Reducing Congestion in Labor Markets: A Case Study in Simple Market Design

John J. Horton
MIT Sloan & NBER

Shoshana Vasserman
Stanford GSB & NBER

Mitchell Watt
Stanford University*

November 2024

Abstract

Many matching markets are suspected to suffer from inefficient levels of congestion. We show this is a real concern in an online labor market and present results of two market-wide experiments designed to reduce congestion. The first intervention introduced a "soft cap" on the number of applications that could be received for a job opening. Despite reducing the number of applications per opening, the intervention did not reduce the hiring probability or reported match quality. A second intervention attempted to price externalities directly, but failed. The application fees introduced by the platform reduced hiring rates and competition among candidates.

*Email: svass@stanford.edu. Thanks to Adam Ozimek, Apostolos Filippas, Dan Walton, Philipp Kircher, Sydnee Caldwell, Mandy Pallais, and Ada Yerkes Horton for helpful comments and suggestions. Thanks also to Michelle Zheng for great research assistance. This paper supersedes "Job-Seekers Send Too Many Applications: Experimental Evidence and a Partial Solution".

1 Introduction

It has long been argued that labor markets suffer from the problem of uninternalized search externalities (Diamond, 1982; Hosios, 1990; Mortensen and Pissarides, 1994): job-seekers do not fully capture the benefits their search has on markets, nor do they bear the screening costs imposed on firms and the costs imposed on competing job applicants due to crowding out. This view is foundational to the matching perspective on the labor market.

In this paper, we (a) document inefficient congestion and (b) report the results of two field experiments designed to ameliorate the problem. We say that an online labor market is **congested** if some employers receive more applications than they are willing or able to screen profitably. Congestion is **inefficient** if overall welfare could be improved by redistributing applications from congested job openings to non-congested ones.

We describe how inefficient congestion can arise even when workers optimize their job searches and employers optimize their hiring searches, due to application and screening costs. This inefficiency stems from a 'missing market" for job applications: employers do not subsidize applications or charge application fees, while job seekers do not compensate firms for screening job applications. As a result, job seekers fail to internalize their effect on the hiring probability of competing applicants or the costs incurred by firms, leading to an excess of applications at high-value jobs and too few at others. Reallocating applications between congested and uncongested job openings can improve social welfare.

To test for inefficient congestion, we study two interventions in an online labor market platform that influenced the number of job applications per job opening.[1] We document evidence of the application and screening costs that drive congestion in the online labor market we study. Around 92% of applicants submitted at most one application on their median day on the platform, while employers receiving at least 30 applications opened fewer than half of the applications received. Applications were unevenly distributed, with the top 25% of jobs receiving 75% of total applications, while the bottom 25% of jobs received just 0.4%. While this pattern reflects underlying variation in job attractiveness, our experimental evidence suggests inefficiency is also present.

In the first intervention, the *autopause* experiment, a randomly selected set of employers had "soft" caps imposed on (a) the number of applications that could be received per opening and (b) the time window for applications. When a job opening received 50

[1] We use the terms "employer," "worker" and "hire" to be consistent with the labor literature and not as a comment on the nature of the relationships created on the platform.

applicants or when five days had passed, the job opening was automatically closed. We describe this treatment as a *"soft"* cap because the employer could easily reopen a job by clicking a single button in the platform's interface. This intervention aimed to prevent job-seekers from applying to jobs where their application would likely be ignored, while still allowing employers to obtain additional applications if needed.

We first document that the autopause intervention caused experimental variation in the number of applications per job opening. Each job opening in the autopause treatment received an average of four fewer applications (an 11% reduction) compared to the control group. This reduction was largest for jobs that would have otherwise attracted many applicants: the quantile treatment effect at the 95th percentile was a reduction of 20 applicants. Employers requested additional applicants for only about 7% of openings.

Our main evidence for inefficient congestion is that, despite reducing the number of applications per job, the autopause treatment did not lower the hiring or the (employer-reported) quality of successful hires. About 41% of openings were filled in *both* treatment and control groups, while the average number of hours worked and employer-reported "star rating" of workers hired were slightly higher in the treatment group (although not statistically significant).[2] Employers whose application pools were capped by the autopause treatment thus hired from their existing application pools with the same probability as employers in the control group, with no discernible reduction in match quality. To rule out an alternative explanation for this finding, we show that later applicants are not adversely selected and irrelevant to the employer: in the control group, job openings receiving at least 50 applications hired an applicant who applied 55th or later 18% of the time.

We do not claim application count is irrelevant: limiting openings to too few applications would reduce hiring probability and match quality. Rather, the autopause experiment suggests that for many employers, the marginal benefit of additional applications decreases with application volume, eventually becoming less than the *de minimis* cost of reopening the job with a single click. Allowing congested job openings to remain open can harm efficiency because applying to jobs is costly for job seekers, and their efforts may be better directed toward less congested job openings.

From the job seeker's perspective, we show that the autopause intervention likely saved significant time—more so than the changes in applicant counts per job opening would suggest. To understand these out-sized effects on job seekers, note that while only about

[2] This fill rate is close to the thirty-day fill rate for jobs in the US reported by the job search engine Indeed in 2015, *see* `https://web.archive.org/web/20170505160128/http://press.indeed.com/wp-content/uploads/2015/01/Time-to-fill-jobs-in-the-US.pdf`.

10% of job openings hit the fifty-applicant cap, these openings were disproportionately important to job-seekers, attracting 43% of applications. That is, jobs with more applications are more likely to receive any given application–a version of the friendship paradox (Feld, 1991). Despite applying to fewer jobs overall, workers applying to treated job openings had a 17% increase in their probability of being hired.

In the second study, the *tokens* intervention, we study a decision by the platform to introduce application fees for certain jobs using a proprietary token currency. Job listings were assigned one of three fee levels (equivalent to about $0.30, $0.60, or $0.90), as a deterministic function of their characteristics. We exploit the staggered timing of the rollout of these fees to different job categories to estimate the effects on employers and workers using the staggered differences-in-differences estimator of Sun and Abraham (2021).

Unlike the autopause experiment, the tokens intervention reduced both the number of applications submitted and the probability of a hire (by 1.4%). Applications to treated openings decreased by 29%, with 19% fewer applications from repeat workers and 5% fewer still for each $0.30 increase in the fee. Average match quality, based on employer feedback, slightly increased. The effect on workers was heterogeneous in worker experience: first-time workers on the platform were less likely to be hired, while repeat workers were more likely, and the wage bids of hired applicants increased, though not significantly. This suggests the tokens intervention reduced competition, benefiting experienced workers and improving match quality but lowering the overall number of successful matches.

Neither the *autopause* nor *tokens* intervention directly addresses the underlying "missing market" problem, which would require more flexible transfers between workers and employers (e.g., application fees, subsidies, or reimbursement for screening costs). Consequently, both interventions are subject to the theory of the second-best (Lipsey and Lancaster, 1956), and may have varying success in different labor markets. We attribute the relative success of the autopause treatment to the simplicity of implementing the "soft" cap: whereas implementing efficient prices would require the platform to access detailed information about costs and benefits,[3] the soft cap is relatively simple to implement and employers could easily opt in to receive additional applications.

The rest of the paper is organized as follows. In Section 3, we describe the empirical context for the autopause and tokens interventions. In Section 4, we offer a theoretical framework for inefficient congestion in labor markets and demonstrate that the necessary

[3] As discussed in Hayek (1945) and formalized in Nisan and Segal (2006), competitive markets typically play this information aggregation role.

preconditions for inefficient congestion in our theoretical framework are present in our empirical setting. In Section 5, we discuss the results of the autopause intervention; in Section 6, we discuss the results of the token intervention. Finally, in Section 7, we further discuss our findings in the context of the literature and conclude.

2 Theoretical Framework

In this section, we study a theoretical model of an online labor market in which inefficient congestion can arise in the decentralized labor market equilibrium.

In our model, each worker chooses a job to apply to based on a probabilistic assessment of its likelihood of being hired and its payoffs conditional on hiring, and then—after observing its match value and labor costs for the job—applies to the job by submitting a wage offer. After receiving all job applications, each employer selects a subset of agents to screen, learning each screened agent's match value, and commits to hiring the applicant that creates the highest match surplus given the applicant's wage offer.

Congestion arises in the equilibrium of this model due to heterogeneity in job attractiveness and the presence of screening costs for employers and application costs for job applicants. Some employers may receive more applications in equilibrium than they are willing to consider, and these excess applications impose an (unpriced) externality on other applicants to the job. We then provide a price-theoretic interpretation of congestion in our model and discuss how the same drivers of congestion may arise in other richer labor market models.

Agents and Payoffs The labor platform consists of a population of N candidate workers (hereafter **workers**) and J job openings (hereafter **employers**). If worker n is hired by employer j, it creates **total value** v_{nj} for the employer and incurs a **labor cost** c_{nj}. We refer to the difference $m_{nj} := v_{nj} - c_{nj}$ as the **match surplus**.

Applying to jobs and screening job applicants are costly. For simplicity, we model application costs by assuming that workers can apply to only one job at a time. This corresponds to the empirical observation that most candidates on our platform apply to at most one job per day, suggesting that the main cost of applying is the shadow cost of time.[4] Each employer faces screening costs associated with assessing applications, assumed to be

[4] As with other simplifications in our stylized model, this restriction can easily be extended to more complex budget constraints and a more complex model of how a worker decides between (or accepts) multiple job offers.

convex in the number of applicants screened. If employer j screens M_j applications, it incurs a screening cost $k_j(M_j)$ for some convex and increasing function $k_j : \mathbb{N} \to \mathbb{R}$.

Each worker seeks to maximize his expected payoff from being hired, which is the difference between his wage and the labor cost incurred. Each employer aims to maximize her expected payoff from hiring, which is the value created by her match with the worker minus the wage she pays and any screening costs.

Information Structure and Timing Neither the worker nor the employer observe values or costs *ex ante*. Instead, we suppose that the vector (c_{nj}, v_{nj}) is a random variable drawn independently for each worker from an employer-specific joint distribution, which is known to the workers and the employers. For simplicity, we assume that the match surplus $m_{nj} = v_{nj} - c_{nj}$ is i.i.d. for each worker conditional on the employer, according to distribution F_j, which is common knowledge.

If worker n decides to apply to employer j, it observes (c_{nj}, v_{nj}) and submits an application containing its wage bid w_{nj} to the employer. We refer to $v_{nj} - w_{nj}$ as the **reported match surplus**.

The employer can decide to screen only a subset of the applications received. If an employer receives N_j job applications and chooses to screen $M_j \leq N_j$ of them, it selects a (uniformly) random subset of M_j applications from the N_j received and observes v_{nj} and w_{nj} for each worker in that set, incurring screening costs $k_j(M_j)$. Only screened applicants can be hired, and the employer may choose not to screen any applications.

We suppose that each employer commits to hiring the worker with the highest reported match surplus and paying the wage bid associated with that worker. Under that assumption, the employer effectively runs a first-price scoring auction among the screened workers with scores equal to the applicants' reports of match surplus.

To summarize the model's timing, there are four main phases:

1. **Auction entry phase**: Workers and employers observe match surplus distributions $\{F_j\}_{j \in J}$. Each worker n selects an employer j to apply to and observes the vector (c_{nj}, v_{nj}) for the selected job opening.
2. **Bidding phase**: Each worker n submits an application to the job opening, consisting of a wage offer w_{nj}.
3. **Screening phase**: Each employer j observes the number of bids it received N_j and determines a number of applications to screen $M_j \leq N_j$. Each employer j chooses a

random subset of applicants of size M_j to screen, observing v_{nj} and the wage bid w_{nj} for each screened applicant.

4. **Auction run and payoffs realized**: Each employer j hires the worker with the highest reported match surplus $v_{nj} - w_{nj}$ among the screened workers. Employer j hiring worker n realizes a payoff equal to the value of the match minus wages paid and the costs of screening, $v_{nj} - w_{nj} - k_j(M_j)$, while worker n realizes payoff equal to the wage minus its labor cost, $w_{nj} - c_{nj}$. Unmatched workers and employers earn zero payoffs.

