Forthcoming, *Management Science* (2019)

Can reputation discipline the gig economy? Experimental evidence from an online labor market

Alan Benson, Aaron Sojourner, and Akhmed Umyarov*
University of Minnesota

Just as employers face uncertainty when hiring workers, workers also face uncertainty when accepting employment, and bad employers may opportunistically depart from expectations, norms, and laws. However, prior research in economics and information sciences has focused sharply on the employer's problem of identifying good workers rather than vice versa. This issue is especially pronounced in markets for gig work, including online labor markets, where platforms are developing strategies to help workers identify good employers. We build a theoretical model for the value of such reputation systems and test its predictions in on Amazon Mechanical Turk, where employers may decline to pay workers while keeping their work product and workers protect themselves using third-party reputation systems, such as Turkopticon. We find that: (1) in an experiment on worker arrival, a good reputation allows employers to operate more quickly and on a larger scale without loss to quality; (2) in an experimental audit of employers, working for good-reputation employers pays 40 percent higher effective wages due to faster completion times and lower likelihoods of rejection; and (3) exploiting reputation system crashes, the reputation system is particularly important to small, good-reputation employers, which rely on the reputation system to compete for workers against more established employers. This is the first clean field evidence of the effects of employer reputation in any labor market and is suggestive of the special role that reputation-diffusing technologies can play in promoting gig work, where conventional labor and contract laws are weak.

Key words: Online labor markets, online ratings, employer reputation, labor markets, economics of information systems, job search

---

* Authorship is equal and alphabetical. We thank our excellent research assistants: Harshil Chalal, Sima Sajjadiani, Jordan Skeates-Strommen, Rob Vellela, and Qianyun Xie. We also thank Panos Ipeirotis for sharing M-Turk tracker data. For their useful feedback, we thank John Budd, Eliza Forsythe, Mitch Hoffman, John List, Colleen Manchester, Mike Powell, David Rahman, Chris Stanton, and workshop participants at the ASSA Meetings, MEA-SOLE meetings, Minnesota Applied Microeconomics Workshop, MIT Sloan Organizational Economics Lunch, MIT Conference on Digital Experimentation, MIT Sloan IWER seminar, LERA meetings, Michigan Ross, Northwestern Law School, Organization Science Winter Conference, Barcelona GSE Digital Economics Summer Workshop, IZA World Labor Conference, NBER Personnel summer institute, GSU, the Advances in Field Experiments conference at BU, and Tennessee-

1. Introduction

Amazon Mechanical Turk (M-Turk), TaskRabbit, Upwork, Uber, Instacart, DoorDash, and other online platforms have drastically reduced the cost of seeking, forming, and terminating work arrangements. This development has raised the concern that platforms circumvent regulations that protect workers. A U.S. Government Accountability Office (2015, p. 22) report notes that such "online clearinghouses for obtaining ad hoc jobs" are attempting "to obscure or eliminate the link between the worker and the business ... which can lead to violations of worker protection laws." Developers of these online platforms argue that ratings systems can help discipline employers and other trading partners who break rules and norms. To what extent can the threat of a bad public reputation prevent employer opportunism and the promise of good reputation encourage responsible employer behavior? To what extent can workers voluntarily aggregate their private experiences into shared memory in order to discipline opportunistic employers? These questions are especially important to gig jobs and online labor platforms, which are characterized by high-frequency matching of workers and employers, and where platform owners have moved fast and often sought to break out of traditional regulatory regimes.

Our setting focuses on M-Turk, which possesses two useful and rare features for this purpose. First, no authority (neither the government nor the platform) disciplines opportunistic employers. In M-Turk, after workers put forth effort, employers may keep the work product but refuse payment for any reason or no reason. Workers have no contractual recourse or appeal process. Because M-Turk also has no native tool for rating employers, many workers use third-party platforms to help them find the best employers. Among the most popular of these platforms is Turkopticon, which allows workers to share information about employers. Second, such opportunism is observable to researchers with some effort. In traditional labor markets, workers would presumably rely on an employer's reputation to protect themselves against events not protected by a contract, such as promises for promotion or paying for overtime. Unfortunately, these outcomes are especially difficult to observe for both courts and researchers alike. The literature on both traditional and online labor markets has instead focused on learning about opportunism by workers, not by employers (e.g., Oyer et al. 2011, Moreno and Terwiesch 2014, Horton and Golden 2015, Pallais 2014, Stanton and Thomas 2015, Filippas et al. 2018, Farronato et al. 2018). These two contextual features enable our empirical study.

Knoxville. We also thank Minnesota SOBACO for funding. Corresponding author: asojourn@umn.edu. Prior version as IZA Discussion Paper 9501.

We begin by outlining a model of employer reputation. The model allows for two forms of employer reputation: a public reputation disseminated by a formal rating system and an informal reputation governed by an individual employer's private visibility. The model yields three key, testable predictions regarding the value of reputation. First, employers with better reputations are rewarded through an enhanced ability to attract workers at a given promised payment level, and, in this sense, a better reputation acts as collateral against future opportunism. Second, workers earn more when they have information that enables them to work only for better-reputation employers. Third, the value of the public reputation system to employers depends on their visibility: less-visible employers with good reputations on the system rely on it to attract workers, while better-known, good-reputation employers are less reliant. This third prediction has important implications for how a credible reputation system changes what types of trading relationships a market with poor enforcement can bear.

Interpreting this prediction, less well-known employers, such as newer or smaller firms, struggle more to earn workers' trust in order to recruit workers and grow. Credible reputation systems especially help these less-visible employers and, thereby, support the entry of firms from the competitive fringe and promote economic competition and dynamism. Our model is consistent with the digitization literature's findings that rating systems are especially important to relatively unknown agents and not as important to agents with an otherwise highly visible reputation. Luca (2016) finds that Yelp.com reviews are especially important for smaller, independent restaurants than for restaurant chains, and indeed, that Yelp penetration in an area is associated with the decline in chains that presumably rely on other forms of reputation. Nagaraj (2016) finds that digitization of magazine articles has a greater effect on Wikipedia entries for less-known individuals than for well-known individuals for whom other information is readily available.

In three tests, we provide the first clean field evidence of employer reputation effects in a labor market. These tests follow the model's three predictions. Specifically the first experiment measures the effect of employer reputation on the ability to recruit workers. We create 36 employers on M-Turk. Using a third-party employer rating site, Turkopticon, we endow each with (i) 8-12 good ratings, (ii) 8-12 bad ratings, or (iii) no ratings. We then examine the rate at which they can recruit workers to posted jobs. We find that employers with good reputations recruit workers about 50 percent more quickly than our otherwise-identical employers with no ratings and 100 percent more quickly than those with very bad reputations. Using M-Turk wage elasticities estimated by Horton and Chilton (2010), we estimate that posted wages would need to be almost 200 percent greater for

bad-reputation employers and 100 percent greater for no-reputation employers to attract workers at the same rate as good-reputation employers. Outside of M-Turk, one might think of the attractiveness of the job as the firm's ability to attract applicants and reputation as a substitute for wage for that purpose. A better reputation shifts the firm's labor supply curve outward, allowing it to recruit and select from more workers for a given wage offer in a given period of time. We also estimate that about 55 percent of job searchers use Turkopticon, suggesting that more complete adoption would magnify these effects. We find evidence that Turkopticon is signaling employer characteristics rather than just task characteristics. These results demonstrate that workers use reputation to screen employers and that reputation affects employers' ability to recruit workers.

The second experiment tests the validity of online reputations from the perspective of a worker. Reputation systems based on individual ratings are vulnerable to inflation and inaccuracy (Filippas et al. 2018). We act as a blinded worker to assess the extent to which other workers' public employer ratings reflect real variation in employer quality. One research assistant (RA) randomly selects tasks from employers who have good reputations, bad reputations, or no reputation and sends each job to a second RA who does the jobs while blind to employers' reputations. This experimental feature ensures that worker effort is independent of employer reputation, and so any observed differences in subsequent employer behavior toward the worker are not due to differences in worker effort. Consistent with a pooling equilibrium, we observe no difference in the average per-task payment promised up front by employers of different reputations. However, effective wages while working for good-reputation employers are 40 percent greater than effective wages while working for bad-reputation employers. We decompose this difference into the shares due to differences in job quality (how long the job takes) versus the probability of being paid at all. This experiment shows that, although the reputation system aggregates ratings that are all voluntarily provided and unverified, it contains useful information for workers.

Lastly, we focus on instances when Turkopticon servers stopped working by using a difference-in-differences design to study effects when the reputation system is removed temporarily. We match data on M-Turk tasks completed across employers over time from Ipeirotis (2010a) with each employer's contemporaneous Turkopticon ratings and compare the change in worker arrival rates for different types of employers when Turkopticon crashes. When Turkopticon crashes, employers with bad reputations are largely unaffected. Taken with the prior evidence, this result suggests that employers with bad reputations do not recruit workers who use Turkopticon and rely only on uninformed workers. However, the effect on employers with good reputations on Turkopticon is heterogeneous by employer

visibility, measured as the number of times an employer posted work in the past and proxying for the likelihood that a worker would have encountered them before. Workers sharply withdraw their labor supply from less-visible, good-reputation employers, who presumably were benefiting from Turkopticon informing workers of their good reputations. In contrast, worker arrival rates increase for more-visible, good-reputation employers, who are presumably already better-known as safe bets.

2. Literature review

Most markets have information problems to some degree. For M-Turk workers, Turkopticon is the Dun & Bradstreet of procurers, the Moody's of bond buyers, the Fair Isaac of consumer lenders, and the Metacritic of moviegoers. Each of these institutions offers extralegal protections against contractual incompleteness based on information sharing and the implicit threat of coordinated withdrawal of trade by one side of a market.

A large literature studies unilateral ratings of suppliers in online markets. In principle, ratings should be unilateral when suppliers (including workers, sellers of goods, and service providers) vary in ways that are difficult to contract upon, creating problems for buyers. In contrast, suppliers do not face many issues in choosing customers; buyers are essentially money on the barrel. For example, prior work (e.g. Dellarocas and Wood 2008, Nosko and Tadelis 2015) has studied sellers on eBay, where buyers rate sellers on the accuracy of product descriptions and timeliness, but sellers only care that buyers submit payment. A similar case could be made for Yelp (e.g., Luca 2016), where the quality of the food and service is difficult to contract upon but restaurants serve all comers.

While the vast majority of work has concerned unliateral rating regimes, Airbnb (e.g., Fradkin et al. 2015) and Uber (e.g., Rosenblat et al. 2017) offer two major exceptions: Airbnb hosts and guests rate each other, as do Uber drivers and passengers. While both sides value the ease of platform matching and microcontracting, both also worry about opportunistic behavior by counterparties. These platforms try to create healthy markets while avoiding regulation by developing two-sided online ratings systems. The digital labor platform Upwork has also used a two-sided rating system, but research there has not focused on employer reputation.

Issues of trust and contract enforcement among trading partners are especially important to online spot markets and other forms of gig work. The nature of the employment relationship, especially in gig work, features bilateral uncertainty. Despite this, the literature on both traditional and online labor has largely focused on the employer's problem of evaluating and rating employees rather than vice versa. Labor economics has given great

attention to how employers interpret educational credentials, work experience, or their experience at that firm to identify the most productive workers (for a review, see Oyer et al. 2011).

This focus on guarding against worker opportunism now extends to the burgeoning literature on online labor markets. Studies, for instance, have examined how online employers infer workers' abilities from their work histories (Pallais 2014), oversubscription (Horton 2018), work through outsourcing agencies (Stanton and Thomas 2015), national origin (Agrawal et al. 2013), or platform endorsements (Barach et al. 2017). Although these studies present a sample of the recent literature on online labor markets, they are characteristic of its present focus on rating workers rather than firms. As such, the literature offers little evidence on the risks that workers face when selecting employers or how platforms can design systems that promote trade absent regulation.

Our work builds on a handful of studies that have sought to empirically examine employer reputation. Turban and Cable (2003) provide the first correlational evidence that companies with better reputations tend to attract more applicants using career services data from two business schools. Hannon and Milkovich (1995) find mixed evidence that news of prominent employer rankings affects stock prices. Using a similar methodology, Chauvin and Guthrie (1994) find small but significant effects. Brown and Matsa (2015) find that distressed financial firms attract fewer and lower-quality applicants. List and Momeni (2017), also in an experiment on M-Turk, find that employers that promise to donate wages to a charity attract more work, albeit with more cheating, suggesting that workers take moral license when performing work for good causes. As such, the prior evidence on employer reputation is either correlational or concerns uncertainty of other characteristics of the employer (e.g. the firm's future success or altruism), which workers may value in themselves or because they also view these as correlated with their quality as an employer. We provide the first field experimental evidence regarding how employers' reputation for treating workers affects their ability to attract work.

