their application decision on the employer's treatment assignment, so all terms are exogenous.[10] The dependent variables in Columns 1–3 of Table 8 are

[10]We might worry that treatment assignment could affect applicant composition by altering employer hiring probabilities which in turn dynamically affect applications. However, there was no detectable effect on hiring probabilities overall and given that employer hiring decisions happen considerably later than application decisions (Horton, 2014), this concern is not likely to be important.

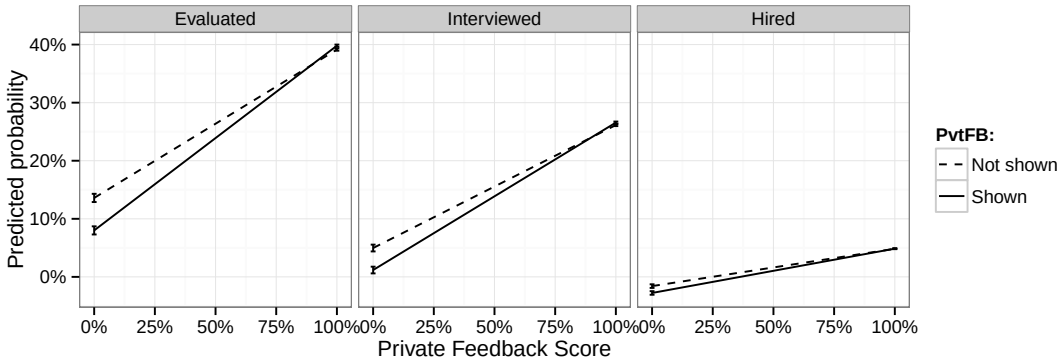
Figure 10: Distribution of aggregated private feedback scores among applicants



Notes: This figure shows the aggregated private feedback scores of applicants, at their time of application.

whether the worker's application was evaluated by the employer; whether the worker was interviewed by the employer; and whether the worker was hired by the employer, respectively.

Figure 11: Fitted regression models of employer screening, interviewing and hiring



Notes: This figure shows the regression results from Table 8. For each point estimate, a 95% CI is shown, based on uncertainty in the fixed effects only.

In each regression, we can see that workers with high aggregate feedback scores are far more likely to be evaluated, interviewed and hired, regardless of whether this information is shown to employers. This is unsurprising, as other measures of worker quality, such as aggregate public feedback, past on-platform experience, test scores and so on are all positively correlated with private feedback. The important coefficients are the treatment indicator, "ShwnPrvtFB" and its interaction with the private feedback score, "PrvtFB." Across regressions, the interaction is positive while the level of treatment indicator is negative. This implies that workers with visible low scores are evaluated/interviewed/hired less often, while the opposite is true for those workers with high private feedback scores. The experimental results provide strong evidence that employers consider the aggregating private feedback as informative.

Table 8: The effects of revealing aggregated private feedback on employer preferences

	Dependent variable:		
	Evaluated	Interviewed	Hired
	(1)	(2)	(3)
Private FB	0.255^{***}	0.212^{***}	0.064^{***}
	(0.004)	(0.003)	(0.002)
FB Shown (Treatment)	-0.056^{***}	-0.038^{***}	-0.012^{***}
	(0.005)	(0.004)	(0.002)
Private FB × FB Shown	0.063^{***}	0.042^{***}	0.012^{***}
	(0.005)	(0.004)	(0.002)
Constant	0.136^{***}	0.050^{***}	-0.016^{***}
	(0.004)	(0.003)	(0.002)
Observations	1,458,131	1,458,131	1,458,131
Log Likelihood	−810,397.000	−540,163.900	392,966.700
Akaike Inf. Crit.	1,620,806.000	1,080,340.000	−785,921.400
Bayesian Inf. Crit.	1,620,879.000	1,080,413.000	−785,848.200

Notes: This table reports a regression where the dependent variables are measures of employer interest: Column (1) is whether the candidate was evaluated, Column (2) is whether the candidate was interviewed and Column (3) is whether the candidate was hired. The sample consists of all applicants to job openings by employer assigned to the experiment. The independent variables in each regression are the applicant's aggegated private feedback score interacted with the treatment indicator, which determined whether the employer saw this score. As there is a many-to-one relationship between applications to openings, an opening-specific random effect to account for the hierarchial nature of the data. Note that a fixed effect estimator would not be appropriate, as treatment assignment does not vary with an opening. Significance indicators: $p \leq 0.05 : *$, $.p \leq 0.01 : **$ and $p \leq .001 : ***$.