Equilibrium A strategy for each employer consists of a **screening rule** $M_j(N_j)$, which is a choice of the number of applications to screen M_j as a function of the number of applications received N_j. A strategy for each worker consists of the worker's **job choice** for its application, based on $\{F_j\}_{j \in J}$, and a **wage offer rule** chosen as a function of its observed labor cost and match value (c_{nj}, v_{nj}). Because the match value v_{nj} is eventually observed by the employer for each screened worker, it suffices to consider a wage offer rule as a function of the match surplus, $w_j(m_{nj})$.

We focus on the following **decentralized labor market equilibrium** concept. In the *screening phase*, we suppose that the employer screens a number of applicants $M_j^*(N_j)$ that maximizes its expected returns from the auction minus the screening costs it incurs. If $M_j^*(N_j) < N_j$, we say that job opening j is **congested**, else it is **uncongested**. In the *bidding phase*, we suppose that each worker takes as given the auction entry choices of other workers and the screening rule chosen by the employer and chooses bids according to the symmetric equilibrium of the first-price auction among the number of applicants that the employer will screen. In the *auction entry phase*, we focus on the pure-strategy equilibria in which each worker applies to a job with an accurate forecast of the number of other workers applying to each job.[5]

The decentralized labor market equilibrium can be calculated by backward induction. Under the assumption that workers choose optimal wage bids in the first-price scoring auction, the revenue equivalence theorem implies that the expected profit to the employer

[5] The decision to focus on the pure-strategy equilibria (as in the auction entry model of Levin and Smith (1994)) shuts down an alternative source of inefficiency due to miscoordination among applicants to job opening, leading to more or fewer job applications per job opening than in the pure-strategy equilibria. We do not expect this kind of miscoordination to be a significant driver of inefficiency in our application because job applicants on the online platform we studied can see the number of applications received when deciding whether to apply.

is

$$\Pi_j(M_j) := \mathbb{E}\left[m_j^{(2;M_j)}\right] - k_j\left(M_j\right),$$

where $m_j^{(2;M_j)}$ is the second-highest draw of M_j draws from the distribution of match surpluses F_j. If F_j has monotone hazard rate,[6] Watt (2022) shows that the employer's expected profit function $\Pi_j(M_j)$ is concave in M_j and has a finite optimizer $\overline{M_j}$, so that the optimal screening rule is a threshold rule, $M_j(N_j) = \min\{N_j, \overline{M_j}\}$. To guarantee equilibrium existence and this threshold property, we make the following assumption in the rest of this section:[7]

Assumption 1. *Each match surplus distribution F_j has monotone hazard rate.*

A worker's expected profits $\Phi_j(N_j)$ from applying to employer j that receives N_j total applications is the probability of being screened multiplied by the *ex ante* expected "winner's rent" from a first-price auction with $M_j^*(N_j)$ participants, which is

$$\Phi_j(N_j) = \frac{M_j^*(N_j)}{N_j} \times \frac{1}{M_j^*(N_j)}\mathbb{E}\left[m_j^{(1;M_j^*(N_j))} - m_j^{(2;M_j^*(N_j))}\right],$$

where $m_j^{(1;M_j^*(N_j))}$ is the highest of $M_j^*(N_j)$ draws from the match surplus distribution F_j. Because each worker is *ex ante* symmetric, we can characterize the equilibrium in terms of the *number* of applicants to each job opening without specifying which worker applies to each job. We summarize the equilibrium of the model in the following proposition.

Proposition 1. *Under Assumption 1, a decentralized labor market equilibrium exists, and in that equilibrium:*

(a) each employer chooses a screening rule $M_j^*(N_j) = \min\left\{N_j, \overline{M_j}\right\}$, where the screening threshold $\overline{M_j}$ *satisfies*

$$\overline{M_j} = \arg\max_M \left\{\Pi_j(M)\right\},$$

[6] Recall that a continuous random variable X with distribution F and density function f has monotone hazard rate if $\frac{f(x)}{1-F(x)}$ is non-decreasing. Examples of distributions with monotone hazard rate are uniform distributions on a convex set, normal, exponential, gamma, and beta distributions, and, more generally, any log-concave distribution. Monotone hazard rate implies Myersonian regularity.

[7] If F_j has bounded support (but not necessarily monotone hazard rate), it is still optimal to screen a finite number of applications, but $\Pi_j(M_j)$ may not be concave, and so the optimal screening rule cannot be written in the same form.

(b) the number of applicants at each job opening $\{N_j\}_{j \in J}$*, satisfies*

$$\Phi_j(N_j) \geq \Phi_{j'}(N_{j'} + 1) \text{ for each } j, j' \in J, \text{ and}$$

(c) each worker chooses a wage bid according to the symmetric first-price auction bidding rule

$$w_j(m_{nj}) = c_j + \mathbb{E}\left[m_j^{(1;M_j^*(N_j))} - m_j^{(2;M_j^*(N_j))} \middle| m_{nj} = m_j^{(1;M_j^*(N_j))} \right].$$

Inefficiency We evaluate **efficiency** with respect to the *ex ante* expected **social surplus**, which is the sum of the *ex ante* expected payoffs of workers and employers.[8] The social surplus when employer j screens M_j applications as

$$S_j(M_j) := \mathbb{E}\left[m_j^{(1;M_j)} \right] - k_j(M_j)$$

and the efficient application screening rule

$$M_j^{**}(N_j) = \max_{M_j \leq N_j} S_j(M_j).$$

An allocation $\{N_j\}_{j \in J}$ of applications to jobs is efficient if and only if for each pair of jobs j and j',

$$S_j(M_j^{**}(N_j)) - S_j(M_j^{**}(N_j - 1)) \geq S_{j'}(M_{j'}^{**}(N_{j'} + 1)) - S_{j'}(M_{j'}(N_{j'})).$$

This is the appropriate (discretized) analog of the Hosios (1990) condition in our model.

There are two possible sources of inefficiency in the equilibrium of this model.

The first is the possibility of **insufficient screening** of applicants by employers: the number of applications screened by the employer may differ from the socially optimal level of screening because the employer does not internalize the benefit to social surplus of a marginal application beyond the direct benefit to the employer. Concretely, the employer cares about increases in the *second-highest* draw from the match surplus distribution as its sample size grows, while the social planner cares about increases in the *highest* draw (both care about screening costs). Under Assumption 1, Watt (2022) shows that the increase in the second-highest draw is always *less* than the increase in the highest draw, so that the employer can only choose to screen *too few* applicants (rather than inefficiently too many).

[8] By *ex ante*, we mean before the vectors (c_{nj}, v_{nj}) are drawn.

The second is the possibility of **misallocation** of job applications by workers: workers may choose to submit applications that, while privately optimal, could be reallocated to an alternative job opening to increase social surplus. Three observations clarify how misallocation – a part of **inefficient congestion** – may arise in the equilibrium of the model described in Proposition 1.

First, in equilibrium, any application to an uncongested job opening is efficient. This follows because the marginal benefit of such an application is the same for the employer and the worker, as a result of the well-known property of order statistics that the expected increase in the highest draw from F_j (the social surplus) associated with an additional screened application equals the marginal applicant's expected winner's rent.[9] Then, under Assumption 1, if the employer values an additional application, the social planner does too.[10] Finally, because both the employer and the social planner assess the costs of screening equally, each application submitted to an uncongested job opening must be efficient.

Second, the expected private benefit of an additional application to a job opening is always positive, while the expected social benefit of the marginal application to a congested job is zero. This follows because any candidate stands a chance of being the most qualified for the job, whereas the social benefit of an application to a congested job is zero because the employer does not screen any additional candidates (and therefore does not increase her expected payoff from hiring).

Third, the expected private benefit of an additional application to a congested job can be larger than the expected private benefit of an additional application to an uncongested job. This implies that there may be applications to congested jobs that, while privately optimal, could be reallocated towards uncongested job openings to increase social surplus.

The following proposition characterizes the circumstances under which the equilibrium is inefficient.

[9] That is, $\mathbb{E}[m^{(1;M_j)} - m^{(1;M_j-1)}] = \frac{1}{M_j}\mathbb{E}[m^{(1;M_j)} - m^{(2;M_j)}]$. See, for example, p. 283 in Krishna (2009).

[10] Because, via the previously-cited result of Watt (2022), the increase in the first-order statistic is at least as large as the increase in the second-order statistic, under Assumption 1.

Proposition 2. *Under Assumption 1, the decentralized labor market equilibrium is inefficient if either*

(a) there is insufficient screening *at a congested job, so that* $M_j^*(N_j) = \overline{M_j} < N_j$ *and*

$$\overline{M_j} < \arg\max_{M_j} \left\{ \mathbb{E}\left[m_j^{(1;M_j)} \right] - k_j(M_j) \right\}, \text{ or}$$

(b) there is misallocation of applications, *so that there is a congested job j opening with an unscreened application, while an uncongested job opening j' has fewer applications than its screening threshold, that is,* $N_j > \overline{M_j}$ *and* $N_{j'} < \overline{M_{j'}}$.

In Appendix B, we illustrate the possibility of inefficient congestion in a parametric example.

Generalizations and Price Theoretic Interpretation Three main features generate inefficient congestion in our model:

1. *Screening costs on employers*: Employers face costs associated with assessing an application. This leads to a tradeoff for employers between the benefits of assessing an application—the expected incremental improvement in the payoff associated with the best-identified applicant—versus the cost of screening an additional application.
2. *Constraint on number of applications per applicant*: Applicants are limited in the number of job openings to which they may apply. This limit on the number of applications may arise due to time costs associated with preparing an application or opportunity costs. This budget implies that workers do not apply to all jobs for which they may be suited but instead focus their applications on jobs that lead to the highest *ex ante* expected payoff.
3. *Heterogeneity between jobs*: Jobs differ in their attractiveness to job applicants and/or their screening costs. This leads applicants to focus their limited application budget towards those job openings with higher expected benefits.

These three features are minimal conditions for the misallocation of applications in the auction model. Without screening costs on employers, additional applications continue to benefit employers, which offsets the negative externality imposed on other applicants. In the auction model, these effects exactly equate because (in the absence of screening costs) the increase in the social surplus associated with an additional application equals the marginal applicant's winner's rent in expectation. Without a constraint limiting the number

of applications submitted by applicants, applicants optimally apply for any positions for which expected benefits exceed costs, and although congestion may arise, this would be the same strategy proposed by a social planner. Similarly, without heterogeneity between jobs, each job opening will receive approximately the same number of applications, and misallocation of applications cannot occur.

Fundamentally, inefficiency arises in our model due to a "missing market" problem.[11] Both employers and employees face a cost (or shadow cost) associated with applying or screening an application that is not priced in the market. One way to observe this is to note that the marginal applicant to a congested job may be willing to pay an employer to screen an additional application: that is, the expected benefit to the applicants to that job of the employer screening one additional application may be large enough in aggregate to compensate the employer for the additional cost of screening, and yet, this trade cannot occur. Similarly, an employer with very few applicants to their job opening might be willing to compensate potential applicants for the (shadow) cost of applying to their job opening, yet this trade does not occur. The absence of a price implies a failure of the market to equate the marginal benefits of an application with the marginal costs both *within* and *between* jobs.[12] Within a job, this failure of the equilibrium condition implies that the market does not attain equilibrium: either more applications are received than can be considered and (inefficient) rationing occurs, or fewer applications are received or screened than socially optimal. The failure of the equilibrium condition between jobs leads to the misallocation of applications.

Without addressing the missing market for applications, alternative modeling assumptions are unlikely to change the general takeaways from our auction model. For example, the above analysis is almost entirely unchanged in the presence of reserves set by employers in the first-price screening auction.[13] This is significant because, in practice, we observe many job openings that receive job applications but do not result in hires. Alternative informational assumptions, like interim entry decisions (applicants choosing which job to apply for after observing a signal of their match values for all jobs) and residual uncertainty about match values, would change agents' expected payoff calculation but not address the missing market problem.