Theoretically, a better public reputation would allow trading partners to extract greater work or higher prices (Klein and Leffler 1981). Moreno and Terwiesch (2014) find that online service providers leverage better reputations to either charge more or increase their probability of being selected for a project. Ba and Pavlou (2002), in a market "similar to eBay," find that buyers are willing to pay a price premium to deal with trusted sellers. Similarly, Banerjee and Duflo (1999) find that supplier reputation is important in the Indian software market, where postsupply service is important but difficult to contract. McDevitt (2011) finds evidence that residential plumbing firms with high records of complaints are

more likely to change their name, suggesting that firms seek to purge bad reputations. However, in their study of eBay sellers, Bajari and Hortacsu (2003) find only a small effect of reputation on prices. While these studies once again focus on rating the supply side, our theory and empirics will speak to the demand side by presenting evidence that employers with good reputations can extract lower prices and greater quantities of work.

We also consider how public ratings of employers substitute for more traditional markers of being an established organization. Stinchcombe and March (1965) famously theorized that new organizations face a credibility problem when attracting trading partners, a phenomenon they refer to as the "liability of newness." Luca (2016) offers perhaps the most closely related test for the relative importance of online ratings versus being established; he finds that Yelp ratings are more important to independent restaurants than to chains. As we will discuss, our setting offers a relatively clean opportunity to exploit exogenous removals of a reputation system, to do so for the demand side, and to do so in a labor market.

3. Theory and Hypotheses

We offer a formal model of job search in which there is no contract enforcement. Under certain conditions, a "public relational contract" emerges, whereby the threat of losing future workers discourages employer opportunism. We explore how workers and different kinds of employers rely on an employer reputation system and test the model's hypotheses in the online market.

The model begins from a basic sequential search model. Workers search for a job to receive a wage offer from a random employer and choose whether to accept each offer and put forth effort. Then, the employer chooses *ex post* whether to renege on payment. Workers' decision to work or to forego an offer depends on their beliefs as to whether the employer will pay, which depends on two exogenous factors: whether the worker is informed by a reputation system, and whether the employer is visible.[1] Specifically, there are two types of workers: workers informed by the reputation system see all employers' past payment histories, and uninformed workers who do not. Being informed, in this sense, signifies access to the collective experience of prior workers as would be diffused by a public reputation system. Second, uninformed workers may nonetheless observe the pay history of an offering

[1] As discussed more later, we estimate at least one-half of the M-Turk labor supply installs Turkopticon, though it remains a puzzle why not all workers do so. Exogenous employer visibility is also interesting: one might imagine that, in small communities, all market participants would have a well-established reputation without the aid of any formal reputation system. M-Turk features both a few large brokerages (e.g., Crowdflower) that constitute a substantial share of labor demand, and a long tail of smaller requesters.

employer with a probability that depends on a property of the employer: its visibility.[2] The visibility of an employer can be thought of simply as it is treated in the model: an employer characteristic that makes its history known to workers, even without the reputation system. Perhaps the most obvious empirical proxy for visibility is simply the size of the employer, which, by definition, signifies the extent of workers' past experience with the employer.

Though the single-shot model would devolve into the classic hold-up problem in which employers never pay and workers never put forth effort, we characterize an interesting, non-unique, steady-state equilibrium where an employer's good reputation acts as collateral against future wage theft. They continue to pay so that they can attract workers and get work done in the future. We explore how the existence of the public reputation system (informed workers) differentially affects highly-visible and less-visible employers. Employers with a good reputation ("high-road employers") continue to pay as long as the share of informed workers is sufficiently high, while employers with a bad reputation ("low-road employers") never pay.[3] We describe conditions that avoid making high-road employers' reneging temptation too great, which would then cause all workers to exit the labor market. This temptation is disciplined by the expected future flow of workers, which depends on the extent of workers' ability to identify high-road employers. As a result, we endogenously characterize the scope of economic activity that this environment can bear, and how it depends on the ability of workers to screen employers through either a public reputation system or private knowledge.[4]

The model makes three key claims. First, the public reputation system deters employer opportunism by creating a credible threat that employers' ability to attract workers in the future will erode (tested in study 1). Good-reputation employers can attract more workers to the same job offer than either no-reputation or bad-reputation employers; a better reputation shifts the labor supply curve inward. Second, it is incentive compatible

[2] Perfect revelation of past payment simplifies the exposition. Board and Meyer-ter Vehn (2013) considers reputation building when learning is imperfect. Their model also yields ergodic shirking, with increasing incentives for noncontractible investments as reputation becomes noiseless.

[3] Workers may face the two standard kinds of information problems with respect to unobserved employer heterogeneity: adverse selection and moral hazard. Employers' technologies or product markets may differ in ways that make low-road practices more or less profitable. In this adverse-selection setting, it is trivial to understand why variation in employment practices emerges. An alternative theory is that there is no essential heterogeneity between employers. Differences in strategic employment practices appear between essentially-homogeneous employers (Osterman 2018). We focus on this, more-interesting case. In all labor markets, both mechanisms are almost certainly empirically relevant. Cabral and Hortacsu (2010) did such an accounting in a consumer-goods market, baseball cards on EBay. We know of no analogous accounting in any labor market. That remains for future work.

[4] Other studies show how reputation systems and credentials can improve efficiency in other online markets including eBay (Nosko and Tadelis 2015, Hui et al. 2016) and Airbnb (Fradkin et al. 2015).

for workers to screen out low-road employers in their job search, a strategy that boosts workers' hourly earnings (study 2). Third, the reputation system is a substitute for the visibility of individual employers. The reputation system matters especially for smaller, less-visible, high-road employers, whose good payment histories would otherwise be unknown to jobseekers (study 3).

Formally, assume a job search environment with measure 1 of workers indexed by $i \in [0,1]$ and measure 1 of risk-neutral employers indexed by $j \in [0,1]$. A share, $s \in [0,1]$, of employers have a high-road history of making promised payments. Workers with $i \leq p \in (0,1)$ are fully-informed to all employers' past play, via a public reputation system. Employers differ in their history's visibility $v_j \in (0,1)$ to other workers, where high values of v_j represent "highly-visible" employers. Let $v \equiv \mathbb{E}[v_j]$. Workers who are indifferent between accepting and rejecting offers choose to accept. Employers indifferent between paying and reneging choose to pay. Normalizing, we assume each earns 0 if they do not participate in this market (e.g., nonwork or the labor market characterized by this environment). The timing of a period of job search is:

1. Worker i chooses whether to search. Those who do incur cost c and receive a wage promise w from a random employer j. Fully-informed workers observe j's past decisions to pay or renege. Other workers observe employer j's history with probability v_j. Non-searching workers receive 0 and proceed to the next period of job search. 0 represents the value of not participating in the online labor market.
2. Worker i decides whether to accept or reject employer j's offer. If the worker accepts, he incurs cost of effort e and employer j receives work product with value y.[5] If the worker rejects, he receives 0 and proceeds to the next period of job search.
3. Employer j decides whether to pay w or to renege and pay 0. Employers discount future periods at rate δ.

We provide a set of parametric restrictions, prove the existence of an equilibrium, and explore its properties. First, promised wages exceed the value of work effort: $w - e > 0$. This is a trivial precondition for anyone wanting to work. Second, an uninformed worker will not accept an unknown employer's offer because the expected payoff for doing so does not justify the certain cost of effort, $sw < e$. This assures not everyone works for anyone. Third, for uninformed workers, promised wages (w) and the chance of being matched to a visible, high-road employer (vs) times its payoff ($w - e$) exceed the certain cost of search, $vs(w - e) \geq c$. The uninformed worker's market participation constraint is met. It implies

[5] Our model abstracts away from the real but very well-studied employer information problem of dealing with workers who may differ in quality or shirk.

a lower bound on wage, $w^{min} = e + (vs)^{-1}c$, necessary to keep uninformed workers from dropping out of the market. It increases in the cost of effort and search and decreases in the share of high-road employers and average employer visibility.

The fourth parameter restriction is less trivial and more interesting. For high-road employers, it requires that the gains from high-road trade $(y - w)$, employer farsightedness δ, and the flow of workers $(p + v - pv)$ outweigh the one-time payoff of reneging (y) and continuing as a low-road employer: $(1-\delta)^{-1}(p+v-pv)(y-w) \geq y$. We will refer to this as the "reputation diffusion" criterion, noting that an employer's past play must be observed either through worker informedness or employer visibility. As both p and v approach zero, workers can't screen employers well enough and employers don't extract sufficient rents on a good reputation to justify maintaining a high-road reputation.[6] Otherwise, high-road employers would renege, the value of market participation for all workers becomes negative, no work is performed, and the labor market unravels.

An adapted Lerner index, $(y - w)/y$, measures the share of worker productivity kept as economic rent by high-road employers in this equilibrium. The reputation-diffusion criterion implies this share cannot fall below $\frac{1-\delta}{p+v-pv}$. Better information for workers, p or v rising to 1, creates space for high-road employers to raise wages and to retain a smaller share of worker productivity because they will have reliable access to a larger share of workers. This property reflects the reputation system literature's view that better reputation systems can expand the scope of economic activity that can be completed online by arms-length trading partners (Jøsang et al. 2007).

First, consider workers, who vary in their informedness. Informed workers encounter a high-road employer in any period with probability s. They accept high-road employers' offers because $w - e - c \geq -c$, which follows from the assumption that trade is profitable: $w - e > 0$. They reject low-road employers' offers because $-c > -c - e$. Therefore, the present value of this strategy for fully-informed workers is $(1-\delta)^{-1}[s(w-e)-c]$. Uninformed workers encounter an employer with an observable pay history with probability v. In this case, they face the same incentives and make the same decisions as fully-informed workers. Uninformed workers who encounter a non-visible employer reject the offer due to the parameter restriction $sw < e$. The present value of this strategy to uninformed workers is $(1-\delta)^{-1}[vs(w-e)-c]$. Both informed and uninformed workers' payoffs satisfy their participation constraint under the condition, $vs(w-e)-c \geq 0$.

[6] Another approach is to let employers exogenously vary in their discount rates, in which case farsighted employers become high-road employers.

Next, consider employers, which vary in their visibility. Low-road employers are supplied no labor, since workers only accept offers from revealed, high-road employers.[7] High-road employers receive workers at a rate of $p+(1-p)v_j$, receiving all informed workers and the share v_j of uninformed workers. In equilibrium, it is incentive compatible for employers to pay if the present value payoff of paying exceeds the immediate temptation of reneging, $(1-\delta)^{-1}(p+v_j-pv)(y-w) \geq y$. This follows from the reputation-diffusion criterion, discussed above. The high-road employer's participation constraint is simply that $y-w \geq 0$.

The first two hypotheses concern the model's key predictions regarding why it is incentive compatible for employers to maintain a good reputation and for workers to screen for jobs on a good reputation.

Hypothesis 1 *For a given job offer, employers with a better reputation will attract workers more quickly.*

Hypothesis 2 *Workers have higher average pay when working for employers with better reputations.*

Our final hypothesis follows from an interesting property of the model: that the reputation system that informs workers and employer visibility are substitutes. This is the key feature of the reputation-diffusion criterion, and yields the surprising result that, when the share of informed-type workers is sent to zero (e.g., due to the crash of the reputation system), the only employers affected are those with good reputations and imperfect visibility.

Hypothesis 3 *If the reputation system is disabled, sending the share of informed workers to zero, then (a) highly-visible employers with good reputation will lose a small or no share of workers, (b) less-visible employers with good reputation will lose a larger share of workers, (c) employers with bad reputation will be unaffected regardless of visibility.*

To see this, recall that the flow of workers to high-road employers is given by $p+v_j-pv_j$. As the share of informed workers decreases (*i.e.* the employer reputation system crashes, sending $p \to 0$, a high-road employer's worker flow converges to v_j. The most visible employers ($v_j \to 1$) are unaffected, while the arrival rate to the least-visible employers ($v_j \to 0$) falls more. Put another way, the reputation system (the mass of informed workers) is least

[7] We could allow low-road employers to receive some work, and some profit, by allowing visibility to yield an imperfect signal. Then s could arise endogenously by allowing employers to exogenously vary in their discount rates; cheap and patient employers renege, rich and impatient employers pay. We omit this complexity because it adds little insight, and our empirics are concerned with whether the signal is valuable, not the degree to which it is precise.

valuable to the most-visible employers, and most valuable to the least-visible employer. In our model, change in the reputation system p does not affect low-road employers. They get the same arrival rate regardless of informedness or visibility.