7.1 Effects of the treatment on non-eligible applicants

The majority of applicants to a job opening were not eligible to have their private feedback score (if any) displayed. We might worry that for openings in the treatment, these non-eligible workers would be disadvantaged. As Pallais (2013) shows, workers without certain background characteristics (such as prior work experience) are substantially less likely to be hired. We see whether the presence of the private feedback score negatively affected applicants without it by estimating the regression

$$y_{ij} = \alpha_j + \beta_0 \text{Eligible}_{ij} + \beta_1 \text{ShwnPrvtFB}_{ij} + \beta_2 \left(\text{Eligible} \times \text{ShwnPrvtFB}_{ij}\right) + \epsilon_{ij} \tag{16}$$

where Eligible_{ij} is an indicator for whether the ith applicant to the jth job was eligible to have his or her private feedback score shown and ShwnPrvtFB_{ij} is the treatment indicator; the outcome y_{ij} is, as before, indicators for whether the applicant was evaluated, interviewed and hired in Columns (1) through (3), respectively.

Table 9 reports estimates of Equation 16. Across the three estimates, we can see that those who are eligible are strongly positively selected in terms of employer interest—the coefficient on Eligible is large relative to the baseline and always significant: for example, in the case of interviewing the baseline interview rate is about 15% for non-eligible workers whereas for eligible workers it jumps to about 19%.

Being in the treatment has no apparent effect on probability of being evaluated or interviewed. It is important to remember that in this regression, we are not conditioning on the private feedback score itself. The results that were conditioned on this score showed that applicants with low scores were hurt and those with high-scores were very slightly helped. The "pooled" effect that we get in Table 9 is essentially a precisely estimated 0. Furthermore, the interaction term is also a precisely estimated 0, indicating that although eligible workers did better that non-eligible workers, this difference was not affected by the treatment.

The pattern changes when we consider hiring: the coefficient on FBShown is now negative and significant, though the magnitude is not enormous: a worker not eligible to have their feedback shown that switches from the control to the treatment would see his or her hire probability drop a little more than 2%. For an eligible worker, the effect from switching would be negligible (the coefficient on the interaction term is a precise 0). Column (3) is some evidence of the new private feedback feature crowding out the hiring of workers not eligible to have their scores shown.

7.2 Effect of private feedback revelation on outcomes measured at the level of the job opening

In addition to examining how the revelation of information affected who was hired, we can also examine whether this change in hiring affected ultimate opening outcomes. We regress several outcomes on the treatment indicator. In Column (1) of Table 10 the dependent variable is the private feedback score of the hired worker. The effect is positive and highly significant but small in magnitude: it is less than a 1% increase. Column (2) reports a regression of an indicator for whether the opening filled on the treatment. The Column (1) sample is all employers, since we can always measure whether an opening filled. It is slightly negative but far from significant (statistically or economically).

The outcome in Column (3) is the feedback received by the hired worker (if any). As feedback is only collected on completed contracts, the sample size in Column (3) is considerably smaller than in Column (2). Although we have strong evidence that the treatment altered who the employer hired, it has no detectable effect on overall outcomes—though this finding is subject to the strong caveat that the measures available are not particularly good proxies for a "good" match being made.