---

[11] The "missing market" interpretation of market failure is classically attributed to Arrow (1969), although he did not use this terminology.

[12] Except in the rare case that the zero vector can be obtained as a competitive equilibrium price in the "missing" market.

[13] The same approach as above applies for reserve prices that are not very high: Watt (2022) shows that the employer's objective is concave when $F(r) \leq 1 - 2/n$, where n is the number of bidders.

Platform Interventions One natural response to address a market failure in the form of a missing market is to introduce that market. In the online labor platform case, this would require a mechanism to facilitate job opening-specific (even applicant-specific) transfers between potential hirers and applicants to compensate for the relevant application and screening costs in order to equilibrize the supply and demand of applications for a given job. However, this mechanism seems challenging to implement practically: it may be difficult for employers to identify a relevant pool of potential applicants, and compensation of either side of the market may induce unwanted behavior on the platform (e.g., frivolous openings and applications).

In this paper, we consider two alternative interventions that aim to reduce congestion on an online platform.

The first intervention we consider is the introduction of a "soft cap" or *autopause* on the number of applications received by each job. For each job opening, after fifty applications are received or five days pass, the job opening is automatically closed unless the employer explicitly asks to receive more applicants. This intervention aims to reduce the probability that applications will be received after the employer's screening threshold $\overline{M_j}$ is met. The choice of threshold is based on the hypothesis that the marginal return to the employer of additional applications (after five days or fifty have already been received) is likely to be low, and so making these congested jobs unavailable on the platform may focus applicants' search towards jobs where their applications have a higher contribution to social surplus. But to avoid harms to employers who value receiving more than fifty applications (for whom $\overline{M_j} > 50$) or more applications after five days, the employer can reverse the autopause.

The second intervention is a *partial* price in the market in the form of modest application cost in *tokens* for applicants. Employers do not collect these charges, so they do not function as a price signal on the demand side of the market to incentivize additional screening. Instead, the goal of these application costs is to induce applicants to partially internalize the costs of congestion in the market. One challenge associated with this intervention is identifying the right application cost: a "real" market for applications would perform this role in lieu of a market designer,[14] but this market is not practically implementable for the reasons discussed above.

Because neither intervention directly addresses the underlying missing market problem, they are both subject to a "theory of the second best" (Lipsey and Lancaster, 1956); that is,

[14] See Hayek (1945) for the classical discussion of a market's role in aggregating such information, and Nisan and Segal (2006) for a related mathematical formalization.

the welfare implications of these new market distortions depend deeply on underlying (measurable and immeasurable) market parameters and require empirical or experimental analysis before adoption. This is the goal of the experimental analysis in the remainder of this paper.

3 Empirical context

Our setting is a large online labor market. In this market, employers post job openings that workers can typically apply to without restriction. The kinds of work offered include tasks that can be done remotely, such as programming, graphic design, data entry, translation, writing, and so on. Jobs can differ substantially in scope, with some formed matches lasting for years and others lasting only a day or two. See Horton, Kerr and Stanton (2017) for roughly contemporaneous details on the distribution of kinds of work, contract structure, and patterns of trade in an online labor market.

Employers can solicit applications by recruiting workers, or workers can apply to openings they find. Most applications on the platform come from workers finding job openings through various search tools and then applying. Applying workers submit a wage bid (for hourly contracts) or a fixed amount (for fixed-price jobs). When applying, the worker can observe the number of applications already submitted. Employers then screen applicants and potentially make one or more hires—although hiring a single worker is by far the most common choice, conditional upon hiring anyone.

Applicants arrive very quickly. One reason for this speed is that workers do not know exactly when the employer will start making a decision. Fast applications also seem to be the case in conventional markets when application behavior is observed (see van Ours and Ridder (1992)).

A burgeoning literature uses online labor markets as a domain for research. Pallais (2013) shows via a field experiment that past on-platform worker experience is an excellent predictor of being hired for future job openings. Stanton and Thomas (2016) shows that agencies (which act as quasi-firms) help workers find jobs and break into the marketplace. Agrawal, Lacetera and Lyons (2013) investigate what factors matter to firms in making selections from an applicant pool and present some evidence of statistical discrimination, which can be ameliorated by better information. Horton (2017) explores the effects of making algorithmic recommendations to would-be employers. Barach and Horton (2020) reports the results of an experiment in which employers lost access to wage history when making hiring decisions.

Although our setting offers a rich, detailed look at hiring, there are limitations. A downside of our context is that it is one marketplace. However, when applications are observable in conventional markets, the success probability also appears to be low and similar to what we observe (Marinescu and Skandalis, 2021). Although our context is unique, the fundamental economic problem—workers not internalizing the externalities of search intensity—is commonplace, and there is emerging evidence that the precise context matters less than we might imagine for generalization (DellaVigna and Pope, 2019). Furthermore, job search on online job boards is quite similar to our setting, even if the resulting jobs are different (Marinescu and Wolthoff, 2020).

We conduct two field experiments on our online labor market platform, and the remainder of this paper focuses on the results from these experiments.

In the first experiment, *autopause*, job openings were randomized at the employer level into *treatment* and *control* cells. Once a treated job opening had 50 applicants or 120 hours (5 days) had elapsed since posting, the job was made "private" and no further applications were accepted. The employer was notified of this change through the platform interface and via email and could, at any time, revert the change from public to private by pushing a single button.

In the second experiment *tokens*, the platform introduced financial application costs for workers. Job openings were assigned an application cost over the course of several months. Openings were targeted for treatment by job category and application costs were introduced across categories in a randomized staggered order. Treated openings were assigned application costs by a coarse deterministic function chosen by the platform, with costs intended to increase in an opening's desirability. Once a category was treated, all new openings in that category were assigned an application cost in this way.

The next section describes how our experimental setting fits an auction theoretic framework with inefficient congestion. The following sections present and provide a discussion of our results from both experiments.

4 Drivers for Congestion

In Section 2, we laid out three conditions under which inefficient congestion may arise in an auction-theoretic model of the online labor market. In this section, we argue first that the auction theoretic framework is appropriate for our setting. We then argue that the conditions for inefficient congestion hold as well.

Most hires on our platform are acquired by a bidding process: an employer posts a listing; prospective employees consider the listing, observe the number of applications currently submitted, and choose whether to submit an offer with their profile and a wage bid. There is heterogeneity across both job listings and applicants. Using the autopause control sample as a baseline, we find that 18% of job listings received no applications at all, while the most popular listing received 595 applications. Among listings with at least one applicant, the average listing in the bottom 90% of listings received 11 applications, whereas the average listing in the top 10% received 82, and the average listing in the top 1% received 206.

Applicants are heterogeneous in the match value that they can offer an employer. Wage bids often vary substantially within the same listing, and it is common for the applicant with the lowest wage not to be hired. In fact, the applicant with the lowest bid was hired in only 7% of the listings in which at least two applications were received and at least one hire was made. Across all job listings leading to a hire, the median wage bid is .358 standard deviations lower than the winning wage bid for the same opening. Figure 7 in Appendix A.2 shows the substantial bid heterogeneity across jobs grouped by the number of applications received.

The heterogeneity in bids for the same listing reflects substantial heterogeneity in match-specific qualifications across candidates. Applicants submit credentials with their applications and employers evaluate wages against the match value of each potential hire. Table 1 presents suggestive evidence of match effects using listings from the control group of the autopause experiment. Without controlling for the identity of individual workers or jobs, we find that workers submitting higher wage bids were significantly more likely to be hired. However, once worker and job fixed effects are included, the relationship is reversed such that a unit increase in the log wage bid corresponds to roughly a 1% decrease in the probability of being hired. That effect is independent of the arrival order of the job applications for a given job opening. Together with the dispersion of bids for the same job listings, these facts suggest that employers evaluate candidates simultaneously and choose their hires based on a balance of price and match quality.

However, heterogeneity in match quality is not the only predictor of wage bids. As we demonstrate in Section 5, applications to job listings that were subject of the "soft cap" yielded higher probabilities of being hired and higher wage bids even after controlling for worker fixed effects or for heterogeneity in the number of applications that each worker submitted in our sample. Together, these two effects suggest that the auction dynamics in the hiring process are meaningfully binding. Applicants—who can see how many other applications have already been submitted before submitting their own bids—respond

Table 1: Worker application wage bidding vis a vis hiring in the *autopause* control group

	Hires (1/0) x 1000		
	(1)	(2)	(3)
Log wage bid	2.018***	-11.019***	-11.018***
	(0.247)	(1.117)	(1.117)
Applicant arrival rank			-0.002
			(0.008)
Intercept	9.492***		
	(0.526)		
N	262,463	262,463	262,463
DV Mean	13	13	13
Worker FE	N	Y	Y
Job Opening FE	N	Y	Y
Worker Cluster SE	Y	Y	Y
R squared	0.00025	0.75011	0.75011

Notes: The table reports regressions of application-level outcomes—namely whether the applicant was hired. In the experiment, employers posting jobs were randomized to a treatment or a control. Employers in the treatment could not receive additional applicants once they received 50 applicants or 5 days had passed since posting. However, the employer could opt out of this cap by clicking a single button. The regressions are weighted by the inverse of the total number of applications sent by the worker. The sample consists of all applications to all job openings assigned to the control group in the *autopause* experiment. Standard errors are clustered at the worker level. Significance indicators: $p \leq 0.05 : *$, $.p \leq 0.01 : **$ and $p \leq .001 : ***$.

strategically to higher competition by offering lower wages. Employers, on the other hand, hire the candidate with the best available value proposition—at least if there is a sufficiently qualified candidate in the applicant pool.

In many cases, no applicant is hired at all. Across all job listings in the autopause sample, only 41% involved any hires. Among listings with at least 50, 75, or 100 applicants, only 52%, 54%, and 55%, respectively, led to a hire. This reflects another type of heterogeneity across jobs: some jobs require more specialized skills than others. Often, these jobs are well compensated and attractive to applicants, but the quality of the employee-job match is so important that competition on wages is not sufficient to generate a hire.

At the same time, both applicants and employers appear to be capacity-constrained. 92% of applications in the autopause sample submitted at most one application on their median day on the platform, and 71% submitted no more than one application on any day in our sample. Figure 1 shows that although the number of applications received varies widely across jobs, the number of applications that are actually considered by the employer is more uniform, and appears to be capped. Employers for jobs receiving anywhere between 30 and 85 applications only viewed (as captured by opening the application tab on the platform) 10-15 applications on average. Even employers receiving over 100 applications for their job viewed only 20 of them on average. On the other hand, employers viewed a higher proportion of applications for jobs with relatively small numbers of applications submitted.

In summary, job applications on our platform have all of the features of inefficient congestion that we laid out in Section 2. Both applicants and employers are limited in the number of jobs and workers, respectively, that they can review at a time. Meanwhile, jobs and workers are highly heterogeneous. The most attractive jobs are so much more appealing than the others that they attract a plurality of applicants. Finding a good match is difficult, and many jobs are ultimately left unfilled. However, this is likely a function of not only quality but also of price. Vertical quality can be compensated with wages, and applications to jobs with fewer other applicants (smaller job auctions) appear to contribute more to finding adequate matches than larger ones. As such, our model suggests that the status quo allocation of applications to the largest job auctions on the platform is too high and, perhaps, to the smallest job auctions, too low relative to the social optimum.