The relative value of the reputation system to more-visible employers versus less-visible employers is a chief contribution of this model, at least beyond labor markets. Conceptually, we can think of the reputation system as any technology that makes it less costly for one trading party to observe the past behaviors of potential partners. In our setting, we observe relatively-brief, unexpected instances when the market's public reputation system crashes. In these instances, workers who remain in the market must rely on other mechanisms to screen employers. Because our setting's reputation system is hosted through a third-party, its outages present a special opportunity to study how workers adapt, and how this change in search behavior affects employers of varying visibility.

Before continuing, it's also important to note some other interesting features of the model. One relates to the substituability of informed workers and visible employers. If $p \to 1$, then v no longer affects search or the flow of workers; if $v \to 1$, then p no longer affects search or the flow of workers. In either case, workers always accept jobs from known high-road employers. In this sense, technologies that give workers a collective memory serve as a scalable substitute for the personal experience that markets have traditionally relied upon to avoid opportunistic trading partners. A second insight relates to the upper bound on wage that this environment can bear before the market breaks down from high-road employer defection. Rearranging the reputation diffusion criterion yields $w \leq y[1-(1-\delta)(p+v-pv)^{-1}] \equiv w^{max}$, which increases in worker informedness p and average employer visibility v. In other words, stronger public reputation systems (higher p's) and better private sources of information (higher v's) both increase the upper bound on wages that are supportable in the absence of enforceable contracts.

4. Setting

4.1. M-Turk

M-Turk is an online labor market that allows employers (these purchasers of labor are called requesters) to crowdsource human intelligence tasks (HITs) to workers over a web browser. Common HITs include audio transcription, image recognition, text categorization, and other tasks not easily performed by machines. M-Turk allows employers to process large batches HITs with greater flexibility and at generally much greater speeds and lower costs than traditional employment.

Amazon does not generally publish detailed usage statistics; however, in 2010, it reported that more than 500,000 workers from over 190 countries were registered on M-Turk.[8] In 2014, Panos Ipeirotis's web crawler found that the number of available HITs fluctuated between 200,000 and 800,000 from January and June 2014.[9] Ross et al. (2009) found that a majority of workers were female (55%) and from the U.S. (57%) or India (32%). Horton and Chilton (2010) estimate that the median reservation wage was $1.38 an hour. M-Turk's platform revenue comes from 10% brokerage fees paid for by employers.

M-Turk specifically has many features making it attractive for studying how workers navigate employer heterogeneity using public employer reputation.[10] First, there is no variation in the terms of contracts. In most labor markets, relationships embody a mix of enforceable and unenforceable elements and the nature of the mix is unknown to the researcher; observed differences between employers may reflect differences in workers' contracts and access to legal recourse. In M-Turk, workers put forth effort, employers acquire the work product, and then employers choose whether to pay workers. Employers may refuse payment for any reason or no reason, and workers have no contractual recourse. This complete lack of contract enforcement is rare and valuable for research, although potentially maddening for workers. Here, one can be sure that all employer behavior is discretionary and is performed absent the possibility of enforcement. Second, M-Turk does not have a native employer-reputation system, a feature it shares with offline labor markets but unlike other online labor markets. This also proves useful by allowing us to decouple worker effort from employer reputation in the audit study.

To help avoid employer opportunism, many M-Turk workers use Turkopticon, a third-party browser plugin that allows workers to review and screen employers (Silberman and Irani 2016). There are several reasons these ratings may be uninformative. First, the system is unnecessary if workers face no information or enforcement problem. Second, the system relies on workers voluntarily contributing accurate, private information to a common pool, which costs time and directs other workers to scarce, high-paying tasks. This distinguishes labor markets from consumer markets where trade is non-rival. Third, ratings systems vary widely in their informativeness due to reputation inflation and other issues (Nosko and Tadelis 2015, Horton and Golden 2015). Anyone can post any review on Turkopticon. It has no revenue and is maintained by volunteers.

[8] Available online at `https://archive.fo/FaVE`

[9] Available online at `http://www.mturk-tracker.com` (accessed June 14, 2014).

[10] In legal terms, M-Turk is a brokerage that facilitates relationships between two contracting parties: one that seeks work for pay and another that performs work. We use "employer" as shorthand for the former.

When an employer posts a task, it appears to workers on a list of available tasks. This list specifies a short description of the task, the number of tasks available in the batch, the promised pay per task, the time allotted for workers to complete the task once they accept it, and the name of the employer. The employer also may restrict eligibility to workers with a sufficiently high approval rating, which requires a history of having submitted work approved and paid for by past employers. Workers may preview the task before accepting. Upon acceptance, a worker has the allotted time to submit the task. The employer then has a predetermined period to approve or reject the task, with or without an accompanying note. If the employer approves the task, the employer pays the posted rate and broker fees to Amazon. The conditions for approval are not contractible; if the employer rejects the task, the worker's submitted work remains in the employer's possession but no payment is made. Moreover, the worker's approval rate will decline, reducing the worker's eligibility for other tasks in the future. There is no process for appealing a rejection.

Opportunism takes many forms in this market. Employers may disguise wage theft by posting unpaid trial tasks, implicitly with the promise that workers who submit work that matches a known, correct answer will receive work for pay, when in fact the trial task is the task itself and the employer rejects all submitted work for being defective. In addition to nonpayment, employers may also advertise that a task should take a set amount of time when it is likely to take much longer. Therefore, although the promised pay for accepted submissions is known, the effective wage rate, depending on the time it takes to complete the task, is not. Employers can also delay accepting submitted work for up to thirty days. Employers may or may not communicate with workers. Employers can also differ in how great they are to workers. Some might not really do much quality control and pay for all work regardless of how bad it is or pay a lot for very easy tasks.

4.2. Reputation on M-Turk

Within M-Turk, there is no tool allowing workers to review employers, and workers cannot observe employers' effective wages or payment histories. However, several third-party resources allow workers to share information voluntarily regarding employer quality. These include web forums, automatic notification resources, and public-rating sites.[11]

We test our hypotheses regarding the value of the employer reputation system using Turkopticon, a community ratings database and web-browser plugin that we estimate is used by a slight majority of M-Turk jobseekers.[12] The plugin adds information

[11] Popular resources include Reddit's HitsWorthTurkingFor, CloudMeBaby.com, mturkforum.com, mturkgrind.com, turkalert.com, turkernation.com, turkopticon.ucsd.edu.

[12] For details on our estimates, see the end of the Study 2 results section. Silberman et al. (2010), Irani (2012), Silberman (2013) provide background on Turkopticon.

to the worker's job search interface, including community ratings of an employer's communicativity, generosity, fairness, and promptness. Ratings take integer values from one to five. As of November 2013, Turkopticon included 105,909 reviews by 8,734 workers of 23,031 employers. The attributes have a mean of 3.80 and a standard deviation of 1.72.[13] Workers can click on a link to read text reviews of an employer. These reviews typically further recommend or warn against doing work for a given employer. Figure 1 provides an illustration.

Figures 1 and 2 illustrate an M-Turk worker's job search process. Figure 1 shows how workers search for tasks for pay. Figure 2 shows a preview of the task that we use for this study.

Turkopticon is remarkable because it relies on voluntary feedback from a community of anonymous workers to provide a signal of employer quality. These reviews are costly in terms of the worker's time and the content of the review is unverifiable to other workers. More importantly, there is wide variation in the effective pay rate of individual tasks. Because employers typically post tasks in finite batches and allow workers to repeat tasks until the batch is completed, the wage-maximizing behavior would be to hoard tasks posted by good employers by misdirecting other workers.[14] Because reviews are anonymous, direct reciprocity and punishment is limited. As such, sharing honest reviews could be thought of as a prosocial behavior that is costly to the worker in terms of time and valuable private information, and in which social recognition or direct reciprocity is limited. Other studies of online reputation systems suggest that reviewers are primarily motivated by a "joy of giving" and fairness (Cornes and Sandler 1994, Resnick and Zeckhauser 2002).

5. Experiment 1

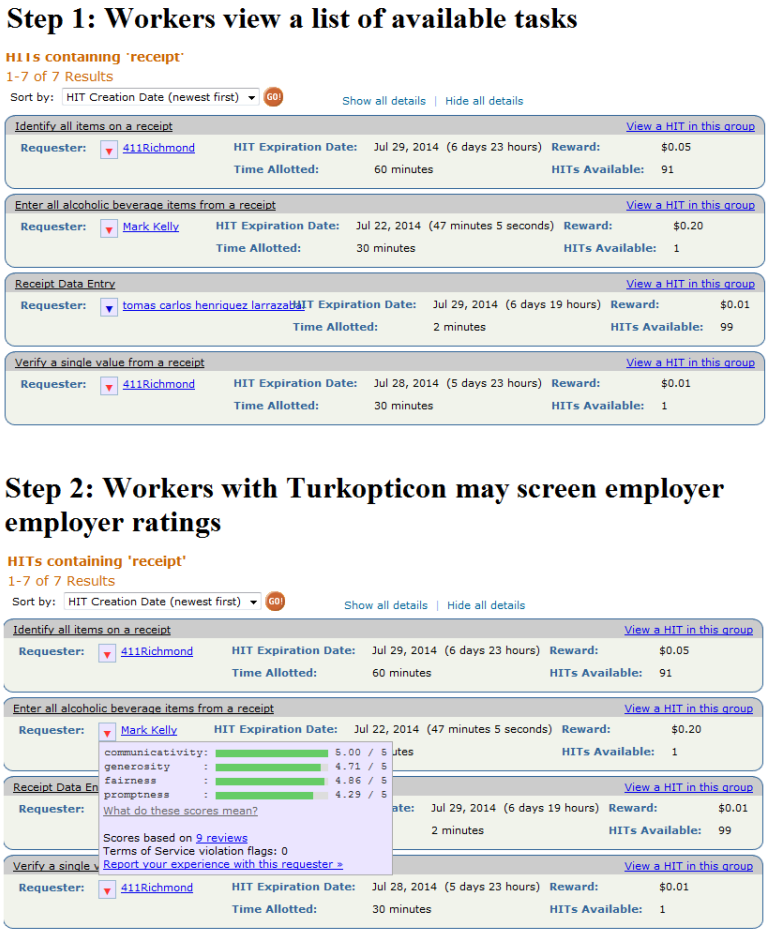
5.1. Setup

The first experiment examines the value of the reputation system to employers. Specifically, we examine whether a good reputation helps employers attract workers. We do so

[13] These statistics are based on our analysis of data scraped from the site. Attribute ratings are determined by the mean from the following questions: (i) for communicativity, "how responsive has this requester been to communications or concerns you have raised?" (ii) for generosity, "how well has this requester paid for the amount of time their HITs take?" (iii) for fairness, "how fair has this requester been in approving or rejecting your work?" (iv) for promptness, "how promptly has this requester approved your work and paid?" Their means (standard deviations) are respectively 4.01 (1.68), 3.98 (1.62), 3.71 (1.68), and 3.18 (1.91), suggesting that ratings are meaningfully spread. Their number of reviews are 93,596, 93,025, 99,437, and 44,298. Reviews are somewhat consistent across dimensions; the correlation between any one dimension and the mean value of the other three dimensions is 0.57. On workers' displays, average ratings are color coded; scores less than 2 are red, scores between 2 and 3 are yellow, and scores greater than 3 are green.

[14] This competition between workers to get the best jobs is the basis of resources such as TurkAlert.com, which allows workers to receive an alert whenever employers of their choosing post new tasks.

FIGURE 1: M-Turk worker's job search process: Turkopticon



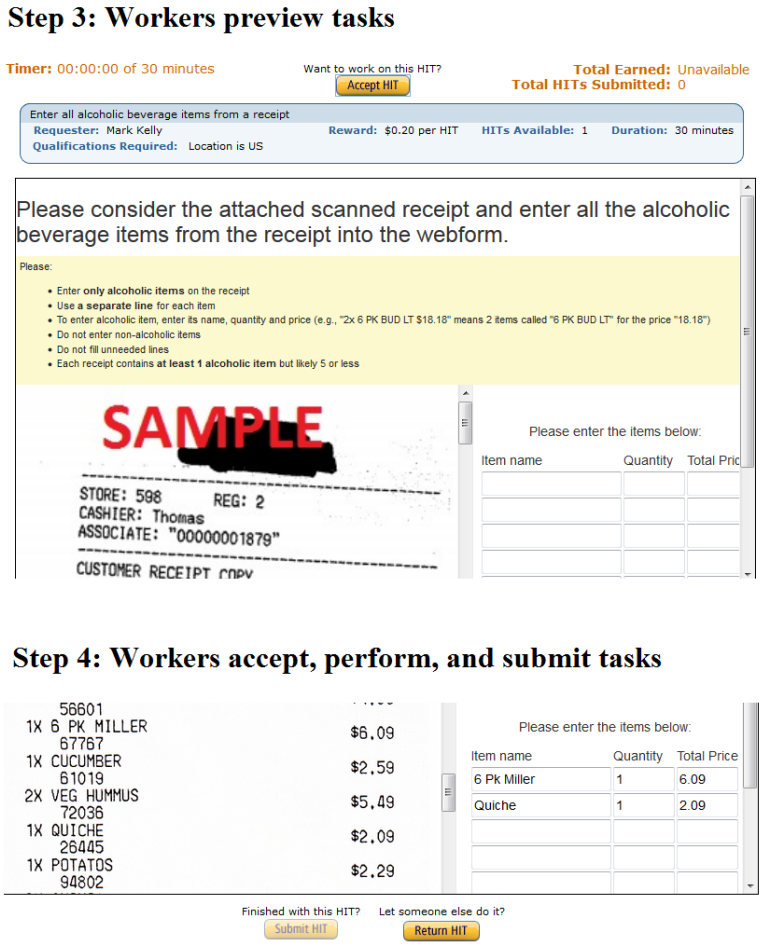
NOTE – Screen capture of a M-Turk worker's job search interface. The tooltip box left-of-center is available to workers who have installed Turkopticon, and shows color-coded ratings of the employer's communicativity, generosity, fairness, and promptness. It also offers a link to long-form reviews.

by creating employers on M-Turk, exogenously endowing them with reputations on Turkopticon, and then testing the rate at which they attract work.