Table 9: The effects of revealing aggregated private feedback

	Dependent variable:		
	Evaluated	Interviewed	Hired
	(1)	(2)	(3)
FB Shown (Treatment)	0.0004	0.0002	-0.001^{***}
	(0.001)	(0.001)	(0.0002)
FB Eligible for Showing	0.043^{***}	0.038^{***}	0.010^{***}
	(0.001)	(0.0004)	(0.0002)
Eligible × FB Shown	−0.0002	0.001	0.00004
	(0.001)	(0.001)	(0.0003)
Constant	0.292^{***}	0.151^{***}	0.023^{***}
	(0.001)	(0.001)	(0.0002)
Observations	4,034,215	4,034,215	4,034,215
Log Likelihood	−1,801,107.000	−781,094.100	2,149,847.000
Akaike Inf. Crit.	3,602,225.000	1,562,200.000	−4,299,681.000
Bayesian Inf. Crit.	3,602,305.000	1,562,279.000	−4,299,602.000

Notes: This table reports a regression where the dependent variables are measures of employer interest: Column (1) is whether the candidate was evaluated, Column (2) is whether the candidate was interviewed and Column (3) is whether the candidate was hired. The sample consists of all applicants to job openings by employers assigned to the experiment. The "Eligible" indicator is whether than applicant was eligible to have their private feedback score shown. Note that a fixed effect estimator would not be appropriate, as treatment assignment does not vary with an opening. Significance indicators: $p \leq 0.05 : *$, $.p \leq 0.01 : **$ and $p \leq .001 : ***$.

Table 10: Effect of treatment assignment on opening outcomes

	Dependent variable:		
	Private FB Score	Anyone Hired?	Public FB on formed match
	(1)	(2)	(3)
Private FB Shown	0.006^{***}	-0.001	-0.018
	(0.002)	(0.002)	(0.014)
Constant	0.932^{***}	0.363^{***}	4.804^{***}
	(0.001)	(0.002)	(0.010)
Observations	9,849	195,966	9,849
R^2	0.001	0.00000	0.0002
Adjusted R^2	0.001	-0.00000	0.0001
Residual Std. Error	0.097 (df = 9847)	0.481 (df = 195964)	0.697 (df = 9847)
F Statistic	10.161^{***} (df = 1; 9847)	0.191 (df = 1; 195964)	1.702 (df = 1; 9847)

Notes: This table reports a regression where the dependent variables are measures of opening outcomes. The dependent variable in Column (1) is an indicator for whether the employer filled the job. The sample for this regression is the first opening posted by an assigned employer. In Column (2), the sample is restricted to filled hourly jobs and the dependent variable is the log of total hours worked to date. Column (3) is the feedback given to hired workers, with the sample restricted to those employers that have closed a filled contract and provided feedback. Significance indicators: $p \leq 0.05 : *$, $.p \leq 0.01 : **$ and $p \leq .001 : ***$.

8 Conclusion

This paper documents that reputation inflation occurs in an online marketplace and argues that the cause is driven by the costs associated with leaving negative feedback. It shows that when these costs are reduced—namely by allowing buyers to give feedback without the seller knowing it—buyers are substantially more candid. Further, the buyers who had the strongest incentive not to be candid—namely those using the marketplace intensively—showed the biggest "candor gap." Through the experimental validation, we show that buyers act upon this information, which suggests they (correctly) believe it has information content. This paper illustrates a market design problem, analyzes its root cause and then experimentally validates a proposed solution mechanism. Because aggregated private feedback cannot be traced back to the employer, this "collect privately and then disseminate aggregates publicly" design is less likely to be prone to inflation.

References

Abowd, John M, Francis Kramarz, and David N Margolis, "High wage workers and high wage firms," *Econometrica*, 1999, *67* (2), 251–333.

Agrawal, Ajay K, John Horton, Nico Lacetera, and Elizabeth Lyons, " Digitization and the Contract Labor Market: A Research Agenda ," in "Economics of Digitization," University of Chicago Press, 2013.

_ **, Nicola Lacetera, and Elizabeth Lyons**, "Does Information Help or Hinder Job Applicants from Less Developed Countries in Online Markets?," January 2013, (NBER Working Paper 18720).