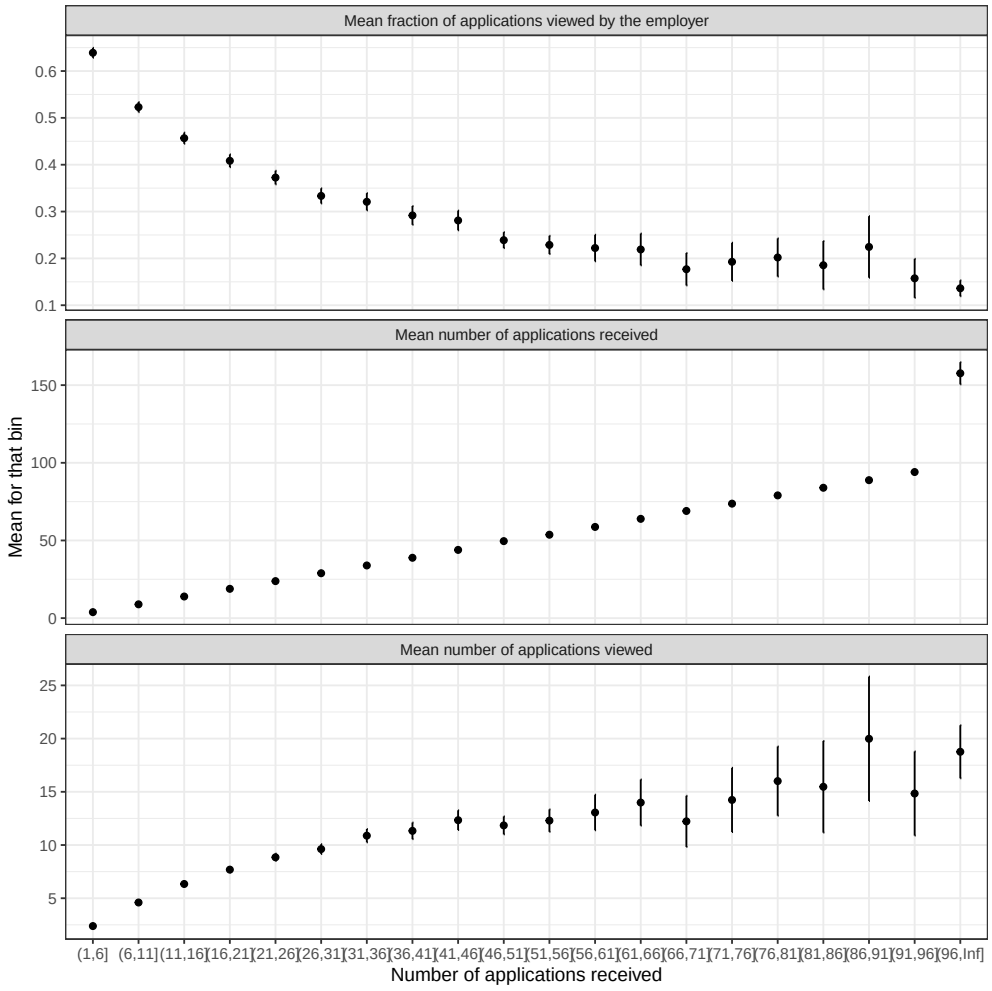


Figure 1: Numbers of applications received and viewed by job opening in *autopause*.

5 Autopause Experiment: Soft Caps as an Indirect Market Intervention

The inefficiencies described in the previous section suggest that there may be scope for welfare improving interventions by the platform. In this section, we describe the impacts of the first experimental intervention under our study: a soft cap on applications.

The experiment ran as follows. A total of 45,742 jobs openings posted by employers were assigned, covering job openings posted between 2013-11-04 and 2014-02-14. Of these, 23,075 job posts were in the treatment and 22,667 were in the control group.[15] Job

[15] While the p-value for the χ^2 test is 0.056, daily counts of allocated jobs show no obvious imbalance and a table of pre-randomization job attributes shows excellent balance, suggesting the low p-value from the χ^2 test is simply due to sampling variation.

openings in the treatment group were given a soft cap: once the opening had received 50 applicants or 5 days had elapsed since posting, the job was made "private" and no further applications were accepted, unless the employer pressed a button requesting to re-open it. Job openings in the control group were untouched regardless of the number of applications or the time elapsed. The experimental sample was itself randomly drawn from all job openings being posted on the platform. We do not report the exact fraction, but it was less than 1% of all job openings posted in the market, which reduces concerns about cross-group interference.[16]

To evaluate the impacts of the intervention, we consider several dimensions of platform activity. First, we examine the efficacy of the intervention: did the soft cap bind, and did it reduce applications to job openings likely to be oversubscribed? Next, we examine the second-order effects of reducing applications: were treated job openings less likely to result in a satisfactory hire? Finally, we consider the equilibrium effects of reducing applications in a competitive labor market: did workers hired for treated jobs receive higher wages?

Figure 2 summarizes the outcomes of the experiment in terms of daily aggregates. As the top panel demonstrates, the number of job openings allocated to the treatment and control groups track closely, suggesting successful randomization. However, as the second and third panels show, the number of applications in the treatment group was substantially lower than in the control group across time. Despite this, as the bottom panel shows, there was not a clear difference in the proportion of openings that made a hire. In the rest of the section, we explore these effects in more depth and discuss their implication for platform design.

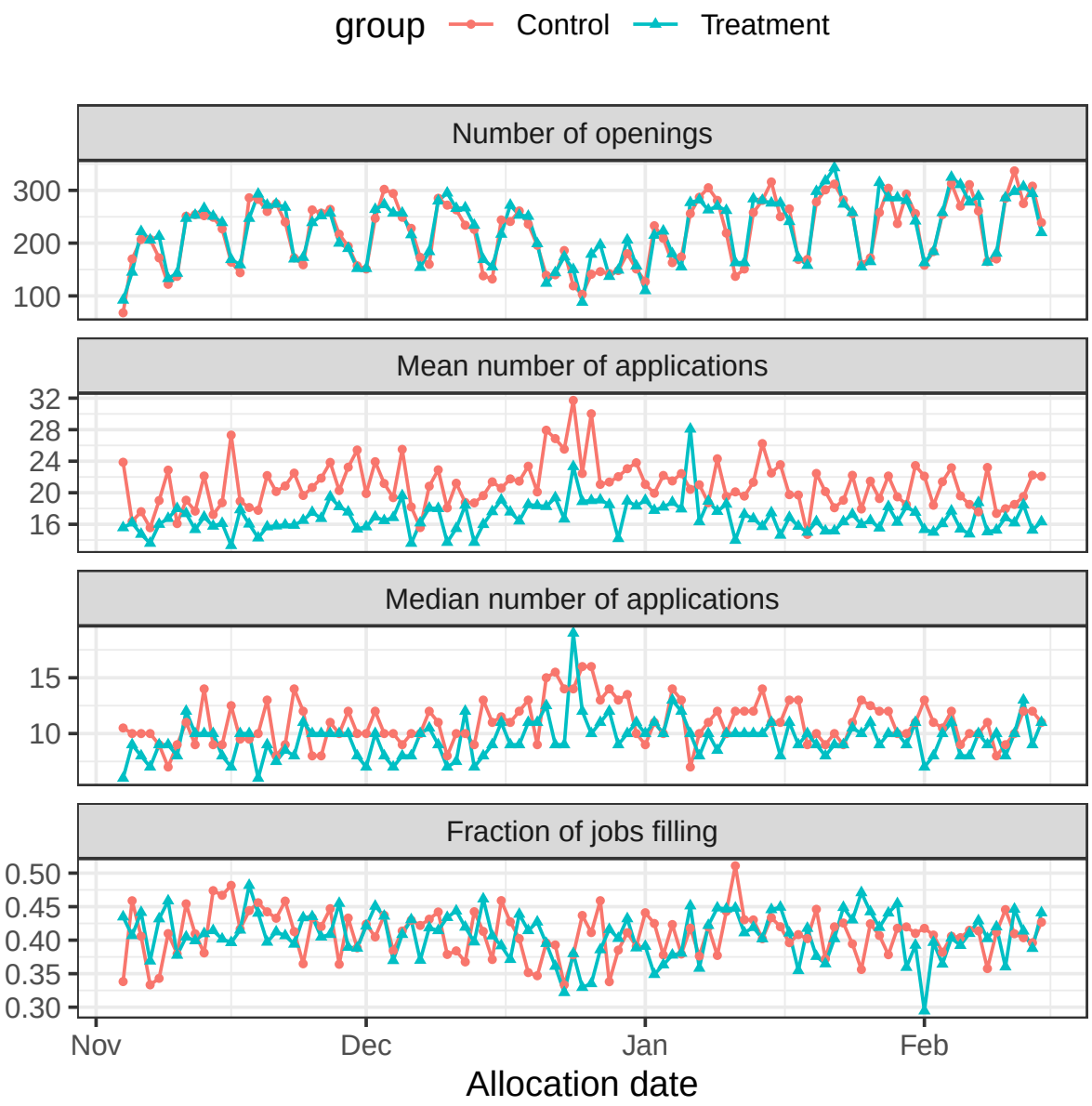
Efficacy In Figure 3, we plot the treatment effect of the soft cap intervention on the number of applications submitted to job openings in our sample.

Figures 3a and 3b plot the kernel density estimates for application counts in the treatment and control samples by application time and number of applicants, respectively. In both cases, the distribution of application counts is nearly the same until near the soft cap threshold (5 days or 50 applicants). After the cutoff, treated jobs experienced a notable decrease in applications, suggesting that the soft cap constraint was binding and was generally not overridden.[17]

[16] After being assigned to a cell, any subsequent job openings by that employer received the same treatment assignment. However, we only use the first job opening in our analysis, as subsequent job openings could have been affected by the experience in the first opening.

[17] Despite the cut-off of 50 applicants, there is actually excess mass at numbers slightly greater than 50—a fact obscured by the density plot. What causes this is that some applicants withdraw their applications,

Figure 2: Group-specific outcomes by allocation date, over time in *autopause*

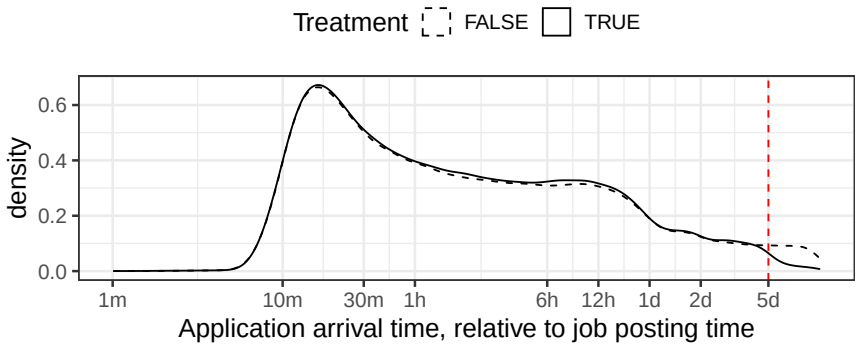


Notes: This plot shows by-day times series for the two experimental groups in the *autopause* experiment. In the experiment, employers posting jobs were randomized to a treatment or a control. Employers in the treatment could not receive additional applicants once they received 50 applicants or 5 days had passed since posting. However, the employer could opt out of this cap by clicking a single button.

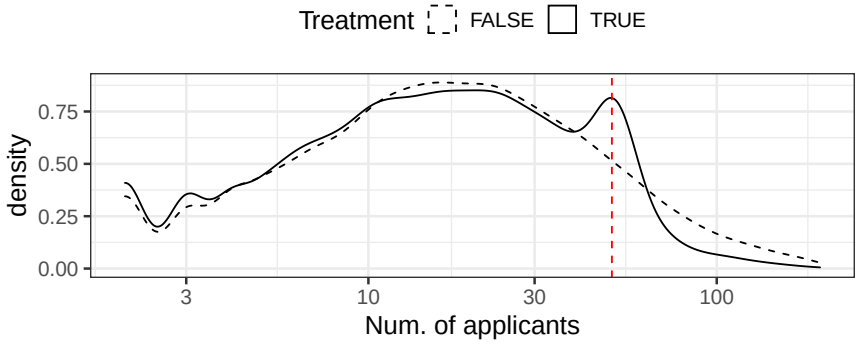
Figure 3c shows precisely where the treatment effects on application counts were concentrated using quantile regressions. The y-axis is log transformed. The x-axis is the associated percentile. Below about the 25th percentile, there is no evidence of an effect. From the 25th

and withdrawn applicants do not count against the cap.