1. We create 36 employer accounts on M-Turk. The names of these employers consist of permutations of three first names and twelve last names.[15] We use multiple employers to protect against the evolution of ratings during the experiment. We choose these names because they are: common, Anglo, male (for first names), and our analysis of Turkopticon ratings find that these names are not generally rated high or low.

[15] The first names are Joseph, Mark, and Thomas. The last names are Adams, Clark, Johnson, Jordan, Kelly, Lewis, Martin, Miller, Owens, Roberts, Robinson, and Warren.

FIGURE 2: M-Turk worker's job search process: previewing, accepting, and submitting tasks



NOTE – Screen capture of a M-Turk worker's job search interface. From the list of tasks, workers must choose to preview a task before accepting the task. They then enter data into the webform and submit their work.

2. We endow 12 employers with good reputations, 12 employers with bad reputations, and leave 12 employers with no ratings or reputation. We create accounts on Turkopticon and post numerical attribute ratings and longform text reviews. Reviews for our bad-(good-)reputation employers are taken as a sample of actual bad(good) reviews of bad-(good-)reputation employers on Turkopticon.[16] Good- and bad-reputation employers receive eight to twelve reviews each. These reviews make our good- and bad-reputations not unusual with regard to their mean reputations, although the bad-reputation

[16] For this purpose, we define bad reviews as those giving a score of 1/5 on all rated attributes and a good review as giving a 4/5 or 5/5 on all rated attributes. The text reviews clearly corroborate the numerical rankings; an RA given only the text reviews correctly identified the employer type in 285 of the 288 reviews.

employers do have an unusual degree of rater consensus about their badness.[17] Because M-Turk workers may sort tasks alphabetically by employers' names, we balance reputations by the first name of the employer so that reputation is random with respect to the alphabetical order of the employer.

3. Our employer identities take turns posting tasks on M-Turk. They do so in seventy-two one-hour intervals, posting new tasks on the hour. At the start of each hour, the employer posted hundreds of tasks, more than were ever done within the hour. Each worker was allowed to do only one task and this was apparent when browsing the job listing. Posts began at 12:00 AM on Tuesday, July 7 and ended at 11:59 PM on Thursday, July 9. For example, the employer named Mark Kelly, who was endowed with a good reputation on Turkopticon, posted tasks at 12:00 AM and ceased accepting new submissions at 12:59 AM, thereafter disappearing from workers' search results. At 1:00 AM, Joseph Warren, who had no reputation on Turkopticon, posted new tasks.

   We balance the intervals so that: (1) in each hour, over three days, the three reputation types are represented once, (2) in each hour, over each six-hour partition of a day, the three reputation types are represented twice. We chose the final schedule (Table 1) at random from the set of all schedules that would satisfy these criteria.

   The tasks consist of image recognition exercises. Workers are asked to enter the names, quantity, and prices of alcoholic items from an image of a grocery receipt that we generated. Receipts are twenty items long and contain three to five alcoholic items.[18] Workers may only submit one task in any one-hour interval. The pay rate is \$0.20, and workers have fifteen minutes to complete the task once they accept it.

4. Simultaneously, we create three employers that post 12-cent surveys requesting information from workers' dashboards. These employers post new batches of tasks each hour for twenty-four hours each. Their reputation does not vary. The purpose of this task is to determine a natural baseline arrival rate that could be used as a control in the main regressions.

[17] At the time of the experiment, of the 23,095 employers rated on Turkopticon, 22.9% met out criteria for being bad-reputation and 48.1% met our definition of being good-reputation. Many of the bottom ratings come from employers with few reviews. Of the 1,564 employers with 8-12 reviews, only 1.1% met our definition of bad and 41.5% met our definition of good. In this sense, our "good" employers had good ratings but not uncommonly so. However, the mean ratings of our bad employers are especially bad given the consensus across so many raters that they merit 1 on all dimensions. As such, our estimate of the effect of bad-reputation (relative to good and no reputation) might be interpreted as an upper bound where about half of workers are using this reputation system to screen employers.

[18] Alcoholic items came from a list of 25 bestselling beers. This task therefore features simple image recognition, abbreviation recognition, and domain knowledge.

TABLE 1: Balanced, random allocation of employer identities to time-slots with reputation

	Tuesday	Wednesday	Thursday
0:00	(g) Mark Kelly	(b) Thomas Jordan	(n) Mark Jordan
1:00	(n) Joseph Warren	(g) Joseph Jordan	(b) Mark Warren
2:00	(g) Thomas Warren	(n) Mark Jordan	(b) Joseph Kelly
3:00	(n) Thomas Kelly	(b) Thomas Jordan	(g) Thomas Warren
4:00	(b) Mark Warren	(n) Joseph Warren	(g) Mark Kelly
5:00	(b) Joseph Kelly	(g) Joseph Jordan	(n) Thomas Kelly
6:00	(g) Joseph Lewis	(n) Thomas Lewis	(b) Mark Lewis
7:00	(n) Mark Roberts	(g) Thomas Roberts	(b) Thomas Clark
8:00	(b) Thomas Clark	(n) Thomas Lewis	(g) Mark Clark
9:00	(g) Mark Clark	(b) Mark Lewis	(n) Joseph Clark
10:00	(n) Joseph Clark	(b) Joseph Roberts	(g) Joseph Lewis
11:00	(b) Joseph Roberts	(g) Thomas Roberts	(n) Mark Roberts
12:00	(b) Thomas Martin	(n) Joseph Johnson	(g) Joseph Martin
13:00	(n) Thomas Adams	(b) Joseph Adams	(g) Mark Adams
14:00	(n) Mark Martin	(g) Mark Adams	(b) Mark Johnson
15:00	(g) Thomas Johnson	(n) Thomas Adams	(b) Joseph Adams
16:00	(b) Mark Johnson	(g) Thomas Johnson	(n) Mark Martin
17:00	(h) Joseph Martin	(b) Thomas Martin	(n) Joseph Johnson
18:00	(n) Thomas Miller	(b) Joseph Robinson	(g) Thomas Robinson
19:00	(g) Thomas Robinson	(n) Mark Robinson	(b) Thomas Owens
20:00	(g) Mark Owens	(b) Joseph Robinson	(n) Mark Robinson
21:00	(n) Joseph Owens	(g) Joseph Miller	(b) Mark Miller
22:00	(b) Mark Miller	(n) Thomas Miller	(g) Joseph Miller
23:00	(b) Thomas Owens	(g) Mark Owens	(n) Joseph Owens

NOTE – Parentheses denote employers endowed with good (g), no (n), and bad (b) reputations.

5. We record the quantity and quality of completed tasks. We do not respond to communications and do not pay workers until the experiment concludes.

Note that employer reputation may affect the labor supply through multiple causal mechanisms. One obvious mechanism is that the workers would just consult public reputation of the employer directly each time and pick or reject the jobs accordingly. The other mechanism would be that some workers who discover a "good" employer would invite others to work for this "good" employer as well[19]. Accordingly, some workers who discover a "bad" employer may warn others not to work for the "bad" employer. In other words, the effect of employer reputation as examined in our study is not limited to just the first-order effect but rather is the overall effect on labor supply.

[19] In both of these cases, the workers have no information of their own and thus, base their decisions only on the available public reputation that they see

Since our experiment was three full days long, we also considered the possibility that the reputation of our employers may start evolving as workers will eventually start adding their own true feedback into the reputation system or may eventually start noticing the similarities and patterns between the employers. In order to account for this, we switched between employers rather quickly – every hour. In addition to that, we also used a balanced randomization procedure such as every six-hour slot[20] of the day has at least 2 good employers, 2 bad employers and 2 neutral employers as presented in our schedule in Table 1. This way none of the treatment groups is affected disproportionally by any potential time evolution of the experiment and each six-hour slot offers equal exposure to all treatment groups.

As we were monitoring the progress of the experiment, indeed, on Thursday[21] at 4:14 PM, an observant worker compiled and publicly announced a list of our 24 employers with good and bad ratings on Reddit forum, noting their similarities and suggesting that the reviews were created by fake accounts. On Thursday at 5:22 PM, we reached out to this worker and to a concerned group of other workers on a Turkopticon discussion board to address the concerns regarding our employers falsifying reviews with the intent of defrauding workers. We disclosed to this group of workers that all workers would be paid. On Thursday at 6:14 PM, the description of the experiment was cross-posted on Reddit. As this possibility was considered in our randomization procedure, we present two sets of the results: one including this last 6 hour shift and one excluding this last 6 hour shift. As can be seen and discussed below in Table 2, our results are qualitatively identical in both cases.

5.2. Results

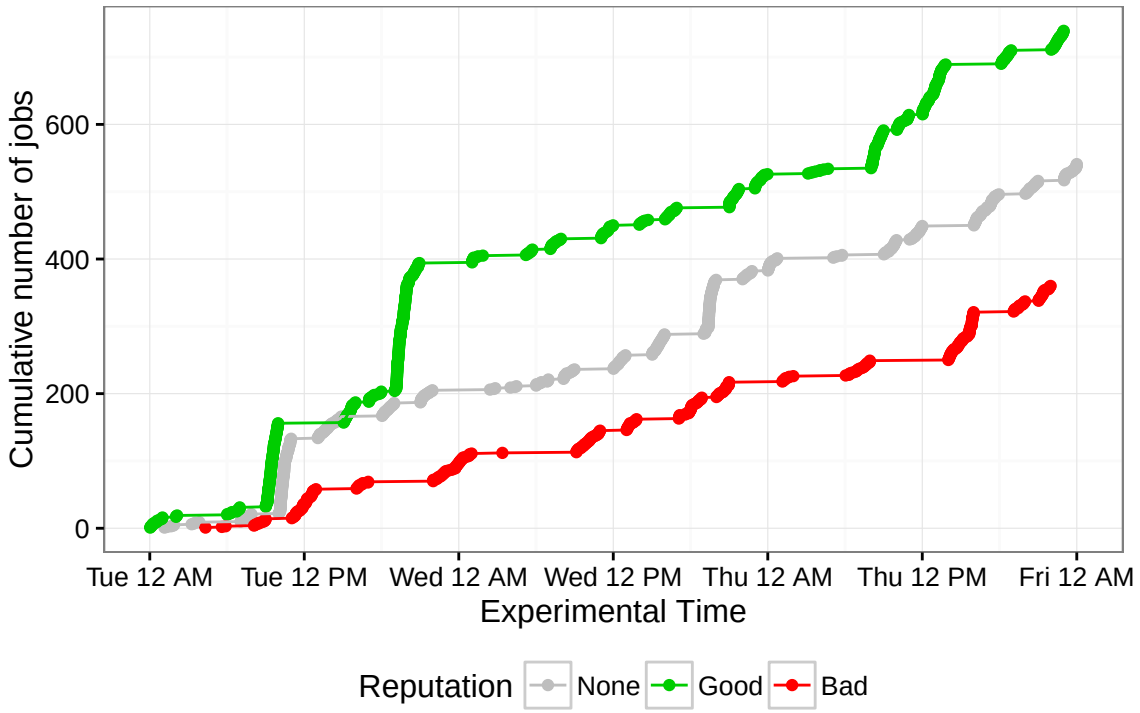
Summarizing the results of the experiment, Figure 3 shows the cumulative distribution of arrivals across the three employer reputation types. As can be seen in Figure 3, our results are qualitatively consistent throughout the time of the experiment starting from the very beginning owing to the balanced time allocation schedule that we discussed above. By the conclusion of each of the twelve six-hour partitions, the employer with good ratings had attracted more work than the employer with neutral ratings, and the employer with neutral ratings had attracted more work than the employer with poor ratings.