Aperjis, Christina and Ramesh Johari, "Optimal Windows for Aggregating Ratings in Electronic Marketplaces," *Management Science*, 2010, *56* (5), 864–880.

Bolton, Gary, Ben Greiner, and Axel Ockenfels, "Engineering trust: Reciprocity in the production of reputation information," *Management Science*, 2013, *59* (2), 265–285.

Bolton, GaryE., Elena Katok, and Axel Ockenfels, "Bridging the Trust Gap in Electronic Markets: A Strategic Framework for Empirical Study," in Joseph Geunes, Elif Akçali, PanosM. Pardalos, H.Edwin Romeijn, and Zuo-JunMax Shen, eds., *Applications of Supply Chain Management and E-Commerce Research*, Vol. 92 of *Applied Optimization*, Springer US, 2005, pp. 195–216.

Cabral, Luis and Ali Hortaçsu, "The Dynmamics of Seller Reputation: Evidence from eBay," *The Journal of Industrial Economics*, 2010, *58* (1), 54–78.

Dellarocas, Chrysanthos, "The Digitization of Word of Mouth: Promise and Challenges of Online Feedback Mechanisms," *Management Science*, 2003, *49* (10), 1407–1424.

_ **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.

Friedman, Jerome, Trevor Hastie, and Rob Tibshirani, "glmnet: Lasso and Elastic-net Regularized Generalized Linear Models," 2009.

Gaure, Simen, "lfe: Linear Group Fixed Effects," *The R Journal*, 2013, *5* (2), 104–117.

Gibbons, Robert and Lawrence Katz, "Layoffs and Lemons," *Journal of Labor Economics*, 1991, *9*, 351–80.

Greif, Avner, "Contract Enforceability and Economic Institutions in Early Trade: The Maghribi Traders' Coalition," *The American Economic Review*, 1993, pp. 525–548.

Hirschman, Albert O., *Exit, Voice, and Loyalty: Responses to Decline in Firms, Organizations, and States*, Vol. 25, Harvard university press, 1970.

Horton, John, "The Effects of Subsidizing Employer Search," *Working Paper*, 2014.

Jurka, Timothy P., Loren Collingwood, Amber E. Boydstun, Emiliano Grossman, and Wouter van Atteveldt, "RTextTools: Automatic Text Classification via Supervised Learning," *R package version 1.3.9*, 2012.

Kreps, David M and Robert Wilson, "Reputation and Imperfect Information," *Journal of Economic Theory*, 1982, *27* (2), 253–279.

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," Technical Report, Harvard Business School 2011.

_ **and Georgios Zervas**, "Fake it Till You Make It: Reputation, Competition, and Yelp Review Fraud," *Working Paper*, 2013, (14-006).

Mayzlin, Dina, Yaniv Dover, and Judith Chevalier, "Promotional Reviews: An Empirical Investigation of Online Review Manipulation," *The American Economic Review*, 2014.

Nosko, Chris and Steve Tadelis, "The Limits of Reputation in Platform Markets: An Empirical Analysis and Field Experiment," *Working Paper*, 2014.

Pallais, Amanda, "Inefficient Hiring in Entry-level Labor Markets," *American Economic Review*, March 2013, (18917).

Resnick, Paul and Richard Zeckhauser, "Trust among strangers in Internet transactions: Empirical analysis of eBay's reputation system," *Advances in applied microeconomics*, 2002, *11*, 127–157.

_ **, 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.

Soujourner, Aaron, Alan Benson, and Ahkmed Umyarov, "The Value of Employer Reputation in the Absence of Contract Enforcement: A Randomized Experiment," *Working Paper*, 2014.

Stanton, Christopher and Catherine Thomas, "Landing the First Job: The Value of Intermediaries in Online Hiring," *Available at SSRN 1862109*, 2012.

Varian, Hal R., "Computer Mediated Transactions," *American Economic Review*, 2010, *100* (2), 1–10.