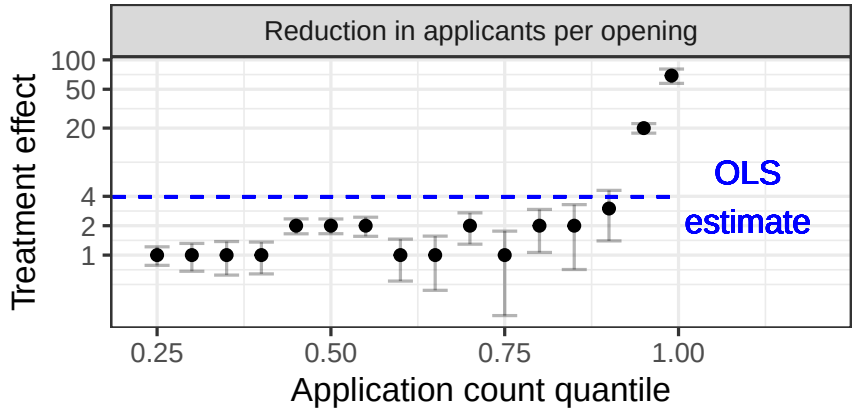
Figure 3: Evidence of the effect of the *autopause* interventions of applicant pools



(a) Distribution of application arrival times, by treatment group, for jobs receiving fewer than 49 applicants



(b) Kernel density estimate of the distribution of applicants per job opening, by treatment and control



(c) Reduction in applicants per job opening by application count quantile

to about the 90th percentile, the reduction is about 1 or 2 applications, but it is much larger above the 90th percentile. For comparison, the OLS estimate of the treatment effect—a reduction of approximately 4 applications—is plotted as a horizontal dashed line.

Externalities Despite reducing the number of applications that were submitted, the soft cap had no discernible effect on the probability that a treated job was filled, but rather shifted the set of applications being considered to primarily those submitted before the cap. Table 2 summarizes the treatment effects on the intensive and extensive margins of hiring. Column (1) shows that the treatment effect on whether any applicant was hired at all is a precise zero, relative to a baseline fill rate of about 41%. Column (2) considers the effect on the *total* number of hires, which accounts for openings with multiple available positions. Here, the treatment effect is also small and not statistically significantly different from zero. However, this is not an indication that late applicants were not considered absent a soft cap. For instance, among openings in the control group that received over 50 applications and hired anyone, an applicant of rank 55 or higher was hired 18% of the time. As Column (3) shows, this reflects the outcome of about 2% of openings in the control group overall. Under the soft cap treatment, this number reduced by 0.7 percentage points (about 33% of the baseline). Combined with the evidence that employers consider a limited number of applications per opening, the substitution towards earlier applicants suggests that without the soft cap, late applicants were crowding out early applicants.

Table 2: ATE on the number of applications and whether the job opening filled in *autopause*

	Any hires?	Total hires	Any hires after 55?
	(1)	(2)	(3)
Treatment	-0.002	-0.016	-0.007***
	(0.005)	(0.010)	(0.001)
Intercept	0.411***	0.528***	0.020***
	(0.003)	(0.007)	(0.001)
N	45,742	45,742	45,742
R squared	0.00000	0.00006	0.00068

Notes: This table reports the effects of treatment assignment on whether a hire wade in Column (1), the total quantity of hires, in Column (2), and whether a hired applicant was greater than the 55th arrival. In the experiment, employers posting jobs were randomized to a treatment or a control. Employers in the treatment could not receive additional applicants once they received 50 applicants or 5 days had passed since posting. However, the employer could opt out of this cap by clicking a single button. Significance indicators: $p \leq 0.05 : *$, $.p \leq 0.01 : **$ and $p \leq .001 : ***$.

In Table 3, we present evidence that the soft cap intervention did not induce a reduction in

Table 3: Match outcomes, conditional upon a hire in *autopause*

	Log hired worker wage	Log hours-worked	Feedback on worker
	(1)	(2)	(3)
Treatment	-0.002	0.042	0.011
	(0.014)	(0.044)	(0.012)
Intercept	2.240***	2.987***	4.667***
	(0.010)	(0.031)	(0.009)
N	9,354	7,082	16,330
R squared	0.00000	0.00013	0.00005

Notes: The sample for these regressions are those job openings where a hire was made. In the experiment, employers posting jobs were randomized to a treatment or a control. Employers in the treatment could not receive additional applicants once they received 50 applicants or 5 days had passed since posting. However, the employer could opt out of this cap by clicking a single button. In Column (1), the sample consists of hourly job openings; in Column (2), job openings where at least 1 hour was billed. In Column (3) the sample is all job openings, including fixed price jobs. Significance indicators: $p \leq 0.05 : *$, $.p \leq 0.01 : **$ and $p \leq .001 : ***$.

match quality for job openings that were filled. Column (1) shows that the average hourly wage (in logs) did not differ between hires made for treated and control openings. Columns (2) and (3) show that worker output in terms of log total hours worked and the average feedback that the employer left for the worker on a 1 to 5 star scale, respectively, increased slightly on average in treated jobs. In both cases, the treatment effect estimates were small and noisy. However, taken together, these indicators suggest that match quality was not adversely affected by the induced decrease in applications received and considered.

Worker Perspective Because randomization for the soft cap intervention was done at the employer level, many applicants were exposed to job openings in both the treatment and control group.[18] In either case, workers did not know the treatment status of the job opening when deciding whether and how to apply, and had the same types of information (*e.g.*, job details, the number of applications already in).

In Table 4, we summarize the difference between outcomes for applications to openings in each group, controlling for worker fixed effects. Each column corresponds to a regression of the form

$$y_{ij} = \beta \cdot \text{TRT}_j + \text{APPCOUNT}_j + \gamma_i + \epsilon \tag{1}$$

[18] The average worker in our sample submitted 5.7 applications in total.

where y_{ij} is the outcome (or choice) for worker i applying to job opening j, TRT_j is the treatment assignment of the applied-to job opening, γ_i is a worker-specific fixed effect and APPCOUNT_j is a fixed effect for the total applicant count for that opening. As workers send different numbers of applications, we weight these regressions by the inverse of the total number of applications sent by the worker.

Table 4: Application outcomes by treatment status of the applied-to job listing

	Hired	Rank	Log wage bid
	(1)	(2)	(3)
Treatment	0.003***	-13.712***	0.012***
	(0.001)	(0.186)	(0.002)
DV Mean	0.02	33.23	1.89
Worker FE	Y	Y	Y
Worker Cluster SE	Y	Y	Y
N	738,861	738,861	466,592
R squared	0.54924	0.80494	0.95144

Notes: This table reports regressions of applicant-level outcomes on the treatment status of the applied-to job opening. The sample consists of all applications sent to assigned job openings. In the experiment, employers posting jobs were randomized to a treatment or a control. Employers in the treatment could not receive additional applicants once they received 50 applicants or 5 days had passed since posting. However, the employer could opt out of this cap by clicking a single button. Regressions are weighted by the inverse total number of applications sent by each worker. Each regression includes a worker-specific fixed effect. Standard errors are clustered at the level of the individual job applicant. Significance indicators: $p \leq 0.05 : *$, $.p \leq 0.01 : **$ and $p \leq .001 : ***$.

As Column (1) of Table 4 shows, workers are 0.3 percentage points more likely to be hired when applying to a treated job opening. Given the baseline hiring probability of about 3%, this difference implies about a 17% increase in hiring probability per application.[19] As the worker fixed effect aims to control for heterogeneity in cross-job worker quality, the increased probability of being hired suggests that holding all else fixed, workers applying to treated jobs were more likely to be considered by the employer. More mechanically, Column (2) shows that applications to treated jobs were more likely to come earlier in the applicant pool. Here, rank 1 corresponds to the first application received; rank 2 corresponds to the second, etc. This results from the fact that later applications (which were hit by the soft cap) are excluded from treatment group openings, since most employers did not lift the cap. Column (3) shows that this lower arrival order (at which point there were fewer competing applications submitted) corresponds to a slightly higher log wage bid,

[19] Note that 2% is nearly identical to the rate of being hired found by Marinescu and Skandalis (2021).

consistent with the predictions of our auction-theoretic framework. Note, however, that this does not indicate that workers earned a higher wage in treated jobs. As we reported above, average log wages for hired workers were statistically unchanged between the treatment and control groups. The average wage bid increasing while the winning bid stays unchanged suggests that marginal applications to congested job openings did not affect the winning bid.

Overall, having a soft cap on the number of applications seems to reduce excess application costs without negatively affecting the probability of hiring or the quality of the hire despite the smaller applicant pool. Although the threshold of 50 was far lower than the number of applications in the most popular jobs, it is possible that a lower threshold might have further reduced the number of excess applications without negatively impacting matches on the platform. Our finding that the number of applications that were viewed was capped at 25 for nearly all jobs, regardless of the number of applicants, suggests that a cap closer to this number might closer to optimal. However, we cannot rule out the possibility that a larger pool of applicants is necessary in order to maintain the same rate of hiring. Although further evidence regarding the optimal soft cap threshold is beyond the scope of our analysis, in the next section, we describe a second experiment, *tokens*, in which we explore the effects of an application cost which affects all applicants across all openings.

6 Tokens Experiment: Application Costs as a Direct Market Intervention

As we discussed in Section 2, the congestion in applications to popular job listings in our data suggests that there may be a "missing market" problem: there is no credible way for candidates (or employers) to signal that they should get special consideration. A natural intervention to mitigate this problem is to introduce additional payments into the market. For instance, if prospective applicants face an additional application fee, then only those who most value applying would continue to do so. If priced efficiently, this may lead to fewer applications, while maintaining or even increasing job fill rates. However, a miscalibrated application cost may backfire. For instance, if costs are too high and match values are correlated across jobs (so that high match-value applicants have many options), it may be that too few qualified applicants will apply and more jobs go unfilled.

In this section, we consider an experimental intervention in which the platform introduced fees for applying to certain jobs. The intervention worked as follows. Over the course of

two years, the platform introduced a financial fee for all applicants applying to a job. The fees were introduced on two separate dates, 5/24/2019 and 6/26/2019, split by category of work (e.g. application fees started on 5/24/2019 for Translation and on 6/26/2019 for Design & Creative). Once a category was affected, all job listings in that category were assigned one of three possible application fee levels—$0.30, $0.60, or $0.90— based on a deterministic function increasing in the job's desirability. Prior to the intervention, there was no effective binding financial cost to applying to any job opening. Following the intervention, all applicants interested in applying to an affected job listing were subject to the fee assigned to that job.

As in the soft cap intervention, we evaluate the application cost experiment through the lens of *efficacy* and *externalities*. Because job openings were treated at the category level, we cannot do a simple comparison of treatment and control as before. However, we can leverage the staggered treatment timing across the different job categories to estimate treatment effects for each treated job based on the pre-treatment outcomes of similar jobs and the outcomes of jobs that were not yet treated. We use the treatment fee levels as a grouping strategy to compare similar jobs, since the levels are assigned based on observed variation in job desirability.

In Tables 5 and 6, we present average treatment effects over the treated period, using the staggered difference-in-differences estimator developed by Sun and Abraham (2021). For robustness, in Appendix A.1, we present dynamic event study plots for the outcomes in Table 5 with estimates following Sun and Abraham (2021), Callaway and Sant'Anna (2021) and Two-Way Fixed Effects.