Table 2 shows results from a negative binomial model. In all samples except for the six-hour partitions, employers with good reputations attract work more quickly than employers with poor reputations with $p < 0.01$. However, if comparing only against no-reputation

[20] There are 4 six-hour slots in each day: 12am-5am, 6am-11am, 12pm-5pm, 6pm-11pm that roughly correspond to night, morning, day and evening shifts

[21] Thursday is the last day of the three days of the experiment.

FIGURE 3: Cumulative accepted jobs by employer reputation



NOTE – Bold points represent active job listings.

employers at a 5% significance level, employers with a good reputation do not receive submitted work significantly faster than those with no reputation, and employers with a poor reputation receive submitted work significantly slower only in the full samples.

While finding workers at a slower pace may not seem like such a major punishment for employers with poor reputation at first, we would like to examine this finding very carefully: the only people who are working for the bad employers are "uninformed" workers. In other words, if the labor market has approximately 50% informed workers and 50% uninformed workers (such as Amazon Mechanical Turk as of the time of the experiment as described below), the bad employers would indeed get the work done in approximately twice the time. However, in the labor market where almost everyone is informed, the already slow pace of work may become too slow for the bad employer to survive in the market.

In addition to the above, we also examine differences in estimated effort and quality. The mean time spent per task for good reputation, no reputation, and poor reputation employers were respectively 136, 113, and 121 seconds. The difference between good reputation and no reputation employers is statistically significant with $p < 0.01$. For each of the three groups, the error-free rates were between 61% and 63% and the major-error rates (e.g., no

TABLE 2: Negative binomial regression for arrival of submitted tasks and other events

	Good Reputation		No Reputation			
Sample	β	SE	β	SE	periods	events
Event: submitted tasks						
Full sample						
(1) All submitted tasks	2.053*	(0.500)	1.503	(0.368)	72	1,641
Subsamples						
(2) Day 1 only	4.104*	(1.969)	2.135	(1.030)	24	695
(3) Day 1-2 only	2.424*	(0.766)	1.76	(0.559)	48	1,125
(4) 12AM-6AM	1.679	(0.823)	1.393	(0.689)	18	114
(5) 6AM-12PM	2.843*	(1.201)	2.157	(0.915)	18	534
(6) 12PM-6PM	1.096	(0.267)	.978	(0.239)	18	415
(7) 6PM-12AM	2.694*	(0.955)	1.648	(0.589)	18	577
Excluding last 12 hours						
(8) No controls	2.466*	(0.704)	1.803*	(0.516)	60	1313
(9) Controls for baseline rate	2.523*	(0.719)	1.808*	(0.515)	60	1,313
(10) Day fixed effects	2.294*	(0.654)	1.778*	(0.498)	60	1,313
(11) Hour fixed effects	1.858*	(0.274)	1.374*	(0.205)	60	1,313
(12) Day and hour fixed effects	1.836*	(0.262)	1.364*	(0.196)	60	1313
Event: other						
(13) Task previews	2.314*	(0.571)	1.495	(0.370)	72	1,837
(14) Task accepts	2.141*	(0.529)	1.551	(0.384)	72	1,799
(15) Error-free submissions	2.018*	(0.548)	1.5	(0.410)	72	1,012
(16) First submissions	2.871*	(0.804)	1.644	(0.465)	72	899
(17) Error-free first submissions	2.88*	(0.928)	1.641	(0.536)	72	508

NOTE – * $p < 0.05$. Each row is a regression. Coefficients are incident rate ratios with bad reputation as the omitted category. Standard errors in parentheses.

alcoholic items identified) were between 3.0% and 5.2%. Differences in the error-free rates and major-error rates are not statistically significant.[22] Mason and Watts (2010) also found that higher payments raise the quantity, but not quality, of submitted work.

In the full sample, 45.2% of the submitted tasks were not the first tasks submitted by an individual worker, and 9.7% of the submitted tasks were the sixth task or greater. The high incidence of repeat submissions may be for a number of factors, including: power-users, correlated task search criteria (e.g., individuals continuously search using the same criteria), automated alerts (e.g., TurkAlert), or purposely searching for the same task across hours.

Table 3 shows results from our preferred specification of the negative binomial regressions to estimate the arrival rates of task previews, acceptances, submissions, first submissions

[22] Differences are for a two-sample t-test for equal means of the log-work time with $\alpha < 0.1$. Error-free receipts are those in which all alcoholic items were identified, no non-alcoholic items were identified, and the prices were entered correctly. Major-error receipts are those in which no alcoholic items were identified, or more than six items are listed.

TABLE 3: Preferred specification: negative binomial regression of arrival rates in the first sixty hours with day and hour fixed effects

	Previews (1)	Acceptances (2)	Submissions (3)	First submissions (4)	Correct first submissions (5)
Good reputation	1.964*	1.909*	1.836*	2.488*	1.855*
	(0.280)	(0.277)	(0.262)	(0.426)	(0.405)
No reputation	1.403*	1.387*	1.364*	1.608*	1.261
	(0.204)	(0.203)	(0.196)	(0.277)	(0.278)
Constant	16.56*	14.10*	13.31*	8.024*	3.54*
	(4.907)	(4.300)	(4.002)	(2.788)	(1.729)
Day FE	Yes	Yes	Yes	Yes	Yes
Hour FE	Yes	Yes	Yes	Yes	Yes
Observations	60	60	60	60	60

NOTE – *p< 0.05. Standard errors in parentheses. Bad reputation is the omitted category. All coefficients for good employers are significantly different from coefficients for bad employers with $p < 0.05$.

(by worker), and correct first submissions. These specifications omit the last twelve hours in which the experiment was disclosed and also include day and hour fixed effects. Arrival rates for good-reputation employers are significantly greater than no reputation employers for all outcomes, and arrival rates for no reputation employers are significantly greater than bad-reputation employers for all outcomes except correct first submissions with $p < 0.05$. Results provide evidence that good reputations produce more previews, acceptances, submissions, first submissions, and correct first submissions.

The point estimates in column (3) suggest that arrival rates for employers with good and no reputations respectively exceed those of employers with bad reputations by 84% and 36%.

Table 3 also provides evidence about the effects of reputation on various steps in the matching process. Conditional on a worker previewing a task, the probability of accepting the task is not significantly different by treatment. If information received by previewing a task (e.g., the type of the task, the intuitiveness of the user interface) were a substitute for reputation information, then good-reputation employers would lose fewer workers during the preview stage than no-reputation employers. In the former, but not latter, workers would already have received the signal prior to previewing the task. This evidence suggests that observable task characteristics do not substitute for reputation information. The reputation system adds information above what workers can otherwise observe.

Turkopticon is not native to the M-Turk interface and must be installed by the worker. As such, the reputations we endow are visible only to a fraction of workers, and so only part of

the "treated" population actually receives the treatment. To estimate the share of M-Turk jobseekers who use Turkopticon, we posted a one-question, free response survey asking, "How do you choose whether or not to accept HITs from a requester you haven't worked for before? Please describe any factors you consider, any steps you take, and any tools or resources you use." Because we posted the survey from a requester account that did not have a Turkopticon rating, and because we require workers to identify Turkopticon specifically, we expected this procedure to yield a conservative estimate of the true portion of jobseekers who use Turkopticon. Of these, fifty-five of the 100 responses mention Turkopticon explicitly, and seven other responses mention other or unspecified websites.[23]

Experiment 1 also offers three additional pieces of evidence that Turkopticon provides information of employer type rather than task type. First, we find that observed probability of accepting a task conditional on previewing a task does not vary significantly by employer type. Second, we find that the elapsed time that workers spend previewing tasks prior to accepting the task does not vary significantly by reputation type. Third, our survey of 100 M-Turk workers featured no workers who reported a belief that certain tasks were inherently more fairly or highly compensated, though nearly all cited observable employer characteristics from past experience or tools like Turkopticon. These suggest that workers screened on Turkopticon ratings and not on information (e.g., task type) gathered during the task previews. This, along with ratings criteria used by Turkopticon and the test in Experiment 1, lead us to conclude that workers use Turkopticon to get information about employers that wouldn't be accessible until after they would have otherwise exerted effort (e.g., time to completion and nonpayment), rather than getting information on task type.

5.3. Alternative dependent variables

Employers, especially on Mechanical Turk, want to get work done quickly, cheaply, and accurately. Our results suggest that better reputations help employers attract more workers at a given price and quality, giving an advantage in speed. What if employers wished to use their reputation to achieve lower prices or greater quality?

Horton and Chilton (2010) estimate that M-Turk workers have a median wage elasticity for recruitment of 0.43. If this point elasticity holds for our sample, a bad-reputation employer that pays \$0.59, a no-reputation employer that pays \$0.37, and a good-reputation employer that pays \$0.20 would attract work at the same rate. This is a conservative

[23] Otherwise, responses emphasize estimated pay, estimated time to completion, and perceived trustworthiness (e.g., from a known organization). To the extent one is interested in the effect of reputation among informed workers, this treatment-on-treated effect is 82% ($=0.55^{-1}$) larger than the estimated effect in the observed equilibrium. The estimated effect is a weighted average of a larger effect among workers who use the reputation system and a zero effect among those who don't.

estimate. Horton & Chilton's estimate of the mean elasticity is lower (0.24), implying that generating as large a difference in worker arrival rates as reputation generates would require even larger differences in promised payments. Dube et al. (2018) synthesize existing evidence, including Horton & Chilton and subsequent experiments, and new evidence they generate to estimate that the mean recruitment elasticity is even lower (0.06), implying that reputation is even more valuable as a substitute for higher wages in generating changes in recruitment and worker arrival rates to an employer for a given job posting. Moreno and Terwiesch (2014) also find that online service providers on vWorker.com substitute between higher prices and greater volume. Our study focuses on recruitment, leaving effects on retention for future work.

To estimate the value of a good reputation for getting work of better quality, consider moving to a majority-rules process. In particular, each alcoholic item costs an average of \$0.030 in our study, and each item was coded correctly with a probability of $p = 0.890$. If a third rater is used only if the first two raters disagree, then the average cost per item will rise to \$0.071, and the probability an item is coded correctly will rise to 0.966.[24] The elasticity estimates above imply that reducing price per worker-task to hold average costs per completed task constant will reduce quantity of work completed by 23.7%, less than the quantity gained by a good reputation relative to no reputation. In other words, a good-reputation employer could implement a majority-rules process, cut the price per worker-task so as to achieve the same cost per completed task, improve accuracy from 0.890 to 0.966, and still get work done more quickly than an employer with no reputation, though doing so may compromise the employer's good reputation (especially "generosity" ratings) in the long run.

6. Experiment 2

6.1. Setup

Our model assumes that the reputation system provides accurate information to workers. In reality, the cheap talk, voluntarily-contributed ratings could be biased or very noisy, such that "informed" who make decisions based on the reputation system are no more informed than others. Our second experiment validates this assumption empirically and examines the value of the reputation system to workers.

Specifically, we examine whether Turkopticon ratings are informative of three employer characteristics that workers value but about which they face uncertainty during the

[24] Assuming errors are independent, the expected cost is $2c[p^2+(1-p)^2]+3c[1-p^2-(1-p)^2]$. The probability of a correct decision is $p^2+2(p^2(1-p))$.

search process: the likelihood of payment, the time to payment, and the implicit wage rate. As reflected in the literature on online ratings, informedness shouldn't be taken for granted. Horton and Golden (2015) show that oDesk, an online labor market with a native bilateral rating system, experiences extensive reputation inflation as employers and workers strategically, rather than truthfully, report experiences. Others report similar biases on eBay (Dellarocas and Wood 2008, Nosko and Tadelis 2015), Airbnb (Fradkin et al. 2015), and Yelp (Luca 2016). The validity of Turkopticon ratings may be even more surprising, given that tasks offered by revealed good employers are rival (unlike, for example, good products on retail markets).