Zervas, Georgios, David Proserpio, and John Byers, "A First Look at Online Reputation on Airbnb, Where Every Stay is Above Average," *Working Paper*, 2015.

A Summary statistics

Table 11 shows summary statistics for the openings, by treatment group.

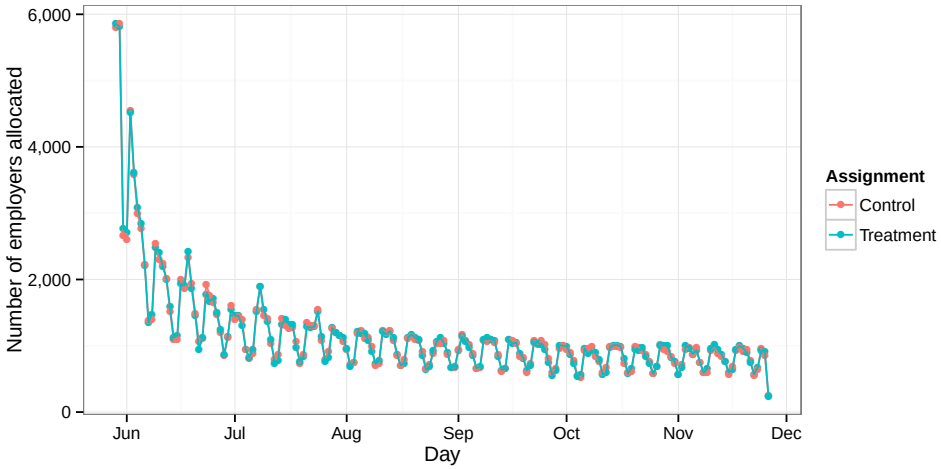
Table 11: Summary statistics for openings

Variable	Levels	n	Min	q_1	$\tilde{x}$	$\bar{x}$	q_3	Max	s	IQR	#NA
total_charge	ctrl	97631	-51.9	0	0	178.0	44.4	147596.5	1287.6	44.4	0
	trt	98335	-35.0	0	0	172.1	44.4	162667.8	1186.1	44.4	0
	all	195966	-51.9	0	0	175.0	44.4	162667.8	1237.7	44.4	0
team_size	ctrl	97631	0.0	0	0	3.0	1.0	1431.0	13.0	1.0	0
	trt	98335	0.0	0	0	3.0	1.0	2591.0	13.6	1.0	0
	all	195966	0.0	0	0	3.0	1.0	2591.0	13.3	1.0	0
recruited_applicants	ctrl	97631	0.0	0	0	1.1	1.0	289.0	3.4	1.0	0
	trt	98335	0.0	0	0	1.1	1.0	241.0	3.6	1.0	0
	all	195966	0.0	0	0	1.1	1.0	289.0	3.5	1.0	0
num_hires	ctrl	97631	0.0	0	0	0.4	1.0	163.0	1.1	1.0	0
	trt	98335	0.0	0	0	0.4	1.0	228.0	1.2	1.0	0
	all	195966	0.0	0	0	0.4	1.0	228.0	1.1	1.0	0
num_applications	ctrl	97631	0.0	1	7	17.0	22.0	737.0	27.6	21.0	0
	trt	98335	0.0	1	7	17.1	22.0	781.0	28.2	21.0	0
	all	195966	0.0	1	7	17.0	22.0	781.0	27.9	21.0	0

B Balance in experimental units

Figure 12 shows the allocation of experimental subjects over time.

Figure 12: Allocation of employers over time



Notes: This figure shows the daily number of employers allocated to the experiment. Note that as both new and old employers were eligible, the rate is declining since the "stock" of unallocated employers is declining.

C Predicting public feedback scores using textual written feedback

We used the RTextTools package from Jurka et al. (2012) to construct the document term matrix for the written feedback. To actually fit the model, we used the Lasso, as implemented in the glmnet R package by Friedman et al. (2009). Cross-validation was used to select tuning parameters. Figure 13 shows a sample of coefficients from the fitted model, ordered by value.

Figure 13: Text predictors of feedback score