Table 5: ATE on the number of applications and whether the job opening filled in *tokens*

	Any Apps? (1)	Log # Apps (if any) (2)	Any Hires? (3)
ATT	-0.013***	-0.291***	-0.014***
	(0.001)	(0.020)	(0.003)
R^2	0.00613	0.11132	0.05202
Observations	2,085,072	2,021,138	2,085,072
Category x Price FE	Y	Y	Y

Table 6: Match outcomes, conditional upon a hire in *tokens*

	Log worker wage (1)	Log hours-worked (2)	Feedback on worker (3)	First time worker? (4)
ATT	0.010	-0.147***	0.018***	-0.010***
	(0.015)	(0.028)	(0.006)	(0.003)
R^2	0.19584	0.17728	0.01802	0.00640
Observations	322,687	303,518	594,216	939,893
Category x Price FE	Y	Y	Y	Y

First order effects By contrast with the soft cap intervention, the application fee intervention lowered not only the number of applications submitted to treated job openings, but also the probability of making a hire. Table 5 summarizes the average treatment effects on a set of first order outcomes that are analogous to those examined in Section 5. As Column (1) shows, the probability that a treated job opening received *any* applications at *all* decreased by about 1.3 percentage points relative to its predicted rate without fees. Applications to openings fell on the intensive margin as well. Column (2) reports the effect on the log number of applications per opening, for openings that received at least one application. This suggests that openings that received *some* applications, received 29% fewer of them on average when they were treated with an application fee. Furthermore, the reduction in applications corresponded to a reduction in hiring rates. As Column (3) reports, the probability that a hire was made in a treated job was 1.4 percentage points lower than in an untreated job. While these effects are relatively small — the pre-treatment average hiring rate is 46% — they are significant and may be economically meaningful.

Externalities In order to interpret the impact of the reduction in applications for openings where a hire *was* made, we consider the wages, hours worked and feedback for applicants who were hired. As Column (1) of Table 6 shows, contracted hourly wages increased slightly on average for treated jobs, but there was no statistically significant effect. Nonetheless, Columns (2) and (3) suggest that the employer might have benefited from a better match. As Column (2) reports, applicants who were hired to treated jobs worked 14.7% fewer hours. While in principle this may be a positive or a negative signal — fewer hours worked could either be a sign of competence or of poor value — Column (3) suggests the former is more likely. Feedback on the hired workers improved modestly for treated jobs: a gain of 0.018 relative to a pre-treatment average of 4.82. Together, these estimates suggest that higher quality applicants may have been more likely to be considered and

subsequently hired under the application fee intervention. Column (4) of Table 6 suggests that this may in part be explained by experience. The probability that a hired applicant was new on the platform decreased by 1 percentage point for treated jobs, suggesting that less experienced applicants were more likely to be crowded out by more experienced workers, or to not apply at all.

Worker Perspective Our analysis above suggests that the application fee intervention induced compositional changes in the applicant pools of affected openings. In Table 7, we examine how the intervention changed application conditions and outcomes for repeat applicants, controlling for their average experience with an applicant fixed effect. Each column corresponds to a regression of the form

$$y_{ij} = \alpha p_j + \beta \text{POST}_j + \gamma \left(\text{POST}_j \times p_j\right) + \delta_i + \epsilon \tag{2}$$

where y_{ij} is the outcome (or choice) for worker i applying to job opening j. POST_j is an indicator that the opening happened after the application fee intervention began for the corresponding job category. p_j is the application fee that was charged (or that would have been charged) for job j. δ_i is a worker-specific fixed effect.

Table 7: Effects of price intervention on repeat applicant outcomes in *tokens*

	Log # Applications	Log Hourly Rate	Applicant Hired?
	(1)	(2)	(3)
Price, p	0.426^{***}	-0.008^{***}	-0.014^{***}
	(0.001)	(0.001)	(0.0002)
POST	-0.189^{***}	0.027^{***}	0.005^{***}
	(0.001)	(0.001)	(0.0003)
POST $\times$ p	-0.159^{***}	0.005^{***}	-0.002^{***}
	(0.002)	(0.001)	(0.0003)
Applicant FE	Y	Y	Y
Observations	19,193,666	19,193,666	19,193,666
R^2	0.383	0.909	0.083

Notes: Significance indicators: $p \leq 0.05 : *$, $.p \leq 0.01 : **$ and $p \leq .001 : ***$.

Because applicants were aware of the fee intervention and may have changed which openings they applied to, the regression results cannot be read as a direct treatment effect. However, they are indicative of the cumulative equilibrium effect that the intervention had on workers' experiences. Column (1) reports the effect on the logged number of total applications received by openings that repeat workers applied to across different price

categories. The coefficient on price is large and significant, suggesting that openings slated to have a $0.30 higher application fee would be expected to have 13% more applications at the baseline, prior to the intervention. This confirms that the formula used to assign prices to job openings captured meaningful heterogeneity in the baseline attractiveness of different jobs. However, applications to openings affected by the intervention had 19% fewer applications on average, and 5% fewer still for each $0.30 jump in the application fee price.

Column (2) reports the effect of the intervention on the logged hourly wage rate requested by workers in their applications. Openings slated to have higher application fees received more applications and lower wage bids on average prior to the intervention. This is consistent with an interpretation of these openings as more competitive. However, wage bids to openings affected by the intervention were at least 2.7% higher on average, and over 3% higher for the jobs with the highest application fees. This increase in wage bids may reflect both a decrease in competition for openings and an increase in the propensity of higher value workers to apply given the application fee.

Finally, Column (3) shows that holding all else fixed, repeat workers were slightly more likely to be hired when applying to openings that were affected by the intervention—about 0.44% more for the jobs with the lowest fees and 0.32% more for the jobs with the highest fees. This is consistent with the compositional effect demonstrated in Table 6 and suggests that the reduction in competition induced by the interventions may have benefited workers with more experience on the platform.

While having application costs slightly increased employer satisfaction upon hiring a worker, the increase in openings which received *no* applications and the decrease in hiring rate suggest that the application costs were perhaps set too high for a significant portion of openings. Even without an application cost, there were job openings which did not get enough applications, so setting any positive cost is clearly inefficient for such openings. In the experiment, the introduction of application costs also induced a compositional shift in the applicant pool towards more experienced workers, who likely offered higher match value and therefore gave higher wage bids. Such selection against first-time workers may not be desirable in the long run, both for the highly desirable jobs which hired a more experienced worker and for the jobs which received few applications and made no hires.

7 Conclusion

We studied two experimental interventions on an online job matching platform to test their impact on efficiency: one implementing a soft cap on the number of job applications, and another introducing application fees based on job attractiveness.

The soft cap (*autopause*) experiment showed that job matching platforms may be able to improve matching and reduce wasted applications by limiting applications. In our study, employers rarely chose to lift the soft cap once they received 50 applications or five days, and hiring rates and match quality were unchanged. This suggests that a per-job application limit could (a) cut down on the costs of wasted applications for job seekers and (b) decrease late applicant crowd-out without decreasing the probability of a high-quality match. The fact that the soft cap did not affect the winning wage bid or the employers' ability to make high-quality hires indicates that additional applications to popular jobs have little effect on hiring decisions. Future experiments could help determine the optimal cap, which might be lower than the 50-application threshold used here.

This is the first experiment we are aware of where the number of applications to a job opening was experimentally varied. The key contribution of the paper is to use this experimental variation to show that many job applicants are inframarginal in the decentralized labor market equilibrium. We also illustrate the crowd-out effect of other applicants in a particularly direct way, compared to the literature (Lalive, Landais and Zweimüller, 2015). Our crowd-out results call into further question the equilibrium justification for job search assistance (Crépon, Duflo, Gurgand, Rathelot and Zamora, 2013; Marinescu, 2017).[20] The soft cap offers a practical (if partial) solution to inefficient congestion in job applications that could be implemented by any online labor matching marketplace.

In our second experiment *tokens*, the platform imposed application costs of $.30, $.60, or $.90 based on observable job characteristics. Priced applications had a lower hiring probability, but the quality of hire was higher, conditional on hiring anyone. Although application fees led to a lower hiring probability, the quality of hires improved, and wage bids were higher. This was driven by more experienced applicants (and fewer first-time users) applying to treated jobs. However, the decrease in job openings filled suggests that the fees were miscalibrated, leading to inefficiencies. Application fees may be harder

[20] Though there is evidence that more targeted recruiting assistance can be helpful without much crowd-out (Horton, 2017,1) and that interventions that have job-seekers consider a wider range of options could be beneficial, as in Belot, Kircher and Muller (2019). The lack of an *increase* in hiring in the treatment is evidence against the "choice overload" hypothesis (Iyengar and Lepper, 2000), which itself has been called into question (Scheibehenne, Greifeneder and Todd, 2010).

to optimize than a soft cap, underscoring that the latter may be a more "robust market design," which can improve platform efficiency without without the complexity of pricing externalities.

References

Agrawal, Ajay K, Nicola Lacetera, and Elizabeth Lyons, "Does information help or hinder job applicants from less developed countries in online markets?," 2013, (NBER Working Paper 18720).

Arrow, Kenneth J, "The organization of economic activity: Issues pertinent to the choice of market versus nonmarket allocation," in "The Analysis and Evaluation of Public Expenditure: the PPB system," Vol. 1 1969, pp. 59–73.

Barach, Moshe E. and John J. Horton, "How do employer use compensation history? Evidence from a field experiment," *Journal of Labor Economics*, 2020.

Belot, Michele, Philipp Kircher, and Paul Muller, "Providing advice to jobseekers at low cost: An experimental study on online advice," *Review of Economic Studies*, 2019, *86* (4), 1411–1447.

Callaway, Brantly and Pedro HC Sant'Anna, "Difference-in-differences with multiple time periods," *Journal of Econometrics*, 2021, *225* (2), 200–230.

Crépon, Bruno, Esther Duflo, Marc Gurgand, Roland Rathelot, and Philippe Zamora, "Do labor market policies have displacement effects? Evidence from a clustered randomized experiment," *Quarterly Journal of Economics*, 2013, *128* (2), 531–580.

DellaVigna, Stefano and Devin Pope, "Stability of experimental results: Forecasts and evidence," *Working paper*, 2019.

Diamond, Peter A, "Aggregate demand management in search equilibrium," *Journal of Political Economy*, 1982, *90* (5), 881–894.

Feld, Scott L, "Why your friends have more friends than you do," *American Journal of Sociology*, 1991, *96* (6), 1464–1477.

Hayek, Friedrich August, "The use of knowledge in society," *American Economic Review*, 1945, *35* (4), 519–530.

Horton, John J., "The Effects of Algorithmic Labor Market Recommendations: Evidence from a Field Experiment," *Journal of Labor Economics*, 2017, *35* (2), 345–385.

____ , "Buyer uncertainty about seller capacity: Causes, consequences, and a partial solution," *Management Science*, 2019.

____ **, William R Kerr, and Christopher Stanton**, "Digital labor markets and global talent flows," in "High-skilled Migration to the United States and Its Economic Consequences," University of Chicago Press, 2017, pp. 71–108.

Hosios, Arthur J, "On the efficiency of matching and related models of search and unemployment," *The Review of Economic Studies*, 1990, *57* (2), 279–298.

Iyengar, Sheena S and Mark R Lepper, "When choice is demotivating: Can one desire too much of a good thing?," *Journal of Personality and Social Psychology*, 2000, *79* (6), 995.

Krishna, Vijay, *Auction theory*, Academic press, 2009.

Lalive, Rafael, Camille Landais, and Josef Zweimüller, "Market Externalities of Large Unemployment Insurance Extension Programs," *American Economic Review*, 2015,

105 (12), 3564–96.

Levin, Dan and James L Smith, "Equilibrium in auctions with entry," *American Economic Review*, 1994, pp. 585–599.

Lipsey, Richard G and Kelvin Lancaster, "The general theory of second best," *Review of Economic Studies*, 1956, *24* (1), 11–32.

Marinescu, Ioana, "The general equilibrium impacts of unemployment insurance: Evidence from a large online job board," *Journal of Public Economics*, 2017, *150*, 14–29.

——— **and Daphné Skandalis**, "Unemployment insurance and job search behavior," *Quarterly Journal of Economics*, 2021, *136* (2), 887–931.

——— **and Ronald Wolthoff**, "Opening the black box of the matching function: The power of words," *Journal of Labor Economics*, 2020, *38* (2), 535–568.