We follow the following procedure:

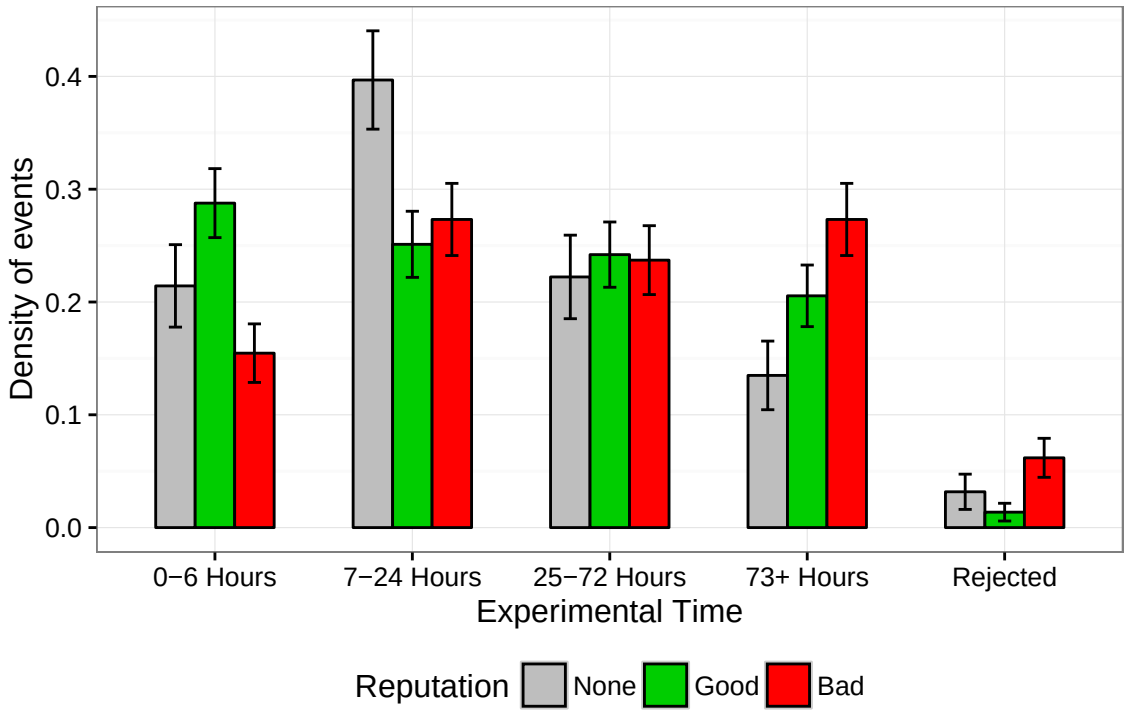
1. We produce a random ordering of three reputation types: Good, Bad, and None.
2. The nonblind research assistant (RA1), using a browser equipped with Turkopticon, screens the list of tasks on M-Turk until finding one that meets the requirements of the next task on the random ordering.[25]
   - If the next scheduled item is Good, RA1 searches the list for a task posted by an employer in which all attributes are green (all attributes are greater than 3.0/5). 26.3% of the 23,031 employers reviewed on Turkopticon meet this criterion.
   - If the next scheduled item is Bad, RA1 searches the list for a task posted by an employer with no green attributes and a red rating for pay (all attributes are less than 3.0/5, and pay is less than 2.0/5). 21.6% of employers reviewed on Turkopticon meet this criterion.
   - If the next scheduled item is None, RA1 searches the list for a task posted by an employer with no reviews.
3. RA1 sends the task to the blinded RA2, who uses a browser not equipped with Turkopticon.
4. RA2 performs and submits the task. RA2 is instructed to perform all tasks diligently.[26]
5. RA1 and RA2 repeat steps 2-4. A web crawler records payments and rejections by employers to RA2's account with accuracy within 1 minute of actual payment or rejection.

The blinding procedure decouples the search process from the job performance process, thereby protecting against the risk that RA2 inadvertently conditions effort on the employer's reputation.

[25] To hold skill constant, the RA omitted any tasks requiring Master's qualification. The task classifications were uncorrelated with Turkopticon scores.

[26] RA2 was not able to complete all jobs sent by RA1. Some expired quickly. Also, bad-reputation employers' jobs were more likely to be so dysfunctional as to be unsubmittable.

FIGURE 4: Time to payment and rejection by employer reputation



NOTE – Whiskers represent standard errors. p-values for a χ^2 test that shares are independent of reputation are respectively: 0.002, 0.011, 0.805, 0.012, and 0.007.

6.2. Results

Figure 4 shows results for rejection rates and time-to-payment by the employer's reputation type. Rejection rates were 1.4 percent for employers with good reputations, 4.3 percent for employers with no reputation, and 7.5 percent for employers with bad reputations.

Table 4 presents further results and significance tests for rejection rates, time-to-payment, and realized hourly wage rates. We define realized wage rates to be payments divided by the time to complete the task if the work is accepted and zero if the work is rejected. We define promised wage rates to be posted payments divided by the time to complete the task; they are not zero if the work is rejected.[27] Employers with good reputations have significantly lower rejection rates and faster times-to-decisions. They do not have statistically different posted pay rates. This distinction is important because the pay for accepted tasks is contractible but the task's acceptance criteria and realistic time requirements are not.

In principle, the ratings on Turkopticon could be orthogonal to employer type, and instead be providing information on task types (e.g., survey or photo categorization) rather than

[27] Counts are lower for wage rates because the blinded RA lost track of time-to-completion for some tasks.

TABLE 4: Rejection and time-to-payment by employer reputation

	Mean	Std. Error	N	paired test p-values: Good	None	Bad
Main outcomes						
1. Rejection rates						
Good Reputation	0.013	0.008	223		0.073	0.003
No Reputation	0.043	0.016	164	0.073		0.246
Bad Reputation	0.071	0.018	211	0.003	0.246	
2. Days to decision						
Good Reputation	1.679	0.146	223		0.132	0.001
No Reputation	2.296	0.433	164	0.132		0.03
Bad Reputation	3.715	0.467	211	0.001	0.03	
3. Realized wage rates						
Good Reputation	2.834	0.228	173		0.011	0.043
No Reputation	1.957	0.259	141	0.011		0.949
Bad Reputation	1.986	0.352	168	0.043	0.949	
Other outcomes						
4. Days to decision, accepts only						
Good Reputation	1.643	0.144	220		0.083	0.001
No Reputation	2.368	0.451	157	0.083		0.023
Bad Reputation	3.943	0.499	196	0.001	0.023	
5. Promised wage rates						
Good Reputation	2.834	0.228	173		0.017	0.098
No Reputation	2.011	0.257	141	0.017		0.771
Bad Reputation	2.142	0.352	168	0.098	0.771	
6. Advertised pay						
Good Reputation	0.277	0.025	223		0.001	0.938
No Reputation	0.159	0.024	164	0.001		<0.001
Bad Reputation	0.28	0.022	211	0.938	<0.001	
7. RA log-seconds to complete						
Good Reputation	5.737	0.228	173		0.372	<0.001
No Reputation	5.639	0.085	141	0.372		0.001
Bad Reputation	6.368	0.069	168	<0.001	<0.001	

NOTE – Rejection rate p-values are from a χ^2 test that rejection rates are the same between the row and column. Time-to-pay p-values are from a two-sample t-test that the mean times-to-pay are the same between the row and column.

employer types. We do not find evidence that this is the case. First, Turkopticon requests workers to rate employers on fairness, communicativity, promptness, and generosity; unlike task type, these are revealed only after workers have invested effort and are subject to hold-up. Textual comments also emphasize information that would only be revealed to prospective workers after investing effort. Second, the RA's task classifications in

experiment 2 are not significantly correlated with employers' Turkopticon scores. We also found evidence against workers screening on task, rather than employer, in experiment 1.

Given the low cost of creating new employers, it is puzzling that employers with poor reputations persist rather than creating new accounts. When the study was conducted, the only cost to creating a new employer was the time filling forms and awaiting approval. Since then, the cost of producing new aliases has grown.[28] If creating new accounts were perfectly costless and employers were informed, we would expect there to be no active employers with poor reputations. However, Turkopticon's textual reviews also suggest that workers are aware that employers with bad reputations may create new identities.

We conclude that the longer work times and lower acceptance rates validate Turkopticon's ratings. In other words, Turkopticon is informative about employer differences that would be unobservable (or at least more costly to observe) in the absence of the reputation system.

To provide an intuition for the magnitude of the value of employer-reputation information to workers, note that our results imply that following a strategy of doing jobs only for good-reputation employers would yield about a 40 percent higher effective wage than doing jobs only no-reputation or bad-reputation employers: \$2.83 versus just under \$2.00 per hour. Results suggest about 20% of the gap in effective pay is explained by nonpayment and 80% is explained by longer tasks. However, this calculation understates the penalties when an employer rejects tasks because the rejected worker is penalized in two ways: nonpayment and a lower approval rating. The latter reduces the worker's eligibility for future tasks from other employers.

7. Natural experiment

Experiments 1 and 2 above demonstrated the effects of the reputation system on employers and workers on the online market. So far, these results suggest that workers can earn substantially more by screening employers with good reputations, and employers with better reputations attract workers more quickly (or alternatively, for a given speed, more cheaply) than those with no or poor reputations. In this section, we address the question of what happens to the job market when the reputation system suddenly disappears.

7.1. Ideal Experiment

Following (Rubin 1974), we begin by describing an ideal experiment that would identify the causal partial-equilibrium effect of the reputation system on the market. Assume a researcher had the ability to (1) shut down the reputation system at will and for any

[28] On July 27, 2014, Amazon began requiring employers to post a legal personal or company name, physical address, and Social Security Number or Employer Identification Number.

TABLE 5: Summary of Turkopticon down time data

Variable	Value
Number of down time episodes	7.00
Average length of a down time episode (hours)	10.53
Average time between down time episodes (days)	61.54
Total range spanned by time episodes (days)	369.27

periods of time and (2) monitor the entire market, including which jobs are being taken, how fast they are finished, and so on. One could randomly assign the time when the reputation system is removed and randomly decide for how long it is absent.[29] Since such a treatment assignment would be independent of market conditions, one could conclude that any changes observed in the market were caused by the treatment. Acknowledging that it is infeasible and unethical to shut down the reputation system website purposefully, we use a natural-experiment approach with observational data, which serves as an approximation to the ideal experiment described above.

7.2. Observational Data

In order to explore the partial-equilibrium effect of reputation system absences, we exploit the seven instances when the Turkopticon servers went down. To accomplish that, we collected the following data:

- *Turkopticon downtimes.* We assembled data on Turkoption's downtime using timestamps from worker and Turkopticon administrative posts on the Turkopticon website, Reddit, Twitter, and Google Groups. These are summarized in Table 5. The chief concern is that Turkopticon's downtimes are correlated with one of our variables, for example, due to especially heavy traffic. However, all administrative posts attributed crashes to unrelated technical issues, like software updates.
- Individual-task level data on the entire market that was collected by the web crawler M-Turk Tracker (Ipeirotis 2010b, Difallah et al. 2015) and is summarized in Table 6. M-Turk Tracker scans the M-Turk market every 6 minutes and records the status of all HITs that it observes, such as the number of tasks left in a particular HIT, the task description, and the reward offered by the task. By studying the changes in the number of tasks still left in each HIT we can explore how fast the jobs are taken and thus, explore shifts in the supply of labor in this market.

Our first goal is to study the total effect of the reputation system shutdown on the labor market. To do that we examine the amount of work done by M-Turk workers at any given

[29] with the control group being the time when the reputation system is randomly up

TABLE 6: Summary of MTracker data

Variable	Value
Number of hit status observations	504,840,038
Number of distinct requesters	65,243
Number of distinct crawls	267,319
Average time between the crawls (min)	12.07

moment in time with respect to whether or not the reputation system is active at the moment. We measure work being done as the "promised" pay rate of a given task multiplied by the number of tasks that were done (Rewards Earned); we prefer this to the number of tasks alone since quick tasks tend to be cheap. We control for time of the day, day of the week, employer, and the episode using fixed effects. More specifically, we use the following model:

$$\log(1+\text{RewardsEarned}_{it}) = \beta_0 + \beta_1 \text{DOWN}_t + \beta_2 H_t + \beta_3 D_t + \beta_4 R_i + \beta_5 E_t + \varepsilon_{it} \quad (1)$$

where $\text{RewardsEarned}_{it}$ is the total "promised" pay to all workers working on task i at time t, DOWN_t is the indicator variable for whether the reputation system is down at time t ($\text{DOWN}_t = 1$) or not ($\text{DOWN}_t = 0$), H_t is the fixed effect for the hour of the day at time t, D_t is the fixed effect for the day of the week at time t, R_i is the fixed effect of the employer who requested task i, E_t is the fixed effect for the down time episode. The analysis sample is restricted to observations occurring between two weeks before and after the start of a down-time episode. Table 7 below presents results.

As shown in Table 7, the overall job consumption on the market actually increases as the reputation system shuts down. From this result, we conclude that the workers tend to stay in the market when the reputation system shutdown, at least in the short term. There are a number of possible explanations for this, not all of which necessarily correspond to higher pay among workers or better allocation of work. For example, workers might speed up work because they're spending less time screening and reviewing employers. In the short term, this might raise the amount of promised pay earned, but less of this promised pay may be realized (given study 1). In the long term, the lack of a reputation system may impair workers' ability to find good but small employers, and may discourage smaller employers from investing in a good reputation.