Mortensen, Dale T and Christopher A Pissarides, "Job creation and job destruction in the theory of unemployment," *The review of economic studies*, 1994, *61* (3), 397–415.

Nisan, Noam and Ilya Segal, "The communication requirements of efficient allocations and supporting prices," *Journal of Economic Theory*, 2006, *129* (1), 192–224.

Pallais, Amanda, "Inefficient hiring in entry-level labor markets," *American Economic Review*, 2013.

Roughgarden, Tim and Éva Tardos, "Introduction to the inefficiency of equilibria," in Noam Nisan, Tim Roughgarden, Éva Tardos, and Vijay V. Vazirani, eds., *Algorithmic Game Theory*, Cambridge University Press, 2007, chapter 17, pp. 443–459.

Scheibehenne, Benjamin, Rainer Greifeneder, and Peter M Todd, "Can there ever be too many options? A meta-analytic review of choice overload," *Journal of Consumer Research*, 2010, *37* (3), 409–425.

Stanton, Christopher T and Catherine Thomas, "Landing the first job: The value of intermediaries in online hiring," *Review of Economic Studies*, 2016, *83* (2), 810–854.

Sun, Liyang and Sarah Abraham, "Estimating dynamic treatment effects in event studies with heterogeneous treatment effects," *Journal of Econometrics*, 2021, *225* (2), 175–199.

van Ours, Jan and Geert Ridder, "Vacancies and the recruitment of new employees," *Journal of Labor Economics*, 1992, *10* (2), 138–155.

Watt, Mitchell, "Concavity and Convexity of Order Statistics in Sample Size," *arXiv preprint arXiv:2111.04702*, 2022.

A Additional Tables and Figures

A.1 Event Study Plots

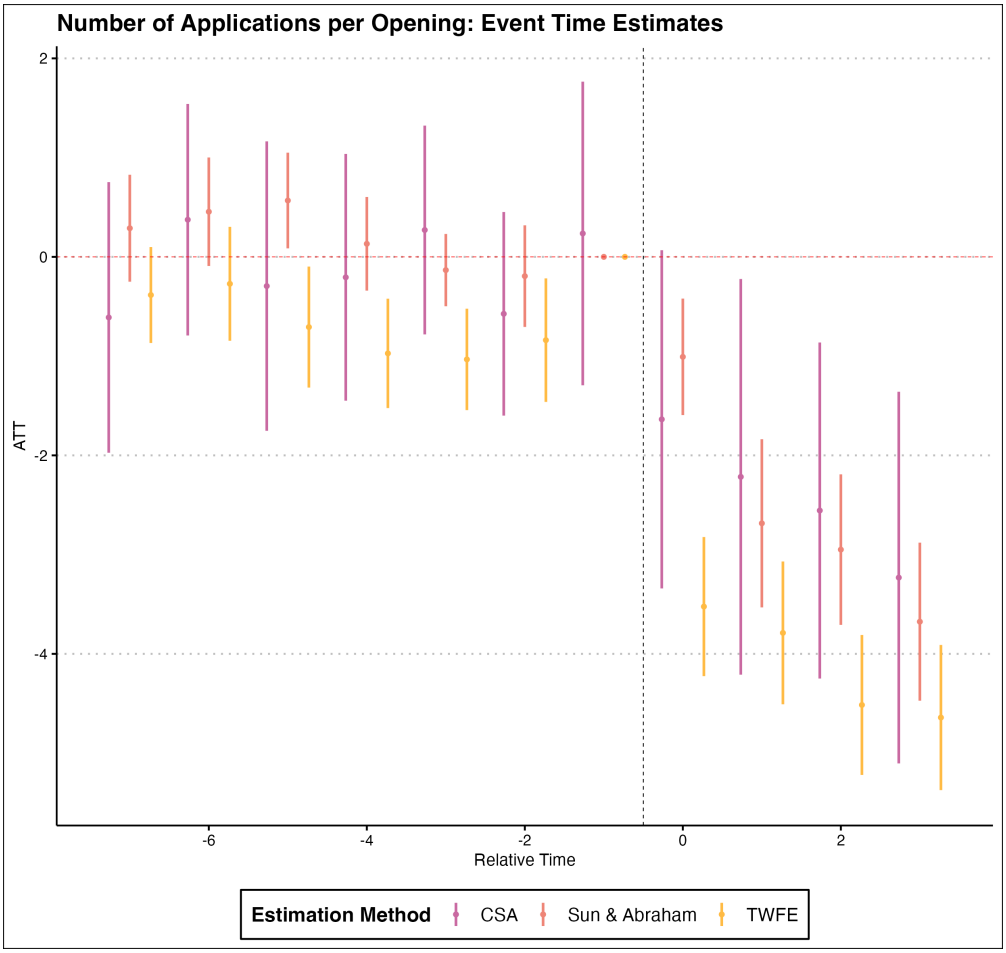


Figure 4: Event study for the number of applications per opening in the application price experiment, using the Callaway and Sant'Anna (2021), Sun and Abraham (2021) and Two-Way Fixed Effects Estimators

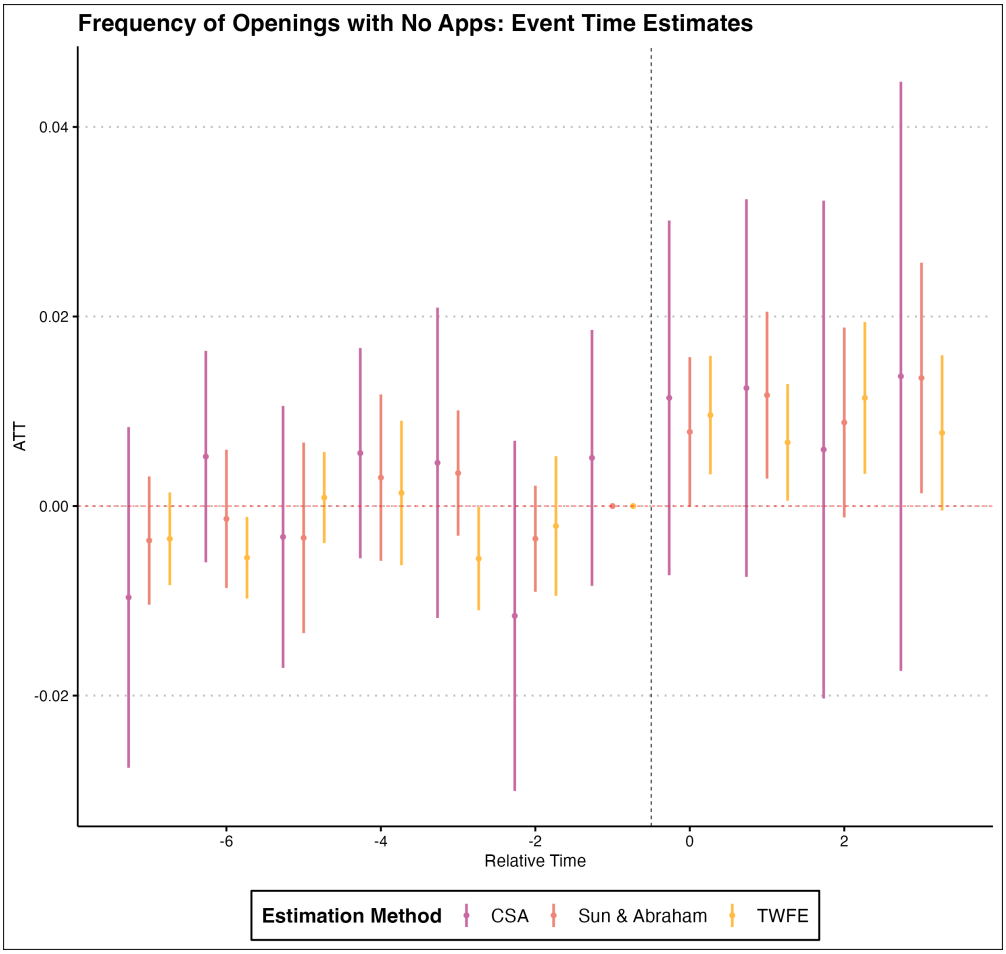


Figure 5: Event study comparison for frequency of openings with no applications submitted in the application price experiment, using the Callaway and Sant'Anna (2021), Sun and Abraham (2021) and Two-Way Fixed Effects Estimators

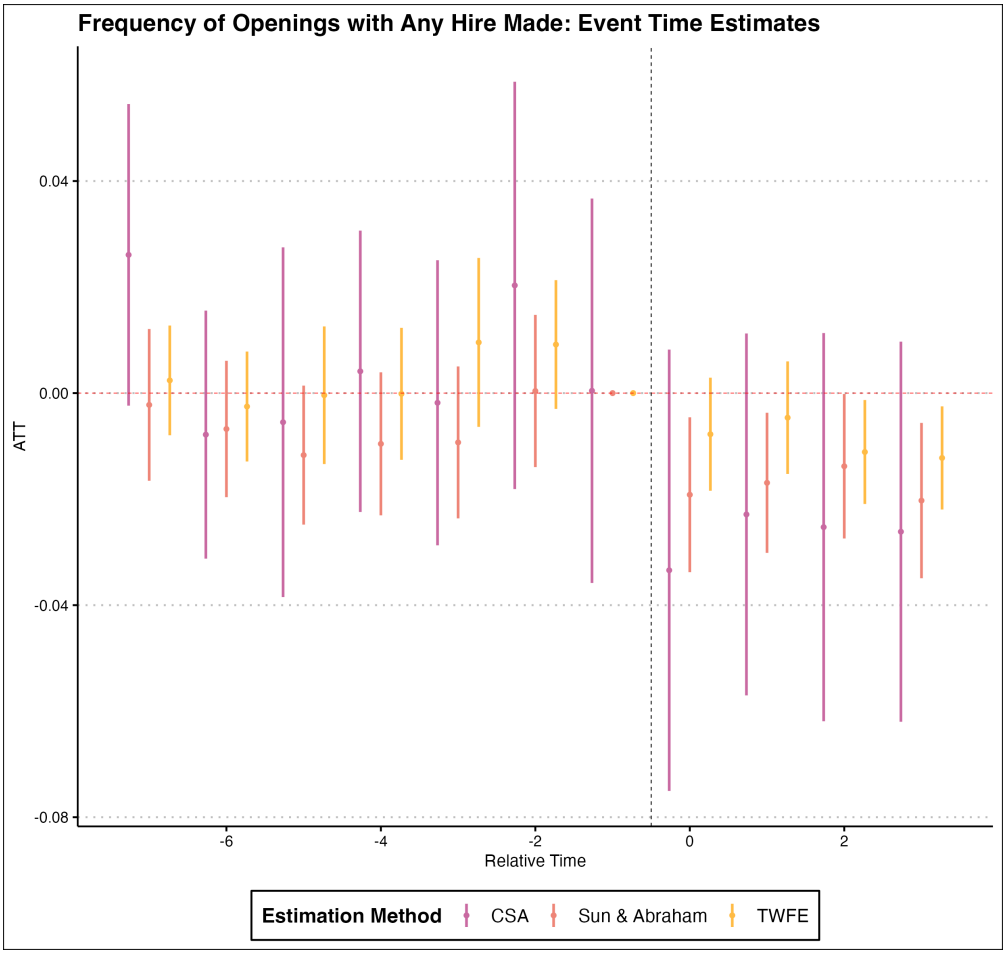


Figure 6: Event study comparison for frequency of making a hire per opening in the application price experiment, using the Callaway and Sant'Anna (2021), Sun and Abraham (2021) and Two-Way Fixed Effects Estimators

A.2 Auction and Bidder Heterogeneity

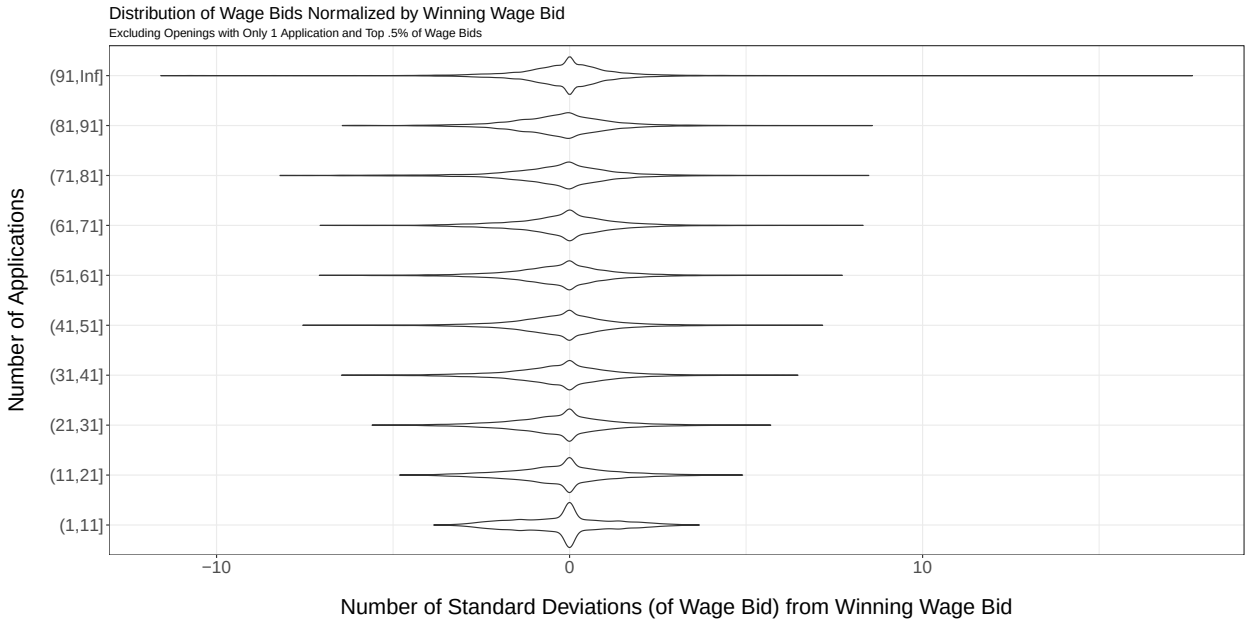


Figure 7: Distribution of wage bids scaled by number of standard deviations from the winning bid, by number of applicants

B Parametric Example of Inefficient Congestion

To build some intuition for the mechanics of congestion in the model introduced in Section 2, we present a simple parametric example. Suppose that there are only two jobs to choose from, $j \in \{1,2\}$, that the employers always hire a candidate and that each candidate's net surplus for each job m_{nj} is drawn identically and independently from the exponential distribution with shape parameter λ_j. Suppose, moreover, that the employer faces a fixed screening cost k_j per application screened, so its expected payoff from screening M_j applications is

$$\begin{aligned}\Pi_j(M_j) &= \mathbb{E}\left[m_{nj}^{(2;M_j)}\right] - M_j \cdot k_j \\ &= \frac{1}{\lambda_j}(H_{M_j} - 1) - M_j \cdot k_j,\end{aligned}$$

where $H_m = \sum_{i=1}^{m} \frac{1}{i}$ is the m^{th} harmonic number. Employer j's screening threshold is then the maximizer of that expression, which is[21]

$$\overline{M_j} = \left\lfloor \frac{1}{\lambda_j \cdot k_j} \right\rfloor.$$

This employer's optimal screening rule is

$$M_j^*(N_j) := \min\{\overline{M_j}, N_j\}.$$

In this example using exponential distributions, this is also the *socially* optimal screening rule: there is no "insufficient screening" effect, as described in Proposition 2.[22] This implies that inefficiency in this example will be caused *only* by misallocation of applications by the job seekers.

Note that $\overline{M_j}$ is increasing in $1/\lambda_j$: the higher the mean (and variance) of the distribution of net surplus values, the more applicants that employer j is willing to screen. If more than $\overline{M_j}$ applications are received, the job is congested, and the employer chooses a uniformly random subset of size $\overline{M_j}$ to screen, discarding the remaining $N_j - \overline{M_j}$ applications.

[21] Notation: here $\lfloor z \rfloor$ is the *integer floor* of z, the largest integer less than or equal to z.

[22] This follows because the expected increase in the first-order statistic from an exponential distribution equals the expected increase in the second-order statistic. .

The expected payoff of an applicant to a job with N_j total applicants is

$$\Phi_j(N_j) = \frac{1}{N_j}\mathbb{E}[m_j^{(1;M_j(N_j))} - m_j^{(2;M_j(N_j))}] = \frac{1}{\lambda_j \cdot N_j}.$$

In a pure-strategy Nash Equilibrium, no candidate wishes to change the job opening she applied to, so that the number of applications submitted to each job opening, N_j^*, satisfies

$$N_1^* + N_2^* = N \text{ and } \Phi_j(N_j^*) \geq \Phi_k(N_k^* + 1) \text{ for all } j \neq k.$$

This implies

$$N_1^* = \begin{cases} \left\lfloor \frac{\lambda_2(N+1)}{\lambda_1+\lambda_2} \right\rfloor & \text{if } \frac{\lambda_2(N+1)}{\lambda_1+\lambda_2} \text{ is non-integer} \\ \frac{\lambda_2 N - \lambda_1}{\lambda_1+\lambda_2} \text{ or } \frac{\lambda_2(N+1)}{\lambda_1+\lambda_2} & \text{if } \frac{\lambda_2(N+1)}{\lambda_1+\lambda_2} \text{ is an integer.} \end{cases}$$
$$N_2^* = N - N_1^*.$$

Whether congestion arises, and how much inefficiency it generates depends on the characteristics of the labor market: the distributions of surplus values $\{F_j\}$, the number of potential applicants and the screening cost k_j for each job. To see how these characteristics affect congestion, in Figure 8, we illustrate the losses due to congestion in the equilibrium of our example with 25 applicants, two employers with screening costs $k_1 = 0.4$ and $k_2 = 0.2$, and for a range of exponential parameters λ_1, λ_2 in $[0.1, 0.8]$.[23] We plot the proportion of the expected social surplus that is lost due to congestion—the inefficient assignment of applications—relative to the expected efficient surplus.[24] Here, the efficient surplus is the sum of the match values of hired applicants minus the costs associated with screening for the allocation of applications that maximizes this objective, holding fixed the screening rule of the employers.[25] So for the parameter values plotted in Figure 8, the reduction in the expected match efficiency of the hired applicants due to congestion is equal to a loss of 0-5.5% of the efficient surplus.

[23] For illustration purposes only, in Figure 8 depicts a continuous approximation of the problem where N may take non-integer values.

[24] That is, we plot (efficient surplus − equilibrium surplus)/(efficient surplus). This is equivalent to $1 - \frac{1}{PoA}$, where PoA is the so-called "price of anarchy" of this congestion game (see Roughgarden and Tardos (2007)).

[25] Here we focus on the strategic behavior of applicants only, because we take the screening decisions of the employers as outside the control of the online labor platform, the market designer. A social planner would prefer the advertiser to process more job applications than M_j^* if $N_j > M_j^*$ because processing more applications has a positive externality on the applicants, but, in this paper and in the experiments, we focus on interventions influencing the actions of applicants.

Several relationships can be observed in Figure 8. First, congestion occurs mainly in the top-left and bottom-right quadrants of parameter space: these are the areas where the attractiveness of the job opening (in terms of the mean of the net surplus distribution) differ most between the job openings. In the bottom-left of the figure, where both employers have high threshold values, the jobs are not congested in equilibrium. In the top-right of the figure, where both employers have low threshold values, significant congestion occurs both in the equilibrium and in the efficient allocation: there is little misallocation of applications. Second, note that the magnitude of the foregone surplus is higher in the bottom-right, where job opening 1 has both a low threshold for applications and is relatively unattractive relative to job two. Although job 1 receives fewer applications in equilibrium, the marginal applications to job 1 are more important to social surplus, so that under-application to job 1 leads to greater inefficiencies. Together, these patterns suggest that congestion occurs most when jobs differ substantially in attractiveness, and the resulting inefficiency is especially pronounced when congested jobs attract applicants away from jobs where marginal applications are most valuable.

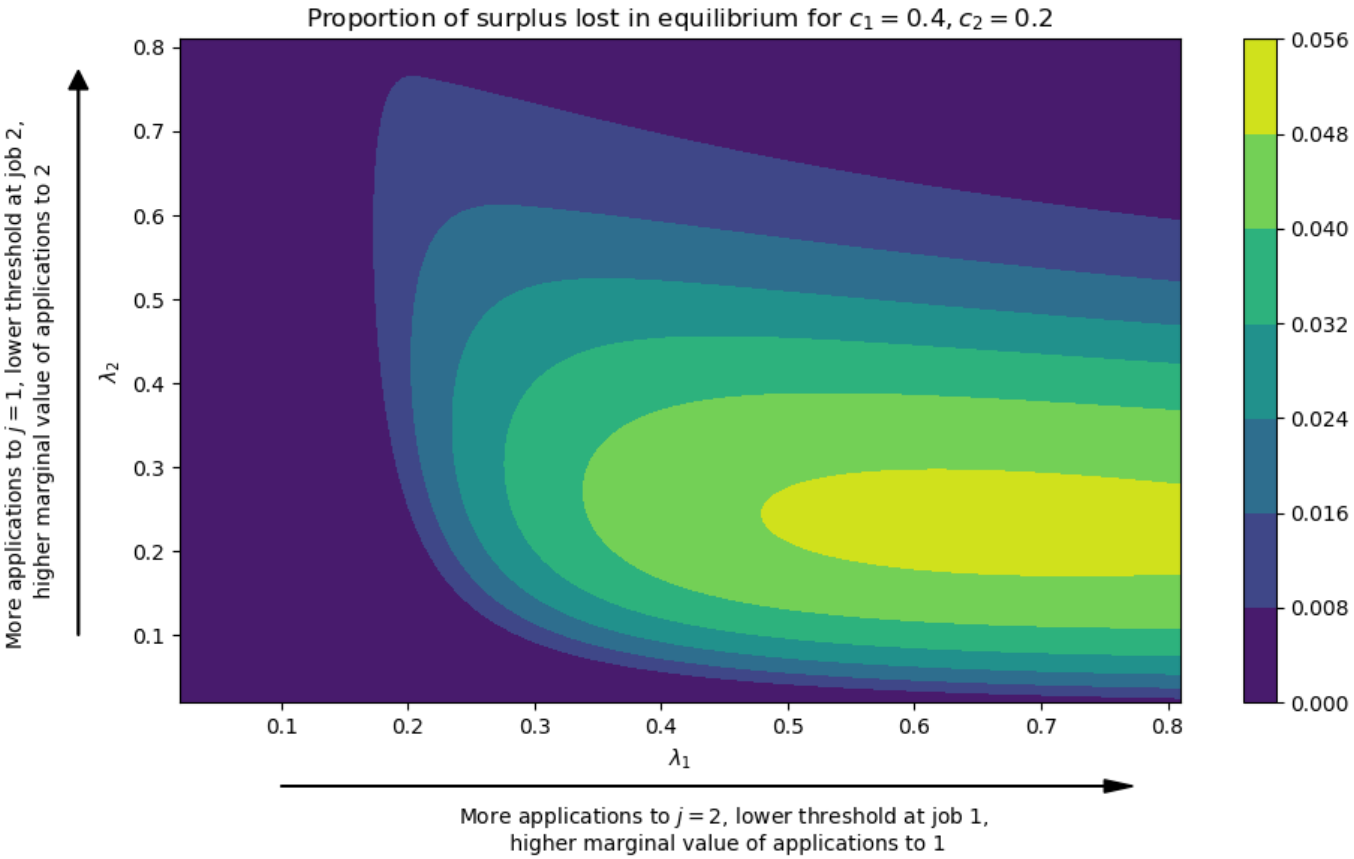


Figure 8: Inefficiency in an example with exponentially distributed surplus values