In order to examine whether workers' job search changes, we study the heterogeneity of the treatment effect. We want to separate reputation into two dimensions: how good an employer's public reputation is and how widely-known or visible an employer is outside

TABLE 7: Overall effect of reputation system shutdown on the job consumption

	Dependent variable:
	Log(Rewards Earned)
DOWN	0.0034***
	(0.0004)
Hour of day fixed effect	Yes
Day of week fixed effect	Yes
Employer fixed effect	Yes
Down time episode fixed effect	Yes
Observations	5,572,840
R^2	0.1422
Adjusted R^2	0.1419
Residual Std. Error	0.1109 (df = 5570882)
Note:	*p<0.1; **p<0.05; ***p<0.01
	Robust standard errors in parentheses

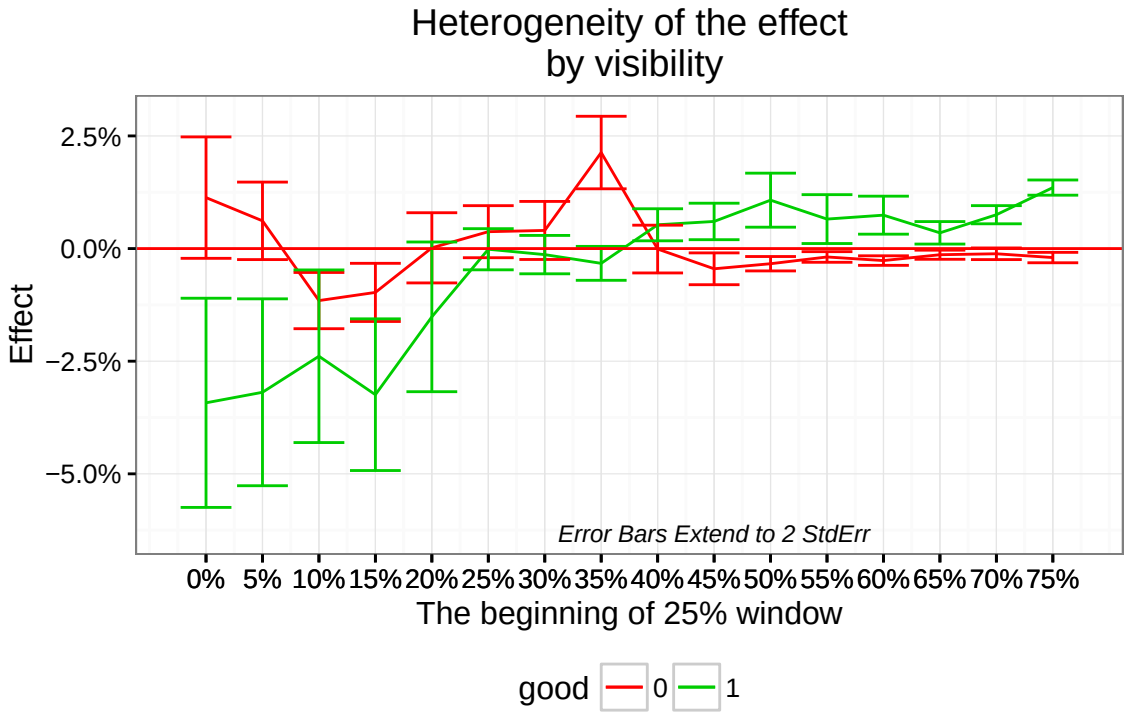
the public reputation system. We measure the quality of an employer's reputation using Turkopticon reviews. We measure the visibility of employer i at time t as the number of times the MTurk-tracker web crawler encountered that employer across all time periods before t. This is designed to capture workers' general familiarity with the employer. Some employers (such as the brokers that use M-Turk to subcontract tasks on behalf of their clients) become well-known among M-Turk workers. However, many employers post jobs only infrequently. Independent of Turkopticon, few workers have private knowledge of these less-visible employers' past behavior. An employer frequently encountered by workers in their day-to-day browsing and work history would also tend to be frequently encountered by the web crawler. On the other hand, if the web crawler (that runs every few minutes) encountered a particular employer only a handful of times then this employer will generally not be familiar to workers.

We perform a semiparametric test to examine heterogeneity by employer reputation and visibility based on the following procedure and plot the results on Figure 5:

1. Pick all the good-reputation employers in the lowest quartile of visibility, whose visibility is in the first 25 percentiles.[30] These are the least-visible, good-reputation employers. Estimate the DOWN coefficient using only jobs for these employers. Denote it $\text{DOWN}_{0\%,\text{good}}$. Plot the estimated coefficient at 0% on the x-axis with a green marker.

[30] Reputations are defined as in experiment 2.

FIGURE 5: The effect of down time depends on visibility of the requester



2. Shift the percentile window by 5%, that is, pick all good-reputation employers between the 5th and bottom 30th percentiles of visibility. Estimate a new DOWN coefficient using only jobs for these employers, denoted by $\text{DOWN}_{5\%,\text{good}}$. Plot the estimated coefficient at 5% on the x-axis and in green.
3. Shift by 5% again and repeat the procedure until $\text{DOWN}_{75\%,\text{good}}$ is estimated, corresponding to the top quartile by visibility of good-reputation employers, that is, the most-visible, good-reputation employers).
4. Repeat the entire procedure for bad-reputation employers in order to estimate $\text{DOWN}_{0\%,\text{bad}}, \text{DOWN}_{5\%,\text{bad}} \ldots, \text{DOWN}_{75\%,\text{bad}}$.

These results suggest that the instantaneous effect of the reputation system varies by employer. Employers with bad reputations are relatively unaffected by the downtime, consistent with these employers attracting only workers who do not use the reputation system. Less-visible employers with good reputations are the most adversely affected. Results are consistent with these employers no longer being discovered by workers who use Turkopticon as a screen. Most-visible employers with good reputations are positively affected, as though workers using Turkopticon stop screening for less-visible, good-reputation employers, and instead use the best-known, good-reputation employers as a

fallback option. In other words, these results suggest that Turkopticon aids in workers' discovery of small, high-road employers and provides these employers an incentive to invest in their reputation. To extrapolate, we might expect that the reputation system promotes competition and prevents the market from devolving into a small, oligopsonistic set of well-known employers, since newer and smaller employers require substantial reputational investments to become sufficiently well-known to attract new workers reliably.

In contrast to the ideal experiment, this study has some caveats. While we would like to study whether the market would reach a new equilibrium (e.g., it would collapse or become an oligopsony) in the absence of a reputation system, we only observe relatively-short, expected-to-be-temporary downtimes that surely don't allow employers to adapt their payment strategies endogenously or workers to adjust their labor market response to such changes. We also cannot observe whether workers are actually paid less for their work. We only observe promised payments and the number of tasks performed. Nonetheless, the results provide relatively-clean evidence for the instantaneous effects of the reputation system on how workers search for jobs and how public and private reputation substitute for one another.

8. Discussion

8.1. Types of reputation and market design

Our experimental design focuses on the value of a publicly available employer reputation system, and not on workers' private signals. To test Hypothesis 1, we allowed workers to complete only one task. To test Hypothesis 2, we submitted only one task for each employer. To test Hypothesis 3, we examined how completed work varied by public ratings and total visibility. Even so, private information remains important in this context, and other tools (e.g., Turk Alerts and DCI New HIT Monitor) are available to notify workers when a employer that they privately flag posts a job. Likewise, employers can privately invite workers to apply for future work.

The coexistence of these more traditional matches, which rely on private information and repeated contracting, is arguably a reminder of the shortcomings of current rating systems . Indeed, the chief value proposition of such crowdsourcing platforms is to provide a quick and efficient method of connecting employers and workers to large numbers of trading partners. The inability to do so is a key concern for both employers and workers.

Amazon might do more to encourage workers and employers to use public ratings. For instance, Amazon could give workers access to historical information on each employer, such as average past wage and rejection rates. It could also create a native, subjective rating

system, as Upwork has done and as Amazon has done for consumer products. The lack of information about employer reputations, coupled with the lack of contract enforcement, may be limiting the market to the small size that a reputation can discipline and to small tasks that are relatively short and well-defined. Relatively few workers would risk investing a week on a task when the criteria for acceptance are poorly defined and payment is nonenforceable (Ipeirotis 2010a). Since our experiment, Amazon has since begun requiring unique tax identification numbers, making it more difficult for employers to reset a bad reputation.

8.2. Why do workers contribute to the public reputation system?

The coexistence of public and private reputation systems also begs the question: why do workers contribute to a collective memory when they can instead hoard private knowledge of the best employers? Prior studies have also noted the collective action problem that this entails (see, for example, Gao et al. 2015, Levine and Prietula 2013). Again, this literature almost exclusively focuses on ratings of sellers and service providers by buyers and clients.

Nonetheless, online labor markets and ratings of employers are also a unique and instructive setting for understanding contributions to public ratings. In a typical product market, goods are nonrival: when a buyer favorably rates a product or seller on Amazon, that buyer's ability to get future products or services is not hindered by an increase in other buyers. However, employers post finite numbers of tasks, and favorable reviews for smaller employers could lead other workers to consume those tasks. In this sense, worker reviews are costly to workers not only in terms of time, but also in that they attract other workers to a rival "good." The ability of Turkopticon to attract large numbers of raters shows that altruism and volunteerism survive even under these conditions. Moreover, our results from our second experiment confirm that these reviews are useful and informative.

8.3. Specific puzzles from our empirical results

Each study features some empirical results that warrant future attention. In experiment 1, we found that good-reputation employers attracted work more quickly with no loss of quality. However, good-reputation employers might also get a reputation for paying, regardless of work quality, leading workers to flock to these employers and then exert minimal effort. Although the analysis of workmanship in experiment 1 finds no evidence for this margin of behavior, it remains an open question whether employer reputation systems can also invite moral hazard.

In experiment 2, why did effective wages for good-reputation employers exceed those for bad-reputation employers? As noted in our review of the literature, studies have generally

(though not always) found the opposite result: good reputations allow trading partners to extract more favorable terms, such as the ability to attract workers at lower pay. Such compensating differentials may be impossible in this setting: although pay-per-task is specified, the time to complete the task is not. Ratings may also capture other aspects of the employer. Following Bartling et al. (2013), some employers may be more altrustic; these employers pay higher wages and also have better reputations. Indeed, Turkopticon ratings include an item for generosity, which intends to capture expected wages. A second alternative is that ratings implicitly capture employers' preferences for getting work done more quickly; impatient employers pay higher wages and maintain good reputations to get work accomplished quickly.

Lastly, in the natural experiment, what would happen to the market if the reputation system remained down? Turkopticon's downtimes suggest that smaller, good-reputation employers are especially dependent on Turkopticon to get work done quickly. Following the logic of our formal model, we may hypothesize that the long-term loss of a reputation system would lead the market to become concentrated among the most visible of the good-reputation employers, while smaller employers are deterred by the cost of establishing a good reputation and may need to go through third-party brokers (such as CrowdFlower) that have established reputations.

8.4. Lessons for reputation systems in offline labor markets

What relevance do these findings have for other markets? M-Turk workers are unconventional in that they are contracted for very small tasks and have minimal interaction with firms. However, the issues that they confront are more general. As Agrawal et al. (2015) describe, "the growth of online markets for contract labor has been fast and steady."

Uber, TaskRabbit, DoorDash, and other online platforms are also blurring the boundaries between offline employment and entrepreneurship (Apte and Mason 1995, Weil 2014, Harris and Krueger 2015). While these platforms are also drawing increasing scrutiny from regulators, wage theft and other forms of opportunism are also pervasive in other settings where legal enforcement is weak, including among independent contractors, undocumented immigrants, misclassified employees, and low-wage employees (Bobo 2011, Rodgers et al. 2014). Wage theft has prompted the U.S. Department of Labor's Wage and Hour Division to award back pay to an average of 262,996 workers a year for the past ten years, and far more cases go unremedied (Bernhardt et al. 2013, Bobo 2011, Lifsher 2014, Bernhardt et al. 2009, U.S. Department of Labor 2016). The value of stolen wages restored to workers through enforcement actions is larger than the total value stolen in all bank, gas station, and convenience store robberies (Lafer 2013).

Krueger (2017) reports that about a third of American workers spent some time in the prior week "working or self-employed as an independent contractor, independent consultant, or freelance worker," including "working on construction jobs, selling goods or services in their businesses, or working through a digital platform, such as Uber, Upwork, or Avon," and 84 percent of these workers report self-employment as their main job. Among these workers, over a third report "having an incident in the last year where someone hired you to do a job or project and you were not paid on time." Over a quarter reported at least one incident of being unable to collect the full amount owed for a job or project that the worker completed. The Freelancers Union has used both reputation and regulatory solutions to address client nonpayment, including their "client scorecard" and a successful effort to lobby New York City to pass the Freelance Isn't Free Act.

As on M-Turk, workers in the broader labor market strive to distinguish which employers will treat them well or ill. Workers have always made decisions with partial information about employer quality, and so these forces have always shaped labor markets. Contracts and bilateral relational contracting are important forces disciplining employer opportunism, but they are certainly incomplete. Workers have always relied on public employer reputations propagated through informal, decentralized, word-of-mouth conversations. Though economists have had theories about how employer reputation would work, the informal system has operated largely outside our view, yielding a very thin empirical literature. As the cost of communications and data storage fell in recent years, employer reputation has become more centralized, systematic and measurable, showing up in general labor market matching sites such as Glassdoor.com and Indeed.com, and in more specialized contexts such as ProjectCallisto.org, which allows workers to share information about sexual harassment and abuse at work, and Contratados.org, which allows migrant workers to review recruiters and employers.

Attention to the worker's information problem also suggests innovative directions for policy and institution building. Can more be done to improve the functioning of the gig economy through helping workers overcome their information problem with respect to employer heterogeneity? Can we improve institutional designs to better elicit, aggregate, and disseminate information about employers? Platform design affects workers' willingness to voluntarily contribute their private information to the public pool (Marinescu et al. 2018). A policy example of this kind of logic in action is that, in 2009, the U.S. Occupational Safety and Health Administration began systematically issuing press releases to notify the public about large violations of workplace safety laws. This effort attempts to influence employer reputation, to improve the flow of information about employer quality, and to

create incentives for providing safer workplaces. Johnson (2016) finds that it also induces competing employers to improve compliance with worker protection laws, though the Trump administration rolled back this effort (Meier and Ivory 2017). Workers have traditionally used labor unions and professional associations as a venue for exchanging information about working conditions and coordinating collective withdrawal of trade in order to discipline employers. The rise of new institutions that facilitate information sharing may be taking up some of this role.

Platforms can better deliver on their promise to reduce matching frictions and increase efficiency to the extent that they help workers distinguish reliable employers. If that goal can be addressed, then the falling costs of information processing and diffusion may move labor markets closer to the competitive ideal.[31] Fulfilling this promise requires designing platforms that help workers to find great employers and avoid bad ones.

9. Conclusion

Online platforms are making it cheaper to connect trading partners, but issues of trust and reliability remain. The empirical literature has focused almost exlusively on sellers, including sellers of products (e.g., eBay brokers), services (e.g., restaurants), and labor (e.g., contract workers on gig platforms). Labor markets have always faced bilateral uncertainty, although the relative absence of regulation has made gig and online labor markets especially prone to opportunistic employers.

This study provides a theoretical and empirical foundation to better understand how employer reputation systems can partially substitute for legal and other third-party contract enforcement. Moreover, the experience of M-Turk and Turkopticon suggests that reputation systems may have an important role to play in providing employers with incentives to treat workers well, giving lesser-known employers direct access to workers, and ultimately expanding the scope of work that can be completed online. Institutions and policies can combat opportunistic employers, but given the complexities of the employment relationship, it seems implausible that opportunism will ever be fully eliminated.

[31] According to Manning (2011), "If one thinks of frictions as being caused by a lack of awareness of where vacancies are... then one might have expected a large effect of the Internet. But if... one thinks of frictions as coming from idiosyncracies in the attractiveness of different jobs... then one would be less surprised that the effects of the Internet seem to be more modest."

References

Agrawal A, Horton J, Lacetera N, Lyons E (2015) Digitization and the contract labor market. *Economic Analysis of the Digital Economy* 219.

Agrawal AK, Lacetera N, Lyons E (2013) Does information help or hinder job applicants from less developed countries in online markets? Technical report, National Bureau of Economic Research.

Apte UM, Mason RO (1995) Global disaggregation of information-intensive services. *Management science* 41(7):1250–1262.

Ba S, Pavlou PA (2002) Evidence of the effect of trust building technology in electronic markets: Price premiums and buyer behavior. *MIS quarterly* 243–268.

Bajari P, Hortacsu A (2003) The winner's curse, reserve prices, and endogenous entry: Empirical insights from ebay auctions. *RAND Journal of Economics* 329–355.

Banerjee AV, Duflo E (1999) Reputation effects and the limits of contracting: A study of the indian software industry .

Barach M, Golden J, Horton J (2017) Skin in the game and platform credibility: Evidence from field experiments. *Working paper* .

Bartling B, Fehr E, Schmidt KM (2013) Use and abuse of authority: a behavioural foundation of the employment relation (vol 11, pg 711, 2013). *Journal of the European Economic Association* 11(5):1230–1230.

Bernhardt A, Milkman R, Theodore N, Heckathorn D, Auer M, DeFilippis J, Gonzalez AL, Narro V, Perelshteyn J, Polson D, et al. (2009) Broken laws, unprotected workers. *National Employment Law Project. New York: NELP* .

Bernhardt A, Spiller MW, Theodore N (2013) Employers gone rogue: Explaining industry variation in violations of workplace laws. *Industrial & Labor Relations Review* 66(4):808–832.

Board S, Meyer-ter Vehn M (2013) Reputation for quality. *Econometrica* 81(6):2381–2462.

Bobo K (2011) *Wage theft in America* (The New Press).

Brown J, Matsa DA (2015) Boarding a sinking ship? an investigation of job applications to distressed firms. *The Journal of Finance* .

Cabral L, Hortacsu A (2010) The dynamics of seller reputation: Evidence from ebay. *The Journal of Industrial Economics* 58(1):54–78.

Chauvin KW, Guthrie JP (1994) Labor market reputation and the value of the firm. *Managerial and Decision Economics* 15(6):543–552.

Cornes R, Sandler T (1994) The comparative static properties of the impure public good model. *Journal of public economics* 54(3):403–421.

Dellarocas C, Wood CA (2008) The sound of silence in online feedback: Estimating trading risks in the presence of reporting bias. *Management Science* 54(3):460–476.

Difallah DE, Catasta M, Demartini G, Ipeirotis PG, Cudré-Mauroux P (2015) The dynamics of micro-task crowdsourcing: The case of amazon mturk. *Proceedings of the 24th International Conference on World Wide Web*, 238–247 (ACM).

Dube A, Jacobs J, Naidu S, Suri S (2018) Monopsony in online labor markets. Technical report, National Bureau of Economic Research.

Farronato A, Fradkin A, Larsen B, Brynjolfsson E (2018) Consumer protection in an online world: When does occupational licensing matter? *Working paper* .

Filippas A, Horton JJ, Golden J (2018) Reputation inflation .

Fradkin A, Grewal E, Holtz D, Pearson M (2015) Bias and reciprocity in online reviews: Evidence from field experiments on airbnb. *Proceedings of the Sixteenth ACM Conference on Economics and Computation*, 641–641 (ACM).

Gao GG, Greenwood BN, Agarwal R, McCullough JS (2015) Vocal minority and silent majority: How do online ratings reflect population perceptions of quality? .

Hannon J, Milkovich G (1995) Does HR reputation affect shareholder value. *Unpublished* .

Harris SD, Krueger AB (2015) A proposal for modernizing labor laws for twenty-first-century work: The independent worker. *the Hamilton project, Discussion paper* 10.

Horton J (2018) Buyer uncertainty about seller capacity: Causes, consequences, and a partial solution. *Forthcoming, Management Science* .

Horton J, Golden J (2015) Reputation inflation: Evidence from an online labor market. *Work. Pap., NYU* .

Horton JJ, Chilton LB (2010) The labor economics of paid crowdsourcing. *Proceedings of the 11th ACM conference on Electronic commerce*, 209–218 (ACM).

Hui X, Saeedi M, Shen Z, Sundaresan N (2016) Reputation and regulations: Evidence from ebay. *Management Science* .

Ipeirotis P (2010a) A plea to amazon: Fix mechanical turk.

Ipeirotis PG (2010b) Analyzing the amazon mechanical turk marketplace. *XRDS: Crossroads, The ACM Magazine for Students* 17(2):16–21.

Irani L (2012) Microworking the crowd. *Limn* 1(2).

Johnson MS (2016) Regulation by shaming: Deterrence effects of publicizing violations of workplace safety and health laws.

Jøsang A, Ismail R, Boyd C (2007) A survey of trust and reputation systems for online service provision. *Decision support systems* 43(2):618–644.

Klein B, Leffler KB (1981) The role of market forces in assuring contractual performance. *The Journal of Political Economy* 615–641.

Krueger A (2017) Independent workers: What role for policy? Moynihan Lecture slides.

Lafer G (2013) The legislative attack on american wages and labor standards, 20112012.

Levine SS, Prietula MJ (2013) Open collaboration for innovation: Principles and performance. *Organization Science* 25(5):1414–1433.

Lifsher M (2014) California cracks down on wage theft by employers. *Los Angeles Times* .

List J, Momeni F (2017) When corporate social responsibility backfires: Theory and evidence from a natural field experiment. *NBER working paper* .

Luca M (2016) Reviews, reputation, and revenue: The case of yelp.com. *HBS NOM Unit Working Paper No. 12-016* .

Manning A (2011) Imperfect competition in the labor market. *Handbook of labor economics*, volume 4, 973–1041 (Elsevier).

Marinescu I, Klein N, Chamberlain A, Smart M (2018) Incentives can reduce bias in online reviews. Working Paper 24372, National Bureau of Economic Research, URL `http://dx.doi.org/10.3386/w24372`.

Mason W, Watts DJ (2010) Financial incentives and the performance of crowds. *ACM SigKDD Explorations Newsletter* 11(2):100–108.

McDevitt RC (2011) Names and reputations: An empirical analysis. *American Economic Journal: Microeconomics* 3(3):193–209.

Meier B, Ivory D (2017) Worker safety rules are among those under fire in trump era. *New York Times* Accessed: 06/10/2017.

Moreno A, Terwiesch C (2014) Doing business with strangers: Reputation in online service marketplaces. *Information Systems Research* 25(4):865–886.

Nagaraj A (2016) Does copyright affect reuse? evidence from the google books digitization project .

Nosko C, Tadelis S (2015) The limits of reputation in platform markets: An empirical analysis and field experiment. Technical report, National Bureau of Economic Research.

Osterman P (2018) In search of the high road: Meaning and evidence. *ILR Review* 0019793917738757.

Oyer P, Schaefer S, Bloom N, Van Reenen J, MacLeod WB, Bertrand M, Black SE, Devereux PJ, Almond D, Currie J, et al. (2011) Handbook of labor economics. *Volume 4b, Chapter Personnel Economics: Hiring and Incentives* 1769–1823.

Pallais A (2014) Inefficient hiring in entry-level labor markets. *The American Economic Review* 104(11):3565–3599.

Resnick P, Zeckhauser R (2002) Trust among strangers in internet transactions: Empirical analysis of ebays reputation system. *The Economics of the Internet and E-commerce* 11(2):23–25.

Rodgers WM, Horowitz S, Wuolo G (2014) The impact of client nonpayment on the income of contingent workers: Evidence from the freelancers union independent worker survey. *Industrial & Labor Relations Review* 67(3 suppl):702–733.

Rosenblat A, Levy KE, Barocas S, Hwang T (2017) Discriminating tastes: Uber's customer ratings as vehicles for workplace discrimination. *Policy & Internet* 9(3):256–279.

Ross J, Zaldivar A, Irani L, Tomlinson B (2009) Who are the turkers? worker demographics in amazon mechanical turk. *Department of Informatics, University of California, Irvine, USA, Tech. Rep* .

Rubin DB (1974) Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology* 66(5):688.

Silberman M, Irani L (2016) Operating an employer reputation system: Lessons from turkopticon, 2008-2015. *Comparative Labor Law & Policy Journal, Forthcoming* .

Silberman M, Ross J, Irani L, Tomlinson B (2010) Sellers' problems in human computation markets. *Proceedings of the acm sigkdd workshop on human computation*, 18–21 (ACM).

Silberman MS (2013) Dynamics and governance of crowd work markets. *International Workshop on Human Computation and Crowd Work*.

Stanton C, Thomas C (2015) Landing the first job: The value of intermediaries in online hiring. *The Review of Economic Studies* rdv042.

Stinchcombe AL, March JG (1965) Social structure and organizations. *Handbook of organizations* 7:142–193.

Turban DB, Cable DM (2003) Firm reputation and applicant pool characteristics. *Journal of Organizational Behavior* 24(6):733–751.

US Department of Labor (2016) Fiscal year statistics (wage and hour division). `https://archive.fo/DGVX1`, [Online].

US Government Accountability Office (2015) Contingent workforce: Size, characteristics, earnings, and benefits. `https://archive.fo/uKbJL`, [Online].

Weil D (2014) *The Fissured Workplace* (Harvard University Press).